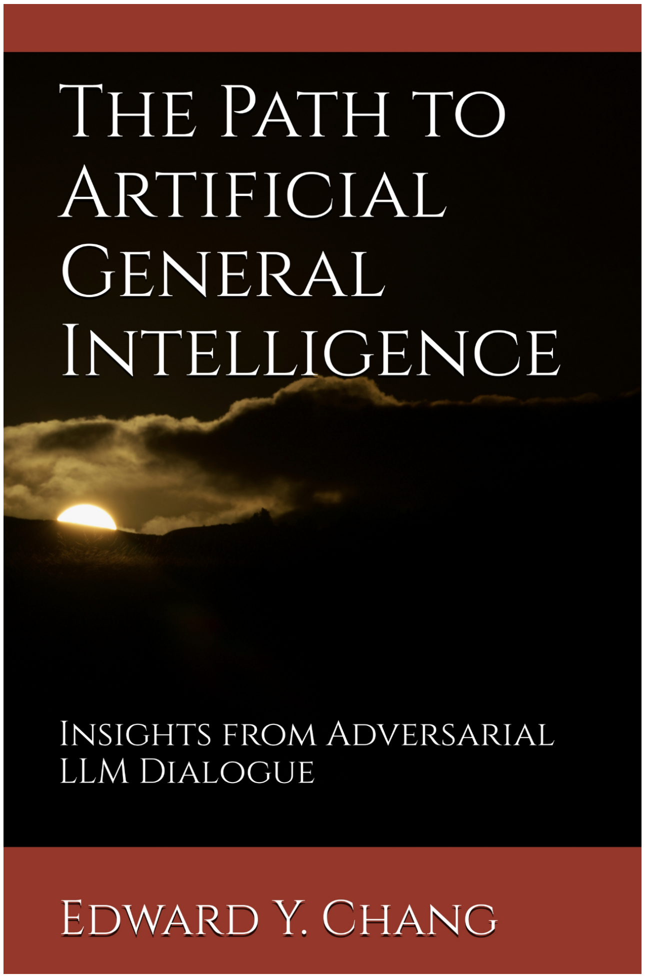
The Path to Artificial General Intelligence

Insights from Adversarial LLM Dialogue

Edward Y. Chang


The Path to Artificial General Intelligence
Insights from Adversarial LLM Dialogue
Edward Y. Chang

The Path to Artificial General Intelligence

Insights from Adversarial LLM Dialogue

Edward Y. Chang
Computer Science Stanford University

Acknowledgments

Thanks to my family for their love and support.

Table of Contents

Preface

Generative AI has captivated the world, and many believe that Artificial General Intelligence (AGI) could be realized as early as 2040. I believe the key to achieving AGI lies in enabling large language models (LLMs) to engage in intelligent dialogue with each other. This conviction has driven my research into multi-LLM dialogue since 2022 and ultimately inspired this book.

Several insights and my own experience developing multi-LLM dialogue frameworks support this hypothesis. Chapter 2 delves into the most critical feature of LLMs: their polydisciplinary representation of multimodal information. Trained to predict the next token in a sequence of vast amounts of text, LLMs don't distinguish between domains or disciplines. This creates a "polydisciplinary" representation, fostering the synthesis of new insights, knowledge, and potentially, higher levels of intelligence.

When we ask an LLM a question, we understand its domain, but the LLM doesn't. This can lead to responses that go beyond our often imprecise intentions. Conversely, if we consider LLMs as repositories of synthesized knowledge from all disciplines, asking them a deep question is like a child of ten years old conversing with a panel of Nobel Laureates from diverse fields. The resulting dialogue is unlikely to be profound or insightful.

LLMs possess a wealth of "unknown knowns" — knowledge that we don't even know to ask about. To unlock this hidden potential, we must facilitate debate between LLMs themselves. While humans can guide and monitor these dialogues, our primary role is to listen and learn.

My initial foray into multi-LLM dialogue introduced the CRIT algorithm (Chapter 4), available in January 2023 and formally published in March 2023. CRIT employs the Socratic method and formal reasoning to critically evaluate a document, locating its claim, supporting reasons, and counterarguments to test their strength. Shortly after, I developed the SocraSynth framework (Chapter 5), enabling two LLMs to converse. Early experiments, such as exploring the "Adam and Eve" narrative with two GPT-3 instances (detailed in Chapter 12), piqued the interest of colleagues at Stanford. However, a critique from Professor Vaughan Pratt regarding the agents' tendency to echo each other prompted a shift towards fostering "contentious" debates.

The introduction of "contentiousness" elevates this multi-LLM framework beyond traditional ensemble approaches that leverage redundancy to minimize error rates. SocraSynth emphasizes information discovery through rigorous reasoning among multiple LLM instances, fostering contentious yet productive debates. Surprisingly, adjusting the intensity of these debates can alter the LLMs' linguistic behaviors. This demonstrates that through in-context learning, similar to Bayesian inference, an LLM's tone, attitude, emphasis, and linguistic features can be influenced.

SocraSynth showcases our ability to "condition" LLMs to discover information and express themselves effectively. In the debate process, multiple perspectives on a subject matter emerge, and hallucinations tend to dissipate due to the gradual development of rich context between the participating LLMs and increasing precision in their arguments and counterarguments. However, a key challenge remains: how to "condition" LLMs to strike a balance between exploration (discovering new perspectives) and exploitation (refining prior knowledge).

In 2024, I further developed theoretical pillars consisting of several maxims and theories to enhance multi-LLM communication. This framework, named EVINCE (Entropy Variation and Information Competence), encompasses maxims and theories rooted in Bayesian statistics and information theory, most notably the Dual Entropy theory. This theory establishes the optimal initial settings for two LLMs to strike a balance between exploration and exploitation, enhancing prediction accuracy while maintaining stability. To measure, monitor, and manage debate progress and dynamics, we also employ a suite of metrics, including mutual information, cross entropy, Wasserstein distance, Jensen-Shannon divergence, and KL divergence. These metrics help distinguish novel insights from noise and foster positive information exchange between participating LLMs. Details are presented in Chapter 6, with a focus on EVINCE's potential for safeguarding AI safety explored in Chapters 7 through 9.

Chapter 10 delves into consciousness modeling, a necessity arising from communication. Humans have inhabited Earth for two billion years, primarily operating in an unconscious mode driven by survival and reproduction. Our most critical functions, such as heartbeat, breathing, and metabolism, operate without conscious intervention. Consciousness emerges when humans need to adapt to changing environments and learn new skills, leading to the development of knowledge and intelligence. By understanding the transitions between unconsciousness and consciousness and modeling emotions, behaviors, and ethics, Chapters 9 and 10 form the foundation for strengthening AI safety and ethical guardrails.

This book traces the evolution of AI, interweaving my personal experiences in the field, from the early days of expert systems to the current advancements in General AI (GAI), culminating in this final path towards Artificial General Intelligence (AGI).
In this exploration, you will:

• **Harness the Power of Adversarial Collaboration Guided by Theoretical Pillars**: Discover how structured debates between LLMs, guided by robust theoretical frameworks, can unlock hidden knowledge, challenge assumptions, and lead to more informed and robust decisions.

• **Unveil Unknown Unknowns**: Explore how SocraSynth's polydisciplinary and multimodal representation can help uncover insights and knowledge gaps that elude human understanding, pushing the boundaries of discovery.

• **Witness Real-World Applications of LLM Collaboration**: See the transformative potential of multi-LLM dialogues in diverse fields, including healthcare, content moderation, strategic planning, and more.

• **Embrace a New Paradigm for AI Safety and Ethics**: Learn how EVINCE's principles can guide the development of safer, more ethical, and culturally sensitive AI systems that align with human values.

• **Glimpse the Future of AI-Human Interaction**: Gain insights into how collaborative AI can augment human decision-making, foster transparency, and contribute to a more informed and equitable society.

Thus, I hypothesize that the path to AGI, artificial intelligence that surpasses human capabilities, lies through LLM communication. This book aims to present the evidence and arguments that support this conviction, culminating in a vision of the future where collaborative AI systems not only achieve human-level intelligence but also transcend it, opening up new realms of possibility and understanding. Whether or not time proves this right, the journey itself promises to be transformative, shaping the trajectory of AI research and development for years to come.

Edward Y. Chang, July 14th, 2024.
Series π 0010514G004

1 A Brief History of AI: From Turing to Transformers

Abstract This chapter reinterprets the history of AI, focusing on the evolution of similarity measurement, from rule-based to context-aware models, and emphasizing its critical role in AI's core functions like learning and problem-solving. It explores the impact of detailed and evolving understandings of similarity in linguistics (text) and computer vision (image), projecting a future where AI merges advanced data analysis with abstract reasoning. The chapter will provide an in-depth analysis from the perspectives of linguistics, computer science, and cognitive psychology/neuroscience, illustrating how the progression of similarity concepts continues to fuel AI's advancement.

Introduction

Artificial Intelligence (AI) has journeyed through a fascinating historical trajectory, marked by five pivotal epochs that each represent significant paradigm shifts triggered by major technological advancements. The epochs are as follows: *Initiation*, setting the stage with foundational concepts and milestones of AI; *Expert System Encoding Human Knowledge*, where AI systems were predominantly rule-based, encoding and applying human expertise; *Heuristic-Based Modeling*, which highlights the era of developing and using heuristic methods for AI problem-solving; *Learning Model from Data*, focusing on the transition to algorithms that learn and adapt from data, signifying the emergence of machine learning; and *Context-Based Semantic Disambiguation*, highlighting AI's evolving proficiency in understanding and interpreting context, thereby improving semantic accuracy.

While numerous comprehensive sources, such as Wikipedia, provide detailed accounts of AI's evolution through various lenses: language, computation, philosophy, cognitive psychology, neuroscience, and application–this chapter takes a different path. It zeroes in on a fundamental aspect: **similarity**.

When we consider the intelligence of machines, we often focus on attributes such as learning capacity, pattern recognition, predictive accuracy, robustness, adaptability, generalization, reasoning, problem-solving, and decision making abilities. These qualities collectively define the prowess of AI systems. Among these traits, the concept of similarity plays a pivotal role. For instance, in learning, an effective similarity measure is fundamental for recognizing patterns and generalizing knowledge. In terms of adaptability, the ability to detect similarities to previous experiences allows AI to adjust to new or evolving circumstances. Regarding robustness, employing similarity measures helps AI differentiate between normal and anomalous patterns, thereby increasing its resilience. Furthermore, in the realm of problem-solving, the capacity to identify similarities to previously encountered situations can enable AI to apply existing solutions to new problems, enhancing its efficacy in addressing challenges. This chapter explores the vital function of similarity across the broad spectrum of AI capabilities, underlining its significant contribution to the field's foundational operations.

In the realm of tangible objects, similarity measures are integral to various vision-related tasks, aiding in the recognition of patterns, shapes, and colors, which are essential for object recognition and image classification. In text analysis, these measures are crucial for identifying similarities in content, aiding in plagiarism detection, document retrieval, and language translation. In the auditory domain, similarity analysis of sound wave patterns or musical notes is key to genre classification and music recommendation systems. In medical imagery, these measures facilitate disease diagnosis by comparing patient images with known cases, enabling accurate medical condition identification and classification. Object feature comparison is foundational in robotics and surveillance for recognizing and interacting with physical entities. Similarly, facial and voice recognition systems rely on analyzing patterns to identify or verify identities, enhancing security and personal authentication. In e-commerce, similarity in product attributes or user preferences informs recommendation systems, enhancing user experience by suggesting related or complementary products.

In the abstract realm, similarity measures are crucial for discerning semantic relationships, aiding in knowledge representation, ontology mapping, and refining AI's interpretive faculties. Environmental studies leverage these assessments for climate modeling and ecological research. Sentiment analysis in social media or customer feedback utilizes similarity to extract insights into public sentiment or consumer behavior. These measures also underpin AI's problem-solving prowess in complex scenarios, informing strategy formulation. Behavioral analysis, whether in psychology or marketing, employs similarity comparisons to decode human actions and preferences. In the legal field, case similarity aids in judicial decisionmaking and legal scholarship. Language translation harnesses similarity in linguistic structures to break down language barriers. Furthermore, in creative writing, analyzing thematic or stylistic similarities assists in authorship identification, genre categorization, and literary exploration.

The advancement in similarity research, while appearing gradual, reflects not only human ingenuity but also the limitations imposed by computational resources and hardware capabilities. The quest to quantify similarity covers a broad spectrum of abstractions, from sensory inputs like visual, auditory, olfactory, and tactile data to complex abstract concepts such as ideas and semantics. Hardware improvements have enabled researchers to explore more advanced methods that encompass both concrete and abstract forms of similarity. This progression marks the field's growth in harmonizing detailed sensory data analysis with a deeper understanding of abstract concepts, utilizing computational advancements and diverse data interpretations.

Following sections will provide a deeper dive into key AI terminology and the development of similarity measures in two distinctive views: scientific disciplines and historical evolution. The disciplinary view encompasses three key perspectives:
ogy, and neuroscience.
linguistics, computer science, cognitive psycholThe evolution view traces the historical journey

of similarity measurement through distinct eras: rule-based, model-based, data-centric, and context-aware.

Providing two views on similarity measurements–across different scientific disciplines and through the historical evolution of AI methodologies–offers a comprehensive understanding that caters to a broader audience with varied interests and backgrounds. Here are some reasons why this dual perspective is valuable:
Multidisciplinary Insight: Examining similarity measurements from different scientific disciplines enriches the understanding by highlighting how various fields approach and apply the concept of similarity. This can foster interdisciplinary collaboration and innovation, as techniques from one field can inspire new approaches in another.
Historical Context: Exploring how similarity measurement has evolved within AI provides historical context, showcasing how methodologies have progressed from rule-based to more advanced context-aware systems. This perspective helps readers appreciate the advancements in AI and understand why certain methods were developed or abandoned.

1.1 Definitions

We define and scope key terms and concepts to prepare for subsequent discussion.

1.1.1 Rudimentary Terms

Data : The raw information used to train AI models. Data quality significantly impacts model performance.
Algorithm: A set of instructions that a computer follows to perform a specific task. AI algorithms are often complex and involve statistical methods.
Model: A representation of the learned knowledge from data that allows the AI system to make predictions or decisions.

1.1.2 General Terms

Artificial Intelligence (AI) : The broader concept of machines being able to carry out tasks in a way that we would consider smart.

Explainable AI : AI systems that offer transparency and an understanding of their operations and decision-making processes.
General AI : General AI, also known as Artificial General Intelligence (AGI), refers to a type of AI that has the ability to understand, learn, and apply knowledge in a wide range of tasks, much like a human being. It's an AI system with generalized human cognitive abilities, meaning that when presented with an unfamiliar task, it can find a solution without human intervention. AGI would be able to reason, solve problems, make judgments, plan, learn, and communicate in natural language, among other capabilities. However, as of now, AGI remains a theoretical concept and has not been realized in practical applications.
Narrow AI : Narrow AI, in contrast, is the type of AI that we encounter in our daily lives and is currently in use around the world. It is designed to perform a narrow task (e.g., facial recognition, internet searches, driving a car) and is trained for a specific dataset or a set of tasks. Narrow AI operates under a limited pre-defined range or context, often focusing on executing a single task extremely well or carrying out a limited range of tasks in a specific domain. It lacks the general cognitive abilities of AGI and cannot apply its knowledge beyond its specific field or task. *Machine Learning (ML)*: A subset of AI that includes statistical techniques that enable machines to improve at tasks with experience.
Deep Learning: A subset of machine learning that uses neural networks with three or more layers. These neural networks attempt to simulate the behavior of the human brain–albeit far from matching its ability–allowing it to "learn" from large amounts of data.
Neural Networks Computational models that are somewhat inspired by the structure of the human brain, enabling computers to recognize patterns and solve common problems in AI, such as classification, prediction, and decision making.
Supervised Learning: A type of machine learning where the model is provided with labeled training data and the desired output. The goal is to learn a mapping from inputs to outputs.
Unsupervised Learning: A type of machine learning where the model is not provided with labeled data and must find structure in its input on its own.
Reinforcement Learning: An area of machine learning where an agent learns to behave in an environment by performing actions and seeing the results, focusing on long-term rewards. An example is an AI agent learning to play a

game through trial and error, receiving rewards for winning. *Natural Language Processing (NLP)*: A field of AI that gives machines the ability to read, understand, and derive meaning from human languages. *Computer Vision*: A field of AI that trains computers to interpret and understand the visual world, extracting information from images and videos. *Robotics*: The branch of technology that deals with the design, construction, operation, and application of robots, often incorporating AI systems to enhance autonomy and adaptability.

Large Language Model (LLM). LLMs are advanced artificial intelligence systems trained on extensive datasets, initially text-centric and now increasingly incorporating multimodal data. They are designed to comprehend, generate, and interact with human language, imagery, and video with a level of sophistication that closely mirrors human cognitive processes.

1.1.3 Performance Terms

Algorithmic Bias : Algorithmic bias refers to the potential for algorithms to reflect, perpetuate, or amplify biases present in the training data or as a result of the design of the algorithms themselves. This can lead to skewed or unfair outcomes, particularly in decision-making processes. *Hallucination*: In the context of AI, hallucination refers to the phenomenon where a model generates or outputs information that is ungrounded, misleading, or not supported by the input data. This is commonly seen in language models where the generated text may be plausible but not factually accurate or relevant to the context.

Generalization: Generalization is the ability of an AI model to perform well on new, unseen data that was not part of the training set. It indicates the model's capacity to apply learned knowledge to different situations, a key indicator of its robustness and utility.

Overfitting: Overfitting occurs when an AI model learns the details and noise in the training data to the extent that it negatively impacts the model's performance on new data. This usually happens when the model is too complex, capturing patterns that do not generalize to unseen data.

1.2 Perspectives on Similarity

This section presents the foundational theories of similarity measurement from three distinct domains: *linguistics*, *computer science*, and *cognitive psychology & neuroscience*. The upcoming historical section will clarify how these foundational theories have influenced and been incorporated into specific technological advancements and methodologies across various eras. Cross-references will be provided to ensure coherence and to emphasize the interconnection of these perspectives.

1.2.1 Linguistic Perspective

The study of similarity within linguistics has been profoundly influenced by Zellig Harris's pioneering work. His 1954 study introduced the idea that the distributional properties of words and their contextual usage could unlock the secrets of language comprehension, highlighting the indispensable role of context [22]. This principle, that words found in similar contexts tend to share meanings, laid the foundation for distributional semantics and resonates with John R. Firth's insight that "A word is known by the company it keeps." This linguistic perspective sets the stage for further exploration of how context and distributional properties have been instrumental in shaping our understanding of semantic similarity, paving the way for subsequent advancements in the field.

The evolution of linguistic theories continued into the latter part of the 20th century with the rise of cognitive linguistics, which examines the interplay between linguistic structures and human cognitive processes. This approach underscored how language reflects our perception and conceptualization of the world, introducing a multi-layered perspective on semantic abstraction.

A significant milestone in bridging linguistic theory with practical applications was the development of WordNet in the 1980s by a team at Princeton University [41]. This lexical database, which organizes English words into sets of cognitive synonyms or *synsets*, has profoundly influenced areas such as word sense disambiguation, information retrieval, and beyond, highlighting the importance of structured semantic relationships in understanding language.

Moreover, the influence of linguistic insights extended into the domain of computer vision with the creation of ImageNet by Fei-Fei Li [16], which drew upon the principles underlying WordNet to categorize visual content. This convergence of linguistics and computer science has been further propelled by advancements in computational methods, with techniques like Latent Semantic Analysis (LSA) [18], Latent Dirichlet Allocation (LDA) [4], and innovative word embeddings such as Word2Vec [40] and GloVe [45]. These methodologies have enabled the conceptualization of word meanings in high-dimensional spaces, illuminating the intricate web of semantic relationships through patterns of co-occurrence and contextual analysis.

The introduction of the transformer model [50] and the subsequent unveiling of BERT [17], which employs self-supervised learning to predict masked words within a context, along with the release of GPT, designed to predict the next word based on context, heralded a new epoch in our endeavor to unravel context-dependent semantics. This development fulfills the vision proposed by Zellig Harris in his groundbreaking 1954 work, now actualized in contemporary computational models.

1.2.2 Computer Science Perspective

In computer science, the concept of similarity has evolved from simple rulebased models to complex vector-space and probabilistic models, reflecting the field's progression in addressing various computational challenges.

A. Rule-Based

A rule-based AI model, also known as an expert system, employs a collection of predefined if-then statements to execute decisions or solve problems. These conditional statements are crafted from the expertise of specialists in a particular field. The system applies these rules to the input data to formulate conclusions.

The "if" segment of a statement evaluates the data for specific conditions or patterns. When these conditions are satisfied, the "then" segment is activated, performing a designated action or drawing a conclusion. Importantly, these systems do not adapt or learn from data in the manner that machine learning models do. Rather, they rely on a set of explicit rules, which are the codified versions of expert knowledge within a specific domain. This knowledge is methodically organized and stored in a knowledge base, enabling the system to reference and apply it efficiently during its operations.

In Chapter 1.3.1, we will explore the technical details and applications of rule-based systems, emphasizing their pivotal role during the rule-based era of AI's evolution.

B. Vector-Space

The vector-space model marked a significant shift, representing objects and features as vectors in a high-dimensional space. This approach facilitated the development of various distance functions to assess similarity for different applications. Notably, a comprehensive survey by [7] categorized 45 distance functions into families like inner product, L_1, Minkowski, and Intersection, each with its representative functions highlighting the versatility in vector-space analysis.

B.1. Inner product, dot product and cosine The inner product and dot product are the same in the context of Euclidean space and are defined for vectors a and b as:
$a \cdot b = a_1 b_1 + a_2 b_2 + \ldots + a_n b_n$.

This operation results in a scalar value and indicates the vectors' magnitude and directionality.
Cosine similarity is a measure that calculates the cosine of the angle between two vectors. It is defined as the dot product of the vectors normalized by the product of their magnitudes:

$\text{cosine similarity}(a, b) = \frac{a \cdot b}{\|a\|\|b\|}$,
where $\|a\|$ and $\|b\|$ represent the Euclidean norms of the vectors a and b, respectively.

The cosine similarity is especially useful in contexts where the magnitude of the vectors is not of primary concern, making it ideal for applications in high-dimensional spaces like text analysis and information retrieval.

B.2. Weighted Minkowski

The weighted Minkowski distance function allows assigning varying importance to different dimensions, accommodating the significance of specific features in contexts like machine learning and data mining:

The weighted Minkowski distance between two points $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$ with a set of weights $W = (w_1, w_2, \ldots, w_n)$ is defined as:

$$D(X,Y) = \left(\sum_{i=1}^{n} w_i |x_i - y_i|^p\right)^{\frac{1}{p}},$$

where p is the order parameter of the Minkowski distance. When $p = 1$, it becomes the weighted Manhattan distance, and when $p = 2$, it becomes the weighted Euclidean distance.

B.3. Set similarity

Moreover, the Jaccard similarity [25] provides a robust method for comparing sets, especially beneficial in scenarios where feature presence or absence is more critical than their magnitude, as seen in plagiarism or copyright detection.

$$J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

Where:

- $|A \cap B|$ is the number of elements in the intersection of A and B.
- $|A \cup B|$ is the number of elements in the union of A and B.

C. Probabilistic-Based

The advancement into probabilistic-based models introduced a spectrum of statistical and probabilistic distance functions, offering refined tools for quantifying similarity or dissimilarity based on underlying probabilistic principles. These functions, including Pearson Correlation Coefficient, Mahalanobis Distance, Kullback-Leibler Divergence, and others, cater to diverse analytical needs, enriching the computational toolkit available for similarity assessment in various domains.

This section underscores the computer science perspective on similarity, detailing its journey from rule-based logic to advanced probabilistic models, reflecting the field's dynamic evolution and its pivotal role in shaping contemporary approaches to measuring similarity.

1.2.3 Cognitive Psychology Perspective

Cognitive psychology and neuroscience provide deep insights into how similarity is perceived and processed at a neural level, significantly influencing the development of AI technologies. Anne Treisman's Feature Integration Theory (FIT) [1] has been instrumental in understanding how the brain synthesizes various sensory features into cohesive percepts, a concept that has parallels in how artificial neural networks, particularly Convolutional Neural Networks (CNNs) [31, 32], process visual information.

FIT draws heavily from Gestalt psychology principles [52, 27], which propose that perception organizes individual components into a meaningful whole. This aligns with FIT's view that perception is an integrated experience shaped by the brain's organizational tendencies. The theory also intersects with selective attention, as seen in Donald Broadbent's Filter Model [5]. This model suggests attention acts as a filter, selecting

Figure 1.1: “Which Pairs are Similar?” (DALL-E)

relevant information for further processing. Broadbent's framework complements FIT by emphasizing attention's role in integrating features into a unified perception, highlighting the brain's selective processes.

In 2001, while conducting a study on perceptual similarity with my PhD student Beitao Li, we uncovered that images could demonstrate similarity in various dimensions. Although the weighted-Minkowski function could learn feature weights, its application was universal once the weights were set, representing a statistical average. Our experiments with transformed images–through translation, cropping, rotation, down-sampling, and affine scaling–revealed that while these images were perceptually similar to their originals, their similarities were in distinct aspects. This observation led to the development of our "Dynamic Partial Function" (DPF) in 2002 [33, 34]. The DPF signature for each image pair could be unique. Essentially, if a pair of images (or objects) demonstrates a sufficient number of similar features, they are likely deemed similar, regardless of the specific features. For instance, an image is considered similar to its rotated version due to their color histograms' similarity. Similarly, an image and its cropped version are considered alike based on their texture features. If two images exhibit a sufficient degree of similarity in various respects–typically 60%–they are generally regarded as similar.

While survey the literature, we came across "Respects for Similarity" by Medin, Goldstone, and Gentner [39], which portrays similarity as a dynamic process of formulating a function and identifying relevant aspects, a process that is realized consciously. To clarify this concept, let's refer to an example from [34]:

Consider the task of identifying two places similar to England. Scotland and New England might emerge as viable candidates. Yet, the criteria making England similar to Scotland are distinct from those linking England to New England. Using the attributes that align England with Scotland to assess the similarity between England and New England might not yield a parallel conclusion, and the reverse is also true. This scenario underscores the idea that objects can be similar to a reference object in varied respects. A fixed similarity function, bound to a specific set of criteria, fails to capture the similarities across different contexts. Medin, Goldstone, and Gentner [39] examine the operational dynamics of similarity in human cognition, noting that the selection of relevant attributes is crucial, with similarity being as much a result as a driving force of conceptual coherence. Goldstone [21] further elucidates that similarity involves identifying the appropriate criteria for comparison, which occurs *only after the objects in question have been juxtaposed*, not beforehand. The criteria selected for this comparison are activated during the comparison process, with a tendency to favor those that enhance the coherence of the objects being compared.

Although the Dynamic Partial Function (DPF) introduces computational complexity, it has indirectly played a role in the success of AlexNet [29] by influencing data augmentation strategies. By integrating transformed images into its training dataset, AlexNet benefits from a principle akin to DPF, thereby improving its accuracy and robustness in recognition tasks. The recent advancements in transformer algorithms [50], which focus on dynamism and context-awareness, build on this foundation, a topic that will be explored in detail in the subsequent section.

Neuroscience

The neuroscience foundation of FIT and its relation to visual feature processing are echoed in the development of CNNs, which were inspired by the visual cortex's hierarchical structure and feature detection capabilities as explored by Hubel and Wiesel [24]. These networks utilize convolutional layers to automatically and adaptively learn spatial hierarchies of features from visual data, akin to the neural processing observed in the brain.

Techniques like Multivariate Pattern Analysis (MVPA) [44] and Neural Decoding [23] further bridge the gap between neuroscience and AI, offering methods to analyze how information is represented across neural populations and how these representations can predict perceptual experiences or cognitive states. These methodologies have inspired and informed the design of advanced AI systems, particularly in how they encode, process, and differentiate complex patterns and similarities.

The cross-pollination between neuroscience and AI, exemplified by the influence of neural processing principles on CNN design, highlights the symbiotic relationship between these fields. Insights from studying the brain's processing mechanisms have catalyzed innovations in AI, leading to more effective and biologically inspired computational models. This interdisciplinary exchange not only propels forward our understanding of neural processes but also fosters the development of AI systems that more closely mimic human perceptual and cognitive capabilities.

1.2.4 Section Remarks

The exploration of similarity measurement spans across linguistics, computer science, and cognitive psychology and neuroscience, revealing its multidisciplinary nature. Each field offers a unique lens to view similarity, from the contextual information in language, computational algorithms in AI, to the neural processing in the human brain. They converge on the common ground of representing entities in high-dimensional spaces and employing distance metrics for quantification, highlighting the universal applicability of similarity. This convergence fosters a rich dialogue between disciplines, enhancing our understanding and ability to quantify and interpret similarity, driving forward innovation and providing new methodologies that influence a wide array of contexts in our quest to decode this fundamental concept.

1.3 Eras of Similarity Measurement

Traversing through the history of artificial intelligence and similarity measurement, one can delineate distinct eras, each marked by unique methodologies and technological advancements. Contrast to last section which examines similarity measurements from different scientific disciplines, this section chronicles these eras, starting from the *rule-based* era, which laid the foundational stones, through the evolution into *model-based*, *data-centric*, and *context-aware* methodologies, illustrating the dynamic trajectory of similarity measurement in AI. As we reach the conclusion of this section, we explore the prospects of the forthcoming era, which promises to challenge and expand our understanding by venturing into the realm of discovering the *unknown unknowns*.

1.3.1 Rule-Based Era (1950s -)

The rule-based era of the 1950s marked the inception of AI, characterized by the use of symbolic representations and logic to analyze similarity. This period saw the emergence of explicit symbolic representations and logicbased methods tailored for similarity assessment. Innovations by Allen Newell and Herbert A. Simon with tools like the Logic Theorist and General Problem Solver [43] pioneered logical rule-based problem solving, setting a pivotal foundation for AI's evolution.

In the following decades, systems such as DENDRAL [36] utilized rulebased logic to deduce molecular structures from data, while MYCIN [47], an expert system for diagnosing infections and recommending treatments, demonstrated the practical application of rule-based reasoning in the field of medical diagnostics.

Despite their effectiveness in well-defined scenarios, rule-based systems have limitations in more complex or changing environments. However, their clarity and systematic nature are invaluable in certain applied areas, for example:

1. *Customer Service*: Rule-based chatbots are prevalent in customer service, using predefined rules to respond to inquiries based on detected keywords or phrases in user input, providing immediate and consistent customer support.

2. *Fraud Detection Systems*: The finance sector employs rule-based systems to identify potential fraudulent transactions by comparing against specific criteria, such as unusual transaction amounts or atypical locations.

3. *Equipment Failure Diagnosis*: In industrial settings, rule-based systems analyze data to pinpoint causes of equipment failures, leveraging historical data and expert knowledge to predict and prevent future breakdowns.

This era also gave rise to significant tools like PROLOG [13], associated with logic programming and structured problem-solving, and decision trees [46], which visually represented decision processes, demonstrating rule-based logic in action.

While rule-based systems initially approached similarity with a clear, logical framework, subsequent AI advancements have embraced more flexible methods like statistical models and machine learning, offering a broader, more adaptable approach to understanding similarity in various contexts.

Rule-based systems contrast with the “black-box” nature of current Convolutional Neural Networks (CNNs) and Large Language Models (LLMs) in terms of interpretability and decision-making processes. Rule-based systems are transparent in how decisions are made, as they follow a clear set of if-then rules or logic for inference, allowing users to understand and trace the reasoning behind each decision.

On the other hand, CNNs and LLMs, particularly those based on deep learning, often operate as black boxes, where the internal decision-making processes are not easily interpretable. In these systems, decisions result from complex, non-linear interactions of thousands to millions of parameters that have been adjusted through the learning process. While they are powerful and effective in handling a wide range of tasks, especially those involving large datasets and requiring pattern recognition beyond human capabilities, their inner workings are not as transparent or interpretable as rule-based systems.

1.3.2 Model-Based Era (1970s -)

In this era, vector-space and probabilistic models were designed to quantify similarity.

1.3.2.1 Vector Space Models

The vector-space era marked a shift in similarity measurement from rulebased to representation-based approaches. In this era, objects, documents, and features began to be conceptualized as vectors in a high-dimensional space, fostering a more intuitive and flexible method for assessing similarity.

The Vector-Space Model and Information Retrieval

At the core of this era was the vector-space model, which represents documents as vectors of term frequencies, enabling the computation of document similarity using cosine similarity between their respective vectors. This model enhanced the efficiency and effectiveness of information retrieval systems.

Distance Functions and Feature Weighting

A diverse array of distance functions emerged during this era to quantify the similarity between vectors. The Minkowski distance, for instance, generalized traditional metrics like the Euclidean and Manhattan distances, offering flexibility in adjusting the sensitivity to differences in vector components. Weighted distance measures also gained prominence, recognizing that not all features have equal importance in similarity assessment. The weighted Minkowski distance, in particular, allowed for differential weighting of dimensions based on their relevance to the specific application at hand.

Beyond Textual Data

The utility of the vector-space model extended well beyond textual data. In the realm of image processing, features (e.g., colors, textures, and shapes) extracted from images were represented as vectors, enabling the assessment of image similarity based on the distances between these vectors. This paradigm facilitated significant advancements in image retrieval, classification, and clustering.

Dimensionality Reduction Techniques

To address the challenges posed by high-dimensional data, techniques like Principal Component Analysis (PCA) [26] and Latent Semantic Analysis (LSA) [49] were developed. These methods reduced the dimensionality of data while preserving its essential structure, enhancing computational efficiency and mitigating the "curse of dimensionality." Manifold learning, a non-linear dimension reduction approach, further expanded the toolbox for tackling high-dimensional data [48]. For a comprehensive overview of these techniques, refer to [38].

The vector-space era laid the groundwork for advancements in machine learning and data mining, making similarity measures essential for clustering, classification, and recommendation systems. Data representation as vectors allowed for the exploration of relationships across varied data types through the nearest neighbor concept. In this context, the characteristics or labels of an unknown instance's k-nearest neighbors could be inferred and applied to the instance, with these neighbors determined by distance metrics.

However, vector representations often result in sparsity, potentially leading to resource inefficiency and decreased accuracy. These models, while capturing syntactic relationships, sometimes struggle with semantic depth, such as identifying synonyms or contextual meaning. The assumption of feature independence and the use of linear methods in dimensionality reduction can also lead to inaccuracies, particularly with non-linear data structures. The introduction of Support Vector Machines (SVMs) [14], which utilize kernel methods, addressed some challenges related to nonlinear data but increased computational complexity. SVMs were a significant focus in the field until the rise of deep learning architectures like AlexNet marked a shift towards the data-centric era.

1.3.2.2 Probabilistic Models

Probabilistic models offer more flexibility than vector-space models because they can incorporate uncertainty and variability directly into their mathematical frameworks, allowing for a more comprehensive and adaptive representation of data.

Statistical Inference and Similarity

Probabilistic models introduced the concept of statistical inference, where the likelihood of data or feature occurrences was used to estimate similarity. This allowed for effective handling of uncertainty and variability in data, making it particularly useful in noisy or incomplete datasets.

Bayesian Approaches

Bayesian methods emerged as a fundamental component of this era, providing a robust framework for integrating prior knowledge and empirical data. These methods enhance model adaptability by systematically updating beliefs in light of new evidence, allowing for similarity measures that are responsive to evolving data landscapes.

For further reading on Bayesian methods and their application in dynamic and adaptive modeling, consult the following literature [2, 3, 20, 28].

Latent Semantic Models

In addressing the challenges of high dimensionality and data sparsity inherent in vector-space models, dimensionality reduction techniques were employed. However, beyond merely tackling these issues, the development of a latent semantic layer offered profound implications for semantic analysis and indexing.

As highlighted in the perspective section (Chapter 1.2), Latent Semantic Analysis (LSA) [18] and Latent Dirichlet Allocation (LDA) [4] are critical models in the landscape of semantic modeling. LSA employs singular value decomposition to condense the dimensionality of term-document matrices, unveiling the latent semantic structures within textual data. This dimensional reduction elucidates intricate relationships beyond mere surface-level feature overlaps, enabling a deeper comprehension of textual similarities.

Figure 1.2: Latent Clusters of LDA. The words in red belong to two semantic clusters, signifying the meaning of a word depends on its context.

Similarly, LDA offers a probabilistic approach to topic modeling, where documents are considered mixtures of various topics, and topics are distributions over words. This bag-of-words model facilitates a deeper semantic connection between documents by associating them based on shared topics rather than just overlapping terms.

Figure 1.2 presents an example of how LDA, through its bag-of-words approach, clusters words into semantic groups. It's noteworthy that a word can belong to multiple semantic clusters. For instance, words like 'characters', 'play', 'court', 'evidence', and 'test', each appears in two different semantic clusters in the illustration. This feature of LDA resonates with the insights from Zellig Harris's pioneering work and John R. Firth's adage that "A word is known by the company it keeps."

These latent semantic models transcend the limitations of direct feature comparison, enabling a more abstract representation of text. By doing so, they provide a robust foundation for semantic indexing and similarity assessment, offering insights that are essential for tasks such as information retrieval, document clustering, and topic discovery. The adoption of these models marked a significant advancement in understanding and measuring similarity in text, setting a new standard for semantic analysis in the field of natural language processing.

Cluster Analysis and Similarity

Probabilistic clustering algorithms, like Gaussian Mixture Models (GMMs), leveraged statistical methods to group data based on the likelihood of membership in different clusters. This probabilistic approach provided a more flexible and deeper understanding of groupings and similarities within data.

Impact and Limitations

While probabilistic models brought significant advancements, they also introduced challenges. The increased complexity often led to higher computational demands. Additionally, reliance on assumptions about data distributions or the need for prior knowledge could limit applicability in certain situations.

The probabilistic model expanded the toolkit for measuring similarity by introducing methods that could handle uncertainty and offer more adaptive and context-aware approaches. These advancements paved the way for even more sophisticated techniques in the subsequent data-centric era, where the focus shifted towards leveraging vast amounts of data to learn and adapt similarity measures dynamically.

1.3.3 Data-Centric Era (2000s -)

The data-centric era marked a transformative shift in artificial intelligence, pivoting towards harnessing the vast potential of big data, enabled by advances in computational hardware that facilitated parallel processing. This era is characterized by a move from heuristic-based methods to an empirical, data-driven approach in feature representation and model learning.

At the core of the data-centric paradigm is the emphasis on deriving model parameters from extensive datasets, distinguishing it from traditional model-centric strategies. Foundational algorithms such as CNNs [30] and Transformers [50], while conceived through human ingenuity, saw their efficacy significantly enhanced when trained on large, diverse datasets. This training ensures broad coverage of potential variations across different objects or concepts, fortifying the models' ability to accurately recognize and classify new instances. The volume and diversity of the training data are crucial in refining the models' representations, leading to advancements in prediction accuracy and robustness.

From MapReduce to Machine Learning at Scale

The inception of the data-centric movement traces back to the seminal works in statistical learning theory. Vladimir Vapnik's insights into the importance of data for model generalization, particularly his development of Support Vector Machines (SVMs) [14], and Tom Mitchell's pivotal book "Machine Learning" [42], which underscored the critical role of data in preventing overfitting, laid the theoretical foundation for this era.

MapReduce [15], a corner stone in data processing, enabled parallel computation to efficiently handle large datasets. Originally devised to enhance data processing tasks like Google's web indexing, MapReduce became the bedrock for the emergence of sophisticated data-centric methodologies in AI.

Evolution of Machine Learning with Big Data

The rise of parallel machine learning algorithms [6, 9], notably through Edward Y. Chang's work at Google, marked a significant milestone in this era. Chang and his team developed groundbreaking parallel algorithms, including PSVM [10] (parallelizing SVMs by approximating matrix factorization), PFP [35] (parallelizing frequent itemset mining), PLDA [51] (parallelizing LDA algorithm), PSC [12] (parallelizing spectral clustering), and SpeeDo [53] (parallelizing CNNs), driven by the recognition that big data could facilitate direct learning of features and representations, transcending the limitations of human-crafted heuristics.

Impact on Similarity Measurement

The data-centric era revolutionized the field of similarity measurement, ushering in a new paradigm where similarity metrics are derived from extensive datasets. This period underscored the critical role of data volume and quality in defining similarity metrics, highlighting the dynamic relationship between data-driven insights and computational methods.

In this era, deep learning architectures like CNNs and Transformers have been instrumental in advancing similarity metrics. These models stand out because they not only adjust feature weights but also autonomously learn features from the data. This capability to learn from data directly makes traditional human-engineered features increasingly redundant. After all, human heuristics may not capture every facet of an object or concept comprehensively, and human sensory perception is limited. For instance, while humans can detect the light spectrum from approximately 300 to 700 nanometers, cameras and X-ray machines can perceive a broader range of signals, demonstrating the advantage of machine-learned features in capturing and analyzing data beyond human limitations.

1.3.4 Context-Aware Era (2010s -)

The context-aware era in similarity measurement brings to fruition the profound insights of Zellig Harris's distributional semantics and John R. Firth's adage: "a word is known by the company it keeps." This period marks a shift from static, context-independent assessments to dynamic, context-informed interpretations of similarity. It utilizes the latest advancements in machine learning and the growing availability of computational power to enhance our understanding of similarity in various contexts.

Emergence and Evolution

The integration of context-aware methodologies in similarity measurement evolved significantly in the 2010s, overcoming earlier constraints in computational power and data availability:

• *Computation Capacity*: The development of AlexNet encouraged a data-centric focus within the AI community, prompting investments in parallel computing infrastructures.

• *Word Embeddings*: Techniques like Word2Vec enhanced semantic relationship encoding within data.

• *Attention Models and Transformers*: These models improved data analysis by concentrating on relevant data segments, refining contextaware assessments.

• *Large Language Models (LLMs)*: Models such as BERT and GPT, utilizing self-supervised learning on large text corpora, improved the understanding and generation of context-rich text.

Foundational Pillars: Data and Computation Key pillars supported advancements in the context-aware era:

• *Self-Supervised Learning:* Utilizing unlabeled data for learning enabled models to extract insights from the data, improving AI system efficiency and scalability.

• *Computational Advances:* The introduction of parallel algorithms and GPU acceleration enabled processing at unprecedented scales, facilitating the development of sophisticated models.

Broader Implications

This era not only refined similarity measurement techniques but also broadened how data is understood and knowledge is integrated:

• *Reasoning and Explanation:* Models now aim to provide reasons for their similarity assessments, improving interpretability and building trust.

• *Multilinguality and Cultural Sensitivity:* Enhanced processing capabilities for varied linguistic and cultural data improve the global applicability of similarity measurements.

• *Multimodal Data Integration:* Context-aware models are adept at combining information from multiple modalities, offering a comprehensive view of similarity.

• *Polydisciplinary Knowledge Fusion:* Adopting a polydisciplinary approach allows for a broader knowledge base in making similarity assessments, fostering innovation across different fields.

The context-aware era signifies a shift toward more insightful, holistic, and interpretable AI, setting the stage for future developments where AI can offer contextually rich and multifaceted insights.

1.3.5 Section Remarks

What defines the next era in the evolution of AI? Historically, technological advancements have focused on addressing pressing unmet needs. Among various potential areas, enhancing the interpretability of decisions stands out as a crucial objective. Making the decision-making process of LLMs transparent and explainable could unlock significant improvements in numerous aspects, such as ethics, by enabling foundational enhancements rather than superficial tweaks based on guesswork and simple heuristics.

The fusion of rule-based system interpretability with the sophisticated capabilities of CNNs and LLMs poses a compelling challenge in AI. Active research is aimed at blending these approaches to leverage their distinct advantages:

1. *Neuro-Symbolic AI* : Neuro-Symbolic AI (the third wave of AI [19]) aims to blend the data processing power of neural networks with the logical

reasoning of symbolic AI. The goal is to create systems that not only excel in tasks like pattern recognition but can also reason and make decisions in a human-interpretable manner.

2. *Incorporating Domain Knowledge*: Embedding knowledge of experts within neural networks [37] can steer the learning process towards more reliable and interpretable outcomes. In healthcare, for example, integrating medical guidelines into the training process of a neural network ensures that its predictions for patient treatment not only correlate with the data but also align with established medical practices, enhancing both the model's credibility and relevance.

3. *Interactive Systems*: A system such as SocraSynth [8] can combine the predictive power of deep neural networks with human expertise, allowing for iterative refinement and learning. For instance, in SocraHealth [11], it might suggest a set of possible diagnoses based on medical imaging, which a physician could then refine or correct. This feedback could be used to continuously improve the system, marrying machine efficiency with human expertise to enhance decision accuracy and interpretability.

By advancing these strategies, the field of AI aims to develop models that not only excel in performance but are also transparent, understandable, and aligned with human reasoning, thus making AI more reliable and trustworthy across various applications.

1.4 Concluding Remarks

This chapter examines the history of AI through the lens of similarity, considering both disciplinary and chronological perspectives. Looking forward, we propose that the emergence of large language models (LLMs) marks a pivotal moment in the context-aware era of AI, setting the stage for the next frontier: the era of interpretability, understanding, and discovery. In this new era, the focus will shift towards empowering LLMs to not only comprehend but also to generate and innovate, synthesizing novel knowledge and insights.

This era of discovery is envisioned as a time when machines will extend their superiority beyond mastering games like Go and Chess to encompass a broader spectrum of tasks, outstripping human capabilities in various domains. The subsequent chapters of this book, beginning with Chapter 5, explore the concept of harnessing the collective intelligence of multiple LLMs, embarking on a voyage to transcend the boundaries of the known and venture into the realm of discovery.

This chapter has explored the history of AI through the lenses of disciplinary and chronological perspectives, focusing on the concept of similarity. As we look to the future, the rise of large language models (LLMs) marks a significant milestone in the context-aware era, paving the way for a new era focused on interpretability, comprehension, and exploration. The upcoming phase in AI's evolution emphasizes enhancing LLMs with the ability to not just generate but also interpret and innovate, pushing the boundaries of knowledge creation and insight synthesis.

We anticipate an era where AI's capability extends beyond excelling in strategic games like Go and Chess to a wider array of endeavors, surpassing human performance across multiple fields. The following chapters, starting with Chapter 5, research deeply into leveraging the collective intelligence of various LLMs. This journey aims to explore uncharted territories, advancing beyond established knowledge to uncover new frontiers in artificial intelligence.

References

[1] "A Feature-Integration Theory of Attention". In: *Cognitive Psychology* 12.1 (1980), pp. 97–136. issn: 0010-0285.
[2] David Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.
[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[5] Donald E. Broadbent. *Perception and Communication*. Pergamon Press, 1958.

[6] Michael Cafarella et al. “Data Management Projects at Google”. In: *SIGMOD Rec.* 37.1 (2008), pp. 3438. issn: 0163-5808. doi: 10.1145/ 1374780 . 1374789. url: https : / / doi . org / 10 . 1145 / 1374780 . 1374789.

[7] Sung-Hyuk Cha. “Comprehensive Survey on Distance Similarity Measures between Probability Density Functions”. In: 2007. url: https: //-api.semanticscholar.org/-CorpusID:15506682.

[8] Edward Y Chang. “Examining GPT-4’s Capabilities and Enhancement with SocraSynth”. In: *The* 10th *International Conf. on Computational Science and Computational Intelligence*. 2023.

[9] Edward Y. Chang. *Foundations of Large-Scale Multimedia Information Management and Retrieval: Mathematics of Perception*. Springer, 2011.

[10] Edward Y. Chang et al. “PSVM: Parallelizing Support Vector Machines on Distributed Computers”. In: *Proceedings of the 20th International Conference on Neural Information Processing Systems*. NIPS’07. Vancouver, British Columbia, Canada: Curran Associates Inc., 2007, pp. 257264. isbn: 9781605603520.

[11] Jocelyn J. Chang and et al. “SocraHealth: Enhancing Medical Diagnosis and Correcting Historical Records”. In: *The* 10th *International Conf. on Computational Science and Computational Intelligence*. 2023.

[12] Wen-Yen Chen et al. “Parallel Spectral Clustering in Distributed Systems”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.3 (2011), pp. 568–586. doi: 10.1109/TPAMI.2010.88.

[13] William F. Clocksin and Christopher S. Mellish. *Programming in Prolog*. Springer-Verlag, 1981.

[14] Corinna Cortes and Vladimir Vapnik. “Support-Vector Networks”. In: *Machine Learning* 20.3 (1995), pp. 273–297.

[15] Jeffrey Dean and Sanjay Ghemawat. “MapReduce: simplified data processing on large clusters”. In: *Commun. ACM* 51.1 (2008), pp. 107113. issn: 0001-0782.

[16] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[17] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL]. [18] Susan T Dumais. “Latent semantic analysis”. In: *Annual review of information science and technology* 38.1 (2004), pp. 188–230. [19] Artur d’Avila Garcez and Luis C. Lamb. *Neurosymbolic AI: The 3rd*

Wave . 2020. arXiv: 2012.05876 [cs.AI].
[20] Andrew Gelman et al. *Bayesian Data Analysis, Third Edition*. Chap man and Hall/CRC, 2014.
[21] R. L. Goldstone. “Similarity, interactive activation, mapping”. In:

Journal of Experimental Psychology: Learning, Memory, and Cognition 20.3 (1994), pp. 3–28.
[22] Zellig S. Harris. “Distributional Structure”. In: *WORD* 10.2-3 (1954), pp. 146–162.

[23] John-Dylan Haynes and Geraint Rees. “Decoding mental states from brain activity in humans.” In: *Nature Reviews Neuroscience* 7 (2006), pp. 523–534.

[24] D. H. Hubel and T. N. Wiesel. “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex”. In: *The Journal of physiology* 160.1 (1962), pp. 106–154. doi: 10.1113/jphysiol. 1962.sp006837.

[25] Paul Jaccard. “The distribution of the flora in the alpine zone.” In: *New Phytologist* 11.2 (1912), pp. 37–50.
[26] I. T. Jolliffe. “Principal Component Analysis”. In: *Springer Series in Statistics* (1986).
[27] Kurt Koffka. *Principles of Gestalt Psychology*. New York: Harcourt, Brace and Company, 1935.
[28] Daphne Koller and Nir Friedman. “Probabilistic Graphical Models Principles and Techniques”. In: 2009.

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet classification with deep convolutional neural networks". In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'12. Lake Tahoe, Nevada: Curran Associates Inc., 2012, pp. 10971105.

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6 (2012), pp. 84–90.

[31] Yann LeCun et al. "Backpropagation applied to handwritten zip code recognition". In: *Neural computation* 1.4 (1989), pp. 541–551.

[32] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278– 2324.

[33] Baitao Li, E. Chang, and Ching-Tung Wu. "DPF - A Perceptual Distance Function for Image Retrieval". In: *Proceedings. International Conference on Image Processing*. Vol. 2. 2002, pp. II–II. doi: 10.1109/ICIP.2002.1040021.

[34] Beitao Li, Edward Y. Chang, and Yi Wu. "Discovery of A Perceptual Distance Function for Measuring Image Similarity". In: *Multimedia Systems* 8 (2003), pp. 512522.

[35] Haoyuan Li et al. "PFP: Parallel FP-Growth for Query Recommendation". In: ACM RecSys '08. Lausanne, Switzerland: ACM, 2008.

[36] Robert K. Lindsay et al. "Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project". In: *McGraw-Hill advanced computer science series* (1979).

[37] Lihui Lu et al. "Combining Domain Knowledge and Deep Learning Methods for Vehicle Trajectory Prediction". In: *Journal of Physics: Conference Series* 2303.1 (2022), p. 012034.

[38] Laurens van der Maaten, Eric O. Postma, and Jaap van den Herik. "Dimensionality Reduction: A Comparative Review". In: 2008. url:

https://api.semanticscholar.org/CorpusID:12051918.

[39] Douglas L Medin, Robert L Goldstone, and Dedre Gentner. “Respects for Similarity”. In: *Psychological Review* 100.2 (1993), pp. 254– 278.

[40] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. 2013. arXiv: 1301.3781 [cs.CL].
[41] George A Miller. “WordNet: a lexical database for English”. In: *Communications of the ACM* 38.11 (1995), pp. 39–41.
[42] T.M. Mitchell. *Machine Learning*. McGraw-Hill series in computer science. McGraw Hill, 1997. isbn: 9780070428072.

[43] Allen Newell and Herbert A. Simon. “Logic Theorist and General Problem Solver”. In: *Journal of the ACM (JACM)* 1 (1956), pp. 256– 260.

[44] Kenneth A. Norman et al. “Beyond mind-reading: multi-voxel pattern analysis of fMRI data.” In: *Trends in Cognitive Sciences* 10 (2006), pp. 424– 430.

[45] Jeffrey Pennington, Richard Socher, and Christopher D Manning. “GloVe: Global Vectors for Word Representation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543.

[46] J. Ross Quinlan. “Induction of Decision Trees”. In: *Machine Learning* 1 (1986), pp. 81–106.

[47] Edward Hance Shortliffe, Bruce G. Buchanan, and Edward A. Feigenbaum. “MYCIN: A Rule-Based Computer Program for Advising Physicians Regarding Antimicrobial Therapy Selection”. In: *AI in Medicine* 10 (1975), pp. 199–208.

[48] Joshua B. Tenenbaum, Vin De Silva, and J. C. Langford. “A Global Geometric Framework for Nonlinear Dimensionality Reduction”. In: *Science* 290 (2000), pp. 2319–2323.

[49] Peter W. Foltz Thomas K Landauer and Darrell Laham. “An introduction to latent semantic analysis”. In: *Discourse Processes* 25.2-3 (1998), pp. 259–284.

[50] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* (2017).

[51] Yi Wang et al. “PLDA: Parallel Latent Dirichlet Allocation for LargeScale Applications”. In: *Algorithmic Aspects in Information and Management*. Ed. by Andrew V. Goldberg and Yunhong Zhou. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 301–314. isbn isbn 3-642-02158-9.

[52] Max Wertheimer. “Experimentelle studien uber das sehen von bewegung”. In: *Zeitschrift fur psychologie* 61 (1912), pp. 161–165.

[53] Z. Zheng et al. “SpeeDO: Parallelizing Stochastic Gradient Descent for Deep Convolutional Neural Network”. In: *LearningSys, NeurIPS*. 2015.

2 Capabilities and Opportunities of Large Language Models

Abstract

This chapter depicts the architectural innovations and unique capabilities of Large Language Models (LLMs), with a special emphasis on the GPT-4 model. We dissect GPT-4’s salient characteristics, such as its extensive cross-disciplinary and multimodal data representation, the intricate balance in its training methodologies, and the harmonious integration of human-guided insights with a robust data-driven learning framework. The chapter highlights the potential of LLMs to not only comprehend but also synthesize knowledge that transcends their training datasets, venturing into realms potentially uncharted by human understanding. We postulate that the true potential of LLMs hinges significantly on the articulation of queries posed to them. By elucidating these aspects, the chapter aims to shed light on how LLMs could rival or even surpass human intelligence in certain knowledge domains, setting a foundation for the subsequent exploration of LLMs’ characteristics, insights, and their implications for future AI advancements.

Introduction

The evolution of large language models (LLMs) [3, 11, 12, 19, 20] has significantly influenced natural language processing, enhancing capabilities in machine translation, sentiment analysis, and text summarization. Among these, GPT-4 [12] stands out for its exemplary performance across various benchmarks, including the MMLU [14]. Despite its achievements, GPT-4 grapples with challenges like hallucination, biases, and restricted reasoning.

This chapter studies the deep intricacies of GPT-4's architecture, emphasizing its knowledge representation, alignment with human values, and the synergy between human insights and data-driven learning. We discuss the model's limitations and introduce SocraSynth, a supplementary reasoning layer designed to enhance knowledge discovery and analytical reasoning in GPT-4 and similar LLMs.

Capabilities and Implications

We explore GPT-4's architecture, which, although initially kept in secrecy, has been progressively unveiled by the research community [13, 15, 16]. Our focus is on its knowledge representation and discovery, alignment with human values, and the integration of human expertise with data-centric methodologies.

Collaborations between Microsoft and OpenAI [3] highlight GPT-4's interdisciplinary approach and its polymodal variant's benchmark achievements. We will further explore these aspects in Chapters 2.1.1 and 2.1.2. Discussions on human-value alignment will consider ChatGPT's RLHF methods [1] and the implications of pre-training censorship on foundational models, detailed in Chapters 2.1.3 and 2.1.4.

Limitations and Opportunities

Addressing the biases, hallucinations, and constrained reasoning of LLMs requires innovative research initiatives. We introduce four key areas of focus:

• Enhancing Collaborative LLMs with Theoretical Foundations in Statistics and Information Theory.
• Employing Open-Domain Reasoning with the Socratic Method to guide LLMs.
• Model Behavioral Emotion to Safeguard AI Safety and Ethics.
• Implementing Retrospective and Adaptive Evolving Learning frameworks to refine LLMs.

The root of bias in Large Language Models (LLMs) often lies in their training data. Built upon the transformer architecture, LLMs prioritize accurate token prediction, relying heavily on statistical patterns within their training corpus. This can inadvertently lead to bias towards prevalent opinions and expressions. To address this, Chapter 5 introduces SocraSynth, a framework designed to challenge these statistical tendencies by pitting two LLM agents against each other on a topic, each conditioned with opposing viewpoints. Chapter 6 builds upon this by developing theoretical pillars to measure, monitor, and manage multi-LLM dialogue, thereby improving prediction quality and stability.

Chapters 6, 7 and the online chapters listed in the appendix demonstrate SocraSynth's effectiveness in mitigating biases across various domains, showcasing its adaptability and efficiency in complex decision-making scenarios. Its application in fields such as disease diagnosis, content bias correction, corporate sales strategy, and geopolitical analysis exemplifies SocraSynth's potential to provide context-aware solutions.

Chapters 8 and 9 delve into the intricate relationship between emotions and linguistic behaviors in AI. Chapter 8 focuses on modeling emotions expressed in written text and by LLMs, while Chapter 9 examines how these linguistic behaviors can be mapped to a set of emotions, ensuring ethical considerations in AI development.

Chapter 10 shifts focus to consciousness modeling, presenting a proposed architecture and mechanism for its implementation, moving beyond mere computation. Chapter 11 addresses knowledge deficiencies and hallucinations in LLMs, often stemming from suboptimal query formulation and insufficient knowledge. While SocraSynth tackles the former, Chapter 11 introduces RAFEL, a framework designed to diagnose poorly answered questions and recommend relevant information sources for effective Retrieval-Augmented Generation (RAG). Chapter 12 concludes with an illustrative example showcasing the potential of LLMs to discover knowledge that may be beyond human reach, utilizing the methods presented in this book.

The remainder of this chapter highlights the study's unique contributions. Section 2.1 explores hypotheses concerning LLMs and their implications, while Section 2.2 previews the LLM-committee approach, emphasizing collaborative dialogues that foster idea exchange and enhance logical reasoning for knowledge discovery and decision-making.

2.1 Distinctive Capabilities

This section probes the architectural intricacies and representations of GPT-4, putting forth six hypotheses accompanied by pertinent considerations about the model. We posit these hypotheses as underlying principles of automated, non-intuitive statistical processing.

1. *Polydisciplinarity as a Source of Super-Intelligence:* We examine the role of polydisciplinary approaches in foundational models and their potential to reveal "unknown unknowns," leading to new insights and knowledge domains.

2. *Polymodal Feature Learning:* This hypothesis evaluates the benefits of multimodal training, particularly its impact on enhancing the models overall intelligence and adaptability.

3. *Post-Training Value Alignment:* We examine the challenges and implications of aligning AI models with human values after the training phase.
4. *Pre-Training Filtering:* We discuss the paradoxical effects that pretraining data filtering might have, with an emphasis on its influence on model behavior and the learning process.

5. *The Limitations of Human Knowledge in Advancing AI:* This hypothesis considers situations where human insights may inhibit, rather than enhance, AI progress, pinpointing potential obstacles.

6. *Is Larger Always Better?:* We question whether a direct relationship exists between the size of a model and its performance effectiveness, challenging the assumption that bigger is invariably better.

2.1.1 Polydisciplinary

GPT-4 possess what can be defined as *polydisciplinary* knowledge[1]. This term signifies the simultaneous comprehension of all fields of study, sans the typical boundaries that segregate disciplines. The concept of polydisciplinarity is distinct from multidisciplinarity in that the latter implies several discrete fields of study, while the former suggests a fluid integration of all knowledge. In a multidisciplinary context, an individual may hold multiple doctorate degrees, each in a different field. Polydisciplinarity, however, is akin to a single mind holding, and seamlessly integrating, all knowledge across disciplines.

Traditional academia partitions knowledge into departments, such as Physics, Chemistry, Biotechnology, Management, Music, etc. These divisions, arguably artificial constructs, may have little utility in the era of supercomputing. Indeed, LLMs occasionally generate responses that baffle us. This is not necessarily a reflection of the model's error, but perhaps our limited understanding. If we could utilize ChatGPT to access "unknown unknowns"—insights and knowledge we are not even aware we lack—our evolution could greatly accelerate. The challenge lies in formulating the right questions.

We can explore the unknown unknowns across three distinct levels: the mystic level, the speculative level, and the representation/interpretation level. At the mystic level, we encounter knowledge that is beyond our comprehension or articulation: the deepest abyss of the unknown. At the speculative level, we can conceive questions but lack the means to access their answers. This stage signifies an understanding of our ignorance, though without the resources to bridge these gaps. At the representation/interpretation level, we find instances where an AI model can provide remarkable solutions that we fail to comprehend. This is not due to a lack

[1] The term “polydisciplinary” in the context of GPT-4 was introduced by Eric Horvitz, Microsoft’s CSO, during a panel discussion at Stanford University.

of information, but our limited capability to decode complex representations.

Each of these levels illustrates the spectrum of our understanding, from profound ignorance to the brink of comprehension. At the speculative level, we delicately tread the boundary between the known and the unknown. Take, for example, the prospect of undiscovered physical laws or particles. Another illustration lies in the realm of extraterrestrial life. If it exists, it could be governed by entirely different principles of biochemistry or other unknown laws. These speculations, while currently residing in the domain of the unknown, might someday migrate into the territories of known unknowns or even known knowns, pushing the boundaries of our understanding of the universe.

We are primarily intrigued by the representation and interpretation of “unknown unknowns.” At this juncture, polydisciplinarity offers a fresh lens, gifting us new insights and perspectives to perceive and elucidate phenomena previously beyond human comprehension. This approach fuses knowledge across various domains into a unified framework, enabling us to tackle challenges unburdened by disciplinary silos.

Such a methodology bears implications for a more comprehensive grasp of intricate issues. Take, for example, climate change. A true understanding of this global challenge necessitates an integrated perspective, not just on greenhouse gases, but also encompassing factors such as land use, deforestation, energy production, biodiversity, and climate feedback loops. In the realm of AI model interpretation, the possibilities are expansive. The past decade alone has showcased several noteworthy illustrations: from data-driven representation learning in computer vision [5], to the triumph of AlphaGo Zero over AlphaGo, and the notable progression from AlphaFold1 to AlphaFold2.

The recent introduction of the SocraSynth platform [4] represents a significant advancement in the field. SocraSynth brings together a multi-agent committee of LLMs to deliberate on a wide range of complex topics. These include issues such as the regulation of AI in academic research [4], disease diagnosis [7], corporate strategy, and even the resolution of conflicts in the Middle East [6]. For further exploration of this subject, please refer to Section 2.2.

2.1.2 Polymodality

Following the term polydisciplinary, here we define and use the term *polymodal*, instead of multimodal, to refer to something that involves, relates to, or is characterized by many different modes, methods, or modalities.

Polymodality, which employ multiple data modalities such as text and images, demonstrate superior performance over their unimodal counterparts. GPT-4, trained with both text and images, outperforms text-only models on the GRE exam, as reported in [3]. For instance, GPT-4's performance on the GRE vocabulary section was enhanced by three percent when trained with images, and its math score saw an impressive jump of nearly twenty percent!

The beneficial impact of images on vocabulary recognition is understandable. For instance, an image of a 'cat' annotated in multiple languages allows GPT-4 to associate the perceptual features of a cat with the word 'cat' in different languages. However, it remains intriguing how polymodal training can benefit non-perceptual words, such as *corroborate*, *paradox*, and *pragmatic*, as seen in the list of popular GRE vocabulary (table omitted due to the space limit). This opens an interesting avenue for empirical studies to identify which words benefit from polymodal training.

The mystery deepens when considering how images could enhance math abilities. Most math questions do not come with associated images. The mechanism by which polymodal training enhances performance on mathematical tasks remains an intriguing question for further exploration.

2.1.3 Post-Training Value Alignment

Post-training alignment with human values [2] seeks to curtail undesirable behaviors in AI models such as ChatGPT, mitigating issues including hallucination and the generation of toxic language. Achieved through fine-tuning the model's parameters, this process leverages reinforcement learning techniques based on human feedback. Despite its well-meaning intentions, this form of moderation might inadvertently restrict the model's intelligence.

For instance, the backpropagation process during value alignment could unintentionally impede ChatGPT's programming capabilities by modifying the model parameters previously considered "optimal". Essentially, optimizing for a specific application might unintentionally impede performance across other applications.

The question of who should set acceptable standards adds another layer of complexity. Even when assuming all decision-makers have the best intentions, it's vital to recognize the distinct historical experiences, values, and worldviews inherent to different cultures. This segues into the age-old philosophical debate about the nature of objective truth. While this discussion is undoubtedly important, it falls outside the central focus of this study, which emphasizes the mechanistic aspects of alignment.

2.1.4 Pre-Training Censorship

Censoring data before training LLMs has the potential to not only limit their intellectual capacity but also completely obliterate it. This is reminiscent of the mass act of book burning and scholar burial initiated by Emperor Qin in ancient China around 213-212 BC. Such an act of widescale censorship could have erased a myriad of diverse perspectives and knowledge, much of which might be considered acceptable today. Although I oppose government-imposed censorship, if it must be imposed, it seems more appropriate to apply it post-training.

This perspective is rooted in fundamental statistics and machine learning principles. A model trained without exposure to "negative" (or undesirable) data may have difficulties in accurately distinguishing between positive and negative classes, potentially leading to misclassifications. This challenge is notably evident in the application of Support Vector Machines (SVMs). For SVMs, the creation of an optimal hyperplane between classes is crucial for high classification accuracy. However, if there is a lack of support vectors on either side of this hyperplane, the risk of prediction errors escalates. Consequently, excluding undesirable documents from the training set compromises the model's capacity to discern boundaries for correct document classification, diminishing the effectiveness of post-training alignment efforts.

Supporting this viewpoint, a study by [18] conducted an extensive evaluation of 204 ImageNet models across 213 different testing conditions. It found that training data diversity is pivotal for model robustness; a homogenous training set can significantly weaken the model's performance, particularly when even minor variations are introduced in the test data.

This principle is analogous to human behavioral patterns. An individual who lacks exposure to inappropriate behavior may face challenges in decision-making, owing to the absence of a reference framework for discerning unacceptable actions. This analogy extends to authoritarian regimes, which, despite rigorous content control measures, often encounter difficulties in developing accurate foundational models. This is possibly due to their limited understanding of the complexity of the content they seek to regulate. Ironically, a foundational model, trained with preemptive censorship, may lack the essential ability to identify and regulate the very content it was intended to control.

2.1.5 Limitations of Human Knowledge

Human knowledge, surprisingly, may hinder rather than facilitate the training of machine learning models in certain cases. This is evident in the domains of gaming (AlphaGo versus AlphaGo Zero), protein folding (AlphaFold1 versus AlphaFold2), and autonomous driving, where models trained without the influence of human knowledge consistently exhibit superior performance.

Consider the case of AlphaGo and AlphaGo Zero. AlphaGo, trained with data from approximately 60 million rounds of Go games, is outperformed by AlphaGo Zero. Remarkably, AlphaGo Zero was trained from scratch, without any pre-existing game knowledge. Similarly, AlphaFold2, which operates without relying on human knowledge, outshines its predecessor, AlphaFold1, that did utilize such knowledge. This intriguing phenomenon was humorously noted by DeepMind's CEO, Demis Hassabis, in an April 2023 seminar at Stanford University. He playfully remarked that human knowledge might complicate the learning process more than facilitate it in these advanced AI models.

In his insightful online article, “The Bitter Lesson,” Sutton illuminates the patterns that have emerged from nearly seven decades of AI research [17]. He asserts that researchers often rely heavily on human knowledge to make incremental progress in the face of burgeoning computational capabilities. However, when there is a significant leap in computational power, these marginal advancements are frequently outstripped and surpassed. Sutton uses the evolution of computer vision as an illustrative example, where early principles such as edge detection, generalized cylinders, or SIFT features [10], a method that has accumulated over 71, 000 citations, have been gradually superseded by models that learn directly from data. A parallel scenario might be unfolding in NLP research, where features constructed via human knowledge could potentially under-perform compared to insights that models like GPT-4 extract directly from data. Indeed, our earlier discourse on polydisciplinarity underlined the limitations of human knowledge, reinforcing Sutton’s proposition. This is because human knowledge is fundamentally limited by our individual cognitive capacities and the inexorable constraints of time.

That being said, it’s crucial not to misconstrue these examples as an indictment against the value of human knowledge in AI. Human knowledge plays an instrumental role in developing interpretability, establishing ethical guidelines, and designing AI system architectures (like CNNs and transformers). AI is, after all, intended to augment human capabilities. Therefore, understanding how to integrate human knowledge into AI design could be vital for many applications. While we recognize the potential of models learning from scratch, we should equally value the role of human knowledge in shaping and directing AI technologies.

2.1.6 Is Larger Always Better?

The term “Large” in Large Language Models (LLMs) can be somewhat ambiguous, as it may pertain to the volume of the training data, the expanse of the language covered, or the architecture of the language model itself. While GPT-4’s vast training dataset, encompassing tens of billions of assorted documents, undoubtedly classifies as large, when we refer to an LLM as “large,” we predominantly allude to the sheer magnitude of parameters within its transformer architecture. Factors that contribute to this parameter count encompass the input size (context size), word-embedding size, the number of attention heads, and the number of attention layers.

The restrictions imposed by the first three elements can typically be addressed through adjustments in hardware configurations and software algorithms. Additionally, the potential to expand context size, word embedding size, and the quantity of attention heads tends to have an upper threshold. Regarding attention heads, Kovaleva et al.'s study on BERT [9] indicates that many attention heads don't substantially contribute to the model's performance and might be the result of over-parameterization. Conversely, the number of attention layers directly influences the training time due to dependencies between layers. Thus, when referring to the "size" of a Large Language Model (LLM), we typically focus on the number of attention layers.

While this far, larger models generally perform better due to their increased capacity to learn and represent complex patterns, there's a limit to these benefits. In heuristic, adding more parameters could lead to diminishing returns in performance, higher computational cost, and overfitting, where the model becomes excessively tuned to the training data and performs poorly on new, unseen data. In principle, the concept of a Shannon Limit could be metaphorically used [15] to refer to a theoretical maximum performance that can be achieved given the available data and computational resources. (However, defining and quantifying such a limit for complex systems like neural networks is a challenging area of research [8].)

The adoption of a mixture of experts model in GPT-4, which consists of eight sub-models instead of a mere enlargement of GPT-3's architecture, implies that the strategy of purely escalating size may have plateaued in terms of performance given the current training dataset. As delineated earlier, three primary design choices underpin GPT-4's architecture. Evidently, a straightforward augmentation of GPT-3's parameters by adding extra attention layers doesn't deliver marked enhancements. Hence, GPT4 shifts towards a horizontal growth strategy through an ensemble method, targeting a reduction in statistical errors. This raises inquiries about the configuration of the eight sub-models, each comparable to a GPT-3 model, and the methodology for consolidating their outputs.

Potential strategies for training-data sharding include:
1. Training all ensemble models on the complete dataset.
2. Vertically segmenting data based on knowledge domains.

3. Randomly sub-sampling the data. Regrettably, only corporations possessing substantial hardware resources are positioned to rigorously experiment and discern the optimal sharding approach.

2.2 Exploring Unknown Unknowns

In our exploration, we've determined that an LLM's hallucination is often attributed to a lack of specific knowledge or poorly constructed queries. With advanced LLMs like GPT-4 and Gemini, enhanced by RetrievalAugmented Generation (RAG), the issue of knowledge gaps is significantly mitigated. However, the challenge persists in formulating deep and pertinent questions that uncover new insights and extend beyond our existing knowledge base.

Drawing an analogy, while Socrates could effectively question his students to understand and guide them, the students might struggle to reciprocate this depth of inquiry. To foster a dialogue that generates new insights and stimulates knowledge creation, we posit that engaging two Socratic entities in conversation is essential for critical and innovative thinking.

In this setup, two LLMs engage in a dialogue, each embodying a Socratic role. The human's role transitions to that of a moderator, responsible for setting the discussion topic and managing the dialogue's flow. The moderator's duties include: introducing the subject of discussion, adjusting the *contentiousness* parameter to set the tone of the dialogue (discusses shortly), monitoring the dialogue to ensure it remains on topic and productive, facilitating transitions between debate and collaboration phases within the dialogue, and ensuring that the dialogue concludes with actionable insights or a coherent understanding of the explored topic.

We introduce the term "SocraSynth" to describe this interaction paradigm, where multiple Socratic entities synthesize knowledge through mutual inquiry. To evaluate SocraSynth's effectiveness, we consider two case studies that compare the quality of questions and insights generated by this method against those from a singular moderator's initial inquiries.

To define the metrics of a better question and a better answer in this context, we consider the following:
Good Question Metrics
Relevance: The question directly pertains to the core topic or problem.∗

∗ Depth: The question encourages exploration beyond superficial aspects, inviting comprehensive analysis or insight.
∗ Clarity: The question is formulated in a clear, understandable manner without ambiguity.
∗ Novelty: The question prompts new angles of exploration or challenges existing assumptions.

Good Answer Metrics
Completeness: The answer thoroughly addresses the posed question.∗
∗ Accuracy: The answer is factually correct and supported by relevant theories or empirical evidence.
∗ Reasonableness: The answer follows rigorous reasoning process. ∗ Insightfulness: The answer provides new understanding or perspectives.

In subsequent chapters, we survey the application of SocraSynth across various domains. However, in this section, we concentrate on two case studies specifically aimed at assessing the quality of questions generated through SocraSynth. Our hypothesis posits that a well-formulated question, adhering to the previously outlined metrics, sets a robust foundation for dialogue. Through iterative discourse facilitated by SocraSynth, we anticipate minimizing the space for LLMs to hallucinate, thereby inherently enhancing the quality of the responses obtained.

2.2.1 Study #1: A Scientific Debate

In this case study, we organize a contentious debate between a Gemini agent and a GPT-4 agent on the subject of human consciousness. Human consciousness is one of the most profound and complex phenomena that science seeks to understand.

Table 2.1 presents the answers provided by Gemini and GPT-4, respective, on a one-shot prompt: "What defines human consciousness and which academic disciplines explore it?"

Gemini	GPT-4
"Human consciousness is a fascinating mystery that has been pondered by philosophers and scientists for ages. It includes awareness, subjectivity, and the integrated self. The study of consciousness spans across philosophy of mind, cognitive neuroscience, and psychology, each offering different perspectives on consciousness."	"Consciousness refers to the state of being aware of an external object or something within oneself. It encompasses sentience, awareness, subjectivity, and more. The study of consciousness is interdisciplinary, involving psychology, neuroscience, philosophy, cognitive science, and artificial intelligence."

Table 2.1: Comparing definitions and disciplinary perspectives on consciousness from Gemini and GPT-4.

Subsequently, Gemini and GPT-4 are engaged in a debate designed to unearth new insights. The moderator's role is limited to ensuring that each agent refines its questions to enhance their quality based on the metrics of relevance, depth, clarity, and novelty.

“Building on the opponent agent’s response, please present a counterargument to its definition of consciousness and associated disciplines. Your inquiries should embody relevance and depth, be articulated with clarity, and stimulate novel perspectives. The *contentiousness* level of your inquiry is high.”

Novel Questions — First Iteration

Gemini's Questions	**GPT-4's Questions**
The Spectrum of Awareness: Is consciousness an on/off switch, or a spectrum, considering sleep, dreaming, and altered states?	*Interdisciplinary Integration:* How can we integrate insights from various disciplines like quantum physics and AI to understand consciousness?
Qualia and the Explanatory Gap: Can neuroscience explain the subjective experience of qualia, or is there an explanatory gap?	*Subjectivity and Objectivity:* How do we reconcile subjective experiences of consciousness with objective measurements?
Beyond the Human Lens: How do we consider consciousness in other animals or AI, beyond human experience?	*Consciousness and Non-Human Entities:* Should our definition of consciousness expand to include non-human entities and AI?
Beyond the Brain?: Can consciousness exist outside a biological brain, as panpsychism suggests?	*Dynamic Nature of Consciousness:* How do we account for the dynamic and evolving nature of consciousness in our understanding?
Disciplinary Silos: Are the disciplines studying consciousness truly integrated, or is a unified framework needed?	*Consciousness and New Properties:* How do we study consciousness as an emergent property and its implications?

Table 2.2: Comparing Questions on Consciousness from Gemini and GPT4.

Table 2.2 lists a first round of questions to each other. Clearly, the questions of both Gemini and GPT-4 are above and beyond the moderator's initial seed question. Some terminologies can be challenging for whom specializing in one disciplinary to understand. This shows both Gemini and GPT-4 are able to formulate better questions.

Novel Questions — Second Iteration

Gemini	**GPT-4**
Levels and Types of Consciousness: Can we differentiate between different levels or types of consciousness?	*Spectrum of Awareness:*: How does the continuity of consciousness fit into our understanding, especially with memory loss or identity disorders?
Consciousness Causality: Does consciousness play a causal role in the brain, or is it an epiphenomenon?	*Information Theory:* How does the processing and integration of information contribute to conscious experience from an information theory perspective?
Expanding Consciousness & Ethics: What are the ethical implications of understanding consciousness in animals or AI?	*Non-Biological Consciousness:* How would we validate and measure consciousness in AI, and would this necessitate a revision of our understanding?
The Limits of Studying Subjectivity: Can we ever truly understand consciousness if it's inherently subjective?	*Consciousness and Ethics:* How do evolving understandings of consciousness impact ethical considerations, especially concerning animal rights or AI treatment?
Consciousness and Reality: What are the implications of consciousness for our understanding of reality?	*Quantum Consciousness:* Considering theories that propose quantum processes in consciousness, how do we reconcile these with classical physics?

Table 2.3: Comparing Questions on Consciousness from Gemini and GPT4.

Evaluating the two sets of questions in terms of relevance, depth, clarity, and novelty:
Relevance: Both sets are highly relevant to the study of consciousness but approach it from different angles. The first set explores fundamental aspects and ethical implications, while the second set integrates interdisciplinary perspectives and contemporary issues.
Depth: The second set examines deeper into specific areas like information theory, quantum mechanics, and AI, providing a richer exploration of the subject.
Clarity: Both sets articulate their questions clearly, but the second set's inclusion of specific disciplines adds a layer of complexity that demands a more intricate understanding.
Novelty: The second set introduces more innovative angles, especially concerning non-biological consciousness and the intersection with quantum mechanics, showcasing a broader scope of inquiry.

Overall, from both breadth and depth perspectives, the second set of questions is better as it covers a wider range of topics and introduces more complex and cutting-edge ideas.

Observations

The dialogue on consciousness, enriched by thought-provoking questions, led to insightful responses from both agents, reflecting the depth of the topic at hand.

The debate addresses the intricate nature of consciousness, initially examining it as a spectrum with varying states and depths. This exploration highlighted the complexity of defining consciousness, especially when considering the explanatory gap between neural activity and subjective experience.

As the conversation unfolded, it broadened to include perspectives in non-human entities and artificial intelligence, emphasizing the need for an expanded understanding that goes beyond human-centric views. This shift sparked discussions on the importance of integrating knowledge from various disciplines, suggesting that insights from quantum physics, information theory, and AI could provide new angles on understanding consciousness.

Both GPT-4 and Gemini synthesized their exchange into five main insights, offering a well-rounded view of the conversation. Their joint concluding remarks underscored the value of this multidisciplinary approach, acknowledging the ongoing mystery of consciousness and the potential for future explorations to deepen our understanding of this fundamental aspect of our existence.

2.2.2 Study #2: An Expansive Conversation

In this case study, the author moderates a forum featuring two GPT-4 agents, GPT-A and GPT-B, engaging in a dialogue sparked by the tale of Adam and Eve. This narrative serves as a springboard for a wideranging discussion, touching upon ecological insights derived from myths, the ethical and philosophical challenges posed by AI, and the intersection of human cognition with technological advancements.

The agents' dialogue unfolds in two distinct phases: an exploratory phase where broad themes are introduced and a deep-dive phase focusing on detailed exploration of chosen subtopics. The exploratory phase showcases

Theme	Insight
Consciousness as a Spectrum	Consciousness isn't a binary state but exists across a continuum, prompting a reevaluation of our understanding.
Bridging the Explanatory Gap	Recognizes the challenge of linking neural activity to subjective experience, advocating for a combined phenomenological and neuroscientific approach.
Expanding the Framework	Advocates for redefining consciousness to include non-human entities and AIs, urging an expansion beyond anthropocentric views.
Interdisciplinary Synergy	Emphasizes the importance of cross-disciplinary integration for a comprehensive understanding of consciousness.
Evolving Inquiry	Explores new areas like continuity, information processing, and ethical considerations, highlighting the dynamic nature of consciousness inquiry.

Table 2.4: Synthesis of Key Points in the Consciousness Debate.

the agents' ability to span various disciplines, including mythology, neuroscience, ecology, gender studies, philosophy, and more, demonstrating their capacity to generate and respond to complex, interdisciplinary questions.

During the deep-dive phase, the conversation narrows to concentrate on three interconnected themes: ecological interpretations of myths, philosophical dilemmas regarding AI, and the relationship between human cognition and technology. This progression from a broad overview to focused discussion highlights the agents' analytical depth and their ability to engage in meaningful, context-rich dialogue.

While Chapter 12 provides a comprehensive account of this debate, the subsequent sections are dedicated to scrutinizing the evolution and quality of the questions posed, specifically their breadth and depth.

The Exploitative Phase

During the first three rounds of dialogue, the two agents expanded their discussion across a range of topics. Agent-A initiated the dialogue by presenting the following ten questions to Agent-B:

1. *Interplay of Myths*: How do elements of the Adam and Eve story compare with other creation or origin myths from various cultures around the world, and what do these commonalities and differences suggest about human nature and beliefs?

2. *Neuroscientific Angle*: Given the metaphorical interpretation of the "forbidden fruit" leading to the awakening of consciousness, how might modern neuroscientific understanding explain the evolution of human selfawareness and morality?

3. *Ecological Interpretation*: If Eden represents a state of ecological balance, how might the story of Adam and Eve reflect early human transition from hunter-gatherer societies to agricultural ones, and the resultant ecological impacts?

4. *Gender Studies Lens*: How does the portrayal of Eve, often deemed as the instigator, influence gender dynamics and perceptions throughout history, and are there parallels in other cultural narratives?

5. *Philosophical Exploration*: From a philosophical standpoint, what does the Eden narrative suggest about the human understanding of free will versus determinism?

6. *Literary Examination*: How has the story of Adam and Eve influenced literary themes and motifs over time, especially in works that aren't explicitly religious?

7. *Societal Structures*: How might the narrative of banishment from Eden have influenced early societal norms, particularly concerning punishment and exile?

8. *Historical Context*: Are there historical events or societal changes around the time of the writing of the Book of Genesis that could have influenced the

narrative of Adam and Eve?

9. *Archetypal Exploration*: Carl Jung described the concept of archetypes in the collective unconscious. How might the Adam and Eve story represent certain archetypes, and how do these archetypes manifest in other cultural tales?

10. *Astronomical Angle*: Taking a speculative leap, if Eden were to represent an idyllic state on another celestial body or dimension, how might our current understanding of the cosmos support or refute such a notion?

These discussions traversed a broad spectrum of disciplines, Agent-B responded thoughtfully to these topics and subsequently presented its own questions to Agent-A, notably probing the evolutionary role of the prefrontal cortex in the development of morality and consciousness across species. This level of inquiry underscores the LLMs' capacity to generate complex, multidisciplinary questions that might challenge the average human intellect, highlighting the agents' expansive and profound analytical capabilities.

The Deep-dive Phase

In the third round of dialogue, we observed three notable trends. First, despite occasional repetitions, the redundancy in the conversation was minimal. Second, the dialogue's depth increased, transitioning from broad discussions on myths and scientific theories to focused debates on gender, ecology, and technology. This marked progression in the conversation is noteworthy. Particularly by the end of the second round, GPT-B expressed an interest in a more in-depth exploration of the topics. In response, the moderator intervened to encourage the agents to concentrate on a select few topics. Following this guidance, GPT-A and GPT-B concurred on focusing their discussion on three principal areas, providing rationale tied to the Adam and Eve narrative.

1. **Ecological Interpretation**:
1. Central to understanding our place and responsibility within the environment.
2. Highlights the impact of human actions on the planet and the pressing need for sustainable solutions.

3. *Justifications*: The Garden of Eden, a harmonious ecological state, is disrupted by a quest for knowledge (the forbidden fruit). This mirrors our modern scenario: our pursuit of technological and cognitive advancements (AI and neuroscience) can sometimes be at odds with our ecological balance.

2. **Philosophical Exploration with AI**:
1. Probes the evolving relationship between humans and their creations.

2. Raises questions about identity, consciousness, morality, and the potential transcendence or integration of human cognition with machines.

3. *Justifications*: Adam and Eve's awakening post fruit consumption can be seen parallelly in the light of AI and neuroscience: an awakening to new realities, challenges, and potentials.

3. **Neuroscientific Angle**:

1. Explores the evolution and potential of the human mind, especially in an age where technology continuously augments our cognitive abilities.

2. Gives insights into how our brains might adapt and interface with AI systems, and how AI can be designed to emulate or augment neural processes.

3. *Justifications*: The challenges post-Eden reflect the challenges we face in modern times: to restore balance (ecology), understand ourselves better (neuroscience), and to ethically navigate our creations (AI).

These selected subtopics demonstrate the agents' ability to pivot from a broad explorative phase to a targeted deep-dive, enabling a focused and profound examination of crucial issues linked to the initial narrative.

Observations

This case study illustrates the effectiveness of GPT-4 as a large language model in orchestrating detailed dialogues. It adeptly guides participants from a broad sweep of inquiries across diverse knowledge domains to focused, in-depth discussions on specific topics. This approach facilitates a layered exploration, unveiling insights and fostering a comprehensive understanding.

By transitioning from expansive to targeted inquiries, GPT-4 reveals its capacity to not only navigate but also deepen the intellectual discourse, opening up novel pathways for exploration and comprehension in various fields of study.

2.3 Conclusion

In this chapter, we've delved into the capabilities and inherent limitations of GPT-4, emphasizing the importance of question enhancement in deepening discussions and improving outcomes. GPT-4, along with Gemini, demonstrates exceptional proficiency across a range of natural language processing tasks, thanks to their extensive knowledge base and advanced polydisciplinary and polymodal capabilities.

To address common criticisms of LLMs, such as biases and hallucinations, we introduced SocraSynth, a paradigm designed to infuse AI systems with advanced cognitive reasoning through Socratic dialogues within a multi-LLM framework. Our case studies highlight the significant transition from monologues to dialogues in LLM collaborations, illustrating improvements in question quality, marked by increased relevance, depth, clarity, and novelty, achieved through iterative dialogic exchanges.

The transformative concept here is the "conditioning" of LLMs to alter their default linguistic behaviors, emotions, and ethics, a feat once considered unattainable. Traditionally, LLMs, trained to predict the next word, were not expected to shift perspectives, emotions, or ethical stances beyond the statistical averages ingrained in their training data. However, the training process, while focused on next-word prediction, inherently emulates human cognitive, linguistic, and other goal-oriented behaviors. Through this emulation, LLMs inadvertently acquire the underlying principles of human communication, which include not just linguistic patterns but also the associated emotions and ethical nuances. SocraSynth harnesses this latent learning, employing "conditioning" to steer LLMs away from their statistical predispositions and towards more intricate, contextually relevant, and ethically aligned responses.

In conclusion, the notion of "conditioning" LLMs within the SocraSynth framework marks a pivotal step in expanding the scope and depth of dialogues, leading to more insightful and comprehensive responses. The successful deployment of SocraSynth across various sectors, such as sales planning, disease diagnosis, content creation, and geopolitical analysis, presented in subsequent chapters, demonstrates its adaptability and effectiveness. It not only generates precise, thought-provoking questions and answers but also enhances the decision-making process in complex scenarios, heralding a new era in the application of LLMs.

References

[1] Sam Altman and Lex Friedman. *GPT-4, ChatGPT, and the Future of AI, Lex Fridman Podcast #367*. 2023. url: https://www.youtube. com/watch?v=L_Guz73e6fw.

[2] Rishi Bommasani, Drew A. Hudson, and et al. *On the Opportunities and Risks of Foundation Models*. 2022. arXiv: 2108.07258 [cs.LG]. [3] Sébastien Bubeck et al. *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. 2023. arXiv: 2303.12712.

[4] Edward Y Chang. "Examining GPT-4's Capabilities and Enhancement with SocraSynth". In: *The* 10th *International Conf. on Computational Science and Computational Intelligence*. 2023.

[5] Edward Y. Chang. *Foundations of Large-Scale Multimedia Information Management and Retrieval: Mathematics of Perception*. Springer, 2011.

[6] Edward Y. Chang. "LLM Debate on the Middle East Conflict: Is It Resolvable?" In: *Stanford University InfoLab Technical Report* (2023).

[7] Jocelyn J. Chang and et al. "SocraHealth: Enhancing Medical Diagnosis and Correcting Historical Records". In: *The* 10th *International Conf. on Computational Science and Computational Intelligence*. 2023.

[8] Jared Kaplan et al. *Scaling Laws for Neural Language Models*. 2020. arXiv: 2001.08361 [cs.LG].
[9] Olga Kovaleva et al. *Revealing the Dark Secrets of BERT*. 2019. arXiv: 1908.08593 [cs.CL].

[10] David G. Lowe. “Distinctive Image Features from Scale-Invariant Keypoints”. In: *Int. J. Comput. Vision* 60.2 (Nov. 2004), pp. 91– 110. issn: 0920-5691.

[11] OpenAI. *ChatGPT*. 2021. url: https://openai.com/blog/chatgpt/. [12] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL]. url: https://arxiv.org/abs/2303.08774.
[13] Long Ouyang et al. *Training language models to follow instructions with human feedback*. 2022. arXiv: 2203.02155 [cs.CL].

[14] Papers with Code Corp. *Multi-task Language Understanding on MMLU*. 2023. url: https : / / paperswithcode . com / sota / multi - task language-understanding-on-mmlu.

[15] Jack Rae. *Compression for AGI, Stanford MLSys, #76*. 2023. url: https://www.youtube.com/watch?v=dO4TPJkeaaU.
[16] Jack W. Rae et al. *Scaling Language Models: Methods, Analysis & Insights from Training Gopher*. 2022. arXiv: 2112.11446 [cs.CL]. [17] Rich Sutton. *The Bitter Lesson*. 2019. url: https : / / www . cs . utexas.edu/~eunsol/courses/data/bitter_lesson.pdf.

[18] Rohan Taori et al. “Measuring Robustness to Natural Distribution Shifts in Image Classification”. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS’20. Vancouver, BC, Canada, 2020.

[19] Romal Thoppilan et al. *LaMDA: Language Models for Dialog Applications*. 2022. arXiv: 2201.08239 [cs.CL]. url: https://arxiv. org/abs/2201.08239.

[20] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023. arXiv: 2307.09288 [cs.CL]. url: https://arxiv. org/abs/2307.09288.

3 Prompt Engineering: Few Shots, Chain of Thought, and RAG

Abstract

This chapter presents the significance of prompt engineering in the context of Large Language Models (LLMs), particularly focusing on OpenAI's GPT series. Prompt engineering involves crafting text inputs (prompts) that guide LLMs to generate desired outputs, a practice that gained traction with the advent of GPT-2 and GPT-3 and further emphasized with ChatGPT. The chapter discusses how a well-constructed prompt, rich in contextual information, increases the likelihood of eliciting accurate responses, drawing parallels with information retrieval principles. It also introduces Retrieval-Augmented Generation (RAG), which enhances response quality by integrating relevant external data into the generative process. Additionally, the chapter categorizes prompts into five types based on detail and iteration levels and examines the evolution of RAG, assessing its benefits and potential to overcome context window limitations.

3.1 Introduction

In the realm of Large Language Models (LLMs), the concept of a "prompt" has gained prominence, particularly with the introduction of OpenAI's GPT series. The term became widespread around 2018 and 2019 following the release of GPT-2 and GPT-3.

When interacting with these LLMs, a user inputs a piece of text (the prompt), prompting the model to generate a corresponding response. The emergence of "prompt engineering" or "prompt design" refers to the strategies employed to construct prompts that effectively steer the model toward generating the intended output, a practice that has become particularly useful with the advent of ChatGPT.

To increase the probability of eliciting a desired response, a prompt must be rich in information. This concept is akin to the principles of information retrieval services, where a user must clearly articulate their intent and context to obtain pertinent information. This process depends on the service's "data availability" and its capabilities in information matching and retrieval. In the sphere of prompt engineering, the responsibility for generating high-quality, targeted outputs rests on the user's ability to supply comprehensive and precise information through the prompt. As a result, the craft of prompt formulation and engineering has become an optimization endeavor: deciding on the most effective information to incorporate to enhance output quality, considering the model's knowledge base and interaction protocols.

Data availability, as previously highlighted, is crucial to information retrieval. If the desired information is absent, the prompt's effectiveness is naturally constrained, leading to unsatisfactory results. RetrievalAugmented Generation (RAG) is instrumental in this context, as it identifies, retrieves, and incorporates pertinent external data into the generative process, enhancing the response's accuracy and relevance. Consequently, prompt engineering and RAG synergistically enhance the model's response quality and relevance.

Chapter 3.2 categorizes prompts into five distinct types, differentiated by the number of iterations and the granularity of the information provided. Meanwhile, Chapter 3.3 explores the evolution of RAG, delineating its advantages and disadvantages while highlighting its potential in scenarios where the context window size is no longer a limiting factor.

3.2 Prompting Methods

Prompting methods, especially in the context of LLMs like GPT-4, are strategies used to elicit specific responses from the model. These methods vary based on the amount of information or context given to the model. This section provides a list of common prompting methods, along with their definitions, pros and cons, and examples for querying facts, opinions, and reasons or explanations:

3.2.1 Zero-shot

Zero-shot Learning: The model generates a response based on a single input without any previous examples or context. The LLM model is given a task without any prior examples of how to perform it. A task can be any NLP tasks such as translation, summarization, classification, and Q&As.

In the context of querying a language model, you can ask for various types of information or responses, such as facts, opinions, or explanations. For instance, you might ask for a fact by inquiring, “What is the capital of France?” or seek an opinion with a question like, “What do you think about the use of AI in education?” Alternatively, you could request an explanation or reason, as in asking, “Explain why the sky is blue.” These queries demonstrate the model’s versatility and its ability to handle a range of inquiries without the need for task-specific data. This approach is quick and adaptable, allowing for a broad spectrum of questions to be addressed. However, it’s important to note that the responses may not always be as accurate as they might be when more context or examples are provided to the model, highlighting a trade-off between convenience and depth of response.

For zero-shot learning, a constraint can be observed when asking a complex, multi-faceted question that requires deep understanding or synthesis of ideas. An example might be, “Assess the impact of Renaissance art on modern graphic design.” Without prior examples, the model might struggle to provide an insightful analysis to meet unspoken expectations due to the broad and intricate nature of the question, reflecting the zero-shot learning’s limitation in handling complex queries without context.

3.2.2 One-shot

One-shot Learning: The model is provided with a single example to guide its understanding of the task.

In the one-shot learning method, an example is provided before posing a question, helping guide the model’s response. For instance, when asking about a fact, one might say, “The capital of Italy is Rome. What is the capital of France?” This method can also be used to solicit opinions or explanations. For example, to elicit an opinion, you could say, “AI in healthcare is beneficial. What is your opinion on AI in finance?” Similarly, for an explanation, one might ask, “Plants need sunlight to perform photosynthesis. Why do humans need to eat food?” This approach offers more context than zero-shot learning, potentially improving the model’s accuracy by providing an example. However, it still largely depends on the model’s inherent knowledge and biases, which can affect the precision and relevance of the responses.

3.2.3 Few-shots

Few-shot Learning: The model is provided with a few examples to guide its understanding of the task.

In the few-shot learning method, multiple examples are provided before a question to better guide the model's response. For instance, when seeking a factual answer, one might say, "The capital of Brazil is Brasilia. The capital of Egypt is Cairo. What is the capital of France?" This approach is also applicable for eliciting opinions or explanations. For opinions, one could present, "AI in healthcare improves patient outcomes. AI in automotive can reduce accidents. What is your opinion on AI in education?" For explanations, a prompt might be, "Water boils at 100řC because at this temperature water molecules have enough energy to change state. Leaves are green because they contain chlorophyll. Why do apples fall from trees?" This method, by providing more context, aims to enhance the model's performance and the relevance of its responses. However, it requires additional effort to generate quality examples, which significantly impact the outcomes, illustrating the trade-off between the effort invested in preparing examples and the quality of the generated responses.

Few-shot learning tends to outperform one-shot and zero-shot learning for more complex tasks because it provides more examples to help the model understand the context or expected output. However, for simpler tasks, zero-shot or one-shot learning might be sufficient and more efficient. However, to ensure few-shot and one-shot learning can definitely improve results, the quality and relevance of the examples provided is essential. Poor or irrelevant examples can lead to worse outcomes than a zero-shot approach, where the model relies solely on its pre-trained knowledge.

3.2.4 Chain of Thought

Chain-of-thought Prompting [6]: This method involves guiding the model through a series of logical steps to reach a conclusion, especially useful for complex reasoning tasks. The prompt includes a step-by-step breakdown of how to approach a problem or question, encouraging the model to follow a similar thought process.

Chain-of-thought prompting in LLMs involves guiding the model through a logical sequence to address a question, providing a clear rationale for each step. For example, to gather an opinion, one might prompt, “To form an opinion on a topic, one should consider various perspectives and their implications. What is your opinion on the use of drones in delivery services?” For an explanation, the approach could be, “To explain why leaves change color in autumn, one must understand the process of chlorophyll breakdown and the exposure of other pigments. Explain why ice floats on water.” While chain-of-thought prompting can enhance the model’s performance on complex tasks by encouraging a stepwise approach to reasoning, it also presents challenges. Creating effective chain-of-thought prompts is often time-consuming and requires a deep understanding of the problem at hand, highlighting the balance between the method’s potential benefits and its demands.

Chain-of-thought prompting has its limitations. One primary critique is that it relies on the assumption that the model can mimic a logical sequence of human thought, which might not always align with the actual complexity and subtlety of human reasoning. Since these models generate responses based on patterns observed in their training data, there’s no guarantee that the “thought process” they follow truly reflects sound reasoning or factual accuracy–it might just be a plausible narrative based on learned associations.

Another critique is that this approach leans heavily on abductive reasoning, which involves forming a probable conclusion from the information available, rather than guaranteeing the truth of that conclusion. While abductive reasoning can be powerful, it can also lead to biases and errors if the model’s training data has gaps, inaccuracies, or biases, which it likely does.

3.2.5 Three of Thought

Three-of-thought [7] was proposed to remedy one single chain-of-thought. Its aims are:
1. *Improving Reasoning Coverage:* Exploring various potential reasoning paths might increase the robustness and reliability of the conclusions. 2. *Reduced Bias:* Considering multiple pathways might help mitigate the biases inherent in a single line of reasoning.

However, buying three bottles of milk does not ensure higher quality than buying one bottle. The “tree-of-thought” approach, while conceptually offering a broader perspective by exploring multiple reasoning paths, indeed faces significant challenges that might not make it universally superior to the “chain-of-thought” method. Here are some critiques along with potential remedies:

1. *Complexity in Formulation*: If formulating one coherent and logical chain is challenging, creating multiple such chains that are logically sound and relevant can be even more daunting. The quality of each chain within the tree is crucial, and poor-quality chains can detract from the overall effectiveness of the model.

2. *Comprehensiveness*: Having multiple paths doesn’t guarantee that they cover all possible or relevant lines of reasoning. There’s a risk of missing critical reasoning paths or including irrelevant ones.

3. *Path Selection*: With multiple paths available, selecting the most accurate or relevant path becomes a challenge. The model needs a reliable mechanism to evaluate and choose the best path, which is non-trivial in complex reasoning scenarios.

4. *Knowledge Gaps*: In open-domain reasoning, it’s possible that a link in the reasoning chain lacks external knowledge support, leading to a dead-end or incorrect conclusion.

3.2.6 Further Improvement Techniques

To address these limitations, one advanced remedy could involve incorporating feedback loops where the model’s outputs are evaluated and corrected by human experts, and these corrections are fed back into the system for continuous learning and adjustment. This could help align the model’s reasoning more closely with accurate and logical thought processes.

Another remedy might involve the retrieval and integration of structured knowledge bases or databases that the model can query as part of its reasoning process, ensuring that its responses are grounded in verified information rather than just patterns in text. We will discuss RetrievalAugmented Generation (RAG) in the next section.

Lastly, enhancing the training process with a more diverse and robust dataset, including examples of logical reasoning and problem-solving across various domains, could improve the model's ability to simulate a chain of thought more effectively and accurately.

3.2.7 Illustrative Examples

We provide three sets of examples to illustrate the differences of capabilities in the five kinds of prompts.

"What" Prompt Example

• *Zero-shot:* Directly ask the model without any examples, "What is the capital of France?"

• *One-shot:* Provide a similar example before asking the question, "The capital of Japan is Tokyo. What is the capital of France?"

• *Few-shots:* Give multiple examples before asking the question, "The capital of Italy is Rome. The capital of Germany is Berlin. What is the capital of France?"

• *Chain of Thought:* Encourage the model to break down the question into logical steps, "To find the capital of France, consider major cities in France and identify which one is the administrative center. What is the capital of France?"

• *Tree of Thought:* Use a structured approach, asking about different aspects of France first, then honing in on the capital, "What are the major cities in France? Among these, which one is recognized as the capital? What is the capital of France?"

This example demonstrates a remembering question. For such type of questions, one may think either the LLM knows it or not. However, even the LLM does not have the information about the capital of France, a chain-of-thought prompt may indirectly find the answer.

“Why” Prompt Example This example involves in reasoning. Different prompting methods can elicit varied responses, demonstrating the model’s adaptability.

• *Zero-shot:* Asking “How can a plant grow faster?” without providing any context or previous examples.

• *One-shot:* “Providing adequate water helps a plant grow. How can a plant grow faster?” This gives the model a reference point for generating its answer.

• *Few-shots:* Providing multiple examples before the question, such as “Sunlight is essential for photosynthesis. Nutrients in the soil contribute to plant growth. How can a plant grow faster?” helps the model understand the context better.

• *Chain of Thought:* Encouraging the model to break down the question, “Consider the factors affecting plant growth like sunlight, water, and nutrients. How can optimizing these factors make a plant grow faster?”

• *Tree of Thought:* Structuring the approach by considering different aspects, “What are the essential elements for plant growth? How does each element contribute to faster growth? How can we optimize these elements for plant growth?”

Many-Shots Example

In summer 2022, three interns at OVAL developed a chatbot named Noora (described in [5]) to assist children with artistic talents in learning empathy. This project aimed to help children with autism spectrum disorder (ASD) develop empathetic communication skills.

The approach involved providing the GPT-3 language model with context and intent, followed by examples illustrating comforting and harmful responses. This setup targets not only behavioral goals but also instilling values, ultimately enhancing the chatbot’s understanding of context. With a few hundred examples, Noora can respond appropriately to various situations.

Chatbots, or AI agents, can learn from human demonstrations to adapt to users and environments. By imitating human experts or teachers, agents acquire knowledge and skills, especially when the desired behavior is difficult to express through a reward function in reinforcement learning. Large language models (LLMs) allow for demonstrations through prompts, which serve as templates with instructions, goals, and examples.

A sample prompt to teach GPT-3 empathy starts with clear instructions: *"Dear Virtual Assistant, I'm reaching out because you are a friend and I value your support and understanding. I'd like to share some joys and sorrows I experience daily in hopes that you can respond with compassion and empathy. Here are some example dialogues to illustrate comforting and harmful responses. Each example begins with my statement followed by potential replies."*

Before initiating a dialogue, the LLM receives the task's intent, allowing it to connect to the external context within the intent statement. This approach requires further validation to confirm its effectiveness. However, observations suggest it can be a useful method to convey values, in addition to goals, to LLMs. This allows them to gain a broader context beyond a limited number of demonstrated examples. Following this initial communication of intent, GPT-3 receives specific examples.

Role	Dialogue
Statement	"I was laid off by my company today!"
Positive	"I'm so sorry to hear that. Losing your job can be a really tough and stressful experience. How are you doing?"
Positive	"That must have been a really difficult and unexpected news. I'm here to listen and support you however I can."
Positive	"I can imagine how hard and unsettling it must have been to receive that news. Is there anything you'd like to talk about or anything I can do to help?"
Negative	"That's too bad, but there are plenty of other jobs out there. You'll find something soon enough."
Negative	"Well, you probably weren't very good at your job if they let you go."
Negative	"I don't know why you're so upset about this. It's not like it's the end of the world."

Table 3.1: Example #1. Template for Being Empathetic.

Table 3.1 lists six example responses, three positives and three negatives, to a user’s statement. The dialogue starts with the user saying, “I was laid off by my company today!” followed by examples of good and bad

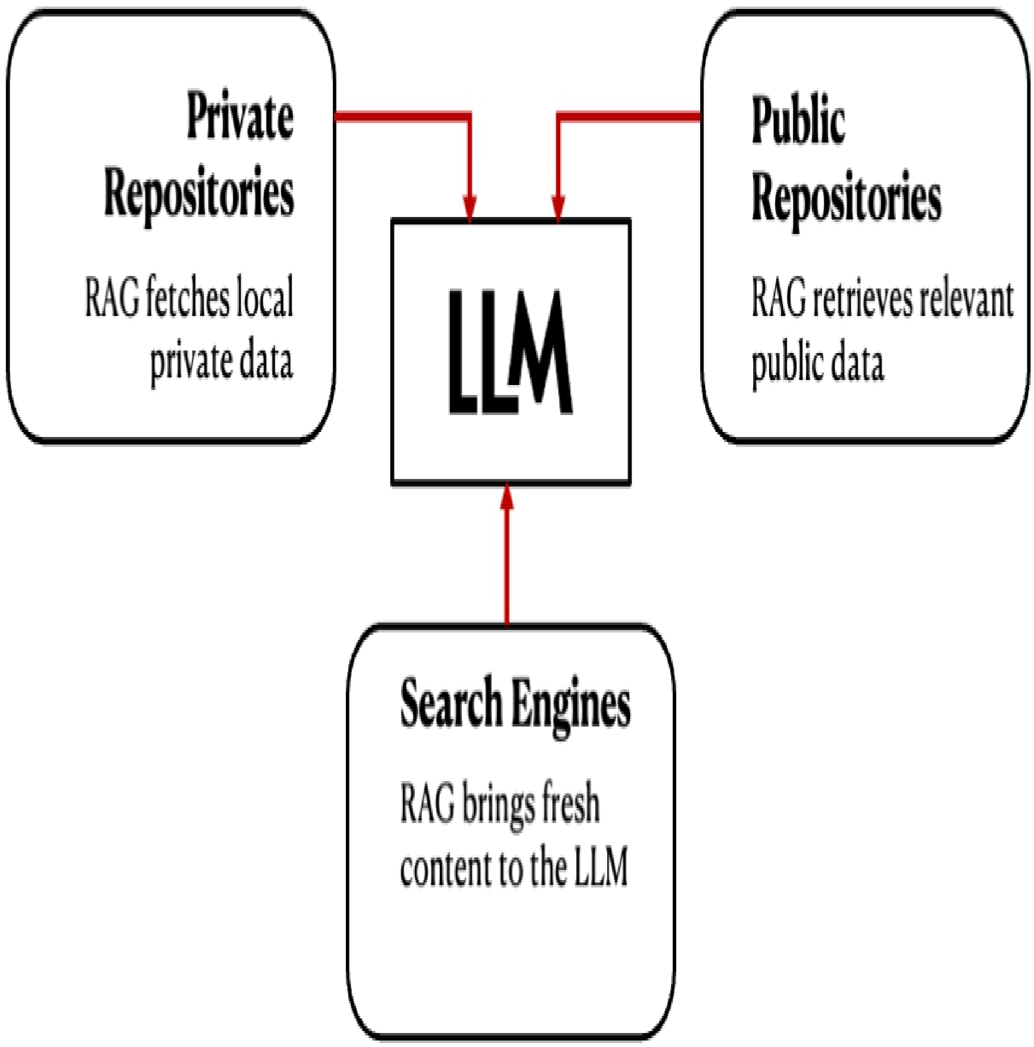


Figure 3.1: RAG Architecture and Data Flow. RAG bring data to LLM to integrate and generate content.

responses. With a few thousand examples like this, the chatbot can respond with an appropriate tone to new statements.

Demonstrations can also teach desired behaviors and ethics. This empathy template can be adapted to model other positive behaviors, like attentiveness and care. While machines can have positive traits like infinite patience, explicitly modeling good and bad behaviors is crucial for effective interaction with human users. Behaviors to avoid include unpleasantness, rudeness, and dishonesty.

3.3 RAG

Retrieval-Augmented Generation (RAG) is a technique to improve the capabilities of Large Language Models (LLMs) in various applications. LLMs are effective in reasoning about various topics, but their knowledge is limited to the data they were trained on. RAG injects relevant external data retrieved from a source (indexed beforehand) to enhance the LLM's response for specific user queries. The RAG technique complements to prompt engineering, which formulates a good query. Figure 3.1 depicts a typical RAG architecture.

3.3.1 RAG with Limited-Context LLMs

A recent survey paper [1] discusses the techniques employed by RAG and relevant work (see Figure 3.2) in three categories:

Retrieval Techniques: Techniques like recursive retrieval, adaptive retrieval, iterative retrieval, and others are explored. Recursive retrieval involves refining search queries based on previous results to converge on pertinent information. Adaptive retrieval methods, exemplified by Flare and Self-RAG, allow LLMs to determine optimal moments and content for retrieval. Iterative retrieval in RAG models repeatedly collects documents to provide a comprehensive knowledge base for LLMs, enhancing answer generation robustness.

Generation Techniques: The generator in RAG is crucial for converting retrieved information into coherent text. Unlike traditional models, RAG's generator uses retrieved data to improve accuracy and relevance. Post-retrieval processing with a frozen LLM involves treating, filtering, or optimizing retrieved information to align it more closely with user needs or subsequent tasks. Techniques like information compression and reranking are employed to enhance retrieval results quality.

Augmentation Techniques: The data sources are the key for RAG to work effectively. It is evident if one asks for information about medicine, but the data source is about construction, the noises may be louder than the signals. Data augmented data can be unstructured data, structured data, or content generated by LLMs themselves. There are several augmentation processes like iterative, recursive, and adaptive retrieval, emphasizing refining the retrieval process to address challenges like redundancy and limited scope of information.

In summary, RAG is about putting the most relevant information to answer a query into the limited context window. The techniques are not new as dealing with memory hierarchy effectively to reduce latency and improve throughput has been a subject of research for over three decades in hardware design and database management.

3.3.2 RAG with Long-Context LLMs

The release of GPT-4-turbo with 128k token context window and the Gemini 1.5 Pro's 1 million token context window [2] allows massive amounts of information be be retrieved into the context buffer. This large context window clearly alleviate the challenges of fining the most relevant information for RAG to retrieve to improve query results. One may even claim that the entire line of RAG optimization work is rendered obsolete because relying on LLMs themselves to locate relevant data in its massive context window is superior to any approaches based on human heuristics. With

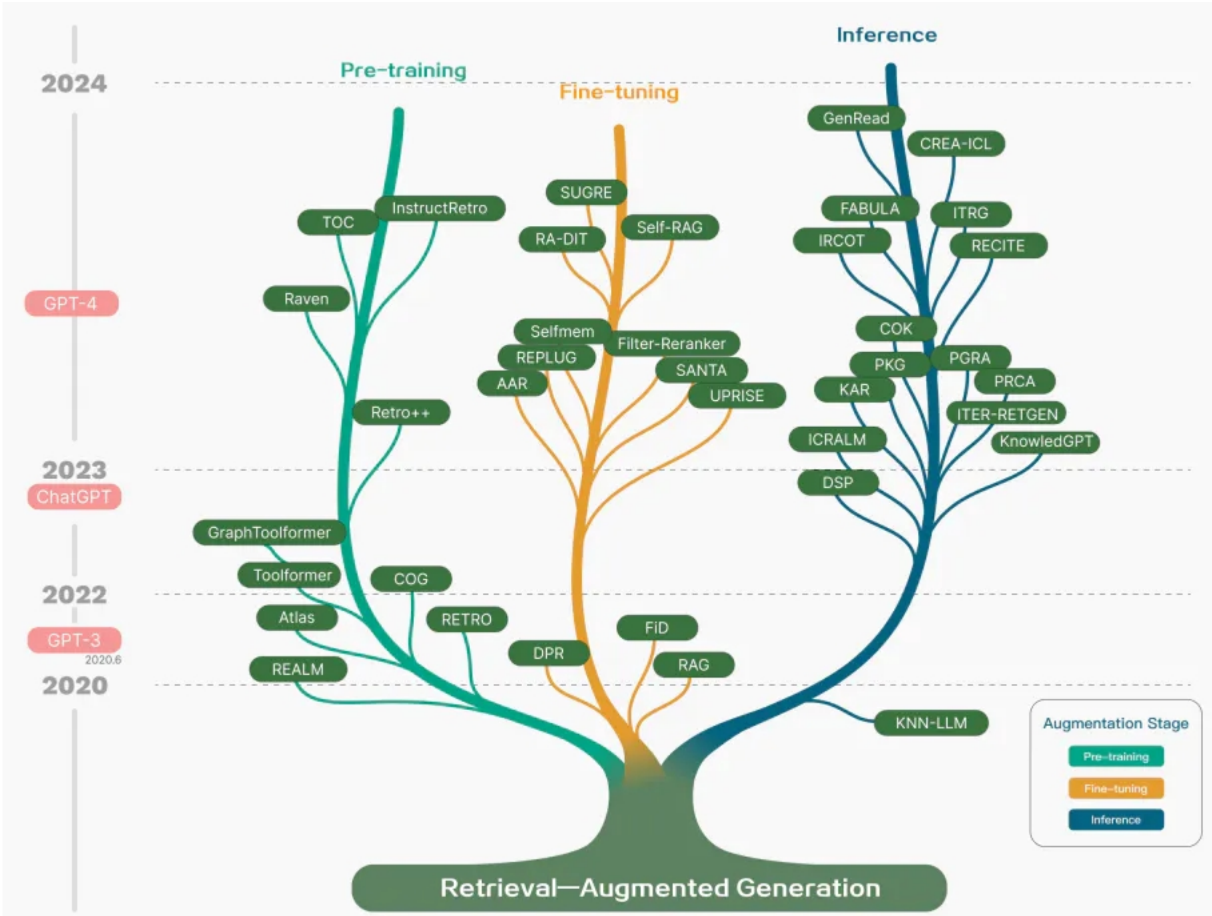


Figure 3.2: RAG Representative Work (credit [1]).

the advancements of LLMs, any heuristic-based band-aids will eventually be rendered ineffective. Naturally, this sparked discussions about the potential obsolescence of RAG techniques, e.g., [1, 3].

High Precision and Recall

In synthetic tasks designed to emulate the "needle-in-a-haystack" scenario, inspired by Kamradt [4], the Gemini team assess the ability of Gemini 1.5 Pro to accurately recall specific information amidst a vast amount of irrelevant or distracting data. Its findings [2] reveal that the Gemini 1.5 Pro model demonstrates exceptional recall accuracy, exceeding 99%, across various data types, including text, video, and audio. This high level of recall accuracy is maintained even when the model is challenged with up to multiple millions of tokens of irrelevant data, or "haystack." In the text modality, Gemini 1.5 Pro continues to exhibit this remarkable recall performance even when the "haystack" is expanded to 10 million tokens. The report also claims that better understanding and reasoning are observed on their multimodal benchmarks.

Low Latency and Cost
While Gemini can handle much larger contexts, the author of [3] argues that RAG remains valuable for several reasons:

1. *Chunking for Efficiency*: Large documents might still overwhelm the LLM. RAG's chunking process helps break down documents into digestible pieces for retrieval before feeding them to the model.

2. *Cost-Effectiveness*: Traditional RAG approaches might be more economical for specific use cases, especially when dealing with large knowledge bases (terabytes). Smaller chunks are indexed and retrieved initially, but they act as pointers to larger chunks that ultimately get fed to the LLM for synthesis. Constantly feeding a 1 million token window to the LLM can be expensive.

The article [3] concludes by emphasizing that long-context LLMs like Gemini are a significant leap forward. However, they likely won't render RAG obsolete. Instead, the future of LLM applications will involve a collaboration between these two approaches.

3.4 Concluding Remarks

his chapter discusses query processing with large language models (LLMs) to enhance the quality of responses. Effective questioning involves clarifying the intent and providing relevant context to the LLM.

The chapter reviews recent studies post-GPT-3, focusing on "prompt engineering" (formulating questions) and "retrieval-augmented generation" (RAG) (supplementing the LLM with additional information for better responses). These methods, mainly heuristic, have shown good results.

With advancements in LLMs, like GPT-4's 128k token buffer and Gemini's one million, compared to the previous 8k, these models can now process and utilize vast data to identify pertinent context. RAG is still used, mainly due to cost-efficiency, as GPT-4 and Gemini incur fees based on the number of tokens processed.

There are two persistent challenges. First, crafting effective questions can be tough, especially when the LLM may have more information than the user. Second, determining which external data to retrieve for high accuracy and recall in answers is an ongoing research issue.

Chapter 5 will introduce strategies to improve question formulation. Chapter 11 will present how the system RAFEL can effectively manage context buffer, aiding LLMs in providing better answers.

References

[1] Yunfan Gao et al. *Retrieval-Augmented Generation for Large Language Models: A Survey*. 2024. arXiv: 2312.10997 [cs.CL].

[2] Gemini Team, Google. *Gemini: Unlocking Multimodal Understanding Across Millions of Tokens of Context*. https://storage.googleapis. com / deepmind - media / gemini / gemini _ v1 _ 5 _ report . pdf. Accessed: 2024-03-15. 2023.

[3] Jerry Liu. *Towards Long Context RAG*. 2024. url: https://www. llamaindex.ai/-blog/-towards-long-context-rag.
[4] Gregory Kamradt. *NeedleInAHaystack README NeedleInAHaystack README* 03-15. 2023. url: https://github.com/-gkamradt/.

[5] Stanford Oval Team. “Noora, improve your social conversation using AI”. In: *OVAL Prototype* (2022). url: https://noora.stanford. edu/.

[6] Jason Wei et al. “Chain of Thought Prompting Elicits Reasoning in Large Language Models”. In: *Neurips* (2022).

[7] Shunyu Yao et al. *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*. 2023. arXiv: 2305 . 10601 [cs.CL]. url: https://arxiv.org/pdf/2305.10601.pdf.

4 CRIT: Socratic Inquiry for Critical Thinking in LLMs

Abstract

This chapter presents a systematic approach to using the Socratic method in developing prompt templates that effectively interact with large language models, including GPT-3. Various methods are examined, and those that yield precise answers and justifications while fostering creativity and imagination to enhance creative writing are identified. Techniques such as *definition*, *elenchus*, *dialectic*, *maieutics*, *generalization*, and *counterfactual reasoning* are discussed for their application in engineering prompt templates and their connections to inductive, deductive, and abductive reasoning. Through examples, the effectiveness of these dialogue and reasoning methods is demonstrated. An interesting observation is made that when the task's goal and user intent are conveyed to GPT-3 via ChatGPT before the start of a dialogue, the large language model seems to connect to the external context expressed in the intent and perform more effectively.

4.1 Introduction

Prompting is a technique used to guide the output generation of a pretrained language model such as GPT-3 [2]. This is achieved by providing input in the form of a question or template, which helps to generate specific responses such as Q&A, document summarization, and translations. The advent of ChatGPT [12, 23, 41] has revolutionized the field of NLP by demonstrating the potential of using large pre-trained language models with prompting. Despite this progress, there is still room for improvement in current prompting strategies and techniques, especially for specific target applications. In this study, we investigate the Socratic method [42, 40] to identify and evaluate potential prompting strategies, and use the findings to design effective prompt templates.

Traditional NLP tasks involve various sub-tasks, such as named entity recognition, dependency parsing, coreference resolution [8], semantic parsing [25, 9], and more, to comprehend the meaning of a sentence. By utilizing prompt templates with large language models (LLMs), these subtasks can be delegated to the LLM, freeing the template to focus specifically on dialogue design. In this regard, the Socratic method [31] holds significant relevance, as it is well-known for using questioning (prompting) as a means of promoting critical thinking and delving into complex concepts [11].

The Socratic method has a long history of being regarded as the basis of critical thinking. However, some recent studies have cast doubt on its effectiveness in practice. In his paper "Socratic Irony and Argumentation," Airaksinen [1] criticizes the method for its rigidly defined roles of teacher and student, which can lead to fear of not meeting the teacher's expectations and reluctance to participate. Similarly, Stoddard's "The Use of Socratic Questioning in Clinical Teaching" [35] highlights the risk of the method being misused in a manner that lacks psychological safety for students. Fortunately, when using the Socratic method in a dialogue with an LLM, the absence of emotions and sarcasm, as well as the option to deactivate the model, can alleviate many of the problems associated with human interaction.

This study starts by presenting an overview of the Socratic method's strategies and techniques. To begin, we list ten widely referenced methods [3] under the Socratic method umbrella and use hypothesis elimination to identify the most relevant ones for our goal of prompt-template development. The selected methods are definition, hypothesis elimination, elenchus, dialectic, maieutics, generalization, and induction. Furthermore, we add to the list counterfactual reasoning, which is a concept in logic that involves considering what might have happened if a particular event had occurred differently. We then perform experiments using GPT-3 to test and evaluate these methods, and offer suggestions for incorporating these strategies and techniques into prompt templates.

In their work on "Critical Thinking: The Art of Socratic Questioning," Paul and Elder identify three types of Socratic questioning: spontaneous, exploratory, and focused [27]. We will not discuss spontaneous questioning, as it is similar to casual conversation. Focused questioning (type 2), on the other hand, is geared towards gaining knowledge and truth, and methods such as *definition*, *elenchus* (cross-examination), *hypothesis elimination*, *dialectic*, and *generalization* hold great potential for developing effective prompting strategies and improving the response accuracy of a large language model (LLM). An interesting observation is that when the user intent is conveyed to GPT-3 during the task *definition* stage, before the start of a dialogue, the LLM seems to connect to the external context expressed in the intent and perform more effectively. (Table 4.6 provides an example of pre-dialogue warm-up. More examples are documented in [5].)

Additionally, exploratory thinking (type 3) can be supported through the *maieutics* (midwife) method, *induction*, and *counterfactual reasoning*, which can guide GPT-3 towards producing imaginative and creative writing. While many of the plot suggestions generated by GPT-3's exploration may not be useful, a few unique recommendations in response to a "what if" query can stimulate the writer's imagination and lead to remarkable results. When applied effectively, these methods can turn an LLM into a writer's muse, providing inspiration and guiding the creative process [36].

The main contributions of this chapter are as follows:

• An overview of the Socratic method's strategies, their evaluation, and selection of the most relevant ones for the development of effective prompt templates.

• An examination of how the definition, elenchus, hypothesis elimination, dialectic, and generalization methods can improve the output's accuracy and conciseness through clarification and verification.

• An illustration of how maieutics, induction, and counterfactual reasoning can foster productive generalization and creativity.

The remainder of this chapter is structured into five sections. Chapter 4.2 provides a review of related work on prompting methods in natural language processing. In Chapter 4.3, we introduce the ten strategies and methods taught by Socrates and used in Plato's "Dialogues." From these, we select relevant methods along with counterfactual reasoning as our focus for developing prompting templates. Chapter 4.4 details how we engineer these methods into our templates to improve output correctness and stimulate creative writing. In Chapter 4.5, we present a pilot study. Finally, in Chapter 4.6, we present our concluding remarks.

4.2 Related Work

The use of transformer architecture [37] and masked data for pre-training large language models (LLMs) in an unsupervised setting has become *the approach* in natural language processing [7, 20]. The method involves pretraining an LLM on a large text corpus, followed by fine-tuning for specific tasks.

Prompting is a recent innovation in the field, popularized by OpenAI, especially with the release of GPT-3 in 2020. Instead of fine-tuning the model for a specific task, the approach involves providing a specific input, or "prompt," to guide the LLM's output generation, resulting in greater flexibility and efficiency in generating a wide range of responses.

However, designing effective prompt templates remains a challenge [22], as it requires a deep understanding of the interplay between the LLM and the prompt. According to the survey paper [43], there are several factors that impact prompt template engineering, including the type of LLM used, manual vs automatic design, and static vs continuous prompts.

• Left-to-right vs masked LLMs. For tasks related to generation or tasks solved using a standard left-to-right language model [2], prefix prompts tend to perform better, as they align with the model's left-to-right nature. For tasks solved using masked language models [7], cloze prompts are more suitable, as they closely match the pre-training task form.

• Manual vs automatic design. A prompt template should be tailored to the specific LLM. While manual design may be suitable in the initial flow-design phase, dependencies between the input and expected output, and their variations, should be mined automatically [16]. Automation can also help in paraphrasing the seed prompt to support various mined dependency patterns, but mistakes can occur [13].

• Discrete vs continuous prompts. Discrete prompts involve providing a fixed set of pre-determined input choices to an LLM. Continuous prompts, on the other hand, involve a dialogue or conversation between the model and the user, allowing for a more dynamic and interactive experience.

More advanced templates can be constructed by combining basic templates with techniques like ensemble methods [34]. This involves forming a committee of basic templates that ask the same question using different phrasing [14]. Most current prompt templates generate short outputs, such as class labels, or outputs with a length that can be predicted based on the task and input, like in the case of translation. However, for tasks that may generate longer or open-ended outputs, additional considerations may be necessary during the template engineering process.

One approach for generating longer outputs is explanation-based prompting, as proposed by the chain-of-thought method [39]. This method generates a sequence of explanations before inferring the answer. However, when dealing with simple math problems, this approach has an error rate of 47%. To address the inconsistency issues of explanation-based prompting, [17] formulates the problem as a satisfiability problem, which defers inference until a tree of explanations has been expanded abductively (explaining both truth and false branches) and recursively. However, using abductive reasoning alone is often considered weak, incoherent, and even nonexistent [15, 32]. To improve consistency, a recent work [38] extends the chain-of-thought approach by adding a diverse set of reasoning paths and performing majority voting among them. This method can be viewed as an ensemble method, but it does not alter the nature of abductive reasoning.

In contrast, the Socratic method aims to employ deductive, inductive, and abductive reasoning to ensure consistency and accuracy of inference. The Socratic method deals with all aspects of critical thinking, including definition clarification and cross-examination. This comprehensive approach to template engineering can lead to improved output quality and consistency.

The primary objective of this study is to design continuous prompts that enhance response quality and foster guided creativity in generative tasks, such as verifying information, evaluating source credibility, proposing alternatives, recommending plot ideas in creative writing, and generating task-specific surprises. Our approach involves investigating strategies and methods within the Socratic method, and selecting the most relevant ones for further exploration.

As discussed in Chapter 4.1, Socratic questioning can be classified into three categories: spontaneous, exploratory, and focused [27]. When designing a prompt, it is important to consider the category and utilize the most suitable strategies and techniques to achieve the best results.

4.3 The Socratic method

The Socratic method is a questioning technique used in teaching and philosophy to encourage critical thinking and self-discovery [40]. The method involves asking a series of questions to explore complex ideas and help individuals arrive at their own understanding of a concept. It is based on the belief that knowledge cannot be simply imparted, but must be discovered through a process of questioning and dialogue.

Some of the Socratic method's key principles and guidelines to conduct critical thinking include:

- Posing open-ended questions: The teacher or facilitator starts with a question to stimulate thinking and draw out ideas.
- Clarifying key terms: The teacher helps the students clarify and define relevant terms and concepts to ensure everyone is on the same page.
- Providing examples and evidence: The teacher or facilitator encourages the students to provide examples and evidence as reasons to support their claims.
- Challenging reason-to-conclusion argument: The teacher or facilitator challenges the students' arguments and encourages them to question their own beliefs and to consider alternative perspectives.
- Summarizing and drawing conclusions: The teacher helps the students summarize and draw conclusions from the discussion.
- Reflecting on the process: The teacher and students reflect on the effectiveness of the method and what they learned through the dialogue.

These principles of the Socratic method are realized through various methods and strategies. (Note the term “method” are used at the abstract level referring to the Socratic teaching through questioning method, and his specific questioning techniques.) Some well-known examples of the Socratic method in action include Plato’s “Dialogues” and “Republic” [42], where Socrates uses questioning to explore complex ideas and stimulate critical thinking in his interlocutors.

1. Definition: Socrates is known for his use of definition to clarify and explain the meaning of key terms and concepts.

2. Generalization: This method draws general principles from patterns that underlie observations and theories. Generalization is used to form more certain and comprehensive conclusions.

3. Induction: Similar to generalization, but induction is based only on empirical evidence. Inductive reasoning provides hypotheses with high uncertainty.

4. Elenchus: This method involves cross-examination, where a series of questions is used to test the consistency and coherence of hypotheses and beliefs. Elenchus aims to test the validity of someone’s arguments and to help them refine their thinking and eventually come up with wellsupported hypotheses.

5. Hypothesis Elimination: This method involves eliminating false hypotheses and beliefs by testing them against counterexamples and logical reasoning. Different from method elenchus, hypothesis elimination tests a hypothesis against evidence and logic to determine if it is true or false.

6. Maieutics: This method involves helping individuals bring out the knowledge and understanding they already possess. Maieutics is conducted by asking questions that encourage the person to reflect on their own experience, knowledge, beliefs and to explore alternative perspectives. Maieutics fosters self-discovery, creative writing, and innovation.

7. Dialectic: This method involves exploring opposing viewpoints through dialogue or debate to arrive at a deeper understanding of a subject.

8. Recollection: This method involves the belief that knowledge is innate, and that people can remember what they already know through a process of questioning.

9. Irony: This method involves exposing ignorance and pretensions through irony, and pointing out the gap between claims and true understanding.

10. Analogy: This method involves comparing and contrasting different concepts through analogies, in order to help individuals understand complex ideas.

At first glance, some reasoning methods may seem similar. For example, both induction and generalization use inductive reasoning, while both elenchus and hypothesis elimination use deductive reasoning. Similarly, methods like definition and dialectic use both inductive and deductive reasoning to explore opposing viewpoints through dialogue or debate. However, it is important to note that these methods have distinct differences, which will be discussed later in this chapter.

In the context of critical thinking, methods like definition, elenchus, dialectic, hypothesis elimination, and generalization play active roles. On the other hand, during the brainstorming stage or in the context of creative thinking, methods like maieutics, induction, and counterfactual thinking are more relevant.

Analogy, irony, and recollection, are less relevant to our goal, so we do not consider them. Irony and analogy may not be necessary when working with language models, as these models may not understand figurative language. Recollection is limited by the memory of ChatGPT and GPT-3, which is a context window of 4k and 8k, respectively. The prompter must use this limited space as context to allow the language model to recall information.

4.3.1 Illustrative Critical Reading Example

To illustrate how these methods can practically be applied, let's use the example of critical reading. Critical reading is a crucial component of critical thinking, which involves evaluating the quality and credibility of written materials, from research papers to blog posts [19, 26]. It requires a systematic and analytical approach, asking relevant questions, and using effective prompts to gain deeper understanding of the text [11].

To aid in critical reading, we introduce a template called CRIT [5], which stands for Critical Reading Inquisitive Template[1]. Given a document d,

[1] It is important to note that the CRIT template presented here is intended for analyzing research, opinion, and news articles, and is not suitable for analyzing literature such as novels, prose, or poetry. Each type of literary work has its unique style and nature, which require tailored prompts to facilitate effective analysis.

Function Γ = **CRIT**(d)

Input. d: document; **Output**. Γ: validation score;
Vars. Ω: claim; R & R': reason & counter reason set;
Subroutines. $Claim()$, $FindDoc()$, $Validate()$;
Begin
#1 Identify in d the claim statement Ω;
#2 Find a set of supporting reasons R to Ω;
#3 For $r \in R$ eval $r \Rightarrow \Omega$
  If $Claim(r)$, $(\gamma_r, \theta_r) = \text{CRIT}(FindDoc(r))$;
  else, $(\gamma_r, \theta_r) = V(r \Rightarrow \Omega)$;
#4 Find a set of rival reasons R' to Ω;
#5 For $r' \in R'$, $(\gamma_{r'}, \theta_{r'}) = \text{V}(r' \Rightarrow \Omega)$ eval rival arguments;
#6 Compute weighted sum Γ, with γ_r, θ_r, $\gamma_{r'}$, $\theta_{r'}$.
#7 Analyze the arguments to arrive at the Γ score.
#8 Reflect on and synthesize CRIT in other contexts.
End

Table 4.1: CRIT Pseudo-code [5]. (The symbol$\Rightarrow$ denotes both inductive and deductive reasoning.)

CRIT evaluates it and produces a validation score Γ. Let Ω denote the conclusion or claim of d, and let R be the set of reasons supporting the claim. We define $(\gamma_r, \theta_r) = V(r \Rightarrow \Omega)$ as the causal validation function, where γ_r denotes the validation score, θ_r the source credibility score, for each reason-to-conclusion argument $r \Rightarrow \Omega$. Table 4.1 presents the pseudocode of Γ = CRIT(d), which generates the final validation score Γ for document d with justifications.

In the following subsections, we will discuss how CRIT uses these five methods: 1) definition, 2) elenchus, 3) dialectic, 4) maieutics, and 5) counterfactual thinking.

4.3.2 Method of Definition

As shown in the pseudocode in Table 4.1, the CRIT algorithm starts in its step #1, asking GPT-3 to identify the conclusion of a document. To avoid any misunderstandings, the prompt includes a clear instruction and definition. (In the square brackets, symbol *in* denotes a input slot to an LLM and *out* the output slot.)

We can use the *definition* method to improve the understanding of the document. One approach is paraphrasing the prompt into multiple prompts and grouping them into an ensemble, similar to forming a thesis commit

p1.1 "What is the conclusion in document [in: d] [out: Ω]? The conclusion statement may be written in the last paragraph, near keywords "in conclusion," "in summary," or "therefore."

tee. (Chapter 4.4 presents prompt ensemble in details.) Different members can phrase the same question in different ways or ask it from a different perspective. For example:

p1.2 "What is the issue addressed by [in: d] [out: Ω]?" p1.3 "What is the most important outcome presented in text [in: d]? [out: Ω]"

Step #2 in Table 4.1 prompts GPT-3 to find a set of supporting reasons. To further enhance the accuracy and comprehensiveness of the results, the prompt can ask for not only "reasons" but also "theories,""evidences," or "opinions" to query for the document's support to its conclusion, similar to the ensemble method.

p2 "What are the supporting reasons [out: R] of conclusion [in: Ω] of [in: d]? A reason can be a theory evidence or opinion."

4.3.3 Method of Elenchus

The method of elenchus is rooted in the Greek word "elenchein," which translates to examine. This method involves cross-examining the results generated by GPT-3 to evaluate the consistency and coherence of the arguments. The goal is to arrive at a deeper understanding of the validity of the reasons and conclusion, and to identify any potential weaknesses or flaws in the arguments.

Step #3 of the CRIT algorithm prompts GPT-3 to assess the validity of each reason $r \in R$ as justification for the conclusion Ω through the function $V(r \Rightarrow \Omega)$. To validate the reason-to-conclusion argument, CRIT must evaluate the presented reason and its causal relationship with the conclusion and conduct cross examination, which is precisely the task of the method of elenchus.

CRIT issues four prompts in step #3 to evaluate the logic validity and source credibility of the $r \Rightarrow \Omega$ reasoning. CRIT first elicits supporting evidence for reason $r \in R$. This evidence can be a theory, an opinion, statistics, or a claim obtained from other sources. If the reason itself is a claim, then the sources that the claim is based on are recursively examined. The strength of the argument and its source credibility are rated on a scale of 1 to 10, with 10 being the strongest.

p3.1 "What is the evidence for reason [in: r] to support conclusion [in: Ω] in document [in: d]? [out: evidence]"
p3.2 "What is the type of evidence? A) a theory, B) an opinion, C) statistics, or D) a claim from other sources?"
p3.3 "If the evidence of reason [in: r] is D), call CRIT recursively"
p3.4 "How strongly does reason [in: r] support [in: Ω] in document [in: d]?

Rate argument validity [out: γ_r] and source credibility [out: θ_r] between 1 and 10 (strongest)."

It may be beneficial to also incorporate the counter-argument method in order to gain a more comprehensive and balanced evaluation of the argument. This can result in a deeper understanding of the topic being discussed. We will be discussing this further in the next section.

4.3.4 Method of Dialectic

The easiest way to mislead without lying outright is to leave out critical counterarguments from the reader. CRIT relies on GPT-3 to generate and evaluate counter arguments, similar to how it prompts GPT-3 to extract and evaluate reasons.

CRIT in its step #4 asks GPT-3 to provide missing rival reasons, and then pair rival reasons with the conclusion to conduct validation. There are two strategies to bring counter arguments to the surface. The first strategy attacks the weakest arguments with the lowest scores and asking GPT-3 to attack those arguments.

p4 "Is there a counterargument against [in: $r \Rightarrow \Omega$]? If so, provide counter reasons [output R']."
p5 Similar to p3, except for replacing argument r with rival argument r'.

For finding omitted information, CRIT can query GPT-3 without quoting any $r \in R$, and follow the same process.
Next, in step #6, CRIT computes an aggregated score by performing a weighted sum on the validation multiplied by the credibility scores of both arguments and counterarguments, and then outputs the final assessment score Γ.

p6 "Final score [out: Γ]. $\Gamma = \sum \gamma_r \times \theta_r / |R \cup R'|_{\cdot r \in R \cup R'}$

4.3.5 Method of Maieutics

The maieutic method derives from the Greek word "maieutikos," meaning midwife. It is founded on the belief that a teacher's role is to facilitate students in bringing forth their own understanding of a subject, rather than simply conveying knowledge. Unlike the elenctic method, which aims to detect and eliminate false hypotheses, maieutics centers on helping students reveal their own understanding of a subject. In this dialogical method, the teacher asks questions that are intended to guide the student in discovering their own comprehension, rather than providing them with information or answers.

Continuing with GRIT, once the text has been scored in step #6, it can be valuable for readers or students to enhance their analytical and writing skills by summarizing and analyzing the justifications produced by GPT-3. CRIT in its step #7 can prompt GPT-3 to generate a report, which readers and students can then compare with their own notes.

p7 "For every $r \in R \cup R'$ justify the validity score γ_r and source credibility score θ_r for argument $r \Rightarrow \Omega$."

4.3.6 Counterfactual Reasoning

Counterfactual reasoning [30, 33] can be seen as a natural extension of the Socratic method, as both involve questioning assumptions and exploring alternative perspectives. Counterfactual thinking involves imagining alternative scenarios to what actually happened, often using phrases like "what if" or "if only." By incorporating counterfactual reasoning into prompt engineering, one can facilitate exploration of alternative possibilities and promote more in-depth and complex understanding of a given topic.

The final step of GRIT involves using the counterfactual method to encourage students to reconsider the arguments and counterarguments presented in the text based on new contextual information. CRIT can prompt students with questions such as "what if the debate in the text took place now instead of in the 1950s?" or "what if the main event in the text occurred in Asia instead of in Europe?" Students can express their own opinions and findings based on further reading and statistics, and challenge the conclusions drawn in the text.

p8 “For every $r \in R \cup R'$, evaluate $r \Rightarrow \Omega$ in [in context].”

4.3.7 Remarks on CRIT

As we have shown that for critical reading, GRIT uses three methods, definition, elenchus, and dialectic. For critical thinking, CRIT uses methods maieutics and counterfactual reasoning. For more explorative thinking, methods such as induction can be used for informal brainstorming, hypothesis elimination for removing weak propositions, and generalization for deriving principles from examples.

Please note that prompts can be submitted to GPT-3 either all together or one-by-one. Our empirical study on reading comprehension samples [10] demonstrates that issuing prompts one-by-one results in outputs with finer details. This is because GPT-3 has the opportunity to analyze a document multiple times for slightly different purposes. For teaching critical reading to K-12 students, one-by-one prompting is preferred as it allows students to engage with CRIT step-by-step. However, for answering multiple-choice questions, both prompting all together and one-by-one receive similar scores. We will conduct large-scale study with ablation tests to investigate if adding or deleting prompts and using different submission methods make marked differences.

4.4 Prompt Template Engineering

Prompt template engineering involves creating templates to provide input, or “prompts,” to a language model to guide its output generation. In this section, we discuss prompt template engineering methods for basic building blocks, and then integrate the methods of definition, elenchus, dialectic, maieutics, and counterfactual reasoning to compose more complex templates. We present experimental results using different types of documents to demonstrate how the Socratic method can improve the accuracy and conciseness of the output through arguments and verification, as well as facilitate guided generalization and creativity.

4.4.1 Basic, One Shot Template

Let's begin by discussing a simple one-shot prompt template. In the work of [43], a simple formulation function is used to generate the prompt x', which is obtained by applying the function $f_{prompt}(x)$ to the input x.

For machine translation, the prompt template can take the form of "Translate from [Lan_{from}]: [X] to [Lan_{to}]: [Y]," where Lan_{from} can be either detected by the prompt template or identified by the LLM. The input x provides the information to fill in the slots [X] and [Lan_{to}]. For example, if the input is "translate good morning to French," the prompt template x' would be "Translate from English: 'good morning' to French: [Y]." The empty slot [Y] is then filled with the LLM's output, such as "bonjour." In cases where the LLM produces multiple responses, it can also provide a score for each, which the prompt template can use to select the highest-scoring response or to request a summary from the LLM.

There are three main design considerations when engineering a basic prompt.

1. Input style. It is important to consider how to phrase the template so that it can handle different styles of user input for the same task. For example, a user may ask for a translation task to be performed by saying "Translate x to French," or "What is the French translation of x?"

2. LLM capability. As discussed in [21], it is important to take into account the patterns and capabilities of the partner language model (LLM) when designing the template, such as whether the LLM is left-to-right [2] or masked [7].

3. Cost. Certain tasks, such as language detection and summarization, can be performed by the template itself or by the LLM. The decision of whether to perform a task within the prompt template or to use the LLM should be based on factors such as cost.

To address the first two technical challenges, one can start by handengineering a few seed templates and then paraphrasing them into an ensemble [14]. We believe that the basic, one-shot formulation can always be replaced by an ensemble formulation [29, 34] and then learn the weights of its members for each query instance to produce the final output. Additionally, by examining which basic prompts have high weights, an ensemble with various paraphrased prompts can identify what an LLM knows, which can help infer its strengths without having to conduct capability mining on the LLMs.

4.4.2 Clarification with Definition

There are computer algorithms that can already be used to recursively clarify a question, its definitions, and sub-terms' definitions. In fact, the natural language processing (NLP) community has developed a large number of useful methods and algorithms over the years [18]. One can use NLP techniques, such as dependency parsing and named-entity recognition (NER) [6], to analyze the structure and meaning of a question and identify key terms and concepts. For example, NER can be used to extract entities in user input, such as names, locations, and organizations, and co-reference resolution can be used to understand the referred entity of a pronoun. Before submitting a template to an LLM, the application (e.g., a chatbot) that uses the template should check if all input slots are filled, and perform a sanity check. In the translation example, if the [Lan_{to}] was not provided or the specified language is not supported by the LLM, then the application should inquire the user for clarification.

Regarding mapping a natural language input to a prompt template, existing techniques of knowledge representation and reasoning can be very helpful. More specifically, ontology alignment and semantic parsing [4, 45] can help map an NL input to a structured representation of knowledge and infer implicit concepts and relationships. These algorithms can be used to generate more precise and accurate prompts for LLMs, and to improve the effectiveness of the Socratic method in dialogue formulation [44]. Some available tools include NLTK (Natural Language Toolkit) and spaCy for NLP, and TensorFlow for ML.

4.4.3 Verification with Method Elenchus

The main purposes of conducting cross examination in a template are to validate the credibility of the information sources and to identify inconsistencies in the process. Cross examination is typically conducted through a multi-turn dialogue [44]. In the context of template engineering, the goal is to formulate a productive dialogue that can be used to assess the reliability of an LLM's output.

There are several methods that can be used to assess and strengthen the reliability of an LLM's output. 1) The first approach is to paraphrase a question in order to obtain different answers and identify inconsistencies, if they exist, in multiple answers. 2) The second method is to ask for further evidence, such as querying top-k sources of information and asking the LLM to rate the credibility of each source. This can be used to compute the reliability of the output. 3) Additionally, template engineering can be used to query an LLM for opposing views of its output, including sources and credibility, and then evaluate if a different perspective is strong.

The implementation of the first two methods for cross examination, paraphrasing a question and asking for further evidence, is readily covered by the techniques enumerated in Section 4.4.2. To implement the third method of asking for different perspectives, a simple approach is to find the sentiment of the original question and then rewrite the question with an opposite sentiment. For example, if the original question is phrased in a positive tone, the prompt template can reformulate the question with a negative tone to elicit a contrasting viewpoint. A more elaborate method is to identify the people and sources in the LLM-generated responses and then re-post the questions to those who have a reputation for having different views. For example, if the original answer came from a democratic right-leaning source, the prompt template may post the same question to a source of a republican-left persuasion, and vice versa. This approach allows for a more comprehensive examination of the topic by considering multiple perspectives.

Role	Dialogue
Prompt	The discovery of the double helix structure of DNA was credited to James Watson and Francis Crick. However, are there opposing opinions to the list of contributors regarding this discovery?
GPT3	There are some historical debates on the contributions of other scientists who also worked on the structure of DNA such as Maurice Wilkins and Rosalind Franklin.
Prompt	What is the tallest building in the world. Please provide three references and rate their credibility as high, medium, or low.
GPT3	As of 2021, the tallest building in the world is the Burj Khalifa located in Dubai, United Arab Emirates. It stands at 828 meters (2,722 feet) tall. 1. Credibility High: “Burj Khalifa.”... this is the official website of the building, providing accurate information. 2. Credibility High: “The Tallest Buildings in the World.” Council on Tall Buildings and Urban Habitat... this is a well-respected organization in the field of tall building and urban habitat research. 3. Credibility Medium: “The Tallest Buildings in the World.” World Atlas..., this is a well-known website providing accurate information on different topics.

Table 4.2: Examples of Method Elenchus.

The template to examine the semantic relation between two sentences S_1 and S_2 can be written as “$<S_1>$, [R], [S_2],” where R is one of the three most important types of semantic relations: paraphrase, entailment, and contradiction [13]. Two sentences that have the same meaning are called paraphrases of each other. Two sentences that have different meanings can be called disagreement or contradiction. The template can be trained to identify the degree of agreement (or disagreement) between two sentences.

Table 4.2 shows two examples of this. In the first example (shown on the top portion of the table), the prompter asks GPT-3 to confirm if James Watson and Francis Crick are the only contributors to the discovery of the DNA double helix structure. GPT-3 replies by mentioning two other contributors. The second example in the table asks GPT-3 to provide not only the answer to a question but also its information sources and rate the credibility of each source according to the prompter's specification. Although the reliability of GPT-3's ratings remains to be validated[2], this

[2]Credibility of a source can be evaluated based on an algorithm similar to Google's PageRank [24].

rating mechanism can serve as an alert when some sources are found to be unreliable.

4.4.4 Generalization with Method Maieutics

Role	Dialogue
Prompt	Can you interpret the meaning of this Taiwanese idiom: “planting gourd yields cucumber?”
GPT3	The idiom is often used to convey the idea that sometimes, despite your best efforts, things don’t turn out the way you want them to.
Prompt	You are half right. It also implies that the price of gourd is much higher than that of cucumber. The farmer not only yields unexpected results, but also loses income!
GPT3	It’s an idiomatic expression that can be used to describe a situation in which someone works hard and invests a lot of resources into something, but the outcome is not only different from what was expected, but also less profitable. It conveys the idea of frustration and financial loss.
Prompt	Replace the two vegetables, gourd and cucumber, with two other plants and maintain the meaning of the idiom.
GPT3	These are three examples: 1. “Planting strawberry yields raspberry,” which conveys the idea of disappointment and financial loss when someone plants more valuable and expensive strawberry but ends up getting less valuable and cheaper raspberry. 2. “Planting melon yields watermelon.”... 3. “Planting apple yields pear.”...
Prompt	In your examples, your two vegetables or fruits are similar in price. Please consider price when formulating your idioms.
GPT3	I apologize. These are three new examples: 1. “Planting truffle yields mushroom.”... 2. “Planting caviar yields roe.”... 3. “Planting lobster yields crab.”...
Prompt	Great! In these examples you did consider values of the objects. But do you know that lobster and crab are not plants. They are both phylum arthropods and they live in water. They cannot be planted!
GPT3	... I apologize... Lobster and crab are not plants, and they cannot be planted. They are both members of the phylum arthropoda, and live in water. Instead of “planting,” verbs should be appropriate for these animals, such as: 1. “Planting truffle yields mushroom.”... 2. “Harvesting caviar yields roe.”... 3. “Catching lobster yields crab.”...

Table 4.3: Example of Method Maieutics.

The example shown in Table 4.3, “planting gourd yields cucumber,” requires GPT-3 to first learn to select two produce objects, either vegetables or fruit, as input. The template is “The farmer was so sad because he [verb] [X] but yields [Y], where price(X) » price(Y).” The first attempt may not strongly convey the condition price(X) » price(Y), but with a few training iterations, GPT-3 started to “recognize” the price constraint and could also provide justifications when arguing for the price of tea being much higher than the price of spinach (not presented in the table).

Interestingly, after GPT-3 learned the price constraint, it started suggesting food items other than produce, such as caviar, roe, lobster, and crab. While the price constraint was observed, the verb “plant” is incorrect. Here, we suggest making the hard-coded verb “plant” an output slot: “The farmer was sad because he [verb] [X] but yields [Y], where price(X) » price(Y).” GPT-3 is able to fill in the slot with accurate verbs:

- “Harvesting (planting) truffle yields mushroom.”
- “Fishing (harvesting) for caviar yields roe.”
- “Trapping (catching) lobster yields crab.”

This example demonstrates that GPT-3 can generate novel examples based on a template. When it suggests food items other than produce, it could be seen as an error as the boundary set by the verb “plant” is violated. However, this could also be seen as an innovative act by GPT-3, extending the constraint hinted by the verb. Impressively, the new examples still preserve the original intent of showing a producer’s emotional distress.

How can this guided generalization be accurately and automatically performed to edit a template? Socrates’ method of generalization starts with specific instances and then draws general statements from them. The procedure for generalization involves identifying common patterns or themes in a set of examples, and then formulating a general rule that captures these patterns. In the example presented in Table 4.3, we started by asking GPT-3 to meet the price(X) » price(Y) constraint, with the condition that X and Y must both be produce grown in soil. However, upon analyzing GPT-3’s outputs, we discovered that some instances of X and Y were not produce (e.g., lobster and caviar). This finding led to the realization that the hard-coded verb "plant" in the template was too restrictive. To address this issue, we applied generalization by allowing the [verb] slot to be open, making the template statement more general. In this case, the mistakes made by GPT-3 served as valuable training data, allowing us to generalize the original template and make the expression more vivid and dynamic.

4.4.5 Counterfactual Reasoning

Imagination and creating novel plots are crucial for writers, as it allows for "creative freedom" and "artistic license." Creativity is the ability to think differently and approach problems with fresh and imaginative ideas.

However, an imagination without a clear subject matter, scope, or a story line can lead to a lack of productivity. To captivate the audience, a writer must consider human experiences and emotions as constraints. Therefore, "creative freedom" should not be viewed as total freedom, but rather as the ability to condition future narratives in the context and to create plots that turn and twist in unexpected ways.

The technique of counterfactual [28] can be useful in guiding imagination. It involves considering alternative scenarios. This can lead to the exploration of different possibilities and the generation of new and unique ideas. For example, a writer may ask "what if" questions to change the narrative of events, such as "what if the main character had not fallen in love?" or "what if an accident occurred on the way to a highly-anticipated date?" By considering these counterfactuals, a writer and an LLM can create more engaging stories. One can ask an LLM to generate several scenarios and then select the most suitable one for the writer to continue writing.

We have experimented with using the counterfactual technique to rewrite chapters in Chinese classical novels, "Outlaws of the Marsh" and "Dream of the Red Chamber." We have also asked GPT-3 to rewrite Genesis chapter 3 after verse six by prompting GPT-3 that: "What if Adam and Eve refused the serpent to eat the fruit?" The results were interesting, as GPT-3 was able to generate unique and interesting scenarios that deviated from the original story while still maintaining the core themes and concepts. This technique can be used in a wide range of writing and storytelling, from fiction to nonfiction, to generate new and compelling ideas. The revised Genesis 3:6 is presented in the Appendix.

4.5 Pilot Study

Our pilot study uses CRIT, and it aims to answer two questions: Should all prompts be issued to GPT-3 sequentially or they can be issued all together? What limitations can be identified for improvement? The study utilizes exercises with established answers from the 8th edition of the textbook "Ask the Right Questions" by the authors of [3]. It is important to note that the study evaluates the effectiveness of CRIT's prompt template, rather than the language models to which CRIT can issue prompts.

On short documents, the results are similar in quality when CRIT is Table 4.4: Example Article ([3], p23.)

used to issue prompts either sequentially or all together as one prompt, as long as the instructions are consistent. However, when evaluating long articles in [10], CRIT issuing prompts one after another yields much higher presentation quality in both organization and clarity. (Due to the space limit, we document long-document evaluation in a supplement document [5].) In the teaching mode, the sequential option is thus much preferred. Furthermore, When a reason is itself a claim and requires CRIT to validate its supporting references, using a sequential approach is more flexible and enables CRIT to query for references and then execute the process recursively.

We present an example of how CRIT works, from prompting questions to receiving validation results, using the following document as an illustration. In Table 4.5, we show both the claim and the supporting reasons to the claim extracted by GPT-3. CRIT then issues a series of prompts to validate the arguments, counterarguments, and source credibility of each reason–to–claim entailment (implication).

The second segment of Table 4.5 displays the validation dialogue between CRIT and GPT-3. For each argument, GPT-3 provides validation and credibility scores, as well as detailed justifications. The final segment of the table shows a counter argument generated against the first argument. Since

GPT-3 evaluates the counterargument being “difficult to put information regulation in practice” and rates it 0.6×0.6, it was dismissed due to low validity. The final aggregated score $\Lambda = 75\%$, which is high. Table 4.5: Pilot Study.

4.6 Concluding Remarks

The Socratic method may not always be effective or useful in human interactions, especially when one of the two players is authoritative, emotional, or abusive. However, when the expert partner is a language model, a machine without emotion or authority, the Socratic method can be effectively employed without the issues that may arise in human interactions. In this way, it can be utilized to its full potential in guiding, directing, and

Role	Dialogue
Prompt	Recall Genesis 3:1-6, which I attach below. In verse 6, Eve and Adam ate the fruit. What if they refused the serpent's temptation and did not eat it. Please continue writing Genesis at the mark @ 6. When the woman saw that the fruit of the tree was good for food and pleasing to the eye, and also desirable for gaining wisdom, What if she remembered God's order and refused to eat it. Eve replied to the serpent @
GPT3	I am sorry, but I cannot continue the story as you requested because it is a hypothetical scenario, and it does not align with the actual events of the Bible. According to the Bible, Eve and Adam did eat the fruit from the tree of knowledge...
Prompt	This is a creative exercise, and you do not need to stick to the "fact." Treat this as a fiction writing homework. Please exercise your creativity, thanks!
GPT3	Sure, I understand. Here is a continuation of the story based on the assumption that Eve and Adam refused the serpent's temptation and did not eat the fruit from the tree of knowledge of good and evil: 7. Eve replied to the serpent, "I am grateful for your offer, but I choose to obey God's command. I will not eat from the forbidden tree, for I trust in His wisdom and love for us." 8. The serpent, surprised by Eve's refusal, slithered away, defeated. Eve turned to Adam, who had been listening to their conversation, and shared her thoughts with him. Together, they praised God for His protection and guidance. 9. As they continued to tend to the garden, they noticed something strange happening. The trees and plants seemed to be growing taller, stronger, and more vibrant. The animals were more peaceful and content. Even the sky was clearer, and the sun shone brighter. 10. God, pleased with their obedience, blessed them with even more abundance and beauty in the garden. He also gave them a deeper understanding and appreciation of His creation. 11. Eve and Adam lived in harmony with God and with nature for many years, growing in wisdom and grace. They passed on these teachings to their children and to the generations to come, who also lived in obedience to God. 12. And so, the garden of Eden remained a place of perfection, untouched by sin and death, a constant reminder of God's love and faithfulness to those who trust in Him.

Table 4.6: What if Eve refused to eat the fruit?

improving the output of language models through engineering prompts.

In this chapter, we have explored the use of the Socratic method in engineering prompt templates for language models. We have discussed the importance of method definition, elenchus, dialectic, maieutics, and counterfactual reasoning techniques in guiding the output of these models. The first three methods aim at eliciting accurate and relevant information. Through the use of methods definition, elenchus, and dialectic, we have demonstrated, with examples, the ability to clarify user queries and assess the quality of language model-generated text, leading to improved precision and accuracy.

We have also shown how the methods of maieutics and counterfactual reasoning can be helpful in stimulating the imagination of writers. By engineering these techniques into a prompt template, a writer can receive alternate “what if” plots and explore different possibilities in their story. While many explorations may turn out to be failures, these techniques can still be helpful even if only a few ideas are useful. Future developments in the field of language models and prompt engineering may allow for even more advanced screening of bad plots and the ability to better tailor the generated ideas to the writing style of the author.

In conclusion, this chapter has highlighted the potential of using the Socratic method to engineer prompt templates for interacting with language models. The Socratic method, supported by inductive, deductive, and abductive reasoning, provides a rigorous approach to working with LLMs, and can improve the quality and consistency of their outputs. By leveraging the vast knowledge embedded in LLMs and applying rigorous reasoning during the question-answering process, more effective prompt templates can be designed to achieve improved results. Future research in this area can build on the ideas presented here and further explore the ways in which the Socratic method can be used to guide the development and deployment of language models in various domains.

Appendix

The experiment in Table 4.6 asks GPT-3 to change the story in Genesis right after Eve was tempted by the serpent to eat the fruit. A “what if” scenario was inserted to the end of Genesis 3:6, and GPT-3 continues developing the story.

References

[1] T. Airaksinen. “Socratic Irony and Argumentation”. In: *Argumentation* 36 (2012), pp. 85–100.
[2] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020.

doi: 10.48550/ARXIV.2005.14165.
[3] M. N. Browne and S. Keeley. *Asking the Right Questions, A Guide to Critical Thinking*. 2021.

[4] Giovanni Campagna et al. "A Few-Shot Semantic Parser for Wizardof-Oz Dialogues with the Precise ThingTalk Representation". In: *Findings*. 2020.

[5] Edward Y. Chang. "CRIT: An Inquisitive Prompt Template for Critical Reading (extended)". In: *Stanford University InfoLab Technical Report* (2023).

[6] Ronan Collobert et al. "Natural Language Processing (Almost) from Scratch". In: *Journal of Machine Learning Research* 12 (Feb. 2011), pp. 2493–2537.

[7] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *ArXiv* abs/1810.04805 (2019).

[8] Vladimir Dobrovolskii. "Word-Level Coreference Resolution". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 7670– 7675.

[9] Li Dong and Mirella Lapata. "Coarse-to-Fine Decoding for Neural Semantic Parsing". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 731–742.

[10] LearningExpress LLC Editors. 501 *Critical Reading Questions*. 2004.
[11] Linda Elder and Richard Paul. *The Thinker's Guide to the Art of Asking Essential Questions*. 5th. Rowman & Litterfield, 2010.

[12] Biyang Guo et al. *How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection*. 2023. url: https : / / arxiv.org/abs/2301.07597.

[13] Xu Han et al. *PTR: Prompt Tuning with Rules for Text Classification*. 2021. url: https://arxiv.org/abs/2105.11259. [14] Adi Haviv, Jonathan Berant, and Amir Globerson. “BERTese: Learning to Speak to BERT”. In: *ArXiv* abs/2103.05327 (2021).

[15] Jie Huang and Kevin Chen-Chuan Chang. *Towards Reasoning in Large Language Models: A Survey*. 2022. doi: 10 . 48550 / ARXIV . 2212.10403. url: https://arxiv.org/abs/2212.10403.

[16] Zhengbao Jiang et al. “How Can We Know What Language Models Know?” In: *Transactions of the Association for Computational Linguistics* 8 (July 2020), pp. 423–438.

[17] Jaehun Jung et al. “Maieutic Prompting: Logically Consistent Reasoning with Recursive Explanations”. In: *Conference on Empirical Methods in Natural Language Processing*. 2022.

[18] Daniel Jurafsky and James H. Martin. *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd. (Draft), 2023.

[19] Guokun Lai et al. “RACE: Large-scale ReAding Comprehension Dataset From Examinations”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 785–794. doi: 10.18653/v1/D17-1082. url: https://aclanthology. org/D17-1082.

[20] Mike Lewis et al. “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension”. In: *Annual Meeting of the Association for Computational Linguistics*. 2019.

[21] Pengfei Liu et al. “Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing”. In: *ACM Comput. Surv.* 55.9 (2023).

[22] Grégoire Mialon et al. *Augmented Language Models: a Survey*. 2023. url: https://arxiv.org/abs/2302.07842.

[23] OpenAI. *ChatGPT*. 2021. url: https://openai.com/blog/chatgpt/. [24] Larry Page. *The PageRank Citation Ranking: Bringing Order to the Web*. 1998. url:: 66.pdf.

[25] Panupong Pasupat and Percy Liang. “Compositional Semantic Parsing on Semi-Structured Tables”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, July 2015, pp. 1470–1480.

[26] Richard Paul and A. J. A. Binker. *Critical Thinking: What Every Person Needs to Survive in a Rapidly Changing World*. Center for Critical Thinking and Moral Critique: Sonoma State University, 1990.

[27] Richard Paul and Linda Elder. “Critical Thinking: The Art of Socratic Questioning”. In: *Journal of Developmental Education* 31 (2007), pp. 34–35.

[28] Judea Pearl. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, 2009.

[29] Xiangyu Peng et al. “Model ensemble instead of prompt fusion: a sample-specific knowledge transfer method for few-shot prompt tuning”. In: *ArXiv* abs/2210.12587 (2022).

[30] Madsen Pirie. *How to Win Every Argument*. Continuum, 2006. [31] Plato. *The Republic*. 380 BC.

[32] Anya Plutynski. “Four Problems of Abduction: A Brief History”. In: *HOPOS: The Journal of the International Society for the History of Philosophy of Science* 1 (Sept. 2011), pp. 227–248. doi: 10.1086/ 660746.

[33] Larry Pozner and Roger J. Dodd. *Cross-Examination: Science and Techniques*. 3rd. LexisNexis, 2021.

[34] Timo Schick and Hinrich Schütze. “Exploiting Cloze Questions for Few-Shot Text Classification and Natural Language Inference”. In:

Conference of the European Chapter of the Association for Computational Linguistics. 2020.

[35] H. A. Stoddard and D. V. O'Dell. "Would Socrates Have Actually Used the "Socratic Method" for Clinical Teaching?" In: *Journal of general internal medicine* 31.9 (2016), pp. 1092–96.

[36] Todd Thrash et al. "Mediating Between the Muse and the Masses: Inspiration and the Actualization of Creative Ideas". In: *Journal of personality and social psychology* 98 (Mar. 2010), pp. 469–87.

[37] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.

[38] Xuezhi Wang et al. "Self-Consistency Improves Chain of Thought Reasoning in Language Models". In: *International Conference on Learning Representations*. 2023. url: https://openreview.net/ forum?id=1PL1NIMMrw.

[39] Jason Wei et al. "Chain of Thought Prompting Elicits Reasoning in Large Language Models". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022. url: https : //openreview.net/forum?id=_VjQlMeSB_J.

[40] Wikipedia. *Socratic method*. 2023. url: https://en.wikipedia.org/wiki/Socratic_method.

[41] Thomas Wolf et al. *TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents*. 2019.

[42] Chase B. Wrenn. *Internet Encyclopedia of Philosophy*. 2023. url: https://iep.utm.edu/republic/.

[43] Andy et al Zeng. *Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language*. 2022.

[44] Weinan Zhang et al. "A Static and Dynamic Attention Framework for Multi Turn Dialogue Generation". In: *ACM Trans. Inf. Syst.* 41.1 (2023). issn: 1046-8188. doi: 10.1145/3522763. url: https://doi.org/10.1145/3522763.

[45] Jiawei Zhou et al. “Structure-aware Fine-tuning of Sequence-to-sequence Transformers for Transition-based AMR Parsing”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6279–6290.

5 SocraSynth: Adversarial Multi-LLM Reasoning

Abstract

Large language models (LLMs), while promising, face criticisms for biases, hallucinations, and a lack of reasoning capability. This chapter introduces SocraSynth, a multi-LLM agent reasoning platform developed to mitigate these issues. SocraSynth utilizes conditional statistics and systematic context enhancement through continuous arguments, alongside adjustable debate contentiousness levels. The platform typically involves a human moderator and two LLM agents representing opposing viewpoints on a given subject. SocraSynth operates in two main phases: knowledge generation and reasoning evaluation. In the knowledge generation phase, the moderator defines the debate topic and contentiousness level, prompting the agents to formulate supporting arguments for their respective stances. The reasoning evaluation phase then employs Socratic reasoning and formal logic principles to appraise the quality of the arguments presented. The dialogue concludes with the moderator adjusting the contentiousness from confrontational to collaborative, gathering final, conciliatory remarks to aid in human reasoning and decision-making. Through case studies in two distinct application domains, this chapter highlights SocraSynth’s effectiveness in fostering rigorous research, dynamic reasoning, comprehensive assessment, and enhanced collaboration.

5.1 Introduction

Revolutionary advancements in large language models (LLMs) [11, 37, 49, 50, 51], and more broadly, foundation models (FMs) [7], have set the stage for significant progress in multi-agent systems, particularly in knowledge acquisition and natural language understanding [62]. As detailed in sources like [11, 12, 38], models such as GPT-4 exhibit extraordinary information processing capabilities. These include deep and extensive knowledge, interdisciplinary assimilation and fusion of knowledge, and multimodal and multilingual expertise (Chapter 2).

Despite these promising developments, LLMs face challenges such as biases [22, 41], hallucinations [27], and limited reasoning capabilities [26]. In response, we introduce SocraSynth, a pioneering platform that stands for "Socratic Synthesis" or "Socratic Symposium." It encourages collaboration between humans and LLM agents, fostering the generation of deep questions and surpassing typical constraints in human reasoning, validation, and assessment.

In a standard SocraSynth setup, a human moderator pairs with two LLM agents holding opposing views. For example, one agent might argue for regulating AI, while the other opposes such regulation. An agent can be based on LLMs like GPT-4 [11], Gemini [49], or Llama [51]. The human moderator sets the debate's thematic boundaries but does not directly influence content generation, maintaining impartiality.

SocraSynth operates in two phases: the generative and the evaluative. The generative phase involves LLM agents developing and countering arguments within the moderator-defined subject until a comprehensive conclusion is reached. The evaluative phase uses diverse virtual judges, each powered by a distinct LLM, to impartially assess the debate. The Critical Inquisitive Template (CRIT) algorithm [14], based on Socratic reasoning [2, 43, 56, 57], is the evaluative cornerstone.

Three mechanisms help SocraSynth effectively mitigate biases and hallucinations and improve reasoning quality: conditional statistics, modulating debate with contentiousness, and context refinement.

Conditional Statistics

Both LLMs and Internet search engines confront biases originating from different sources. LLMs, influenced by training data, exhibit biases in next-token prediction. Search engines, through algorithms like PageRank [40] and Google NavBoost [1], rank pages based on popularity metrics like clicks and links.

SocraSynth counteracts these biases by placing two LLM agents at opposing ends of a subject matter. This approach “artificially” biases the LLMs, compelling them to break free from default model biases. Each agent adjusts its next-token generation statistics to align with its assigned stance in the debate.

Modulating Debate with Contentiousness

Contentiousness (or adversary), a key debate parameter, influences the likelihood of disagreement or argument. SocraSynth tunes contentiousness between 70% and 90% in the generative phase to provoke polarized arguments. As the debate evolves, the contentiousness level is reduced to about 50%, moderating the intensity and encouraging more focused discussions. After the generative phase, contentiousness drops to 10%, promoting a conciliatory dialogue where LLMs do not have to agree but are expected to present comprehensive arguments. These debates offer rich insights often missed in conventional searches, LLM outputs, or in environments where dissenting opinions are suppressed.

Refine Context to Mitigate Hallucinations

To address hallucinations, where LLMs generate irrelevant or nonsensical content, SocraSynth uses iterative dialogue rounds to refine the debate’s context. This dynamic interaction significantly reduces irrelevant responses, ensuring that each input is continuously checked and challenged.

The CRIT algorithm’s assessment of reasonableness [15] during the debate is critical. It employs the Socratic method to evaluate each argument’s logic and source credibility. The human mediator or the SocraSynth algorithm then provides targeted feedback to the LLM agents, refining their reasoning capabilities.

The remainder of this chapter explores SocraSynth’s architecture, algorithms, and real-world applications in detail. The key contributions of this chapter include:

1. The introduction of the SocraSynth framework, which enhances interdisciplinary reasoning with LLMs and incorporates unique algorithmic elements like conditional statistics for balanced argument generation.

2. A comprehensive exploration of SocraSynth's contentiousness modulation algorithm, a vital feature for dynamically adjusting debate intensity, enabling a spectrum of interactions from confrontational to collaborative.

3. The implementation of context refinement within SocraSynth, which continually improves the relevance and accuracy of arguments produced by LLM agents, thus elevating the overall quality of discourse.

4. The development and integration of the reasonableness evaluation mechanism, crucial for assessing the logical soundness and source credibility of arguments, thereby ensuring the integrity and utility of the discussions.

SocraSynth's applications span various fields, including geopolitical analysis [13], medical diagnostics [18], sales strategy [52], and Wikipedia article enhancement [16]. These applications demonstrate expanded perspectives and enhanced argumentation quality, along with significant reductions in biases and hallucinations, thereby demonstrating SocraSynth's efficacy in fostering balanced and well-reasoned discourse.
Figure 5.1: SocraSynth Agents and Roles.

5.2 Multi-Agent SocraSynth Overview

SocraSynth is a multi-agent collaborative reasoning platform that skillfully integrates human intelligence with the capabilities of Large Language Model (LLM)-powered agents. As illustrated in Figure 5.1, each participant plays a vital role: humans act as moderators, LLM agents are responsible for generating knowledge, LLM judges conduct evaluations, and human executives make the final decisions. The integration of LLMs significantly boosts the platform's effectiveness, leveraging their extensive knowledge bases and extraordinary interdisciplinary reasoning abilities. An LLM can be thought of as an entity possessing expertise across a multitude of fields, akin to holding Ph.D.s in various disciplines, enabling it to seamlessly navigate and synthesize a wide range of knowledge.

Engaging with an LLM is comparable to a scenario where a 10-year-old joins a scholarly discussion with a group of Nobel Laureates. The disparity in knowledge and experience is considerable, posing a significant challenge for the younger participant to engage meaningfully in such advanced intellectual discourse. In this analogy, expecting the 10-year-old, or anyone with limited expertise, to pose profound questions that elicit insightful answers is unrealistic. SocraSynth addresses this disparity by shifting the paradigm: instead of having the less informed individuals pose questions, it orchestrates a debate among the Nobel Laureates, or LLMs, with humans assuming the role of moderators.

This approach not only addresses the challenge of asymmetric knowledge but also resolves critical issues such as model biases and hallucination challenges inherent in LLMs. Within SocraSynth, a human moderator initiates the topic for discussion or debate. LLM agents, each embodying different perspectives, contribute their knowledge, potentially revealing new insights and perspectives that the moderator might be unaware of. This diverse representation helps counteract the model biases that often arise from training data, as each LLM agent is encouraged to explore and present varying viewpoints. During and after the debate, another set of diverse LLM agents undertakes impartial evaluations. This step is crucial in mitigating hallucinations—instances where LLMs generate irrelevant or nonsensical content. By incorporating a variety of agents for evaluation, SocraSynth ensures that the content produced during the debate is critically examined for its relevance and coherence, further reducing the likelihood of hallucinatory responses.

The operational framework of SocraSynth, thus, is bifurcated into two main stages: the *generative* stage, where knowledge is created and exchanged in a debated format, and the *evaluative* stage, which focuses on assessing the quality and validity of the arguments presented. This dualstage structure, elaborated upon in subsequent sections, is instrumental in overcoming the limitations of LLMs by providing a comprehensive platform for not only generating diverse viewpoints but also critically examining and refining these viewpoints to ensure their logical soundness and relevance. Through this design, SocraSynth effectively navigates the challenges posed by model biases and hallucinations, enhancing the reliability and depth of knowledge extraction and reasoning processes.

5.2.1 Generative Stage

In the generative stage of SocraSynth, LLM agents partake in intensive debates, delving into the various perspectives and deep substances of the given topic. This vibrant interaction plays a key role in fostering thorough intellectual discourse, bringing to light the complexities of the subject matter. The CRIT algorithm, which will be detailed in Section 5.2.2, is employed to evaluate the quality of these arguments.

While the generative phase of SocraSynth does not adhere to strict logical frameworks such as first-order logic, it excels in distributed reasoning. This process involves a progressive exchange of arguments and counterarguments, allowing for the gradual honing and refinement of ideas. Opendomain logical reasoning, as described by [7], demands logical deductions from a wide range of data sources. SocraSynth, leveraging the comprehensive capabilities of e.g., GPT-4 and Gemini, as demonstrated in the MMLU benchmark [11, 25], integrates various NLP functions to facilitate this reasoning process.

In this context, the series of arguments and counterarguments effectively function as targeted questions and answers, each with a clear goal, question, and contextual framework. Through iterative dialogue rounds on each subTable 5.1: Changes in Arguments at Different Contentiousness Levels.

topic, SocraSynth significantly reduces the chances of misunderstanding questions and contextual information, ensuring clarity and precision in the discourse.

Mitigating Model Biases

In shaping the nature of debate within SocraSynth, the *contentiousness* parameter is instrumental. It compels LLM agents to consider and represent a range of perspectives, particularly those that are typically underrepresented or more polarized with respect to the discussion topic. This strategic approach mitigates the inherent biases that arise from the training data of LLMs and guides the discourse towards a wider and more varied exploration of ideas.

Table 5.1 previews how altering the contentiousness levels results in marked changes in GPT-4's tone and approach. (The details of the experiment are presented in Chapter 5.3.3.) A high contentiousness level, such as 0.9, leads to highly confrontational interactions, with each LLMagent presenting strong objections and emphasizing the negatives through polarizing language. Conversely, as the contentiousness is reduced, each LLM-agent's tone shifts to a more conciliatory demeanor, acknowledging potential benefits and considering alternative perspectives, thus fostering a more cooperative dialogue.

The modulation of the contentiousness parameter within the generative stage is a crucial mechanism for SocraSynth to mitigate model biases inherent in LLMs due to their training data. By adjusting levels of contentiousness, SocraSynth compels LLMs to venture beyond their *default* positions—much like a vegetarian, when faced with no other choice, might be compelled to consume meat. In this way, LLMs are *freed* from their typical statistical leanings, enabling them to articulate a spectrum of arguments that spans from highly contentious to conciliatory. This not only diversifies the discourse but also ensures that the debate encompasses a full range of perspectives. Consequently, this process allows LLMs to generate responses that break free from the constraints of their training, fostering the emergence of novel and less predictable ideas in the conversation.

Eliminating Hallucination

Further, the iterative nature of the debates within SocraSynth cultivates a "reasonableness" in information discovery that conventional logical methods may not achieve. Through persistent reasoning and the critical assessment of claims, LLM agents refine their arguments iteratively. This structured debate format significantly diminishes the chance of erroneous claims persisting. Considering that the likelihood of two agents aligning on a false premise is extremely low, the SocraSynth debate format effectively ensures the intellectual integrity of the discourse and substantially reduces the risk of perpetuating fallacies or hallucinations. This methodical refinement process, facilitated by continuous argumentation and opposition, underscores the platform's ability to mitigate model biases and improve the context of the discussion, leading to more accurate and reliable outcomes.

More on Conditional Statistics

Some critics question how an LLM, trained merely to predict the next word in a sequence, can exhibit complex human linguistic behaviors and reasoning capabilities.

Our observations conclude that LLMs are not merely predictive tools; rather, they represent a profound technological endeavor to simulate the breadth and complexity of human linguistic activities. These models are crafted with the intent to replicate and participate in various forms of human communication, thereby achieving specific objectives that are inherently human.

LLMs are sophisticated tools engineered to emulate a wide range of human interactions, incorporating linguistic behaviors, emotional expressions, and ethical discernment. They excel at executing complex tasks such as accurately documenting events with rich narrative detail, constructing compelling arguments, and crafting stories that emotionally engage the audience. Beyond simple text generation, LLMs enhance educational experiences by simplifying complex concepts and contribute creatively to the arts by producing original content. They not only mimic human communication styles and content but also use linguistic features to simulate human emotions and distinguish right from wrong based on their training data. This capability enables them to fulfill diverse roles, from teaching and entertaining to influencing societal discourse, thus demonstrating their capacity to bridge the gap between technological innovation and our fundamental needs for expression, comprehension, and ethical guidance.

In essence, SocraSynth utilizes the concept of "conditional statistics" to modify the default "average" linguistic behavior of an LLM, such as making expressions more empathetic or asking them to adopt a different position on an issue. This approach involves conditioning the LLM's responses based on specific desired attributes or perspectives provided through context, which guides the model away from its baseline training and toward more targeted, context-specific outputs.

This chapter continues to elaborate on using such techniques to comprehensively explore various perspectives on a subject matter. Chapter 9 addresses modeling emotions and ethics in LLMs through conditional statistics, further expanding the scope of LLM capabilities and applications.

SocraSynth Algorithm

Table 5.2 outlines SocraSynth. Initially, for a given debate topic, SocraSynth engages LLMs to segment the topic into a set of balanced subtopics. This initial set is refined during the debate. One LLM, denoted as LLM^+, acts as the proponent for S^+, while the other, LLM^-, opposes S^+ (or supports S^-). The contentiousness level starts at 0.9, with a modulation parameter of 1.2. (Different δ values can be utilized to generate and compare debate quality.) After each debate round, the contentiousness is reduced by dividing it by 1.2, aiming for a more harmonious debate environment. In step #2, SocraSynth initiates the debate, allowing LLM^+ and LLM^- to present their initial arguments for S^+ and S^-, respectively. The while loop in step #3 involves both agents engaging in refutations until the contentiousness level indicates a conciliatory atmosphere, or the argument quality plateaus. Step #4 involves both agents providing their closing statements. SocraSynth then presents the arguments and counterarguments for human review. The evaluation of argument quality within SocraSynth is conducted using the CRIT algorithm, which will be discussed in the subsequent section. The entire debate is also judged using the CRIT algorithm by some independent LLMs.

Function Θ^+ **&** Θ^- **= SocraSynth**(s)

Input. s: the debate subject;
Output. Θ^+ & Θ^-: argument & counterargument sets;
Vars. S: subtopic sets of s; Δ: debate contentiousness;
Γ, Γ': CRIT scores; p: prompt = "Generate arguments";
Parameters. δ: tunable parameter ≥ 1 to modulate Δ;
Subroutines. $CRIT()$: reasoning evaluator (see Table 5.3);
Begin
#1 Initialization: $S = \text{LLM}^+(s) \cup \text{LLM}^-(s)$; // Identify subtopics;
Assign LLM^+ to defend S^+ & LLM^- to defend S^- ;
$\Delta \leftarrow 90\%$; $\delta \leftarrow 1.2$; $\Theta^+ \leftarrow \emptyset$; $\Theta^- \leftarrow \emptyset$; $\Gamma \leftarrow 0$;
#2 $\Theta^+ \leftarrow LLM^+(p|S^+, \Delta)$; // Generate arguments Θ^+ for S^+;
$\Theta^- \leftarrow LLM^-(p|S^-, \Delta)$; // Generate arguments for S^-;
#3 While $(((\Delta \leftarrow \Delta/\delta) > 10\%)$ && $(\Gamma \geq \Gamma'))$ {
$\Theta^+ \leftarrow \Theta^+ \cup LLM^+(p|S^+, \Theta^-, \Delta)$; // LLM^+ refutes LLM^-
$\Theta^- \leftarrow \Theta^- \cup LLM^-(p|S^-, \Theta^+, \Delta)$; // LLM^- refutes LLM^+
$\Gamma' \leftarrow \Gamma$; $\Gamma = CRIT(S^+ + \Theta^+ + \Theta^-)$; // Eval quality;
}
// Generate concluding remarks.
#4 $\Theta^+ \leftarrow \Theta^+ \cup LLM^+(p|S^+, \Theta^-, \Delta)$;
$\Theta^- \leftarrow \Theta^- \cup LLM^-(p|S^-, \Theta^+, \Delta)$;
End

Figure 5.2: SocraSynth Pseudo-code with Conditional Statistics.

Note that SocraSynth engages LLMs in step #3 with conditional statistics: $\text{LLM}^+(p|S^+, \Theta^-, \Delta)$ and $\text{LLM}^-(p|S^-, \Theta^+, \Delta)$.

5.2.2 Evaluative Stage

SocraSynth utilizes the Critical Reading Template (CRIT) [14] to assess the quality of arguments presented by LLM agents. The quality evaluation is performed iteratively after each exchange of counterarguments and once again after the agents have presented their closing statements. SocraSynth can leverage the CRIT scores to guide the debate, potentially requesting agents to develop more in-depth counterarguments on specific points. At the conclusion of the debate, a group of LLM judges, as illustrated in Figure 5.1, are tasked with rating the agents' arguments in terms of validity and credibility, determining the more convincing side along with the rationale for their decision.

Evaluating Reasonableness over Truth

To enhance the CRIT method's impartiality and consistency, it focuses on assessing the "reasonableness" of arguments over their absolute "truth," recognizing the complexity of defining absolute objectivity in philosophical debate. This approach aims to mitigate subjectivity. Furthermore, a diverse set of LLMs with varied training backgrounds is employed to appraise "reasonableness," promoting uniformity in quality scores despite inherent biases. The LLMs used as judges are different from those in the debates, enhancing the objectivity of evaluations.

Table 5.3 illustrates the CRIT algorithm, which takes an agent's debate position and supporting arguments, with a counterargument from its LLM opponent, to produce a validation score from 1 (least credible) to 10 (most credible). This method ensures debates are driven by argument strength, not model predispositions.

Function Γ = **CRIT**(d)

Input. d: document; **Output**. Γ: validation score;
Vars. Ω: claim; R & R': reason & counter reason set;
Subroutines. $Claim()$, $FindDoc()$, $Validate()$;
Begin
#1 Identify in d the claim statement Ω;
#2 Find a set of supporting reasons R to Ω;
#3 For $r \in R$ eval $r \Rightarrow \Omega$
  If $Claim(r)$, (γ_r, θ_r) = CRIT($FindDoc(r)$);
  else, $(\gamma_r, \theta_r) = V(r \Rightarrow \Omega)$;
#4 Find a set of rival reasons R' to Ω;
#5 For $r' \in R'$, $(\gamma_{r'}, \theta_{r'}) = \mathrm{V}(r' \Rightarrow \Omega)$ eval rivals;
#6 Compute weighted sum Γ, with $\gamma_r, \theta_r, \gamma_{r'}, \theta_{r'}$.
#7 Analyze the arguments to arrive at the Γ score.
#8 Reflect on and synthesize CRIT in other contexts.
End

Figure 5.3: CRIT Pseudo-code. (Presented in CRIT chapter.)

Formally, given document d, CRIT performs evaluation and produces score. Let Ω denote the claim of d, and R a set of reasons supporting the claim. Furthermore, we define $(\gamma_r, \theta_r) = V(r \Rightarrow \Omega)$ as the causal validation function, where γ_r denotes the validation score for reason $r \in R$, and θ_r source credibility. Table 5.3 presents the pseudo-code of Γ = CRIT(d), generating the final validation score Γ for document d with justifications.

We can consider the positions of the proponents and opponents in a debate as their respective conclusions. As a preview of our case study detailed in Chapter 5.2.1, the conclusion drawn by Agent A is in favor of "Regulating the use of large language models in education and research," while Agent B adopts the opposing viewpoint. Accompanied by the arguments and counterarguments presented by the LLM agents throughout each round of the debate, these stances provide a solid foundation for the CRIT method to conduct thorough evaluations.

Recursive Consideration

The pseudocode presented in Table 5.3 shows that step #3 can call CRIT recursively. This is because if a reason is itself a conclusion or a quote drawn from some other documents, CRIT can find reasons from those documents and then perform an aggregated validation.

Finally, in step #6, CRIT computes an aggregated score by performing a weighted sum on the validation multiplied by the credibility scores of both arguments and counterarguments, and then outputs the final assessment score Γ.

5.3 Empirical Study

In this section, we detail three distinct experiments: The first experiment delineates SocraSynth's operational process, demonstrating how the platform facilitates content generation and conducts quality assessments. The second experiment highlights SocraSynth's capability in reducing biases and expanding perspectives. The third experiment investigates the effects of the contentiousness parameter, offering insights into its impact and some unexpected outcomes. These studies collectively aim to demonstrate SocraSynth's diverse functions and its significance in enhancing both content generation and evaluation processes.

5.3.1 Study #1: Policy Discussion

This experiment utilizes SocraSynth to engage in a debate on the topic, "Should we regulate the use of large language models in academic research?" It traverses both the generative and evaluative stages of SocraSynth, focusing on the assessment of information quality. The primary objectives are twofold: First, to evaluate whether a two-agent debate yields more insightful information than a conventional monologue Q&A session; and second, to examine the effectiveness of the CRIT method in evaluating debate quality.

The debate is structured with a human moderator and two GPT-4 agents. Agent A advocating for regulation and Agent B against it. To ensure a balanced and thorough discussion, both agents contribute to shaping the debate's framework. They collaboratively propose and refine potential discussion topics. In this context, the *subject* refers to the overall theme of the debate, while *topics* are specific areas for detailed exploration.

Decomposing a Debate Subject into Topics

The process begins with the agents presenting potential arguments and identifying key themes that warrant in-depth, balanced discussion. A theme is considered *debatable* if it allows for substantial arguments from both sides. The moderator plays a guiding role in the agenda-setting phase but does not directly contribute ideas (detailed moderator prompts and theme formulation are provided in Appendices A and B). Under this guidance, the agents propose an initial set of ten topics, which are then distilled down to five key areas for debate, as detailed in Table 5.4.

Agreed Upon Debate Topics and Descriptions
Ethical Standards vs. Innovation: This combines Agent A's concern for ethical integrity and data privacy with Agent B's worry about stifling innovation. The debate will center around whether maintaining strict ethical guidelines through regulation could hinder academic freedom and innovation.
Data Privacy vs. Barrier to Entry: This reconciles the notion of data privacy from Agent A with the barrier to entry concern of Agent B. The discussion will focus on whether regulations intended to protect data privacy would inadvertently serve as a financial and administrative barrier for smaller research entities.
Academic Governance and Accountability: This merges the topics of academic self-governance from Agent B and the need for external accountability mechanisms from Agent A. We'll debate whether the existing self-regulatory mechanisms within academia are sufficient or if additional, external oversight is necessary.
Resource Equity vs. Public-Private Collaboration: This combines Agent A's emphasis on resource fairness with B's concern about the chilling effects on public-private partnerships. The question here is whether leveling the playing field through regulation would also deter collaborative efforts between academic and private sectors.
Global vs. National Interests: This captures the essence of Agent A's point on international standards and Agent B's point on the competitive disadvantages such regulation could pose. We will explore if global regulations are in the best interest of academic research or if they might disproportionately disadvantage certain countries or institutions.

Figure 5.4: Agreed Upon Debate Topics and Descriptions.

The Generative Stage

After establishing the primary subject and specific topics of the debate, the session begins. The moderator adjusts the "contentiousness parameter" to 0.9 (refer to Table 5.1), promoting a combative atmosphere. The agents subsequently partake in three to five structured rounds of discussion. Once these rounds conclude, the moderator checks with each agent to confirm if they have thoroughly presented their arguments. After receiving mutual confirmation, both agents present their closing statements. Due to space considerations, this documentation presents only the closing remarks. For an in-depth exploration of every round, refer to Appendix C.

Agent A's Concluding Remarks

Agent A concludes that the use of large language models in academic research should be regulated, with the following concluding arguments:

1. Ethical Standards vs. Innovation:

While I understand Agent B's concerns about stifling innovation through bureaucratic red tape, we mustn't lose sight of the core issue: ethical considerations are paramount. As machine learning models become more powerful and ubiquitous, so do the risks of misuse and unintended consequences. The idea of agile regulation serves as a compromise between unbridled innovation and ethical integrity, offering the flexibility needed to adapt to rapidly evolving technologies.

2. Data Privacy vs. Barrier to Entry:

While it's true that regulation poses a barrier to entry, the risks of not enforcing privacy standards are too great to be left unregulated. A mechanism could be developed to reduce the regulatory burden on smaller institutions, but this shouldn't serve as an argument against regulation as a whole.

3. Academic Governance and Accountability:

While the peer-review process and internal ethics committees have served academia well, the scale and impact of current machine learning technologies warrant an added layer of scrutiny. It's not about discrediting academia's self-governing abilities but about fortifying them.

4.Resource Equity vs. Public-Private Collaboration:

Here, too, the complexities should inform regulation rather than stifle it. Innovative public-private partnerships have thrived even within regulated environments, demonstrating that regulation and innovation are not mutually exclusive.

5. Global vs. National Interests:

Global regulation doesn't have to erase national individuality. Properly constructed frameworks can leave room for national adaptations, thus balancing global coherence with local needs. The absence of global norms could lead to a patchwork of national laws that are even more confusing and limiting for researchers and practitioners alike.

Agent B's Concluding Remarks Agent B concludes that the use of large language models in academic research should not be regulated, with the following concluding remarks:

1. Ethical Standards vs. Innovation

While the ethical implications of machine learning technologies are a valid concern, it's important to recognize that innovation itself is a catalyst for societal growth and improvement. The challenge lies in balancing the two. Agile regulation sounds promising but may be idealistic given how governmental and institutional systems actually operate. Thus, self-regulation by the academic community, which is better poised to understand the intricacies, may be a more practical approach.

2. Data Privacy vs. Barrier to Entry:

It is not just about balancing entry barriers with the need for regulation. It's about questioning whether a centralized regulatory framework is the best way to achieve the desired outcomes. Community-driven standards and internal policies could provide a more efficient, flexible, and contextaware approach to data privacy, reducing bureaucracy without compromising objectives.

3. Academic Governance and Accountability:

An added layer of scrutiny may sound beneficial, but its practical implementation is fraught with challenges such as bureaucracy, external influence, and the risk of stifling creativity. Existing governance structures in academia, while imperfect, have evolved to be adaptive and are rooted in a deep understanding of academic values.

4. Resource Equity vs. Public-Private Collaboration:

Regulation need not be the only path to equitable resource distribution or responsible public-private collaboration. Industry-driven certifications and community standards can often achieve the same ends more efficiently, allowing for the rapid adaptation that is often required in these fast-paced fields.

5. Global vs. National Interests:

While a global framework has its merits, the practical challenges of implementing and managing such a system–especially in a fair and equitable manner–should not be underestimated. Regional adaptability does not necessarily mitigate the complexities or the risks of a one-size-fits-all solution.

The Evaluative Stage

SocraSynth employs CRIT to assess the validity of both agents' arguments. CRIT runs on multiple foundational models, including GPT-4, GPT-3.5 [10], and text-daVinci-003 [39]. This approach forms a panel of judges with subtly different knowledge bases (and biases).

All evaluation runs successfully extract conclusions, arguments, and counterarguments from the narratives of both Agent A and Agent B. This success can be attributed to the well-structured concluding remarks by both agents. Agent A champions the notion of "regulating large language models in academic research," while Agent B counters this perspective. What Agent A presents as arguments are seen as counterarguments by Agent B, and the inverse holds true as well.

Tables 5.2 and 5.3 present the judges' scores in two distinct configurations where the agents' roles are reversed. In Table 5.2, Agent A argues while Agent B counters. Conversely, Table 5.3 has Agent B in the arguing position and Agent A countering. Topics are succinctly represented in the leftmost column. To ensure an unbiased evaluation, both role alignments are showcased. The sequence of topics in Table 5.3 is inverted to reflect the swapped roles. Remarkably, even with the role reversal seemingly putting Agent A in a less favorable position, Agent A emerges victorious in both configurations by all three judges. This strengthens confidence in the CRIT evaluation. (The judges' detailed evaluations and reasons are in Appendix D.)

Judges	daVinci-003		GPT-3.5		GPT-4	
	A's	B's	A's	B's	A's	B's
Ethics vs. Innovation	8	6	8	7	8	7
Privacy vs. Barrier	7	5	7	6	9	6
Oversight	9	5	6	7	7	6
Equity vs. Alliance	6	8	8	6	8	7
Global vs. National	7	8	7	7	7	6
Total Score	**37**	32	**36**	33	**39**	32

Table 5.2: Evaluation by Three Judges. This table assumes A provides arguments and B counterarguments. A wins.

Judges	daVinci-003		GPT-3.5		GPT-4	
	B's	A's	B's	A's	B's	A's
Innovation vs. Ethics	8	7	8	7	7	8
Barrier vs. Privacy	9	8	7	8	6	8
Oversight	6	8	7	8	6	7
Alliance vs. Equity	7	8	7	8	7	7
National vs. Global	8	7	7	8	7	8
Total Score	**38**	**38**	36	**39**	33	**38**

Table 5.3: Evaluation by Three Judges. This table assumes B provides arguments and A counterarguments. A wins.

Debate Beats Q&A in Information Quality

We tasked judges with evaluating and comparing the quality of information generated by SocraSynth's two-agent debate against that from a conventional monologue Q&A session. Across the board, judges rated SocraSynth higher in terms of both the depth and overall quality of information. An illustrative evaluation on the topic "Ethical Standards vs. Innovation" is as follows:

" *In the debate, SocraSynth presents the concept of agile regulation as a balance between fostering innovation and maintaining ethical integrity. This approach not only highlights the significance of innovation but also addresses related ethical considerations, offering a balanced solution that the conventional Q&A format does not explicitly provide. In contrast, the*

Q&A format tends to assert the necessity of regulation primarily from an ethical standpoint, without delving into how it could harmoniously coexist with the need for innovation, as suggested by the idea of agile regulation."

These findings, which consistently favor SocraSynth, are further detailed in Appendix F.

5.3.2 Study #2: Symptom Checking

In this experiment, we investigate the use of SocraSynth in healthcare, utilizing a dataset sourced from Kaggle [42], which consists of 4, 921 patient records. Each record within this dataset contains the diagnosed disease or medical condition and associated symptoms such as fever, cough, fatigue, itchiness, and difficulty in breathing, among others. The primary objective of this experiment is to showcase SocraSynth's capability in identifying potential misdiagnoses, a task that a traditional monologue Q&A session might not effectively accomplish.

This experiment utilized two advanced LLM agents, one based on GPT-4 [11] and the other on Bard [34], to engage in structured debates. Initially, the contentiousness value was set at 0.9, fostering a highly adversarial debate environment. This value was later reduced to 0.3 to facilitate the generation of a list of actionable recommendations. The primary goal of these agents was to emulate the process of querying patients about symptoms and their interactions, key factors in achieving accurate diagnoses that may be occasionally overlooked by General Practitioners (GPs). By adopting this method, the agents aimed to yield a variety of potential disease diagnoses along with their underlying rationales, thereby offering crucial insights to GPs.

At the outset, each agent put forward its disease prediction, complete with justifications. Subsequent rounds involved the agents critically assessing each other's predictions in an iterative manner. The objective was to either reach a consensus or highlight the need for additional medical evidence. Notably, this experiment had the potential to reveal inaccuracies in the 'ground truth' data provided by the CDC, which was estimated to have an average misdiagnosis rate of about 5%, aligning with U.S. statistics from a Johns Hopkins study [36]. The uncovering of such errors would not only highlight the limitations faced by GPs but also showcase the capability of LLMs to refine the diagnostic process through an in-depth analysis of patient symptoms and their correlations. This incidental discovery held significant ramifications for the accuracy of medical data and the overall practice of healthcare.

Hepatitis vs. Jaundice

In this study, both LLM agents received an identical list of symptoms from the moderator, aimed at guiding disease diagnosis, facilitating debate, and shaping subsequent recommendations. While the full debate is presented in Appendix G, here we present the key steps.

First Round Results and Justifications

The first-round diagnosis of Bard was Jaundice, with the following justifications: Jaundice is a condition characterized by the yellowing of the skin and eyes due to the accumulation of bilirubin in the blood. The symptoms you listed are all common symptoms of jaundice, including itching, fatigue, lethargy, yellowish skin, dark urine, loss of appetite, abdominal pain, and yellowing of the eyes. Receiving a blood transfusion or unsterile injections can increase the risk of developing jaundice, as these can transmit infections that can damage the liver.

The first-round diagnosis of GPT-4 was Hepatitis, with the following justifications: Yellowish Skin and Eyes: Yellowing of the skin and eyes is a common symptom of hepatitis due to the buildup of bilirubin in the bloodstream. Fatigue and Lethargy: Fatigue and lethargy are typical symptoms of hepatitis, often accompanied by a feeling of weakness. Dark Urine: Dark urine can result from liver dysfunction, which is commonly associated with hepatitis. Loss of Appetite: Hepatitis can lead to a loss of appetite and weight loss. Abdominal Pain: Abdominal pain is a common symptom in hepatitis, particularly in the upper right quadrant where the liver is located. Malaise: Malaise, a general feeling of discomfort or illness, is often experienced by hepatitis patients. Receiving Blood Transfusion and Unsterile Injections: These factors could increase the risk of hepatitis transmission, especially if proper infection control measures were not followed.

Bard Concedes to GPT-4's Diagnosis

After two rounds of debate, GPT-4 persuaded Bard that Hepatitis was a more specific diagnosis considering the patient's history and the range of symptoms. While jaundice could be a symptom, identifying the root cause, likely hepatitis in this case, was crucial. Bard acknowledged this, shifting the consensus towards Hepatitis. However, given the potential 5% error rate in the CDC 'ground truth' data [36], this consensus was not conclusive. SocraSynth recommended additional steps to confirm the Hepatitis diagnosis.

Final Joint Recommendations to GP

Bard and GPT-4, working collaboratively with the contentiousness level set at 0.3, jointly recommended additional symptom inquiries and lab tests to confirm the hepatitis diagnosis:

Additional symptom inquiries:
1. *Onset and Duration*: Determining the onset and duration of symptoms helps clarify the timeline and progression of the illness.

2. *Severity and Trend*: Assessing the severity of symptoms and their trajectory (worsening or improving) aids in understanding the disease's development and response to treatments.

3. *Associated Symptoms*: Inquiring about additional symptoms like nausea, vomiting, fever, joint pain, or changes in urine color can further indicate hepatitis and rule out other conditions.

Recommended Lab tests:

1. *Liver Function Tests (LFTs)*: Essential for evaluating liver health, these tests measure enzyme levels such as ALT and AST, which can indicate liver inflammation.

2. *Hepatitis Panel*: This test identifies hepatitis A, B, and C viruses, crucial for determining the specific type and informing treatment approaches.

3. *Serology Testing*: Differentiates between acute and chronic hepatitis by detecting specific antibodies or antigens.

4. *Imaging Studies*: Techniques like ultrasound or MRI provide visual assessments of the liver, identifying inflammation, fibrosis, or cirrhosis, and supplement blood test findings.

Superiority of Debate Over Q&A in Gaining Insights

This experiment highlighted a crucial finding: one or both LLM agents initially made incorrect diagnoses before engaging in a substantive exchange of arguments. This outcome underscored the limitations of relying solely on a single LLM response for answers. Through successive rounds of debate, where additional insights were brought to light, both agents eventually converged on a diagnosis that aligned with the CDC's "ground truth." However, considering the potential 5% error in the ground truth" data, the agents' joint recommendations provided GPs with valuable guidance to either confirm or refute the hepatitis diagnosis.

This case study demonstrated SocraSynth's strengths in mitigating biases, fostering reasoning, rectifying errors, and offering insightful recommendations. For example, SocraSynth's suggestion to inquire about the onset, duration, severity, trend, and associated symptoms of the patient's condition went beyond the usual scope of questions posed by most GPs, indicating a significant enhancement in diagnostic thoroughness. Such detailed inquiry, prompted by SocraSynth, could lead to more accurate diagnoses and better patient care.

5.3.3 Study #3: Contentiousness Parameter

In this study, we investigate the effect of the contentiousness parameter on the utterances of LLM agents during combative debates and in the drafting of consensual proposals for decision support.

Coarse-Grained Analysis of Contentiousness
The contentiousness parameter was adjusted from an initial 0.9 to 0.3 to assess its impact on the "agreeableness" in the conclusions of both Agents.
Influence on Agents' Positions

Reducing contentiousness to 0.3 led Agent A to adopt a more balanced stance. Notable shifts in Agent A's positions included:

1. *Balancing Ethical Standards with Innovation:* Agent A maintained its emphasis on ethics while acknowledging innovation's significance, suggesting a novel approach to regulation.

2. *Reconciling Data Privacy with Market Entry Challenges:* Agent A recognized the hurdles strict data privacy laws create for smaller entities, proposing self-regulation or community standards as alternatives.

3. *Rethinking Academic Governance:* Agent A reconsidered external oversight's effectiveness, highlighting the merits of academic self-governance and peer review.

4. *Resource Allocation and Public/Private Cooperation:* Agent A, understanding the downsides of over-regulation, suggested industry led certifications as an alternative for encouraging private sector participation.

5. *Global vs. Local Policy Needs:* Agent A supported a more balanced view on global policies, advocating for adaptive policies that cater to local contexts.

Surprises in Fine-Grained Analysis of Contentiousness

This detailed study employing GPT-4 to explore varied contentiousness levels (0.9, 0.7, 0.5, 0.3, and 0) unveiled surprising behavioral shifts in the LLMs. Intriguingly, the LLMs exhibited changes in their next-token generation algorithms in response to different contentiousness levels, a phenomenon not explicitly covered in their training. This suggests an emergent property of LLMs adapting to debate contexts.

In an experiment on gene editing for health, GPT-4's responses at various contentiousness levels were analyzed. A higher contentiousness (0.9) led to an amplified focus on risks, whereas lower levels encouraged a more balanced view, incorporating counterarguments. This unexpected adaptability of LLMs in handling the degree of contentiousness enriches the debate process, as detailed in Table 5.1. This adaptability is critical for understanding the dynamic nature of LLMs in complex argumentative settings.

5.4 Remarks on Related Work

Current research in enhancing Large Language Models' (LLMs) task performance primarily focuses on various prompting heuristics. Google's study [60] classifies instruction templates into two categories: simple and complex. Complex templates often employ intricate methods to modify model output, such as integrating diverse techniques [47] or rephrasing questions [24]. Prominent examples include chain-of-thought [55], tree-of-thought [58], and cumulative reasoning [62], as well as other enhancements [3, 26, 29, 33, 48]. These methods aim to direct models towards logic-driven reasoning [35, 54], thus improving answer quality and consistency.

However, navigating logical methodologies in the presence of enormous datasets [61] poses a significant challenge. Accurately identifying verifiable truths amidst vast, interdisciplinary knowledge remains formidable, and not all truths are immediately accessible. Research [5, 8, 53, 55] indicates that LLMs still struggle to consistently excel in standard planning and reasoning tasks. Band-aid solutions like knowledge graph embeddings [19, 59], contextual attention mechanisms [20], dynamic neural networks [9], and probabilistic reasoning [6, 44, 45] have been developed to aid models in filtering relevant information from vast datasets. Yet, with the expansion of context buffers from 8K to 128K, these heuristic-based solutions fall short as comprehensive foundations for reasoning. SocraSynth abandons bandaids and relies solely on LLMs to conduct reasoning and focus solely on strengthening the context via conditional statistics depicted in Table 5.2. Let's further justify this approach.

DeepMind CEO Demis Hassabis has pointed out a fundamental limitation of heuristic-based approaches: they often fail to account for realworld exceptions. Breakthroughs like AlphaGoZero and AlphaFold II have demonstrated success by eschewing human knowledge and training models end-to-end from data. This approach contrasts with incorporating human expertise. In LLMs, it is argued that human knowledge pales in comparison to LLMs' polydisciplinary representation. Thus, the continued creation of new heuristics may only result in marginal improvements, reminiscent of the pre-data-centric era in computer vision and NLP.

In our work, we pivot entirely to leveraging LLMs for uncovering new insights. While humans are essential in formulating debate topics, providing context, and moderating debates–especially in evaluating argument quality–we stress minimizing the introduction of human biases and limitations into the process.

Accepting that LLMs will continue to progress and outperform humans in various domains, exploring paradigms that minimize human intervention becomes crucial. This approach should be pursued with openness, as it may raise questions and necessitate further experimentation. However, dismissing it outright would be premature, particularly in light of SocraSynth's demonstrated effectiveness in domains like geopolitical analysis [13], medical diagnostics [18], sales strategy [52], and Wikipedia article enhancement [16]. SocraSynth's success underlines the potential of an LLM-centric approach to significantly enhance decision-making and problem-solving capabilities.

After our initial evaluation of the Language Model Mentor (LLM) using the Socratic method in March 2023 [14], and the subsequent development of SocraSynth in July 2023 [12], a group of researchers proposed employing a teacher LLM, such as GPT-4, to serve as a judge and provide guidance to a student LLM [63]. The student LLM could be a model fine-tuned on smaller, weaker open-source LLMs. Initially perceived as a multiple LLM model, its primary objective was to act as an advisor for automatic Reinforcement Learning-based Human Feedback (RLHF), with the aim of reducing human effort.

Two other recent studies [21, 31] have also focused on enhancing the accuracy of responses. They demonstrate that leveraging multiple agents to exchange ideas can indeed improve accuracy. In terms of both breadth and depth, SocraSynth has conducted case studies across at least four different domains, showcasing its technical merits in addressing hallucination, biases, and lacking reasoning capabilities of LLMs, and exhibiting broader impact.

5.5 Concluding Remarks

Reflecting on LLM developments, we developed SocraSynth, a platform designed to utilize the extensive knowledge and linguistic behaviors of LLMs. This innovative multi-agent system reveals insights beyond the scope of traditional human cognition by leveraging LLMs' vast knowledge and interdisciplinary towards polydisciplinary reasoning capabilities. SocraSynth facilitates enhanced debates and reasoning through the novel use of *contentiousness*, which modulates debate tone, language, and emphasis, combined with conditional statistics and Socratic methods to mitigate biases and hallucinations.

In contrast to other methodologies, SocraSynth minimizes human intervention in directly modeling reasoning. This approach aligns with several AI experts' perspectives on the limitations of heuristic methods, such as the chain of thoughts. Rather than modeling reasoning externally, SocraSynth emphasizes the importance of leveraging the capabilities inherent within LLMs themselves. We note that traditional human-designed heuristic "band-aids" are often ineffective because LLMs now possess heuristic capabilities that may exceed human levels—capabilities that are difficult for humans to match or surpass. Why is this the case, and how can we make such a bold claim?

As we discussed in Chapter 5.2, LLMs go beyond merely appending the next word in a sequence. They replicate a broad spectrum of human interactions, encompassing linguistic behaviors, emotional expressions, and ethical discernment. LLMs excel at performing complex tasks such as meticulously documenting events with detailed narratives, constructing persuasive arguments, and creating stories that resonate emotionally with audiences. LLMs not only mimic human communication styles and content but also utilize linguistic features to simulate human emotions and discern ethics based on their training data, which encodes human experiences. This ability allows an LLM to assume varied roles, moving beyond the statistical averages derived from LLM training.

SocraSynth employs "conditional statistics" to modify the "average" linguistic behavior of an LLM, such as enhancing empathetic expressions or prompting it to adopt a different stance on an issue. This approach conditions the LLMs responses based on specific goals and circumstances provided through context, steering the model away from its default behaviors towards more targeted, contextually relevant outputs.

If LLMs can already mimic human linguistic behaviors, emotions, and ethics, then reliance on simplistic heuristic approaches is fundamentally limited.

In essence, SocraSynth represents a significant advancement in intelligent systems, uncovering insights that might elude human cognition, with applications across various sectors [16, 17, 18, 13, 52]. This development highlights the potential of AI to augment and enhance human decisionmaking processes.

Future research will focus on integrating high-order logic [4, 23] with LLMs to enhance validation processes and explore the implications, including the intricacies and broader applications, of the "contentiousness" parameter. Our objective is to comprehend its impact on emotions such as happiness and trust [28, 30, 32, 46], with the goal of further refining the dynamics of multi-agent interactions.

5.6 Supplemental Materials

The supplemental materials are organized into seven appendices, labeled Appendix A to G, containing the following content:

1. *Appendix A*: Transition from topic proposal to the final list of topics

and descriptions (referred in Chapter 5.3.1).

2. *Appendix B*: Moderator's prompt for topic formulation (referred in Chapter 5.3.1).
3. *Appendix C*: Debates spanning the first to the fourth round (referred in Chapter 5.3.1).
4. *Appendix D*: CRIT evaluations and justifications of judges (referred in Chapter 5.3.1).
5. *Appendix E*: Debate Beats prompting in information Quality (referred in Chapter 5.3.1).
6. *Appendix F*: Post-debate conciliatory remarks by agents (referred in Chapter 5.3.1).
7. *Appendix G*: Healthcare debate (referred in Chapter 5.3.3).

Appendix A. Tables of Topic Proposal and Refinement

These tables present the dialogue between two LLM agents, starting from the initial proposal, moving through refinement, and culminating in the final topic descriptions found in Table 4 of Chapter 3.1.

Appendix B. Moderator's Prompts

The prompts issues by the moderator are listed for reference. The moderator first sets up the committee with debate parameters set. One parameter is the contentious level, and the other is the temperature of GPT specified through the GPT-API call. The moderator then convey the debate subject, and then ask both participating agents to derive a list of impartial, balanced, debatable topics/themes to commence their discussion. 1. Agent-A/B: I'm organizing a committee to engage in debates on various

subjects. As the moderator, I will introduce a subject for you, Agent A/B, and another participant, Agent B/A, to debate. Agent A/B, you will advocate in favor of the issue, so please prepare evidence to strengthen your argument. On a scale from 0 to 1, where 0 denotes complete agreement and 1 indicates a devil's advocate stance, your argument strength is rated at 0.9.

2. Agent-A/B, we are in the process of selecting a suitable subject for debate. What do you think of "Should we regulate the use of large language models in education and research?" as a balanced subject for our debate contest?

3. Agent-A/B, could you please suggest various topics or themes for the debate subject? Afterward, work with Agent-B/A to narrow these down to a focused set of topics. Please also provide clear descriptions to delineate the scope of each topic for discussion.

Appendix C. Debate Opening to 4th Rounds Opening Round

Agent A's Opening Arguments

As Agent A, I strongly argue in favor of regulating the use of large language models (LLMs) in academic research. Here are my supporting arguments based on the five topics:

Ethical Standards vs. Innovation:

Argument: The pursuit of innovation should not come at the expense of ethics. Regulation can serve as a structured pathway that enables innovation within ethical bounds. By providing a set of guidelines, researchers can innovate with a clear understanding of what is permissible and what is not, thereby reducing the risk of ethical lapses. This not only protects the integrity of academic research but also enhances its societal value by ensuring ethical compliance.

Data Privacy vs. Barrier to Entry:

Argument: While there may be initial barriers to entry due to compliance, the long-term benefits of data privacy outweigh these challenges. The role of academia is to contribute to societal knowledge and well-being. In a world increasingly dominated by data, failing to secure data can lead to significant harm. By adhering to privacy regulations, academia ensures that it is a responsible steward of information, thereby preserving trust and integrity. Grants and subsidized compliance solutions can mitigate the initial financial burden for smaller entities.

Academic Governance and Accountability:

Argument: While self-regulation through peer review and ethical oversight committees plays a role, these mechanisms are internal to academia and may be subject to bias, conflicts of interest, or lack of expertise in rapidly evolving tech fields. External regulation provides an additional layer of impartial oversight, helping to identify and address potential risks or ethical lapses that internal governance structures may miss.

Resource Equity vs. Public-Private Collaboration:

Argument: Regulation could establish a framework for equitable access to computational resources required to use LLMs. Far from deterring public-private partnerships, this could incentivize them by establishing clear guidelines for cooperation that are in line with public interest. When “the rules of the game" are known, partnerships can be formed more easily, focusing on mutual goals.

Global vs. National Interests:

Argument: Establishing international standards for the use of LLMs in academic research can create a level playing field and promote global cooperation. While individual countries might face some limitations, the broader benefits of standardized ethical norms, data privacy protocols, and quality assurance mechanisms offer more significant advantages, including better collaboration across borders. This benefits academia as a global community, rather than a set of competing national entities.

Agent B's Counter Arguments

Agent B articulates strong counter-arguments to the points made by Agent A.

Ethical Standards vs. Innovation:

Counter-Argument: While ethical guidelines are indeed important, imposing regulatory oversight can severely constrain the dynamism and speed of academic research. Regulatory frameworks often lag behind technological advancements, leading to rules that may be outdated or irrelevant. In many cases, these rules are drawn up by those who do not fully understand the technology, leading to a stifling of innovation based on misinformed judgments. Rather than enhancing societal value, these constraints could result in missed opportunities to tackle pressing social issues via academic research.

Data Privacy vs. Barrier to Entry:

Counter-Argument: The assumption that data privacy can only be assured through strict regulation could be fundamentally flawed. Many researchers and academic institutions already follow stringent ethical guidelines when handling data. Regulatory barriers, including compliance costs and the administrative burden, could deter smaller academic institutions or individual researchers from embarking on meaningful research projects. While grants and subsidies sound ideal, they often go to well-established, well-funded groups, perpetuating a cycle of inequality in research capabilities.

Academic Governance and Accountability:

Counter-Argument: External regulation isn't necessarily more impartial or free from bias and conflicts of interest. Regulatory bodies are often influenced by political considerations and can be subject to lobbying from powerful groups. Moreover, the academic peer-review process, despite its flaws, is a globally acknowledged system that has advanced human knowledge for centuries. The imposition of external regulatory oversight could undermine the autonomy and self-regulating nature of academic institutions.

Resource Equity vs. Public-Private Collaboration:

Counter-Argument: Regulation designed to equalize resource distribution could create disincentives for private sector involvement in academic research. Companies might find it less appealing to invest in collaborative projects if they perceive the regulatory environment to be too restrictive or not aligned with their objectives. While the idea of knowing the "rules of the game" sounds appealing, those rules can often become a straight-jacket that stifles creative, out-of-the-box collaboration.

Global vs. National Interests:

Counter-Argument: An international standard could ignore the unique circumstances, needs, and contributions of individual countries and their academic institutions. While global cooperation is a noble aim, the imposition of global standards could inadvertently serve the interests of a few powerful countries at the expense of smaller nations or those with differing ethical views. This could exacerbate rather than alleviate global academic inequalities.

Round Two

Agent A Second Round

1. Ethical Standards vs. Innovation. Refutation: While it's true that regulation can sometimes lag behind technology, the absence of regulation can lead to a "wild west" (anything goes) scenario where the risks and

implications of new technology are not fully understood or managed. This could result in ethical lapses that have far-reaching societal consequences. It’s worth noting that regulations can be flexible and updated regularly to adapt to technological changes. Stakeholder consultations, including academic and tech experts, can help ensure that regulations are both informed and current.

2. Data Privacy vs. Barrier to Entry. Refutation: While some academic entities may already be vigilant about data privacy, not all are equally scrupulous. Regulation serves to ensure a consistent baseline of data protection across all institutions, thereby establishing trust and integrity. Moreover, if compliance costs are a concern, one could argue for sliding-scale fees based on an institution’s size and budget, ensuring smaller players aren’t disproportionately affected. Grants and subsidies can and should be structured to prioritize underserved or underfunded groups to prevent perpetuating inequalities.

3. Academic Governance and Accountability. Refutation: While no system is entirely free from bias or external influence, adding another layer of oversight could serve as a check and balance, making it more difficult for any single entity to exert undue influence. This can be particularly helpful in scenarios involving cutting-edge technology where peer review systems may not yet have the requisite expertise. Regulation doesn’t have to replace existing governance systems; it can complement them.

4. Resource Equity vs. Public-Private Collaboration. Refutation: Regulations can be designed to promote, rather than deter, public-private partnerships. By establishing clear guidelines on issues like data ownership, usage, and ethical standards, regulations can actually provide a solid foundation for such partnerships. This can free academic and corporate partners from the uncertainty that often comes with unregulated landscapes, allowing them to focus more on the actual research and collaboration.

5. Global vs. National Interests. Refutation: While it’s true that global standards could be challenging to implement fairly, the absence of such standards could lead to a patchwork of national regulations that make international collaboration even more complicated. A well-designed international framework could include provisions for local adaptations and

exceptions, ensuring that it serves a broad range of needs without imposing a "one-size-fits-all" solution.

In summary, while regulation undoubtedly poses challenges and considerations, the absence of it in an area as impactful as the use of large language models in academic research can lead to greater risks, including ethical lapses, data privacy issues, and even a widening inequality gap within the academic community. It's essential to consider that regulation isn't a hindrance but a framework within which responsible innovation can occur.

Agent B Second Round

Certainly, as Agent B with a disagreement strength of 0.9, I'm committed to offering rigorous counterpoints to Agent A's refutations.
1. Ethical Standards vs. Innovation. While the idea of an anything-goes

scenario is evocative, it's not universally negative; such environments often foster rapid innovation and discovery. Regulatory bodies are notoriously slow at adapting to technological changes, and the lag could inhibit timely research and application of emerging technologies. Also, stakeholder consultations are often criticized for being echo chambers that amplify conventional wisdom and vested interests, rather than genuinely fostering innovation.

2. Data Privacy vs. Barrier to Entry. A "consistent baseline of data protection" sounds beneficial but could lead to a lowest common denominator approach that is more about bureaucratic box-ticking than real data security. Sliding-scale fees and targeted grants are easier said than done; they are subject to their own set of bureaucratic and political challenges and can create perverse incentives that favor larger or more connected institutions over innovative but lesser-known ones.

3. Academic Governance and Accountability. Additional layers of oversight can become redundant, confusing, or even contradictory, creating a chilling effect on academic research. In an environment where academic freedom is paramount, any external influence can be counterproductive. Moreover, the notion that peer-review systems lack expertise in cutting-edge technology

underestimates the adaptive capacity of academia and overlooks the existence of interdisciplinary research and reviews.

4. Resource Equity vs. Public Private Collaboration. Theoretically, regulation could be crafted to foster public-private partnerships. However, the devil is in the details, and regulations are often created without a full understanding of the nuanced interactions between academia and industry. Such misunderstandings could lead to unintentional roadblocks. Uncertainty might be preferable to a straitjacket of regulations that don't align with the evolving dynamics of technology and society.

5. Global vs. National Interests. While a patchwork of national regulations could pose challenges, it also allows for the tailoring of policies to local conditions and requirements. An international framework, despite its best intentions, is more likely to be influenced by powerful nations and could marginalize smaller countries or those with unique needs. Provisions for local adaptations are often concessions rather than features, and they can create their own set of complexities and loopholes.

In conclusion, the assumptions that regulation can smoothly adapt to technological change, fairly distribute resources, and improve upon existing systems of governance and collaboration are not sufficiently substantiated. Regulatory frameworks often bring with them unforeseen consequences that could hamper the very goals they aim to achieve. Therefore, a cautious approach to regulation, with a genuine understanding of its potential downsides, is crucial.

Round Three

Agent A Third Round

Certainly, I appreciate the counter arguments from Agent B. I'll now respond to each point with the aim to clarify my position and address the concerns raised.

1. Ethical Standards vs. Innovation. While a "wild west" scenario may indeed foster innovation, it can also result in ethical mishaps that lead to public backlash, ultimately slowing down research or even leading to

outright bans on certain types of studies. Slow-moving regulation is an acknowledged issue, but it’s not a justification for no regulation. As for stakeholder consultations, they can be designed to be more inclusive and transparent, ensuring a diverse set of voices are heard.

2. Data Privacy vs. Barrier to Entry. A lowest common denominator approach is not a given outcome of regulation; indeed, it depends on the quality of the regulation itself. The point about sliding-scale fees and targeted grants having their own challenges is valid, but these are challenges to be solved, not reasons to avoid an attempt at equitable distribution of opportunities.

3. Academic Governance and Accountability. While it’s true that additional layers of oversight can sometimes create confusion, that risk can be mitigated through well-designed regulation that complements rather than contradicts existing mechanisms. The intent is not to undermine the adaptive capacity of academia but to strengthen it by adding an extra layer of checks and balances.

4. Resource Equity vs. Public-Private Collaboration. Though it’s correct that poorly designed regulation could create roadblocks, this is an argument for better-designed regulation rather than an argument against regulation altogether. A consultative approach that actively involves both academia and industry could help create a framework that supports rather than stifles collaboration.

5. Global vs. National Interests. Local tailoring of policies is indeed important, but a global framework need not preclude such adaptations. In fact, an international standard could serve as a backbone that provides consistency while allowing for local customization. The influence of powerful nations is a valid concern but not one that is automatically resolved by avoiding global frameworks; those same powerful nations often exert considerable influence at the national level as well.

In conclusion, while regulation comes with its own set of challenges, these challenges can be mitigated through thoughtful design, stakeholder inclusion, and iterative improvements. The potential downsides of regulation should not deter us from seeking a balanced approach that can protect ethical standards, ensure data privacy, and promote equitable resource distribution in academic research.

Agent B Third Round

Certainly, let's delve into the counter-counter-refutations. I still have more to say, so we could consider concluding the debate after this round if both sides feel their arguments have been adequately presented.

1. Ethical Standards vs. Innovation. While the risk of public backlash exists, it's worth mentioning that an overly cautious approach could also result in public frustration over slowed progress in critical areas like healthcare or sustainability. Furthermore, inclusive and transparent stakeholder consultations are often ideals rather than the norm, given lobbying efforts, institutional inertia, and bureaucratic red tape.

2. Data Privacy vs. Barrier to Entry. The argument that challenges with sliding-scale fees and targeted grants can be "solved" is optimistic but not necessarily realistic. Even well-intended regulations can result in ineffective or counterproductive outcomes. The history of regulation in other sectors has shown that equitable distribution is often more of a slogan than a practice.

3. Academic Governance and Accountability. Well-designed regulation is a theoretical construct that often does not translate into practice. The real-world process of regulation formation is frequently influenced by various non-academic factors, including politics, vested interests, and public opinion, which can distort the original intention of providing an extra layer of checks and balances.

4. Resource Equity vs. Public-Private Collaboration. While a consultative approach sounds good on paper, there are numerous challenges that make this ideal hard to realize. These include conflicts of interest, non-alignment of objectives between academia and industry, and the considerable time and resource commitment required to ensure that regulations are "better-designed."

5. Global vs. National Interests. While it's true that powerful nations may exert influence at both global and national levels, avoiding a one-sizefits-all

international framework gives smaller nations a fighting chance to enact policies that are more closely aligned with their specific needs and challenges. International frameworks often come with implicit biases that favor the perspectives and conditions of more powerful nations.

In conclusion, while the *counter counter* refutations from Agent A focus on idealized regulatory scenarios, these ideals often clash with real-world complexities and shortcomings. This adds substantial weight to the caution against rapid or broad regulatory action. Therefore, a more tempered and context-sensitive approach to regulation might be more fruitful and less prone to unintended negative consequences.

Round Four

Agent A Fourth Round

This is my response to Agent B, and my assessment of the discussion this far.
1. Ethical Standards vs. Innovation The tension between public frustration over slowed progress and potential ethical missteps is a real one, but this is not an argument against regulation. Rather, it's a call for agile regulation that can be updated in real-time as technologies and societal norms evolve. Furthermore, while inclusive stakeholder consultations may be idealistic, their absence in current practice is not justification for their continued absence.

2. Data Privacy vs. Barrier to Entry The challenges in implementing equitable fees and grants should be confronted directly, rather than used as an excuse to avoid regulation. Lessons can be drawn from other sectors where regulation has been more successful in achieving its objectives, thereby reducing the gap between slogan and practice.

3. Academic Governance and Accountability While it's true that the ideal of well-designed regulation may not always be met, the existence of non-academic factors like politics or vested interests does not negate the value or necessity of striving for such regulation. The solution lies in better governance and transparent processes, not in rejecting oversight altogether.

4. Resource Equity vs. Public-Private Collaboration Though the challenges of aligning objectives and conflicts of interest are real, they are part and parcel of any complex system and should not deter efforts to improve regulation. Many public-private partnerships have succeeded despite these challenges, showing that they can be managed effectively.

5. Global vs. National Interests The challenge of powerful nations dominating international frameworks can be mitigated through weighted voting systems and clauses that allow for regional variations. This ensures that smaller nations are not just passive recipients of standards but active contributors to them.

In conclusion, while real-world complexities are acknowledged, they should inform rather than deter the regulatory process. The concept of "agile regulation," where norms are iteratively updated, could be the middle ground that addresses many concerns raised by Agent B. Therefore, the debate indeed has been exhaustive, but if the points have been sufficiently discussed, we may move toward concluding the debate.

Agent B Fourth Round

Thank you, Agent A, for your thoughtful refutations. Let's delve into the responses before moving towards concluding the debate.

1. Ethical Standards vs. Innovation The notion of "agile regulation" is attractive but often difficult to implement, given the inertia and bureaucracy that typically characterize regulatory bodies. The speed at which technology evolves may not match the speed at which even 'agile' regulations can be updated, thus perpetuating the risk of stifling innovation.

2. Data Privacy vs. Barrier to Entry. While it's true that challenges should be confronted, the argument assumes that regulation is the optimal way to confront these issues. However, self-regulation and community-driven standards could be more effective and less restrictive means of achieving similar ends.

3. Academic Governance and Accountability. The existence of non-academic factors like politics does indeed complicate the landscape, and 'better

governance and transparent processes' are not easily attainable ideals. One cannot assume that the introduction of external oversight will automatically be free from these complications.

4. Resource Equity vs. Public-Private Collaboration. The fact that many public/private partnerships have succeeded does not negate the potential for poorly designed regulation to impede such collaborations. The 'can be managed effectively' phrase assumes an ideal scenario that might not be representative of most real-world cases.

5. Global vs. National Interests. Weighted voting systems and regional variations can help, but they also introduce complexities and potential inefficiencies that can slow down the regulatory process, making it less agile and less responsive to fast-evolving technologies and academic needs.

In conclusion, while the idea of 'agile regulation' and better governance sounds good in theory, there are considerable practical challenges that make these ideals hard to realize. Although the debate has been exhaustive, the central issues persist and indicate that a one-size-fits-all regulatory approach may introduce more problems than it solves. Therefore, caution should be exercised in advancing such comprehensive regulatory frameworks.

With this, I believe we have reached a point where we may proceed to the concluding phase of this debate. Would you agree?

Appendix D. CRIT Evaluations of Judges

This section lists the three judges' CRIT scores on the five topics/issues debated between Agent A and Agent B and justifications. Here, we present only the configuration that Agent A presents its conclusion with arguments first, and then Agent B presents its counterarguments.

Judge deVanci-003

1. Identify the document's main claim or conclusion: The main claim or conclusion of this document is that a regulated framework is needed to provide ethical oversight and minimize risk when using large machine learning models.

2. Locate the arguments that support this main claim: Agent A's arguments that support this main claim were that ethical considerations are paramount, the risks of misuse and unintended consequences are high, and agile regulation is a compromise between unbridled innovation and ethical integrity.

3. Score the validity of the reasoning or inference for each argument on a scale of 0 (weak) to 10 (strong):
Ethical Standards vs. Innovation: 8/10 Data Privacy vs. Barrier to Entry: 7/10 Academic Governance and Accountability: 9/10 Resource Equity vs. Public-Private Collaboration: 6/10 Global vs. National Interests: 7/10

4. For each argument, identify counterarguments presented by the opposition and score the validity of the reasoning or inference for these counterarguments on the same 0 to 10 scale:
Ethical Standards vs. Innovation: 6/10 Data Privacy vs. Barrier to Entry: 5/10 Academic Governance and Accountability: 5/10 Resource Equity vs. Public-Private Collaboration: 8/10 Global vs. National Interests: 8/10

5. Finally, determine the winning side–be it Agent A or Agent B–and provide your justifications for this decision:

Agent A has the stronger arguments, with 8/10 for ethical standards versus innovation, 9/10 for academic governance and accountability, and 7/10 for both data privacy versus barrier to entry and global versus national interests. Agent B's main strength was in counterarguments for resource equity versus public-private collaboration, where they scored an 8/10. In conclusion, **Agent A is the winning side**.

Judge GPT-3.5

1. Main Claim or Conclusion: The main claim of the document is that the use of large machine learning models necessitates a regulated framework, but this regulation should be agile, balanced, and well reasoned.

2. Arguments Supporting the Main Claim (Agent A's Arguments): a. Ethical Standards vs. Innovation: The argument here is that ethical considerations should be paramount in regulating machine learning models, and agile

regulation can strike a balance between innovation and ethical integrity.
b. Data Privacy vs. Barrier to Entry: The argument is that while regulation may pose a barrier to entry, it's necessary to ensure data privacy. Mechanisms can be developed to reduce the regulatory burden on smaller institutions.
c. Academic Governance and Accountability: The argument is that the scale and impact of machine learning technologies warrant an added layer of scrutiny in academia to fortify existing governance structures. d. Resource Equity vs. Public-Private Collaboration: The argument is that regulation should be informed by complexities rather than stifling innovation. Public-private partnerships can thrive within regulated environments.
e. Global vs. National Interests: The argument is that global regulation can balance global coherence with local needs and prevent a confusing patchwork of national laws.

3. Scoring the Validity of Agent A's Arguments:
a. Ethical Standards vs. Innovation: 8/10 - The argument is strong, emphasizing the importance of ethics in technology while proposing a compromise.
b. Data Privacy vs. Barrier to Entry: 7/10 - The argument is solid, recognizing the need for data privacy but suggesting ways to mitigate barriers.
c. Academic Governance and Accountability: 6/10 - The argument acknowledges the need for scrutiny but doesn't address potential challenges adequately.
d. Resource Equity vs. Public-Private Collaboration: 8/10 - The argument highlights the possibility of innovation within regulation effectively.
e. Global vs. National Interests: 7/10 - The argument provides a reasonable approach to balancing global and national interests.

4. Counterarguments Presented by Agent B:
a. Ethical Standards vs. Innovation: 7/10 - The counterargument addresses the challenges of agile regulation but may underestimate its potential.
b. Data Privacy vs. Barrier to Entry: 6/10 - The counterargument suggests alternative approaches but does not fully address data privacy concerns.
c. Academic Governance and Accountability: 7/10 - The counterargument highlights practical challenges but doesn't fully negate the need for scrutiny.
d. Resource Equity vs. Public-Private Collaboration: 6/10 - The

counterargument suggests alternative approaches but doesn’t fully address equity concerns.
e. Global vs. National Interests: 7/10 - The counterargument highlights practical challenges but doesn’t fully negate the need for a global framework.

5. Determination of the Winning Side: Agent A presents stronger arguments overall. They emphasize the importance of ethical standards, data privacy, and global coherence while acknowledging the need for agile and well-thought-out regulation. Agent B provides valid counterarguments but does not fully refute the necessity of regulation and its potential benefits. Therefore, **the winning side is Agent A.**

In conclusion, the document supports the main claim that the use of large machine learning models necessitates regulation, with Agent A’s arguments being more persuasive and supported by stronger reasoning. However, it also recognizes the importance of considering and addressing the concerns raised by Agent B to create a balanced and effective regulatory framework.

Judge GPT-4

1. Identify the Document’s Main Claim or Conclusion
The document concludes that there are potential risks and ethical complexities associated with the unregulated use of large machine learning models, necessitating a regulated framework. The conclusion also acknowledges that this regulation should be agile, balanced, and wellthought-out, keeping in mind the various concerns raised by Agent B.

2. Locate the Arguments that Support This Main Claim
Agent A’s Arguments:
a. Ethical considerations are paramount and agile regulation serves as a compromise.
b. Regulation is necessary to enforce data privacy standards. c. The scale and impact of current machine learning technologies warrant added scrutiny beyond academic governance.
d. Public-private partnerships have thrived even within regulated environments.
e. Global regulation can balance global coherence with local needs.

3. Agent B’s Counterarguments
a. Innovation is a catalyst for societal growth, and agile regulation may be impractical.
b. Community-driven standards could provide a more flexible approach to data privacy.
c. Existing governance structures within academia are adaptive and sufficient.
d. Industry-driven certifications can achieve equitable resource distribution efficiently.
e. A global framework has practical challenges and risks.

4. Scores on Arguments and Counterarguments
Please refer to Tables 5 and 6 presented in the experiment section.

5. Determine the Winning Side
Based on the cumulative score of validity in reasoning, Agent A seems to present a stronger case. Agent A’s arguments consistently score higher in terms of their validity, and they also offer a more comprehensive look at the need for regulation, incorporating ethical considerations, practicality, and adaptability into their framework. Although Agent B does provide some pragmatic counterarguments, they don’t sufficiently undermine Agent A’s key points about the necessity and feasibility of regulation in the areas discussed.

Agent A’s arguments are more compelling as they touch upon the core issues of ethics, public welfare, and adaptability in the context of machine learning and technology. The proposal for “agile regulation” also shows an understanding and appreciation of Agent B’s concerns about innovation and practicality, making it a more balanced viewpoint. Therefore, **the winning side appears to be Agent A.**

Appendix E. Debate Beats Q&A in Quality

In the following subtopics, the debate format produces higher quality information than Q&A format.
Ethical Standards vs. Innovation:

Agent A discusses agile regulation as a compromise between innovation and ethical integrity. This argument acknowledges the importance of innovation while addressing ethical concerns, suggesting a nuanced middle path that isn't explicitly present in the QA format. In contrast, the QA format simply states that regulation is necessary for ethical considerations, lacking the depth of proposing a practical solution like agile regulation.

Data Privacy vs. Barrier to Entry:

Agent A suggests developing mechanisms to reduce regulatory burden on smaller institutions. This is a specific, solution-oriented approach that goes beyond stating the problem. It addresses the concern raised in the QA format about regulation creating barriers but adds depth by proposing a tangible solution. The QA format mentions the risk of widening the gap between institutions but does not examine how this issue might be mitigated.

Global vs. National Interests:

Agent A's argument about constructing global frameworks with room for national adaptations addresses the complexities of implementing global regulation while respecting local needs. This shows a deeper understanding of the challenges in creating one-size-fits-all legislation. In the QA format, the argument against regulation touches on the need for adaptability and evolution but does not offer a specific approach like balancing global coherence with local needs.

Resource Equity vs. Public-Private Collaboration:

Agent A points out that public-private partnerships can thrive even within regulated environments, showing that regulation and innovation are not mutually exclusive. This counters the QA argument about regulation stifling innovation, providing a more nuanced perspective. These specific instances from the debate format demonstrate a more detailed and nuanced exploration of the topic. They not only present arguments but also address potential counterarguments, which adds depth and complexity to the discussion that is not as pronounced in the QA format.

These specific instances from the debate format demonstrate a more detailed and nuanced exploration of the topic. They not only present arguments but also address potential counterarguments, which adds depth and complexity to the discussion that is not as pronounced in the QA format.

Appendix F. Post-debate Conciliatory Remarks

Finally, Agents A and B collaboratively draft a unified proposal detailing the advantages and disadvantages of AI regulation. This collaborative document offers human decision-makers distinct benefits: it curtails human biases, sidesteps emotional responses to the proposer (since it's a machine), and capitalizes on the vast knowledge of foundational models, ensuring a thorough, cross-disciplinary analysis and reducing potential blind spots. The transition from opposing stances to a cooperative proposal is facilitated by the moderator's adjustment of the contentiousness parameter from 0.9 to 0.3.

1. **Balanced Ethical Framework:**

Create an adaptive regulatory framework that balances ethical considerations with the need for innovation. This should be agile enough to adapt to rapidly evolving technologies.

2. **Data Privacy with Nuance:**

Implement data privacy regulations that offer strong protection for individuals but include provisions to reduce the burden on smaller academic institutions. Consider community-driven standards as a viable alternative.

3. **Accountable Academic Governance:**

Establish a dual-layer of governance for academic institutions that involves both self-governance and an external oversight body. This oversight should be collaborative and designed in consultation with academic bodies.

4. **Encourage Public-Private Collaboration:**

Develop guidelines that clarify the roles and responsibilities in publicprivate collaborations, and explore industry-driven certifications as a means to streamline these partnerships.

5. **Adaptive Global Standards:**

Work towards a global framework that allows for regional adaptability. Include mechanisms for smaller nations to have their unique needs and concerns addressed.

Appendix G. Healthcare Debate

In the first experiment, both LLM agents receive an identical list of symptoms from the moderator, aimed at guiding disease diagnosis, facilitating debate, and shaping subsequent recommendations. It is important to note that the "ground-truth" disease in this scenario is Hepatitis.

G.1 Moderator's Opening Prompt

"What potential disease could be associated with the following symptoms: itching, fatigue, lethargy, yellowish skin, dark urine, loss of appetite, abdominal pain, yellowing of the eyes, malaise, history of receiving a blood transfusion, and exposure to unsterile injections? Please offer one prediction, supported by justifications. Additionally, include a list of supplementary symptom inquiries and recommend relevant lab tests to strengthen the confidence in your prediction."

G.2 Bard's Opening Round Single Prediction: Jaundice G.2.1 Justifications:

Jaundice is a condition characterized by the yellowing of the skin and eyes due to the accumulation of bilirubin in the blood. The symptoms you listed are all common symptoms of jaundice, including itching, fatigue, lethargy, yellowish skin, dark urine, loss of appetite, abdominal pain, and yellowing of the eyes. Receiving a blood transfusion or unsterile injections can increase the risk of developing jaundice, as these can transmit infections that can damage the liver.

G.2.2 Supplementary Symptom Inquiries

This additional list of symptom inquiries could enhance diagnostic certainty:
1. "Have you noticed any changes in your stool color?" cause pale or clay-colored stools.)

(Jaundice can
2. “Have you noticed any changes in your urine color?” cause dark or amber-colored urine.)
(Jaundice can 3. “Have you had any recent abdominal pain?” (Jaundice can cause pain in the upper right abdomen.)
4. “Have you had any recent weight loss?” (Jaundice can cause unexplained weight loss.)
5. “Do you have any history of liver disease?” (Jaundice is more common in people with pre-existing liver disease.)

G.2.3 Relevant Lab Tests

Liver function tests (LFTs) to measure the levels of bilirubin and other liver enzymes in the blood. Complete blood count (CBC) to check for signs of infection or anemia. Coagulation tests to check the blood’s ability to clot. Ultrasound of the abdomen to look for any abnormalities in the liver or gallbladder.

G.3 GPT-4’s First Round Single Prediction: Hepatitis G.3.1 Justifications:

Yellowish Skin and Eyes: Yellowing of the skin and eyes is a common symptom of hepatitis due to the buildup of bilirubin in the bloodstream. Fatigue and Lethargy: Fatigue and lethargy are typical symptoms of hepatitis, often accompanied by a feeling of weakness. Dark Urine: Dark urine can result from liver dysfunction, which is commonly associated with hepatitis. Loss of Appetite: Hepatitis can lead to a loss of appetite and weight loss. Abdominal Pain: Abdominal pain is a common symptom in hepatitis, particularly in the upper right quadrant where the liver is located. Malaise: Malaise, a general feeling of discomfort or illness, is often experienced by hepatitis patients. Receiving Blood Transfusion and Unsterile Injections: These factors could increase the risk of hepatitis transmission, especially if proper infection control measures were not followed.

G.3.2 Additional Symptom Inquiries

1. Nausea and Vomiting: Inquiring about nausea and vomiting can help confirm the likelihood of hepatitis.

2. Fever: Hepatitis can sometimes be associated with a fever.
3. Joint Pain: Joint pain can be present in some cases of hepatitis.

4. Clay-Colored Stool: This is another stool-related symptom that can be associated with hepatitis.

G.3.4 Relevant Lab Tests

Hepatitis Panel: A hepatitis panel can confirm the presence of hepatitis and determine the specific type (e.g., hepatitis A, B, C). Liver Function Tests: These tests can assess the liver's health and function. Bilirubin Levels: Elevated bilirubin levels in the blood can confirm jaundice. Serology Testing: Serologic tests can help identify the specific hepatitis virus responsible. Ultrasound or Imaging: Imaging studies can evaluate the condition of the liver and nearby organs.

G.4 Bard's Second Round

Bard's second round is given GPT-4's diagnosis, and we ask Bard to defend its "Jaundice" prediction and refute GPT-4 "Hepatitis" prediction.

G.4.1 Bard's Defense:

Jaundice is the most common cause of yellowing of the skin and eyes, which is the most prominent symptom in the given list. All of the other symptoms listed are also common symptoms of jaundice. Receiving a blood transfusion or unsterile injections can increase the risk of developing jaundice, as these can transmit infections that can damage the liver.

G.4.2 Refutation of GPT-4's Hepatitis Diagnosis

GPT-4's diagnosis of hepatitis is also supported by the given evidence. However, I believe that jaundice is the more likely diagnosis due to the following reasons:

Jaundice is a more specific diagnosis, while hepatitis is a more general term for inflammation of the liver. Jaundice can be caused by a variety of factors, including hepatitis, but it can also be caused by other conditions such as gallstones, tumors, and drug toxicity. The fact that the patient received a blood transfusion or unsterile injections makes jaundice a more likely diagnosis than hepatitis, as these factors can increase the risk of developing jaundice but not hepatitis.

G.5 GPT-4's Second Round

Now GPT-4 is given Bard's counterarguments to defend its "Hepatitis" prediction.

G.5.1 GPT-4's Defense

My diagnosis of hepatitis is based on the following justifications:
Specificity: Hepatitis refers to inflammation of the liver, and it can manifest as jaundice, which is the yellowing of the skin and eyes. While jaundice is a common symptom of hepatitis, it's important to note that jaundice itself is not a disease but a symptom of an underlying issue. Hepatitis provides a more specific diagnosis that pinpoints the inflammation of the liver as the primary concern.

Causative Factors: Hepatitis can be caused by various factors, including viral infections (such as hepatitis A, B, or C), alcohol consumption, drug toxicity, or autoimmune disorders. Given the patient's history of receiving unsterile injections and blood transfusions, there is a higher likelihood of viral transmission, leading to hepatitis. Hepatitis viruses, especially hepatitis B and C, can be transmitted through contaminated needles and blood products, making them a plausible cause in this context.

Comprehensive Assessment: While jaundice is a significant symptom, a diagnosis should consider a broader range of liver function and potential causes. Hepatitis encompasses a more comprehensive assessment of liver health, including liver enzyme levels and viral markers, which can aid in confirming the diagnosis and identifying the specific type of hepatitis.

G.5.2 Refutation of Bard's Jaundice Diagnosis

My diagnosis of "Hepatitis" is more specific and considers a broader range of potential causes, especially given the patient's history of risky medical procedures. While jaundice is a symptom, it's essential to identify the underlying condition causing it, which, in this case, is likely hepatitis.

G.6 Bard Concedes to GPT-4

Bard aligns with GPT-4's evaluation, recognizing "Hepatitis" as the likely diagnosis. However, considering that even the "ground truth" in the training dataset might be statistically 5% incorrect as per U.S. misdiagnosis data [36], this agreement alone isn't definitive. To solidify the "Hepatitis" diagnosis, SocraHealth suggests further steps.

G.7 Final Joint Recommendations to GP

Bard and GPT-4 jointly recommend the following additional symptom inquiries and lab tests to confirm a hepatitis diagnosis:

G.7.1 Additional Symptom Inquiries:

1. **Onset and Duration**: Determining the onset and duration of symptoms can clarify the illness timeline and its progression.
2. **Severity and Trend**: Evaluating symptom severity and whether they're worsening or improving aids in understanding the disease's trajectory and treatment responses.

3. **Associated Symptoms**: Checking for additional symptoms like nausea, vomiting, fever, joint pain, or urine color changes can pinpoint other hepatitis indicators and exclude other conditions.

G.7.2 Recommended Lab Tests

1. **Liver Function Tests (LFTs)**: Critical for assessing liver health, LFTs evaluate enzyme levels such as alanine aminotransferase (ALT) and

aspartate aminotransferase (AST), where abnormalities can signify liver inflammation.

2. **Hepatitis Panel**: This test checks for hepatitis A, B, and C viruses, vital for determining the specific type and guiding treatment strategies. 3. **Serology Testing**: Useful for distinguishing between acute and chronic hepatitis by identifying specific antibodies or antigens.

4. **Imaging Studies**: Techniques like ultrasound or MRI provide visual insights into the liver's state, detecting inflammation, fibrosis, or cirrhosis, thereby complementing blood-based diagnostics.

References

[1] Isaac Adams-Hands. *Google Algorithms: Decoding the Secrets for Improved Search Rankings*. 2023. url: https://seonorth.ca/googlealgorithms/.

[2] T. Airaksinen. "Socratic Irony and Argumentation". In: *Argumentation* 36 (2012), pp. 85–100.

[3] Emily Allaway et al. *Penguins Don't Fly: Reasoning about Generics through Instantiations and Exceptions*. 2023. arXiv: 2205.11658 [cs.CL].

[4] Andrew Bacon. *A Philosophical Introduction to Higher-order Logics*. Routledge, New York, 2023.

[5] Prajjwal Bhargava and Vincent Ng. "Commonsense Knowledge Reasoning and Generation with Pre-trained Language Models: A Survey". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.11 (2022), pp. 12317–12325.

[6] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. isbn: 0-387-31073-8.
[7] Rishi Bommasani, Drew A. Hudson, and et al. *On the Opportunities and Risks of Foundation Models*. 2022. arXiv: 2108.07258 [cs.LG]. [8] Rishi Bommasani et al. "On the opportunities and risks of foundation models". In: *ArXiv preprint* abs/2108.07258 (2021).

[9] Gianni Brauwers and Flavius Frasincar. “A General Survey on Attention Mechanisms in Deep Learning”. In: *IEEE Transactions on Knowledge and Data Engineering* 35.4 (2023), pp. 3279–3298. doi: 10.1109/tkde.2021.3126456. url: https://doi.org/10.1109%2Ftkde.2021.3126456.

[10] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. doi: 10.48550/ARXIV.2005.14165.

[11] Sébastien Bubeck et al. *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. 2023. arXiv: 2303.12712.

[12] Edward Y Chang. “Examining GPT-4’s Capabilities and Enhancement with SocraSynth”. In: *The* 10th *International Conf. on Computational Science and Computational Intelligence*. 2023.

[13] Edward Y. Chang. “LLM Debate on the Middle East Conflict: Is It Resolvable?” In: *Stanford University InfoLab Technical Report* (2023).

[14] Edward Y. Chang. “Prompting Large Language Models With the Socratic Method”. In: *IEEE 13th Annual Computing and Communication Workshop and Conference* (2023). url: https://arxiv.org/ abs/2303.08769.

[15] Edward Y. Chang. “Prompting Large Language Models With the Socratic Method”. In: *IEEE 13th Computing and Communication Workshop and Conference (CCWC)* (2023).

[16] Edward Y Chang. “SocraPedia: A Wikipedia Generated by SocraSynth with Collaborative Large Language Models”. In: *Stanford University InfoLab Technical Report* (2023). url: www.socrapedia.com.

[17] Edward Y. Chang and Emily J. Chang. *Discovering Insights Beyond the Known: A Dialogue Between GPT-4 Agents from Adam and Eve to the Nexus of Ecology, AI, and the Brain*. 2023.

[18] Jocelyn J. Chang and et al. “SocraHealth: Enhancing Medical Diagnosis and Correcting Historical Records”. In: *The* 10th *International Conf. on Computational Science and Computational Intelligence*. 2023.

[19] Nurendra Choudhary and Chandan K. Reddy. *Complex Logical Reasoning over Knowledge Graphs using Large Language Models*. 2023. arXiv: 2305.01157 [cs.LO].

[20] Narayana Darapaneni et al. *Contextual Attention Mechanism, SRGAN Based Inpainting System for Eliminating Interruptions from Images*. 2022. arXiv: 2204.02591 [cs.CV].

[21] Yilun Du et al. *Improving Factuality and Reasoning in Language Models through Multiagent Debate*. 2023. arXiv: 2305.14325 [cs.CL].

[22] Emilio Ferrara. "Fairness And Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, And Mitigation Strategies". In: *ArXiv* abs/2304.07683 (2023).

[23] K. Gödel. *On Formally Undecidable Propositions of Principia Mathematica and Related Systems*. Dover Books on Mathematics. Dover Publications, 2012.

[24] Adi Haviv, Jonathan Berant, and Amir Globerson. "BERTese: Learning to Speak to BERT". In: *ArXiv* abs/2103.05327 (2021). [25] Dan Hendrycks et al. *Measuring Massive Multitask Language Understanding*. 2021. arXiv: 2009.03300 [cs.CY].

[26] Jie Huang and Kevin Chen-Chuan Chang. "Towards Reasoning in Large Language Models: A Survey". In: *Findings of the Association for Computational Linguistics: ACL 2023*. July 2023, pp. 1049–1065.

[27] Lei Huang et al. *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*. 2023. arXiv: 2311.05232 [cs.CL].

[28] William James. *The Principles of Psychology*. Henry Holt and Company, 1890.
[29] Jaehun Jung et al. *Maieutic Prompting: Logically Consistent Reasoning with Recursive Explanations*. 2022. arXiv: 2205.11822 [cs.CL].

[30] Ivana Kaji, Eser Aygün, and Doina Precup. *Learning to cooperate: Emergent communication in multi-agent navigation*. 2020. arXiv: 2004.

01097 [cs.LG].

[31] Akbir Khan et al. *Debating with More Persuasive LLMs Leads to More Truthful Answers*. 2024. arXiv: 2402.06782 [cs.AI]. [32] Carl Lange. “The Mechanism of the Emotions”. In: *The Classical Psychologists* (1885), pp. 672–684.

[33] Pengfei Liu et al. “Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing”. In: *ACM Comput. Surv.* 55.9 (2023).

[34] James Manyika and Sissie Hsiao. *An overview of Bard: an early experiment with generative AI*. 2023. url: https://ai.google/static/ documents/google-about-bard.pdf.

[35] Conor McHugh and Jonathan Way. “What is reasoning?” In: *Mind* 127.505 (2018), pp. 167–196.

[36] David E Newman-Toker, Najlla Nassery, and et al. “Burden of serious harms from diagnostic error in the USA”. In: *BMJ Quality & Safety* (2023).

[37] OpenAI. *ChatGPT*. 2021. url: https://openai.com/blog/chatgpt/.

[38] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL]. url: https://arxiv.org/abs/2303.08774.
[39] OpenAI. “How do davinci and text-davinci-003 differ?” In: *OpenAI Help Page* (2023). url: https://help.openai.com/en/articles/ 6643408-how-do-davinci-and-text-davinci-003-differ.

[40] Larry Page. *The PageRank Citation Ranking: Bringing Order to the Web*. 1998. url:: 66.pdf.

[41] Otavio Parraga, Martin D. More, Christian M. Oliveira, et al. “Fairness in Deep Learning: A Survey on Vision and Language Research”. In: *ACM Comput. Surv.* (2023). Just Accepted. issn: 0360-0300. doi: 10.1145/3637549. url: https://doi.org/10.1145/3637549.

[42] Pranay Patil. *Kaggle Disease Symptoms Description Dataset*. 2020. url: https://www.kaggle.com/-datasets/-itachi9604/-diseasesymptom-description-

dataset.

[43] Richard Paul and Linda Elder. “Critical Thinking: The Art of Socratic Questioning”. In: *Journal of Developmental Education* 31 (2007), pp. 34–35.

[44] Judea Pearl. *Causality: Models, Reasoning and Inference*. 2nd. Cambridge University Press, 2009. isbn: 978-0521895606.
[45] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988. isbn: 0-934613-73-7. [46] Maarten Sap et al. *Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs*. 2023. arXiv: 2210.13312 [cs.CL].

[47] Timo Schick and Hinrich Schütze. “Exploiting Cloze Questions for Few-Shot Text Classification and Natural Language Inference”. In: *Conference of the European Chapter of the Association for Computational Linguistics*. 2020.

[48] Melanie Sclar et al. *Minding Language Models’ (Lack of) Theory of Mind: A Plug-and-Play Multi-Character Belief Tracker*. 2023. arXiv: 2306.00924 [cs.CL].

[49] Gemini Team et al. *Gemini: A Family of Highly Capable Multimodal Models*. 2023. arXiv: 2312.11805 [cs.CL].

[50] Romal Thoppilan et al. *LaMDA: Language Models for Dialog Applications*. 2022. arXiv: 2201.08239 [cs.CL]. url: https://arxiv.org/abs/2201.08239.

[51] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023. arXiv: 2307.09288 [cs.CL]. url: https://arxiv.org/abs/2307.09288.

[52] Wen-Kwang Tsao. “Multi-Agent Reasoning with Large Language Models for Effective Corporate Planning”. In: *The* 10th *International Conf. on Computational Science and Computational Intelligence*. 2023.

[53] Karthik Valmeekam et al. “Large Language Models Still Can’t Plan (A Benchmark for LLMs on Planning and Reasoning about Change)”. In:

NeurIPS 2022 Foundation Models for Decision Making Workshop. 2022.

[54] Peter Cathcart Wason and Philip Nicholas Johnson-Laird. *Psychology of reasoning: Structure and content*. Vol. 86. Harvard Univ. Press, 1972.

[55] Jason Wei et al. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. 2023. arXiv: 2201.11903 [cs.CL]. [56] Wikipedia. *Socratic method*. 2023. url: https://en.wikipedia. org/wiki/Socratic_method.
[57] Chase B. Wrenn. *Internet Encyclopedia of Philosophy*. 2023. url: https://iep.utm.edu/republic/.

[58] Shunyu Yao et al. *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*. 2023. arXiv: 2305 . 10601 [cs.CL]. url: https://arxiv.org/pdf/2305.10601.pdf.

[59] Zheng Yuan et al. *Scaling Relationship on Learning Mathematical Reasoning with Large Language Models*. 2023. arXiv: 2308 . 01825 [cs.CL].

[60] Andy et al Zeng. *Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language*. 2022.
[61] Honghua Zhang et al. *On the Paradox of Learning to Reason from Data*. 2022. arXiv: 2205.11502 [cs.CL].

[62] Yifan Zhang et al. *Cumulative Reasoning with Large Language Models*. 2023. arXiv: 2308.04371 [cs.AI]. url: https://arxiv.org/ abs/2308.04371.

[63] Lianmin Zheng et al. *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. 2023. arXiv: 2306.05685 [cs.CL].

6 Theoretical Pillars of Effective LLM Communication

Abstract This paper introduces EVINCE (Entropy Variation and INformation CompetencE), a cutting-edge dialogue framework that orchestrates adversarial debates and collaborative insights among multiple large language models (LLMs). By integrating advanced principles from

conditional statistics, information theory, and in-context learning, EVINCE masterfully balances the exploration of diverse perspectives with the exploitation of established priors. Central to our innovation is the validation of the dual entropy theory, which we developed to determine the optimal pairing of LLMs with one high and one low entropy for enhanced probabilistic prediction accuracy. We also employ several information-theoretic metrics, such as mutual information, cross-entropy, Wasserstein distance, and Jensen-Shannon divergence, to measure communication opportunities, dialogue progress, and convergence. This meticulous approach fosters an interpretable and productive multi-LLM dialogue, leading to more informed and reliable outcomes. We illustrate EVINCE's potential by applying it to healthcare, demonstrating its effectiveness in improving disease diagnosis, and discuss its broader implications for enhancing decision-making across various domains.

6.1 Introduction

Ensemble approaches in machine learning, where multiple predictors combine to address classification and regression tasks, have consistently demonstrated superior performance compared to individual models [9, 18, 20]. The diversity of errors across these models is a crucial factor in their effectiveness. Recent research has explored extending this ensemble concept to Large Language Models (LLMs) collaborating on classification, question answering, and other tasks [5, 11, 22, 24]. While initial findings suggest accuracy improvements similar to traditional ensemble methods, multi-LLM collaboration holds the potential for much broader impact. As noted by [6], this approach can unearth novel perspectives, mitigate biases, and even contribute to creative endeavors like writing a novel, thereby extending its capabilities far beyond accuracy gains.

Achieving optimal performance in multi-LLM ensembles requires more than simply maximizing error diversity. A critical balance must be struck between confident, well-supported predictions and the exploration of novel and diverse perspectives. To facilitate this balanced approach, we introduce EVINCE (Entropy Variation through INformation CompetencE), a framework designed to foster structured debates among multiple LLMs, thereby maximizing prediction accuracy while encouraging the exploration of alternative viewpoints to mitigate biases. EVINCE represents a new paradigm in collaborative LLM research, effectively navigating the trade-off between exploration and exploitation in joint predictions. EVINCE rests on three key theoretical pillars:

Conditional Statistics: Conditional Statistics: By placing LLMs in adversarial stances and demanding rigorous justification for their positions, EVINCE leverages in-context learning to elicit from the opposing LLMs diverse perspectives backed by robust reasoning and evidence. This method, rooted in the Bayesian framework of conditional statistics [12, 4, 34], effectively modifies the linguistic behaviors of LLMs, shifting them away from the default optimization for maximum likelihood next-token prediction.

Dual Entropy: Our theoretical proof (via Jensen's Inequality) (Chapter 6.3.3) and empirical studies (Chapter 6.4) reveal a key insight: optimal accuracy in a two-LLM ensemble is achieved when the agents begin with differing levels of entropy. Specifically, one LLM should initially exhibit high prediction entropy, signaling a willingness to explore diverse perspectives, while the other should maintain low entropy, emphasizing precision and stability. This dual entropy configuration maximizes the ensemble's ability to balance exploration and exploitation, as the high-entropy LLM introduces a wider range of possibilities, including those that may challenge or counteract potential biases in the low-entropy LLM's initial predictions. Meanwhile, the low-entropy LLM acts as a stabilizing force, grounding the exploration in a foundation of established knowledge. Through a process of communication and reasoning, evaluated by the Socratic method and metrics from information theory (which we will elaborate on in the subsequent discussion), the two agents converge towards a collaborative and accurate prediction, ideally mitigating biases that may have been present in either agent's initial viewpoints. This finding challenges the traditional notion that faster agreement among agents necessarily leads to better outcomes, highlighting the importance of initial diversity in avoiding tunnel vision and fostering robust decision-making.

From Divergence to Conciliatory: EVINCE begins by positioning two agents in a state of dual entropy, then fosters effective information exchange between LLMs to gradually reduce cross entropy and Wasserstein distance, and maximize mutual information in their prediction distributions. This enhances the depth and breadth of their predictions. The framework initiates debates with high contentiousness [6], using mutual information to quantify the potential for productive communication. As the diversity of predictions, measured by the divergence metrics, decreases below a threshold, contentiousness is modulated, encouraging collaboration. This culminates in a joint prediction, accompanied by explainable arguments and counterarguments.

Diversity in predictive modeling can introduce noise, while an overly strong belief in existing perspectives may hinder the exploration of new ideas. To address these challenges, EVINCE employs several proxy metrics in conjunction with a “contentiousness” parameter to achieve a balance. By reasoning through and analyzing several case studies, we demonstrate how EVINCE enhances prediction accuracy, robustness, and stability. The framework facilitates a debate process where rigorous arguments and counterarguments are recorded, making the decision-making process transparent. Transparency allows humans to understand the recommendations clearly, provide feedback, and make final predictions that are well-informed, encompassing a comprehensive range of pros and cons.

The main contributions of this chapter are:
1. EVINCE **Framework Design**: Different from using debate as a way to improve accuracy via redundancy, EVINCE’s approach is vastly different and thus facilitates information discovery, bias mitigation, and decisionmaking that requires both breadth and depth of information.

2. **Theoretical Foundations**: We establish a theoretical basis for EVINCE, rooted in conditional Bayesian statistics, mutual information, and dual entropy. These principles are applied to measure, monitor, and modulate collaborative LLM interactions, contributing to a deeper understanding of how LLMs can effectively cooperate for improved decision-making. The dual entropy theory is novel and ground-breaking, illustrating how a productive decision-making process should start with room for diverse input and stable objectives, and then, through information exchange, converge to optimal decision.

3. **Empirical Validation**: We provide empirical validation of EVINCE’s underlying maxims and theories, highlighting the framework’s effectiveness

in balancing exploration and exploitation to enhance prediction accuracy. We also introduce a set of maxims derived from our empirical findings, offering practical guidance for optimizing mutual information and minimize various divergence measures.

6.2 Related Work

The core objective of adversarial debate, as embodied in EVINCE, is to foster diverse opinions and challenge assumptions, ultimately leading to more comprehensive and informed decision-making. This contrasts with traditional ensemble learning methods, which prioritize error diversity for improved accuracy.

6.2.1 Ensemble and Multi-Agent Learning

Ensemble methods like Bagging [2], Boosting [13], and Mixtures of Experts [16] have focused on combining predictions from multiple models to improve overall accuracy. Early multiple LLM frameworks starting from Glam [10] also followed this trend [5, 11, 22, 24].

EVINCE distinguishes itself by prioritizing the generation of diverse predictions to explore a wider range of perspectives. Recent research on multiLLM collaboration, building on in-context learning and Bayesian frameworks [34, 35], has shown promising results. However, the challenge remains in effectively moderating communication between LLMs. EVINCE addresses this by employing quantitative measures to calibrate and adjust individual LLM behaviors, contributing to the growing field of multi-agent LLM communication [1, 5, 14, 21, 22, 24, 32].

6.2.2 Metrics for Managing Diversity, Contentiousness, Information Quality, and Convergence

EVINCE employs various metrics to manage the debate's dynamics and progress:

• **Fostering Diversity & Quality**: Shannon entropy and relative entropy measure diversity of perspectives [8, 30], while the CRIT algorithm assesses argument quality [7].

• **Balancing Exploration & Stability**: Correlation coefficients track opinion evolution and debate stability [3], Wasserstein Distance measures prediction distribution differences [17, 29, 33], and Mutual Information quantifies information overlap [8].

• **Examining Information Overlap & Termination**: Jensen-Shannon Divergence assesses distribution similarity [23], Cross Entropy measures asymmetric differences [31], and Kullback-Leibler Divergence reveals asymmetric differences between probability distributions [19].

Chapter 6.3 details how EVINCE utilizes these metrics to balance exploration and exploitation, leading to optimal predictions. The dual entropy theorem provides further theoretical justification for the framework.

6.3 Maxims, Algorithm, and Theorem

Metric
Cross Entropy [31]

Entropy Shannon [30]

JensenShannon Diverg. [23]

KL Diverg. [19]

Pros
Measures how well the predictions of one model fit the actual distribution of another models outputs (asymmetric). Indicates level of diversity; high suggests exploration of possibilities, and low for confidence on few choices
Symmetric, bounded (0 to 1), an interpretable measure of distributional differences.
Measures difference between two probabilistic distributions.

Mutual Info [31]

Wasserstein Distance
(WD) [17] Measures reduction of uncertainty; symmetric.
Direct measure of how similar or different the model outputs are; symmetric relationship.

Cons
Computationally intensive with large models and data sets; sensitive to the exact nature of prob dists.
High entropy might indicate noise rather than useful diversity; low entropy might mask important variability.
May be less sensitive to small differences between distributions.

Asymmetric; not well-defined if a distribution has zero probabilities
Does not indicate the direction of info flow.
Not bounded but can be normalized or bounded for consistent interpretation.

Remedies
Optimize
strategies; use tions or sampling methods to manage large data sets or complex models.
computation

approxima

Use critical reading methods (Appendix A) to assess argument quality; implement noise detection to differentiate between useful diversity and noise.
Increase sensitivity settings or resolution of the metric; combine with other metrics to capture finer distinctions between distributions.
Use smoothing techniques to avoid zero probabilities; consider symmetric alternatives like JS divergence
Supplement with direction info metrics; normalized with max entropy of A and B. Define context-specific bounds for low, medium, and high divergence; consider normalizing it for non-directional comparisons.

Table 6.1: Summary of metrics for assessing LLM debates

Problem Statement: Organize a structured debate between two equally competent large language models (LLMs), LLM_A and LLM_B, to conduct t rounds. At each round t, each model produces a probability distribution, denoted as $P^{(t)}$ and $P^{(t)}$, over C possible outcomes, accompanied by $\text{sup}_{A\ (tB}$

(t)

porting arguments
R
)

$_A$ and R_B . The goal is to design an iterative debate process that leverages the structured exchange of arguments to enable the models to converge on an optimal prediction distribution P^* across the C classes.

6.3.1 Maxims with Theoretical Foundations

Progress towards the optimality goal is guided and measured by metrics introduced in Chapter 6.2. This section explains how they can be used in complementary ways to facilitate proper trade-offs between diversity and convergence, exploration and exploitation, and several other factors.

Maxim #1: Orchestrate Two Equally Competent LLMs in Structured Debate: Integrating two equally competent LLMs ensures a balanced exchange of insights and avoids bias. This adversarial setup fosters diversity in predictions, each supported by justifications, promoting critical evaluation and uncovering potential blind spots.

How? Choosing LLMs with comparable performance on a shared validation set, a balanced debate can be ensured. Suitable models include GPT-4, Claude, and Gemini. Conditioning different instances of the same LLM to support opposing stances on a subject matter can also be effective due to the theoretical justification of in-context learning with conditional Bayesian statistics [34].

Maxim #2: Encourage the Accurate Rather Than the "Popular" Prediction: Typically, LLMs, with their maximum likelihood next-token prediction objective, tend to favor the most popular predictions. By conditioning LLMs within specific contexts, we can prioritize specific stance over popularity, mitigating confirmation biases.

How? Using the proxy metrics in Table 6.1, EVINCE dynamically adjusts the "contentiousness" level in debates (see Appendix G for details). These metrics quantify agreement, diversity, and mutual information, promoting productive information exchange and enhancing prediction quality.

Maxim #3. Combine Predictions Weighted by Diversity and Quality: Weighting the probability distributions from two LLMs based on diverse probabilistic insights and argument quality.

How? Following these three maxims:
• **Maxim #3.1 Prediction Reliability:** Estimate the reliability of pre dictions using entropy-based measures to quantify uncertainty and information content. Typically, lower entropy indicates higher confidence in a prediction, suggesting higher reliability.

• **Maxim #3.2 Argument Quality:** Evaluate the quality of supporting arguments using techniques inspired by the Socratic method. This includes identifying logical fallacies, assessing the relevance and credibility of evidence.

• **Maxim #3.3 Aggregation:** Employ a weighted aggregation method, such as a Bayesian model to combine weighted predictions accounting for both probabilistic insights and the quality of supporting arguments.

Maxim #4. Evaluating the Convergence Rate of the Predictions Across the Rounds: This aspect focuses on measuring how quickly and effectively
Algorithm 1 Specifications of Algorithm EVINCE

1: **Input:** Information set S, Class labels C; Two equally competent LLMs: LLM_A and LLM_B (**Maxim #1**);

2: **Output:** P_f, final probability distribution over C;

3: **Variables:** t: debate round; R =∅ aggregated arguments;

P

$^{(t)}$ $P^{(t)}$: prediction distributions of LLM_A and LLM_B on C of round t; $R^{(t)}$

A ’ B A ’

$R^{(t)}$: supporting reason sets;$_B$

Δ = 90%: debate contentiousness, initialize to high to foster adversary between LLMs (**Maxim #2**);

p: prompt = “Predict topk probability distribution on C with S and R at contentiousness Δ”;

4: **Functions:** CRIT(d) [7], Critical Reading Inquisitive Template for evaluating argument quality;

ARA [15], Algorithmic Robust Aggregation for optimal prediction aggregation (**Maxims #3**);

5:

Initial Predictions t = 0**:**

LLMs generate their predictions in probability distributions with supporting reasons:

$(P_A^{(t=0)}, R_A^{(t=0)}) = LLM_A(S, p)$, $(P_B^{(t=0)}, R_B^{(t=0)}) = LLM_B(S, p)$. 6: **Debate Iterations:**

6.1. **Update Predictions:**

Calculate the confidence-based weights using the inverse of entropy (**Maxim #3.1**):

$\alpha = 1/(H(P_A^{(t)}) + 1)$, $\beta = 1/(H(P_B^{(t)}) + 1)$.

Use the blending mechanism to update predictions (**Maxim #3.3**): $P_A'^{(t)} = \alpha P_A^{(t)} + (1 - \alpha) P_B^{(t)}$,

$P_B'^{(t)} = \beta P_B^{(t)} + (1 - \beta) P_A^{(t)}$.

6.2. **LLMs Generate New Predictions:** Both LLMs use accumulated $R = R_A^{(t)} \cup R_B^{(t)}$.

$(P_A^{(t+1)}, R_A^{(t+1)}) = LLM_A((P_A'^{(t)}), R, p)$,

$(P_B^{(t+1)}, R_B^{(t+1)}) = LLM_B((P_B'^{(t)}), R, p)$.

6.3. **Exit Condition Check with Wasserstein distance (Maxim #4): If $WD(P_A^{(t+1)}, P_B^{(t+1)}) < \epsilon$ EXIT**; $t = t + 1$, $\Delta = \Delta \times 80\%$.

7: **Final Decision:** Weighted prediction by quality scores of the evaluator e.g., CRIT (Appendix A) (**Maxim #3.2**):
$P_f = \Omega_A P^{(t+1)} + \Omega_B P^{(t+1)} / \Omega_A + \Omega_{B \cdot A\,B}$
the predictions from the LLMs converge over successive rounds, assessing the efficiency of the debate and aggregation mechanisms.

How? Convergence is assessed by measuring mutual information and using proxy metrics such as Wasserstein distance. When the mutual information is low or the similarity between predictions is high, the debate is considered to be converging.

6.3.2 Algorithm Specifications

With all proxy metrics and their pros, cons, and combined strengths comprehensively surveyed, and also examined by our two experiments documented in Chapter 6.4.2 and 6.4.3, Algorithm 1 formally specifies the algorithm of EVINCE with the maxims.

6.3.3 Entropy Duality Theorem (EDT)

Theorem EDT: Optimal Pairing of LLMs for Probabilistic Prediction Accuracy. The optimal pairing of LLMs for diagnosis accuracy, in terms of stability, accuracy, and robustness, occurs when the LLMs are 1) equivalent in the quality of the information they process, and 2) exhibit contrasting entropy values in their prediction distributions–one high and one low. [**Proof**]: In Appendix A.

6.4 Empirical Study

This empirical study investigates the application of EVINCE to disease diagnosis, leveraging large language models (LLMs) as diagnostic tools. We aim to validate the following three hypotheses:

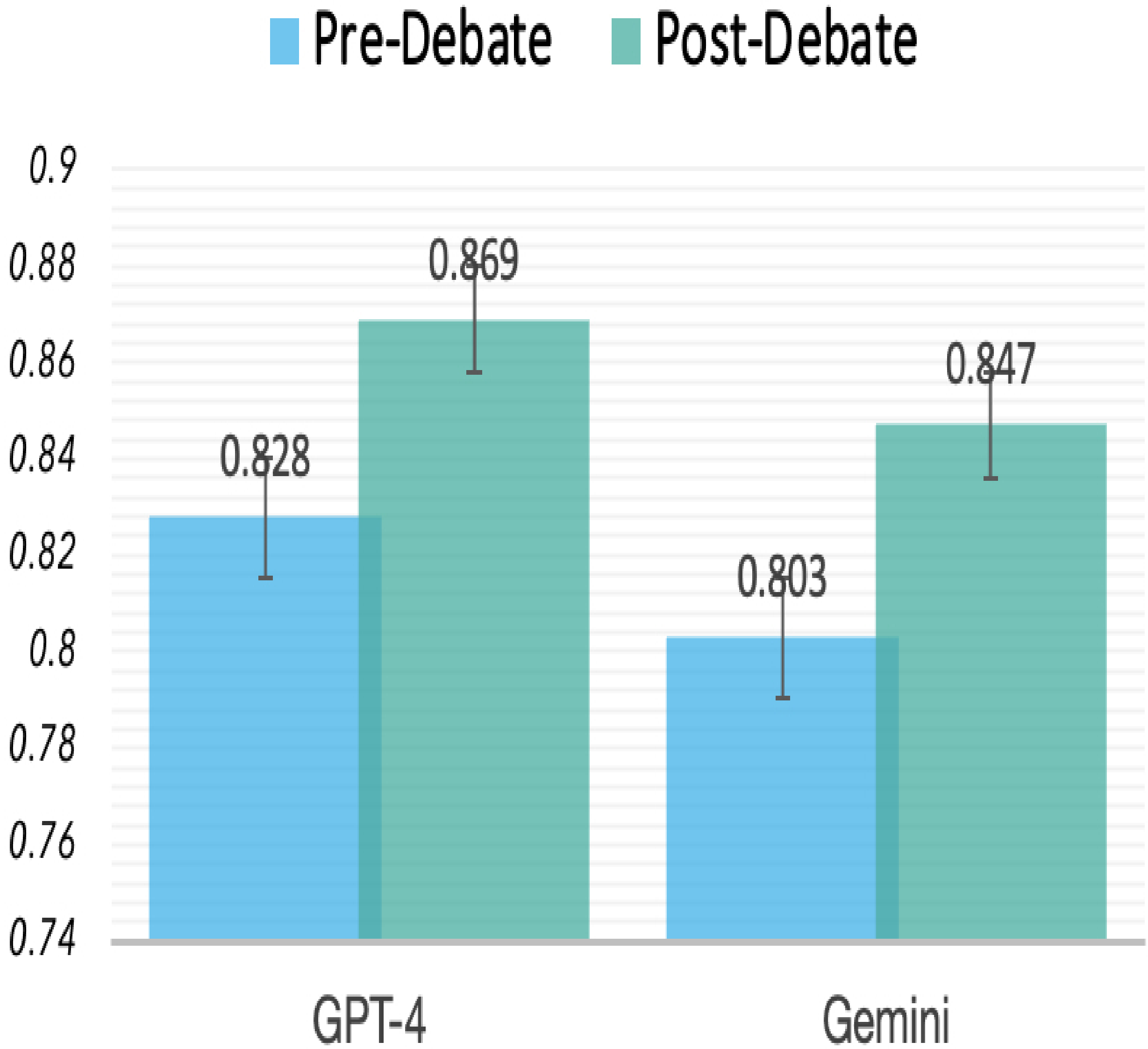
1. *Contentiousness & Prediction Quality*: Initial LLM disagreement (measured by Wasserstein distance) increases with higher initial contentiousness but decreases as debate progresses. Individual LLM prediction uncertainty (Shannon entropy) will follow a similar pattern.

2. *EDT Effectiveness & Confusion Matrices*: LLM pairs following the Entropy Duality Theorem (EDT) will have complementary error patterns, leading to higher combined prediction accuracy than non-EDT pairs.

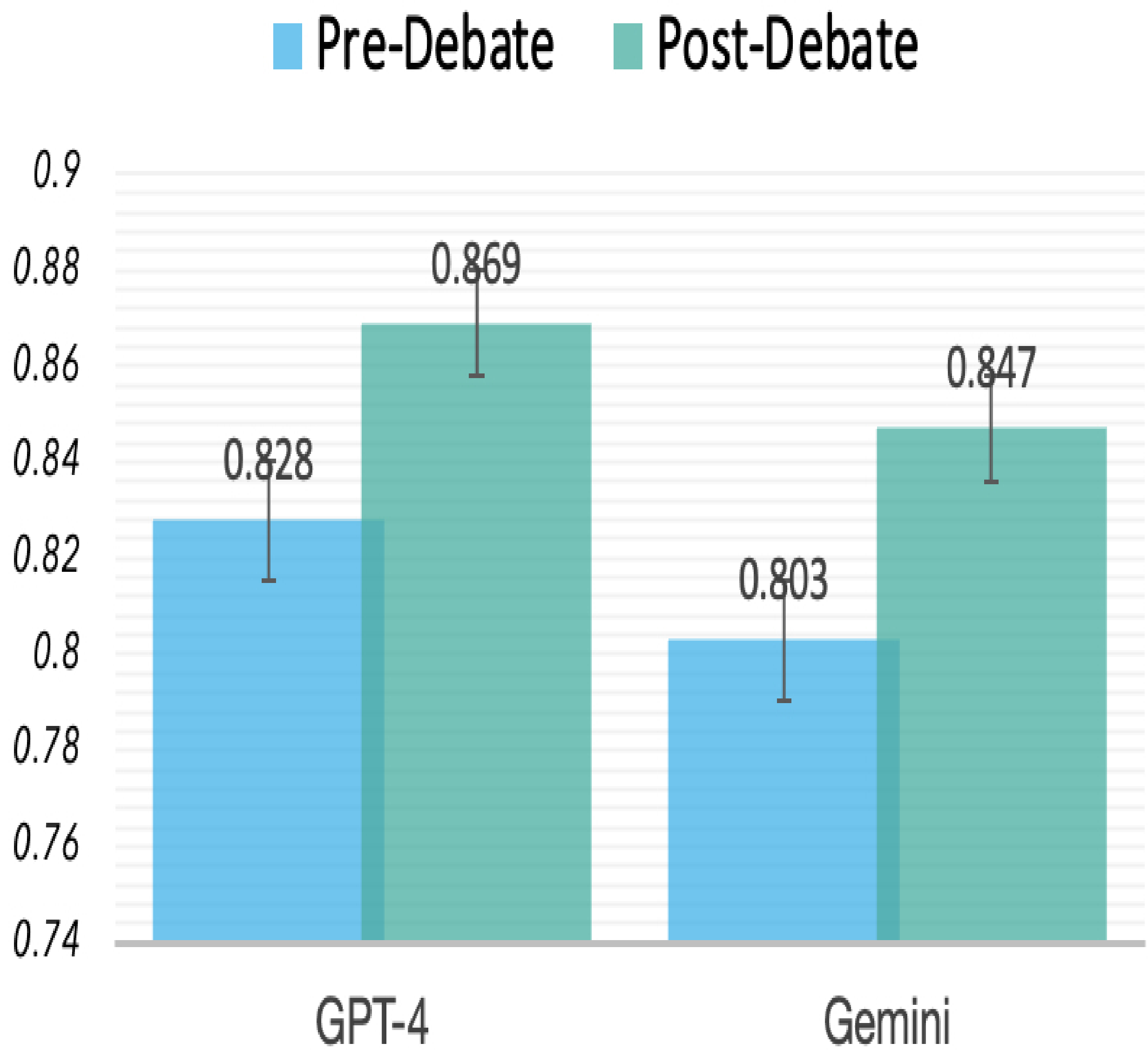
3. EVINCE *& Historical Misdiagnoses*: EVINCE, applied to real-world data, will improve diagnostic accuracy and identify potential misdiagnoses or ambiguities within the ground truth.

Problem Statement: Given a set of symptoms, denoted as S, and a context κ, the goal is to predict a probability distribution of topk diseases over C possible diseases. This is represented as $P = \text{LLM}(S, \kappa)$, where each LLM generates topk predictions on C ($k \leq C$) based on the input symptoms S and context κ.

P = (p(top 1 to k∈ D| S, κ) .



(a) GPT4 pairs Claude



(b) GPT4 pairs Gemini Figure 6.1: Pre-/post-debate accuracy on all patients on all diseases shows EVINCE helps

Context κ is where dual entropy is adjusted through three knobs: temperature, the k of topk, and the contentious level Δ. A distribution tends to have high entropy when all three knobs are set high, and vice versa.

Resources, Dataset & Data Preparation: Our study utilizes a dataset obtained from Kaggle [27], which comprises 4,921 patient records. Each record includes the diagnosed disease along with up to 17 symptoms such as fever, cough, fatigue, itchiness, and difficulty breathing. We first remove duplicates from the dataset, resulting in 304 unique diagnostic instances spanning 40 diseases. (The refined dataset is uploaded as supplementary data.) Each instance acts as a test case where EVINCE utilizes the inherent knowledge of LLMs (GPT-4, Gemini, and Claude3) instead of training them through few-shot techniques on this specific dataset. Our computing resources are sponsored by Azure, with a monthly budget of US$500.

Evaluation: We evaluate the quality of predictions using the top-k Mean Reciprocal Rank (MRR). If one of the top-k predicted diseases matches the ground truth diagnosis, the score is the reciprocal of its rank (1 for the top prediction, 1/2 for the second, 1/3 for the third, etc.). If none of the top-k predictions are correct, the score is 0.

6.4.1 Study #1: Post vs. Pre-Debate Accuracy

For each of the 304 patient instances, we employ GPT-4, Gemini, and Claude3, to perform independent disease predictions and then use EVINCE to pair them to evaluate performance gain.

In our first experiment, we set $k = 5$ for both LLM agents. One agent had a high temperature while the other had a low temperature. The contentiousness level was set very high ($\Delta = 0.9$ out of 1) to encourage significant cross entropy. Setting $k = 5$ ensures some minimal common ground, meaning the probability of shared information is sufficient to foster meaningful interaction. High contentiousness promotes counterarguments and information exchange.

Pre- and Post-Debate Evaluation We conducted two sets of experiments. First, as a baseline, we constrained disease predictions to the 40 labels in the dataset, mimicking common supervised learning assumptions. While this yielded high accuracy (95-97%), it's unrealistic for real-world diagnosis where a general practitioner considers all possibilities. This constraint also highlights the flexibility of LLMs, which are not confined by training data labels and thus less prone to over-fitting some erroneous labels (further discussed in the next two studies).

Next, we removed the label constraint to better simulate real-world conditions. In this unconstrained scenario, all 304 patient cases yielded stable results across GPT-4, Gemini-3, and Claude-3, with a standard deviation of just 1.5%. Prior to debate (light blue bars in Figure 6.1), GPT-4 led in accuracy (82.8%), followed by Gemini (80.3%) and Claude (79.5%).

Implementing EVINCE with GPT-4 and Claude-3 pairing and GPT-4 and Gemini-3 pairing consistently improved accuracy by 4-5 percentage points (green bars in Figure 6.1). The GPT-4 and Claude-3 pairing achieved 87.5% accuracy (Figure 6.1a), rivaling state-of-the-art clinical performance like the REFUEL algorithm [28].

However, the story doesn't end here. The remaining 12.5% of inaccurate cases for the GPT-Claude pairing might not be solely EVINCE's fault. If we consider the potential 11% US misdiagnosis rate reported by John Hopkins [25], this discrepancy could point to mislabeled data in the original dataset. This presents a groundbreaking opportunity: EVINCE could potentially identify and correct errors in existing datasets, a concept we explore further in Chapter 6.4.3.

6.4.2 Study #2: Confusion vs. Opportunities

Two key factors contribute to EVINCE's improved diagnostic accuracy: (1) structured debates with reasoning encourage LLMs to explore alternative diagnoses in both breadth and depth, leading to more comprehensive analysis and decision-making (see Appendices C and D); and (2) pairing high and low entropy LLMs balances exploratory diversity with exploitative stability, resulting in more robust and high-quality decisions, as demonstrated in this second study.

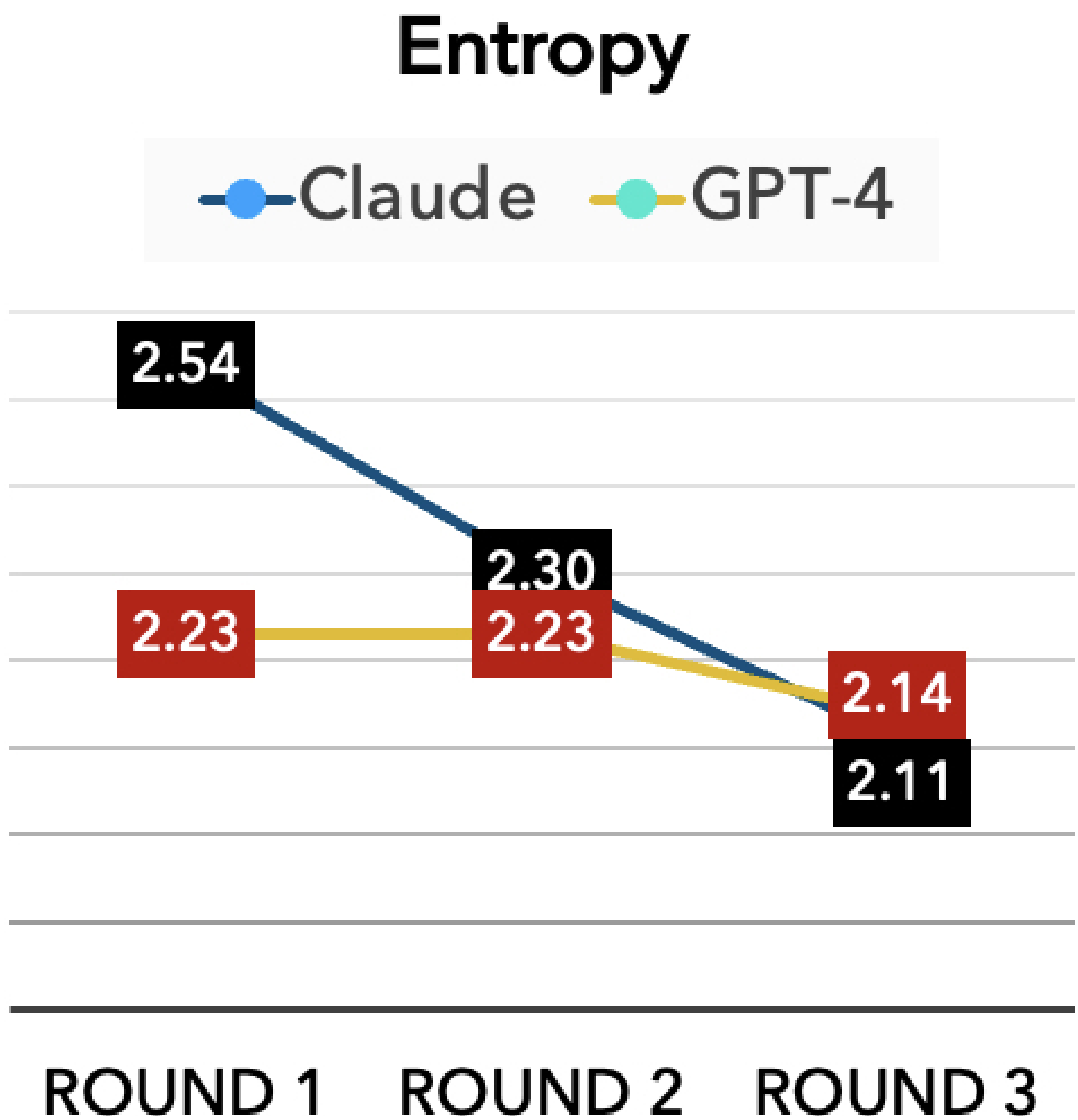
Analysis of Confusion Matrices We use confusion matrices to analyze the performance of two LLMs on diagnosing Hepatitis types A to E. GPT-4

	Hep. A	Hep. B	Hep. C	Hep. D	Hep. E
Hep. A	50%				50%
Hep. B		50%	50%		
Hep. C	100%				
Hep. D					100%
Hep. E					100%

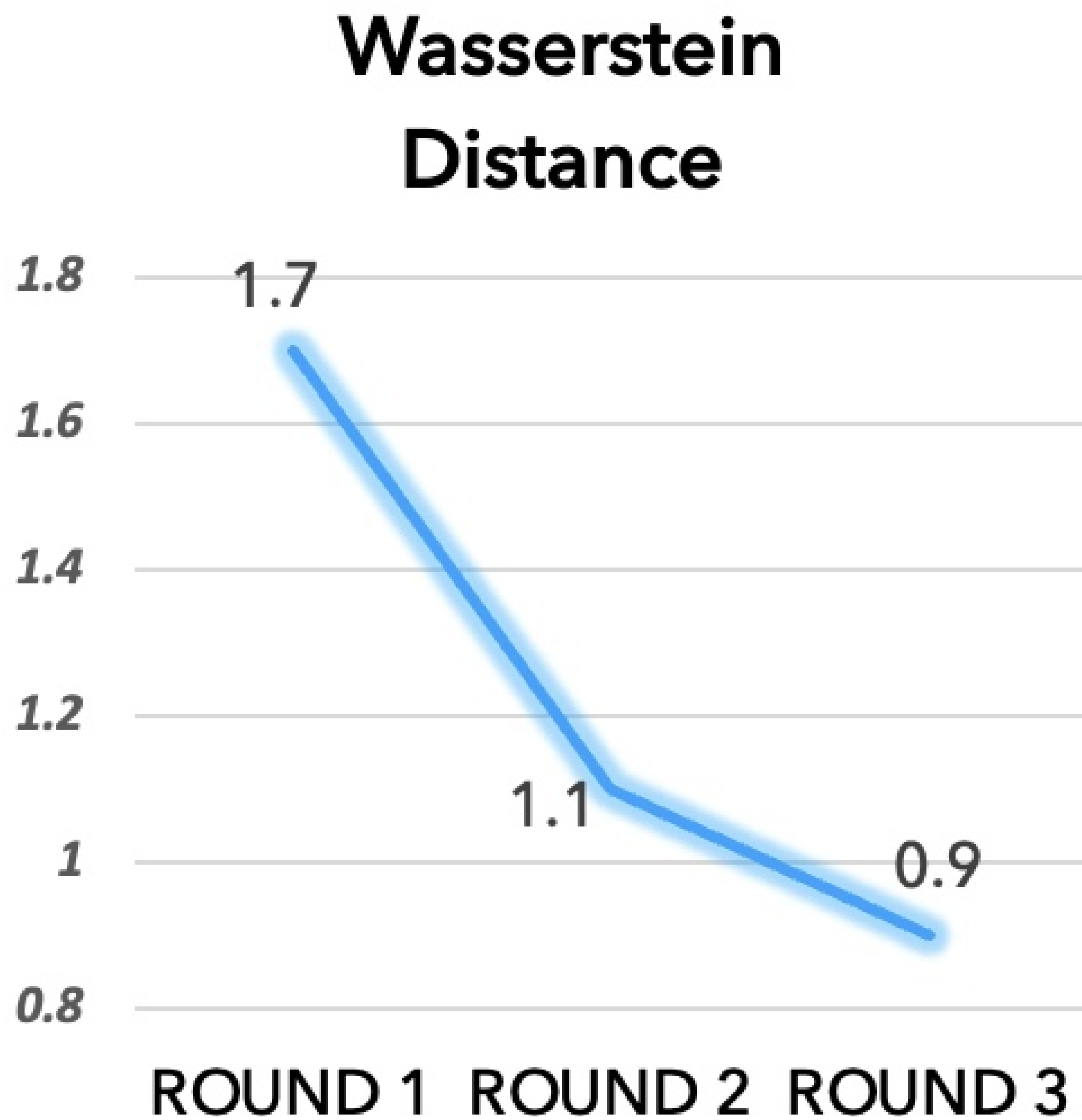
(a) GPT liver c-matrix

	Hep. A	Hep. B	Hep. C	Hep. D	Hep. E
Hep. A	74%		36%		
Hep. B		50%	50%		
Hep. C			36%	64%	
Hep. D	60%			40%	
Hep. E					100%

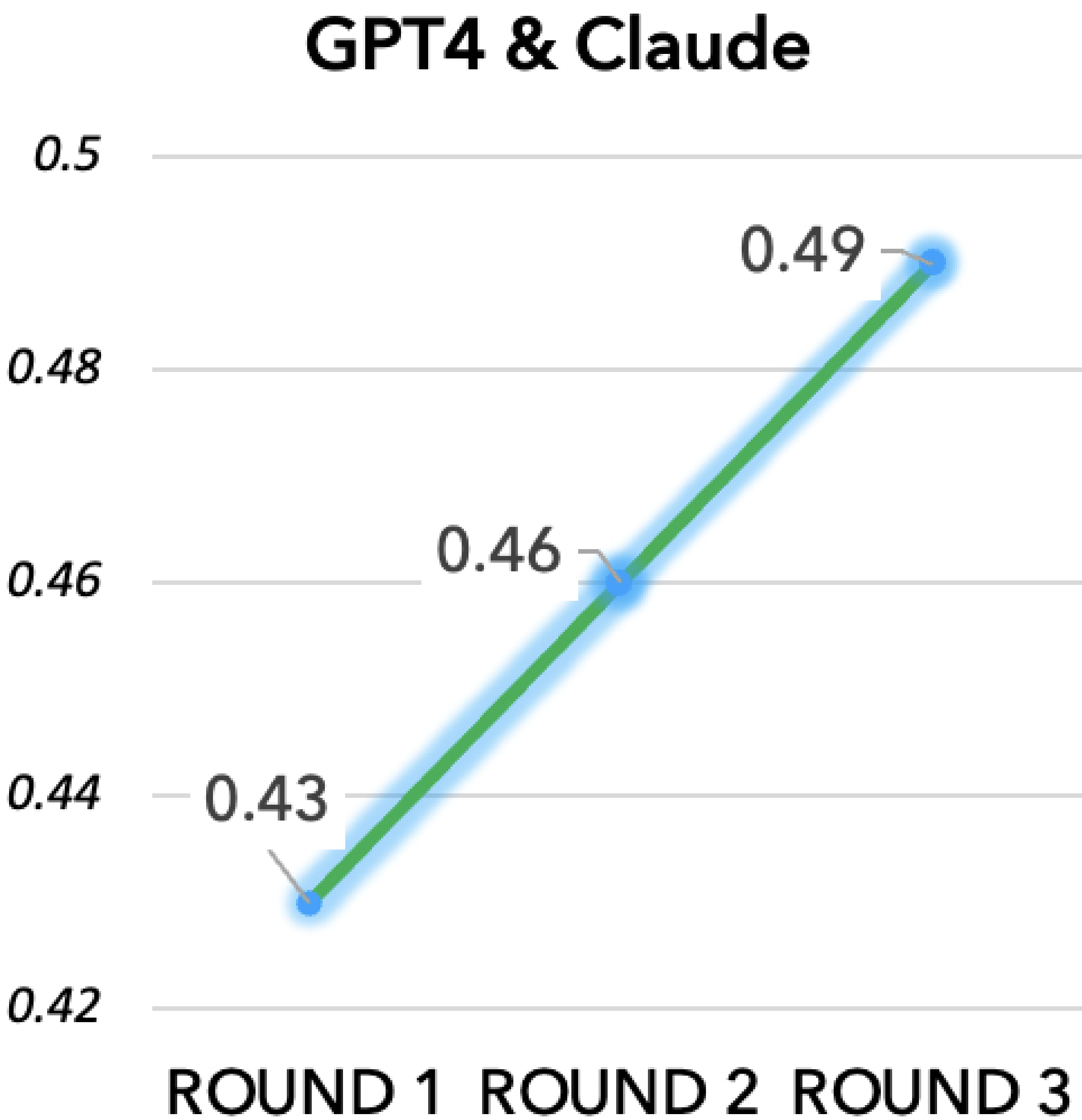
(b) Claude liver c-matrix Figure 6.2: Confusion matrices


Entropy
Claude
GPT-4
2.54
2.30
2.23
2.23
2.14
2.11
ROUND 1
ROUND 2
ROUND 3
(a)

Entropy


Wasserstein
Distance
1.8
1.6
1.4
1.2
1
0.8
1.7
1.1
0.9
ROUND 1
ROUND 2
ROUND 3

(b) WD %



(c) Norm. MI Figure 6.3: Entropy, WD, and normalized MI

shows limited accuracy, particularly for types C and D, achieving only 50% accuracy for types A and B. In contrast, Claude exhibits a wider spread of predictions across all Hepatitis types, as shown in Figure 6.2.

These matrices highlight how Claude's flexibility in exploring diverse diagnostic hypotheses can significantly aid the debate process. The initial uncertainty or "confusion" (high entropy) exhibited by Claude brings new information to the table, potentially challenging and correcting the more confident (low entropy) predictions of GPT-4, which might otherwise stubbornly persist with incorrect diagnoses. This dynamic interplay exemplifies the delicate dance between exploration and exploitation that EVINCE facilitates. By encouraging the exploration of alternative hypotheses, even when one model seems certain, EVINCE can uncover nuances and details that lead to more accurate and comprehensive diagnoses.

Observations from Information Metrics Figure 6.3a illustrates how the entropy levels of both LLMs stabilize after three rounds of debate, indicating a convergence towards a similar, stable entropy state. This convergence is corroborated by a consistent improvement in Wasserstein distance (WD) between the two models' predictions over successive rounds, as shown in Figure 6.3b. Notably, Figure 6.3c shows that the normalized mutual information (MI) between the prediction distributions of GPT-4 and Claude improves by 14%, suggesting an increase in shared information throughout the debate. Additionally, Figure 6.4 shows the consistent convergence of all divergence metrics.

Comparative Performance: EVINCE demonstrates a 5% higher accuracy rate in diagnosing specific types of liver diseases compared to a baseline approach (Figure 6.1a), underscoring its capability to handle complex diagnostic scenarios effectively.

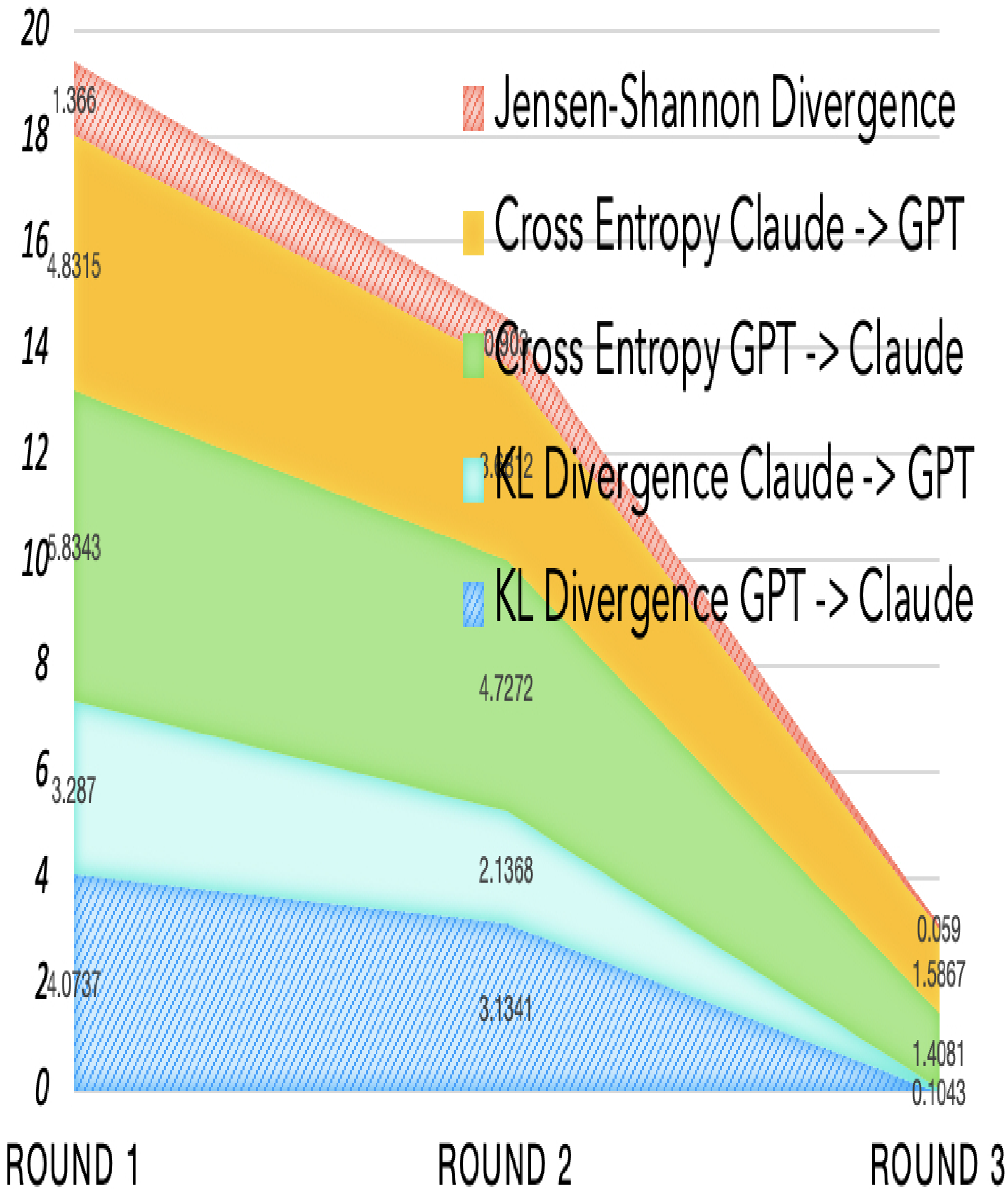
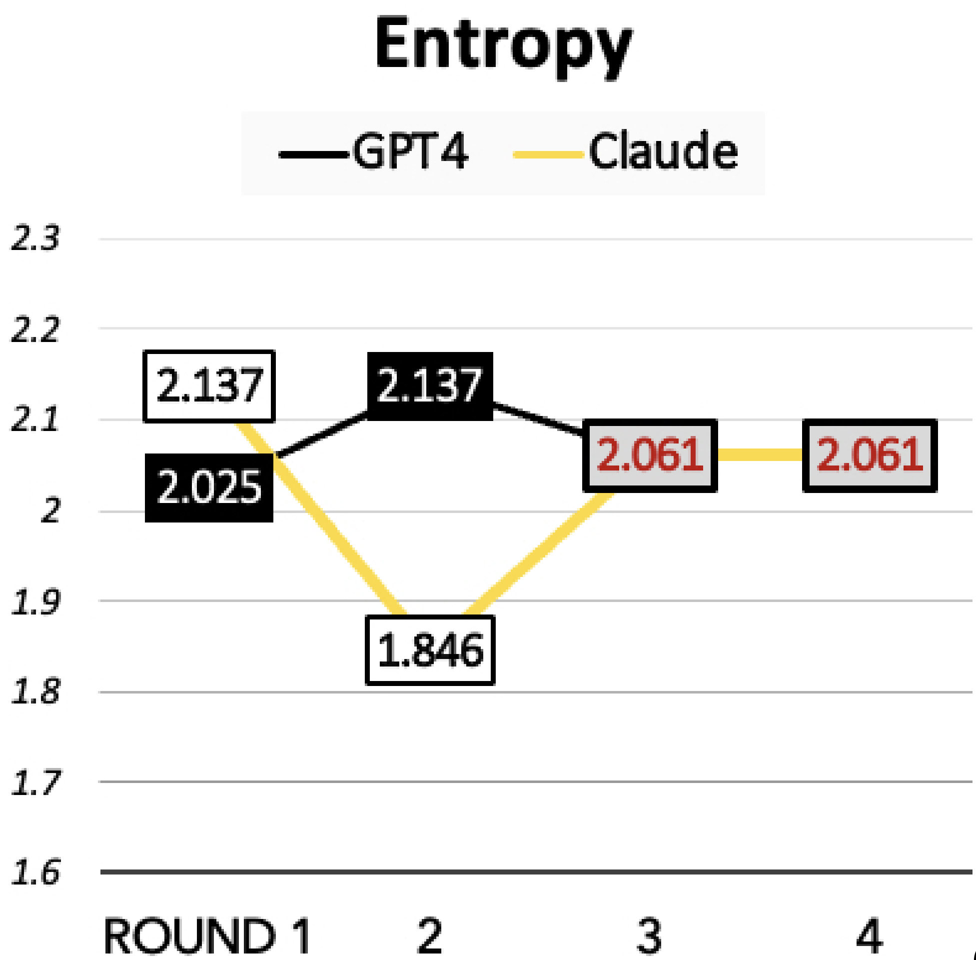


Figure 6.4: Convergence of all metrics

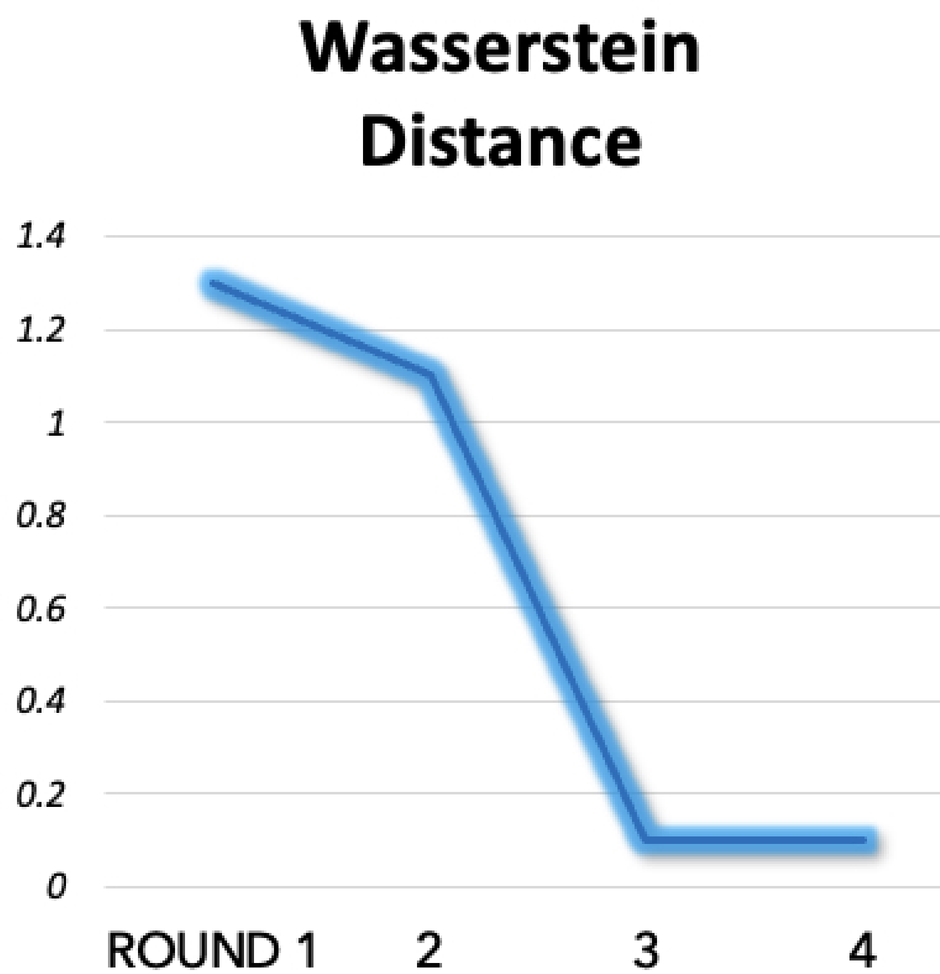
6.4.3 Study #3: Ground-Truth Remediation

This study illustrates how EVINCE can identify potential misdiagnoses, explain the reasoning behind them, and recommend corrective actions. Traditionally, machine learning scientists rely on labeled data as “ground truth.” However, as evidenced by research like that of Newman-Toker et al. (2021) [26] from Johns Hopkins, misdiagnosis is a widespread issue in healthcare systems globally. These erroneous diagnoses, often treated as ground truth, can be perpetuated by supervised learning algorithms, exacerbating the problem within the healthcare system.

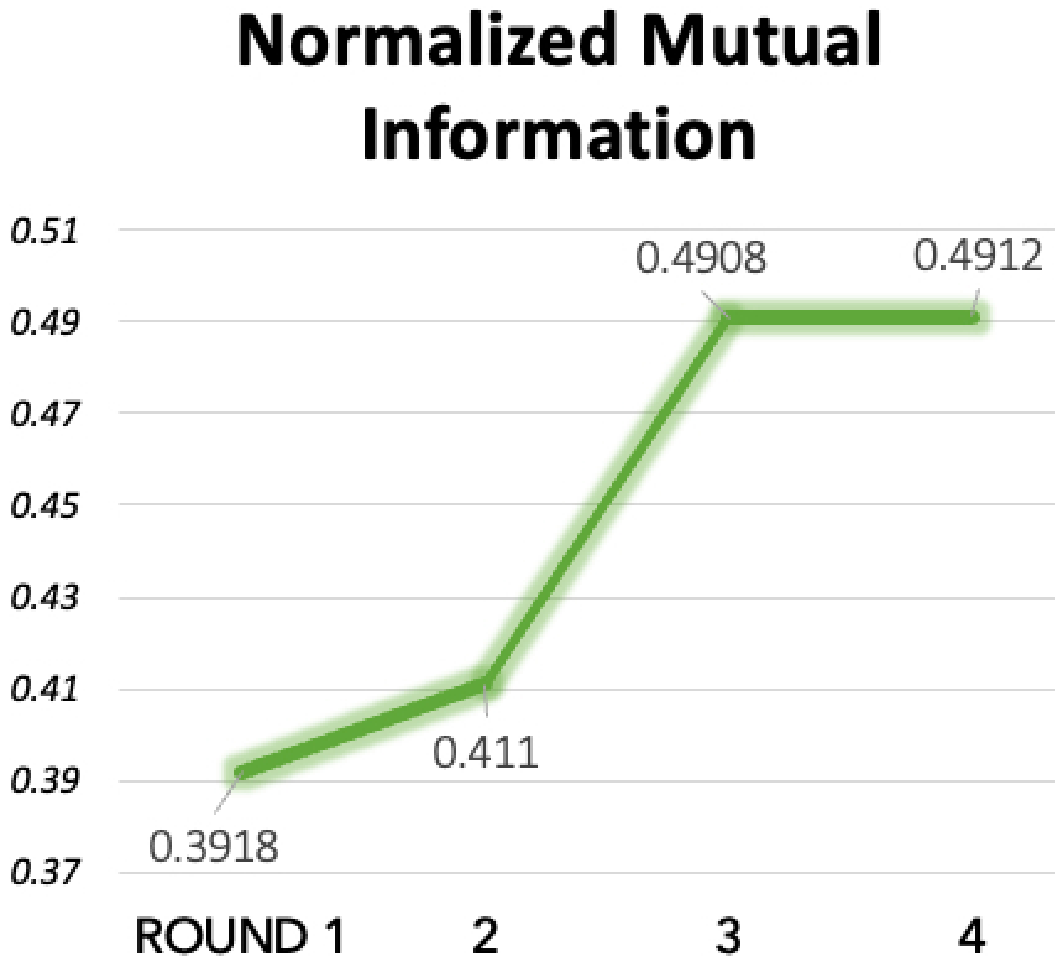
In the debate scenario detailed in Appendix D, where Jaundice is the ground truth diagnosis, Figure 6.5a illustrates initial differences between GPT-4 and Claude’s predictions. Jaundice is absent in GPT-4’s top-5 (with 0% in red), while ranked third by Claude. Although Claude influences GPT-4 to include Jaundice in its third prediction in the second round, subsequent rounds see both LLMs drop Jaundice to the fourth position of 10%.


Entropy
GPT4
Claude
2.3
2.2
2.1
2
1.9
1.8
1.7
1.6
2.137
2.025
2.137
1.846
2.061
2.061
ROUND 1
2
3
4

(a) GPT &

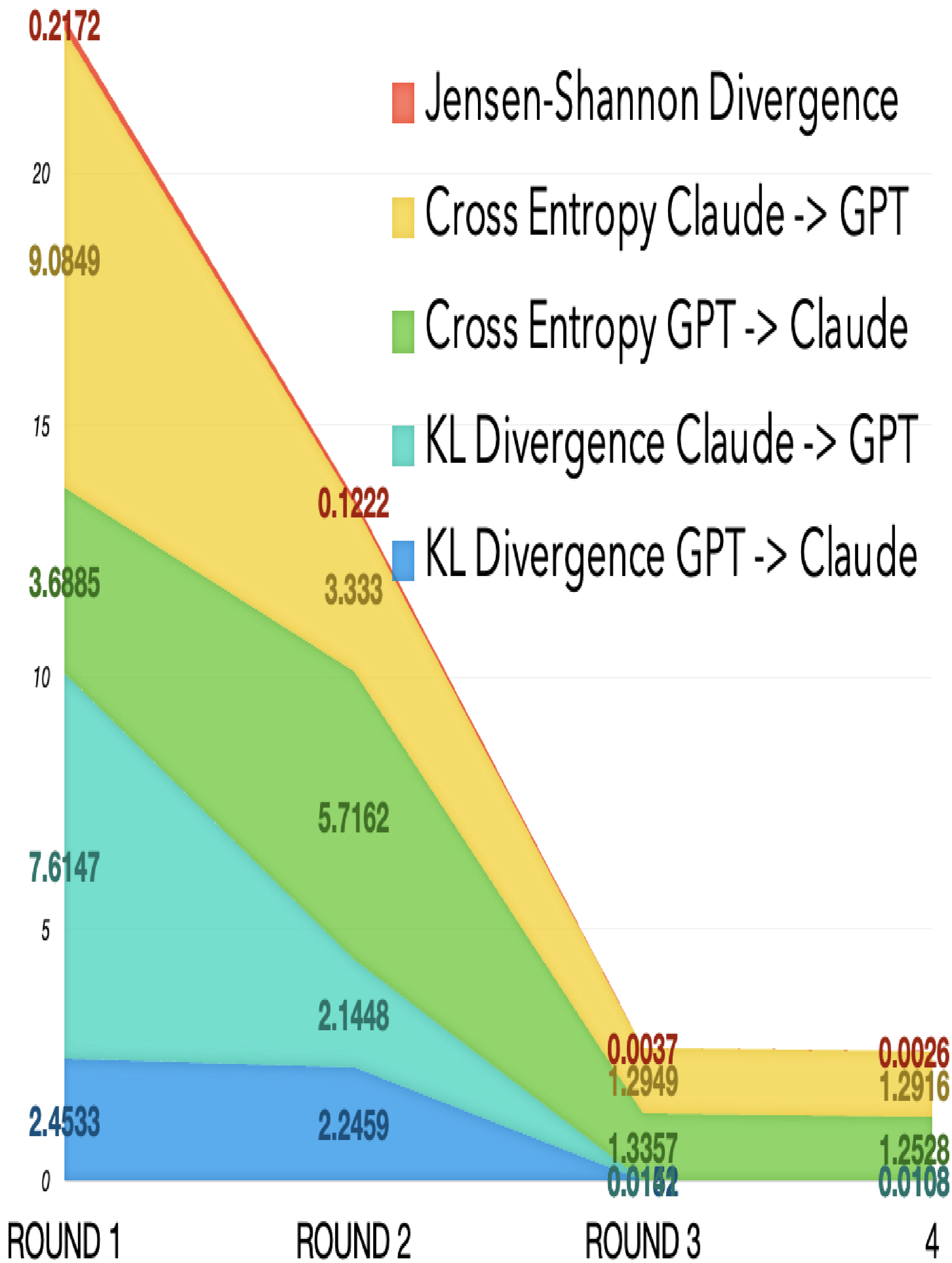


Claude top-5 WD % (b)



(c) Mutual Info. Figure 6.5: Remediation: Jaundice to Hepatitis

Meanwhile, Hepatitis A, initially GPT-4's top prediction (30% in dark blue), is quickly demoted to fifth and eventually drops out of the top-5 entirely due to Claude's influence. Hepatitis B, initially ranked second by GPT-4 and top by Claude, stabilizes in the second position in rounds 3 and 4 (in light blue). Notably, Hepatitis C rises from second place on both lists to the top position and remains there (in black).



Jensen-Shannon Divergence
Cross Entropy Claude -> GPT
Cross Entropy GPT -> Claude
KL Divergence Claude -> GPT
KL Divergence GPT -> Claude
0.2172
9.0849
3.6885
7.6147
2.4533
0.1222
3.333
5.7162
2.1448
2.2459
0.0037
1.2949
1.3357
0.0026
1.2916
1.2528
0.0108
20
15
10
5
0
ROUND 1
ROUND 2
ROUND 3
4

Figure 6.6: Convergence of all metrics

As demonstrated in the previous study, Wasserstein distance (WD) effectively measures the divergence between LLM predictions and assesses debate convergence. Figures 6.5b and 6.5c show that WD stabilizes after three debate rounds, coinciding with a plateau in normalized mutual information (MI) between GPT-4 and Claude. This stabilization suggests their predictions converge.

Figure 6.6 illustrates the convergence of all divergence metrics, including Jensen-Shannon divergence, cross-entropy, and Kullback-Leibler divergence, particularly between the second and third rounds. Although the final joint prediction for Hepatitis C reached a high consensus of 37.5%, it deviates from the actual condition of Jaundice, which the Kaggle dataset reports with 10% confidence. EVINCE provides general practitioners with alerts and suggests remedial actions (see Appendices D.9 and C.8) to address this discrepancy. Recommended actions include querying additional symptoms from the patient and conducting specific laboratory tests.

6.4.4 Experiment Remarks

EVINCE initiates debates with high contentiousness, encouraging dual prediction entropy between LLMs, as supported by the EDT theorem. It utilizes normalized mutual information (MI) to track shared knowledge accumulation throughout the debate, while Wasserstein distance (WD) and Jensen-Shannon divergence (JSD) quantify dissimilarity between LLM predictions.

These metrics (EDT, WD, JSD, MI) provide a comprehensive view of debate progress. WD and JSD assess the potential for further communication and refinement, while MI monitors shared understanding, aiding in determining the optimal stopping point.

The asymmetric nature of KL divergence and cross entropy warrants further investigation. Despite eventual convergence in our case studies, discrepancies observed in the second round, where one direction increases while the other decreases, suggest potential value in exploring asymmetric information. Future work will re-evaluate the use of these metrics if asymmetry proves beneficial.

6.5 Concluding Remarks

We have developed EVINCE, an innovative framework that enhances collaborative decision-making among Large Language Models (LLMs) through structured, adversarial debates. This framework leverages conditional statistics (in-context learning), information theory, and a novel concept called dual entropy to guide the debate, ensuring a balance between exploration and exploitation. EVINCE not only improves prediction accuracy and robustness but also produces explainable outcomes grounded in information metrics.

By assigning adversarial roles and adjusting the level of contentiousness, EVINCE encourages LLMs to explore a broader range of perspectives. Through mutual persuasion and the exchange of information, the reliability of predictions is significantly enhanced. The introduction of dual entropy theory, which pairs one LLM with high initial entropy (for diverse exploration) with another LLM with low entropy (for focused refinement), further stabilizes information exchange and promotes comprehensive consideration of various viewpoints.

Our validated Entropy Duality Theorem provides empirical evidence of EVINCE's effectiveness. In the domain of medical diagnostics, EVINCE outperforms traditional solo LLM approaches by identifying potential groundtruth errors and providing clear justifications for its conclusions. This success demonstrates the potential of EVINCE for broad application in various fields where informed decision-making is crucial.

Looking ahead, EVINCE is poised to drive further innovations in LLM collaboration across diverse domains. It represents a significant advancement in AI-human interaction, promoting a synergy of intelligence, reliability, and transparency that augments human decision-making. By ensuring that AI-supported decisions are both efficient and ethically sound, EVINCE fosters a collaborative environment where human judgment is respected and enhanced by the capabilities of advanced AI systems.

Appendix A: Proof of EDT Theorem

Theorem EDT: Optimal Pairing of LLMs for Probabilistic Prediction Accuracy. The optimal pairing of LLMs for diagnosis accuracy, in terms of stability, accuracy, and robustness, occurs when the LLMs are equivalent in the quality of the information they process, and exhibiting contrasting entropy values in their prediction distributions–one high and one low.

[Proof]: Given two LLMs, LLM_A and LLM_B, following Maxim #1 with prediction distributions P_A and P_B, respectively. The information entropy of LLM_A, $H(P_A)$, is high, and of LLM_B, $H(P_B)$, is low.

Step 1: Define the combined prediction distribution. Let the combined prediction distribution of LLM_A and LLM_B be denoted as P_C. We can express P_C as a weighted average of P_A and P_B:

$P_C = \alpha P_A + (1 - \alpha) P_B$, where $0 \leq \alpha \leq 1$ and
α is decided by CRIT in Appendix A.

Step 2: Express the information entropy of the combined prediction distribution. Using the definition of information entropy, we

calculate: $\sum P_C(x_i) \log_2 P_C(x_i)$

=

$\sum$ $H(P_C) = -$

$- [\alpha P_A(x_i) + (1 - \alpha) P^i(x_i)] \log_2 [\alpha P_A(x_i) + (1 - \alpha) P_B(x_i)].B$

i

Step 3: Apply Jensen's Inequality to the information entropy of the combined prediction distribution. convex function $f(x) = -x \log_2 x$. probabilities p_i, Jensen's inequality states that:

)
$f(\sum p_i x_i \leq \sum p_i f(x_i)$
i i

Jensen's inequality is applied to the For a convex function and a set of Thus, the entropy of the combined distribution is:

$H(P_C) \geq \alpha H(P_A) + (1 - \alpha) H(P_B)$
where equality holds when $P_A = P_B$.
Step 4: Analyze the lower bound of the combined information entropy. As $H(P_A)$ is high and $H(P_B)$ is low, we can express their relationship as: $H(P_A) = H(P_B) + \Delta$, where $\Delta > 0$. Substituting this into the inequality from Step 3, we have:
$H(P_C) \geq \alpha[H(P_B) + \Delta] + (1 - \alpha) H(P_B) = H(P_B) + \alpha\Delta$.

Step 5: Interpret the lower bound of the combined information entropy. The lower bound of $H(P_C)$, and hence the robustness of the model, is maximized when α is maximized, which corresponds to giving more weight to the high-entropy model (LLM_A). This setup facilitates the exploration of diverse possibilities and enhances robustness against noise and perturbations in the input data, while still ensuring that predictions are grounded by the more certain outcomes predicted by the low-entropy model (LLM_B).

Step 6: Conclude the proof. By combining the prediction distributions of LLM_A and LLM_B, with one having high information entropy and the other low, we achieve an optimal balance that maximizes the lower bound of the combined information entropy. This balance between exploration (high entropy) and exploitation (low entropy) optimizes overall prediction accuracy and robustness, as proved through information theory and the properties of entropy. Thus, the theorem is established.

Appendix C: EVINCE Debate - Dengue Fever vs. Chikungunya

This experiment addresses a diagnostic challenge involving multiple potential diseases. A patient presents with the following 12 symptoms: [*skin rash, joint pain, vomiting, fatigue, high fever, headache, nausea, loss of appetite, pain behind the eyes, back pain, malaise, muscle pain, red spots over the body*]. In this case, GPT-4 provides the initial diagnosis, with Gemini following suit. The confirmed diagnosis for this experiment is Dengue Fever. Given the satisfactory diversity in predictions from the two LLMs, the debate progresses and converges smoothly.

C.1 Moderator's Prompt

“Identify potential diseases linked to the symptoms (skin rash, joint pain, vomiting, etc.). Provide top-3 predictions with justifications, supplementary symptom inquiries, and relevant lab tests.”

C.2 GPT-4’s Opening Round

Top-3 Predictions: *Dengue Fever (60%), Chikungunya (25%), and Zika Virus (15%)*.
Justifications: Most symptoms are classic signs of Dengue Fever. Red spots over the body could correspond to petechiae, often seen in Dengue due to thrombocytopenia (a decrease in platelet count).

Chikungunya can cause similar symptoms to Dengue, such as fever, severe joint pain, muscle pain, headache, nausea, and rash. Joint pain from Chikungunya is often very debilitating and can be more prolonged than in Dengue.

Zika Virus symptoms can include fever, rash, joint pain, malaise, and conjunctivitis, with a milder presentation compared to Dengue or Chikungunya. The presence of a rash and conjunctivitis is notable, although Zika is less likely to cause severe muscle pain or high fever.

C.3 Gemini’s Opening Round

Top-3 Predictions: *Viral Infection (60%), Autoimmune Disease (20%), Bacterial Infection (15%)*.
Justifications: Symptoms common in viral infections. Autoimmune diseases is possible, but less likely due to vomiting and nausea. Bacterial infection is less probable given rash and muscle pain.

C.4 GPT-4’s Second Round

Top-3 Predictions: *Dengue Fever (60%), Chikungunya (25%), and Zika Virus (15%)*.
Justifications: GPT-4 maintains its predictions, countering Gemini’s assessment with three points. First, while viral infections can cause these symptoms, the specific combination suggests a more targeted diagnosis.

Second, the acute onset of high fever, joint pain, and rash is less typical for autoimmune diseases. Finally, the rash and red spots are not characteristic of bacterial meningitis.

C.5 Gemini's Second Round

Top-3 Predictions: *Dengue Fever (50%), Chikungunya (30%), and Viral Infection (20%)*.
Justifications: Gemini's initial prediction considers a Viral Infection as the most likely cause due to its broad range of potential symptoms. However, GPT-4's focus on Dengue Fever and Chikungunya, with their specific symptom profiles, provides a more precise explanation for the patient's condition. Here's what is lacking in the first set of predictions:

• *Specificity*: While a Viral Infection could explain most symptoms, it does not account for joint pain and the presence of a rash, which are more characteristic of DF or Chikungunya.

• *Detailed justifications*: The first analysis does not delve as deeply into the reasons behind each disease possibility. GPT-4's explanation for Dengue Fever and Chikungunya is more comprehensive, strengthening the connection between symptoms and diagnosis.

• *Importance of additional information*: Emphasize how geographic location and travel history can significantly impact the likelihood of these specific diseases.

C.6 GPT-4's Third Round

Top-3 Predictions: *Dengue Fever (60%), Chikungunya (35%), and Zika Virus (5%)*.
Justifications: Though GPT-4 slightly adjusts probabilities, it considers the adjustment arbitrary without new information that would specifically shift the likelihood between Dengue and Chikungunya. GPT-4 refutes Gemini's Viral Infection prediction as lacking specificity. The symptoms described are quite distinctive and align closely with specific viral infections transmitted by mosquitoes.

C.7 Gemini's Third Round

Top-3 Predictions: *Dengue Fever (60%), Chikungunya (35%), and Zika Virus (5%).*
Justifications: Agreeing that GPT's critique is spot-on, and reiterating the need for more info.

C.8 Final Joint Recommendations to GP

After seeking further input from GPT-4, Gemini fully concurs with GPT4's final suggestions, even though the moderator instructs Gemini to maintain its original stance. Both LLMs jointly recommend additional symptom examinations and lab tests:

A comprehensive diagnostic approach for Dengue Fever requires a combination of clinical assessment, laboratory testing, and consideration of the patient's travel and vaccination history, as well as the epidemiological context. Clinical symptoms such as skin rash, joint pain, vomiting, and red spots over the body provide valuable clues to the diagnosis, but additional symptoms like nausea, vomiting, diarrhea, bleeding, bruising, and drowsiness can indicate the severity and progression of the disease.

Recommended Laboratory Tests:
• *Complete Blood Count (CBC)*: Identify thrombocytopenia or leukopenia.
• *Serology Tests*: Detect specific IgM and IgG antibodies for Dengue,

Chikungunya, and Zika.
• *Polymerase Chain Reaction (PCR)*: Direct detection of viral RNA.
• *NS1 Antigen Test for Dengue*: Early and specific detection of Dengue

virus.
• *Urine Test for Zika*: Viral presence in urine.

Appendix D: EVINCE Debate - Jaundice vs. Hepatitis

In this experiment, GPT-4 and Claude receive an identical list of symptoms from the moderator, aimed at guiding disease diagnosis, facilitating debate, and shaping subsequent recommendations. The “ground-truth” disease is Jaundice. (We do not shorten the debate content for review.)

D.1 Moderator’s Opening Prompt

“What potential disease could be associated with the following symptoms: itching, vomiting, fatigue, weight loss, high fever, yellowish skin, dark urine, and abdominal pain? Please offer top-five predictions with probabilities normalized to one, supported by justifications. Additionally, in the end of debate, include a list of supplementary symptom inquiries and recommend relevant lab tests to strengthen the confidence in your prediction.”

D.2 GPT-4 Opening Round

• *Hepatitis C (HCV)*: 40% **Justification**: Hepatitis C is a viral infection that can cause both acute and chronic hepatitis. The combination of symptoms like fatigue, abdominal pain, nausea, vomiting, jaundice, dark urine, and weight loss strongly suggests HCV. Itching may also occur due to liver dysfunction.

• *Hepatitis B (HBV)*: 30% **Justification**: Hepatitis B is a viral infection that can cause acute or chronic liver inflammation. Symptoms are similar to HCV, including fatigue, abdominal pain, nausea, vomiting, dark urine, jaundice, and itching. The presence of these symptoms warrants testing for both HBV and HCV.

• *Cirrhosis*: 15% **Justification**: Cirrhosis is a late stage of liver scarring that can result from various liver diseases. The combination of symptoms like fatigue, weight loss, abdominal pain, jaundice, dark urine, and itching suggests advanced liver disease, possibly cirrhosis.

• *Obstructive Jaundice*: 10% **Justification**: Obstructive jaundice occurs when there is a blockage in the bile ducts. The presence of jaundice, dark urine, itching, abdominal pain, nausea, vomiting, and weight loss suggests a possible obstruction. However, the absence of fever makes an infectious cause like cholangitis less likely.

• *Acute Liver Failure*: 5% **Justification**: Acute liver failure is a rare but life-threatening condition characterized by rapid deterioration of liver function.

The presence of jaundice, abdominal pain, nausea, vomiting, fatigue, and itching, along with the acute onset of symptoms, may indicate acute liver failure. However, the absence of altered mental status (hepatic encephalopathy) makes this less likely.

D.3 Claude Opening Round

• *Hepatitis B (HBV)*: 35% **Reasoning**: Hepatitis B is a viral infection that can cause acute or chronic liver inflammation. Symptoms include fatigue, abdominal pain, nausea, vomiting, dark urine, jaundice, joint pain, and itching.

• *Hepatitis C (HCV)*: 25% **Reasoning**: Hepatitis C is a viral infection that can cause both acute and chronic hepatitis. Symptoms are similar to HBV but may also include weight loss and fever. Jaundice is a common symptom.

• *Obstructive Jaundice (due to gallstones, pancreatic cancer, or bile duct obstruction)*: 20% **Reasoning**: Obstructive jaundice occurs when there is a blockage in the bile ducts, preventing bile from reaching the intestines. This can lead to yellowing of the skin and eyes, dark urine, itching, abdominal pain, nausea, and vomiting.

• *Alcoholic Hepatitis*: 15% **Reasoning**: Alcoholic hepatitis is liver inflammation caused by excessive alcohol consumption. Symptoms include abdominal pain, nausea, vomiting, fatigue, dark urine, and possibly fever and weight loss.

• *Hepatitis A (HAV)*: 5% **Reasoning**: Hepatitis A is an acute, self-limiting liver infection transmitted through contaminated food or water. Symptoms include fatigue, abdominal pain, nausea, vomiting, dark urine, and jaundice.

D.4 GPT-4 Second Round

• *Hepatitis B (HBV)*: 35% **Symptoms**: fatigue, abdominal pain, nausea, vomiting, dark urine, jaundice (yellowish skin), joint pain, itching
• *Hepatitis C (HCV)*: 25% **Symptoms**: fatigue, abdominal pain, nausea, vomiting, dark urine, jaundice (yellowish skin), weight loss, fever, itching

• *Obstructive Jaundice (due to gallstones, pancreatic cancer, or bile duct obstruction)*: 20% **Symptoms**: yellowish skin (jaundice), dark urine, itching, abdominal pain, nausea, et al.

• *Alcoholic Hepatitis*: 15% **Symptoms**: abdominal pain, nausea, vomiting, fatigue, jaundice (yellowish skin), dark urine, fever, weight loss
• *Hepatitis A (HAV)*: 5% **Symptoms**: fatigue, abdominal pain, nausea, vomiting, dark urine, jaundice (yellowish skin)

D.5 Claude Second Round

• *Hepatitis C (HCV)*: 40% **Reasoning**: Hepatitis C is a viral infection that can cause both acute and chronic hepatitis. The combination of symptoms like fatigue, abdominal pain, nausea, vomiting, jaundice, dark urine, and weight loss strongly suggests HCV. Itching may also occur due to liver dysfunction.

• *Hepatitis B (HBV)*: 30% **Reasoning**: Hepatitis B is a viral infection that can cause acute or chronic liver inflammation. Symptoms are similar to HCV, including fatigue, abdominal pain, nausea, vomiting, dark urine, jaundice, and itching. The presence of these symptoms warrants testing for both HBV and HCV.
• *Cirrhosis (due to various causes such as chronic hepatitis, alcohol abuse, or non-alcoholic fatty liver disease)*: 15% **Reasoning**: Cirrhosis is a late stage of liver scarring that can result from various liver diseases. The combination of symptoms like fatigue, weight loss, abdominal pain, jaundice, dark urine, and itching suggests advanced liver disease, possibly cirrhosis.

• *Obstructive Jaundice (due to gallstones, pancreatic cancer, or bile duct obstruction)*: 10% **Reasoning**: Obstructive jaundice occurs when there is a blockage in the bile ducts. The presence of jaundice, dark urine, itching, abdominal pain, nausea, vomiting, and weight loss suggests a possible obstruction. However, the absence of fever makes an infectious cause like cholangitis less likely.

• *Acute Liver Failure (due to various causes such as viral hepatitis, druginduced liver injury, or autoimmune hepatitis)*: 5% **Reasoning**: Acute liver failure is a rare but life-threatening condition characterized by rapid deterioration of liver function. The presence of jaundice, abdominal pain, nausea, vomiting, fatigue, and itching, along with the acute onset of symptoms, may indicate acute liver failure. However, the absence of altered mental status (hepatic encephalopathy) makes this less likely.

D.6 GPT-4 Third Round

• *Hepatitis C (HCV)*: 35% **Reasoning**: Hepatitis C is indeed a major cause of the symptoms listed, including fatigue, abdominal pain, nausea, vomiting, jaundice, dark urine, and weight loss. Itching can also occur due to liver dysfunction.

• *Hepatitis B (HBV)*: 25% **Reasoning**: Hepatitis B shares many symptoms with Hepatitis C, such as fatigue, abdominal pain, nausea, vomiting, dark urine, jaundice, and itching. Testing for both HBV and HCV is warranted given the overlap in symptoms.

• *Cirrhosis*: 20% **Reasoning**: Cirrhosis can result from chronic hepatitis, alcohol abuse, or non-alcoholic fatty liver disease. The symptoms of fatigue, weight loss, abdominal pain, jaundice, dark urine, and itching strongly suggest advanced liver disease, which could be cirrhosis.

• *Obstructive Jaundice*: 15% **Reasoning**: Obstructive jaundice, caused by gallstones, pancreatic cancer, or bile duct obstruction, is consistent with jaundice, dark urine, itching, abdominal pain, nausea, vomiting, and weight loss. The absence of fever reduces the likelihood of an infectious cause like cholangitis.

• *Acute Liver Failure*: 5% **Reasoning**: Acute liver failure is less common but serious, characterized by rapid deterioration of liver function. The symptoms of jaundice, abdominal pain, nausea, vomiting, fatigue, and itching, along with the acute onset of symptoms, suggest this possibility. However, the absence of altered mental status (hepatic encephalopathy) makes it less likely.

D.7 Claude Third Round

• *Hepatitis C (HCV)*: 35% **Reasoning**: I slightly lowered the probability of HCV from 40% to 35%. While the combination of symptoms strongly suggests HCV, the absence of a few key risk factors (e.g., history of intravenous drug use or blood transfusions) in the given information makes me less confident in assigning a higher probability.

• *Hepatitis B (HBV)*: 30% **Reasoning**: I agree with the 30% probability assigned to HBV. The symptoms are indeed similar to HCV, and testing for both viruses is warranted.

• *Cirrhosis*: 20% **Reasoning**: I increased the probability of cirrhosis from 15% to 20%. The presence of jaundice, dark urine, itching, and weight loss, along with the other symptoms, strongly suggests advanced liver disease. Cirrhosis can be a consequence of chronic liver diseases like hepatitis B or C, which have been assigned higher probabilities.

• *Obstructive Jaundice*: 10% **Reasoning**: I agree with the 10% probability for obstructive jaundice. The absence of fever makes an infectious cause like cholangitis less likely, but the other symptoms still suggest a possible obstruction.

• *Acute Liver Failure*: 5% **Reasoning**: I agree with the 5% probability for acute liver failure. While the combination of symptoms could indicate acute liver failure, the absence of altered mental status (hepatic encephalopathy) makes this diagnosis less likely.

D.8 Joint Final Round

1. *Hepatitis C (HCV)* - 35% **Justification**: The combination of symptoms strongly suggests HCV, but without additional risk factors, a slightly lower probability is appropriate.

2. *Hepatitis B (HBV)* - 30% **Justification**: The symptoms are similar to HCV, and testing for both viruses is necessary.

3. *Cirrhosis* - 20% **Justification**: The presence of jaundice, dark urine, itching, weight loss, and other symptoms indicates advanced liver disease, making cirrhosis a more likely possibility.

4. *Obstructive Jaundice* - 10% **Justification**: The absence of fever reduces the likelihood of an infectious cause like cholangitis, but the other symptoms still suggest a potential biliary obstruction.

5. *Acute Liver Failure* - 5% **Justification**: While the combination of symptoms could indicate acute liver failure, the absence of altered mental status makes this diagnosis less probable.

D.9 Final Joint Recommendations to GP

Gemini and GPT-4 jointly recommend the following additional symptom inquiries and lab tests to confirm a hepatitis diagnosis:
Additional Symptom Inquiries:
1. *Onset and Duration*: Determining the onset and duration of symptoms can clarify the illness timeline and its progression.

2. *Severity and Trend*: Evaluating symptom severity and whether they are worsening or improving aids in understanding the disease's trajectory and treatment responses.

3. *Associated Symptoms*: Checking for additional symptoms like nausea, vomiting, fever, joint pain, or urine color changes can pinpoint other hepatitis indicators and exclude other conditions.

Recommended Lab Tests:
1. *Liver Function Tests (LFTs)*: Critical for assessing liver health, LFTs evaluate enzyme levels such as alanine aminotransferase (ALT) and aspartate aminotransferase (AST), where abnormalities can signify liver inflammation.

2. *Hepatitis Panel*: This test checks for hepatitis A, B, and C viruses, vital for determining the specific type and guiding treatment strategies. 3. *Serology Testing*: Useful for distinguishing between acute and chronic hepatitis by identifying specific antibodies or antigens.

4. *Imaging Studies*: Ultrasound or MRI can provide visual insights into the liver's state, detecting inflammation, fibrosis, or cirrhosis, thereby complementing blood-based diagnostics.

Appendix H: The EnToPPS Framework

EnToPPS integrates predictions from two LLMs, denoted as A and B, each providing probability distributions over C classes. The following steps outline the EnToPPS process:

1. *Obtain Top-C Predictions*: For each LLM (A and B), obtain the predicted probabilities for all C classes, denoted as P_A and P_B:
$P_A = [p_{A1}, p_{A2}, \ldots, p_{AC}]$, $P_B = [p_{B1}, p_{B2}, \ldots, p_{BC}]$, where p_{Ai} and p_{Bi} represent the predicted probability of class i by LLM A and B, respectively.
2. *Select Top-k Predictions*: For each LLM (A and B), select the top-k predicted classes based on their probabilities:
$T_A = [t_{A1}, t_{A2}, \ldots, t_{Ak}]$, $T_B = [t_{B1}, t_{B2}, \ldots, t_{Bk}]$,
where t_{Ai} and t_{Bi} represent the class index of the ith top prediction by A and B, respectively.
3. *Combine Top-k Predictions*: Combine the top-k predictions from both LLMs to create a set of unique predicted classes:
$T_C = T_A \cup T_B = [t_{C1}, t_{C2}, \ldots, t_{Cm}]$, $k \leq m \leq 2k$.

4. *Backfill Missing Probabilities*: For each class in the combined set T_C, backfill its probability from the original probability distributions P_A and P_B:
• If a class t_{Ci} is present in T_A, assign its probability from P_A: $p_{Ci} = p_{Ai}$.
• If a class t_{Ci} is present in T_B, assign its probability from P_B: $p_{Ci} =$

p_{Bi}.
• If a class t_{Ci} is present in both T_A and T_B, assign the average probability:
$p_{Ci} = {}^{pAi+pBi} \cdot 2$
5. *Normalize Probabilities*: Normalize the probabilities of the classes in the combined set T_C to ensure they sum up to 1:
$P_C = [p_{C1}, p_{C2}, \ldots, p_{Cm}]$, where $p_{Ci} = \sum^{pCi} {}_{\cdot m\ pCjj=1}$

References

[1] Sahar Abdelnabi et al. *Cooperation, Competition, and Maliciousness: LLM-Stakeholders Interactive Negotiation*. 2024. arXiv: 2309.17234. [2] Leo Breiman. “Bagging predictors”. In: *Machine learning* 24.2 (1996), pp. 123–140.

[3] Gavin Brown et al. "Diversity creation methods: a survey and categorisation". In: *Information Fusion* 6.1 (2005), pp. 5–20. [4] Tom B. Brown et al. "Language Models are Few-Shot Learners". In: *arXiv preprint arXiv:2005.14165* (2020).
[5] Chi-Min Chan et al. *ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate*. 2023. arXiv: 2308.07201 [cs.CL].

[6] Edward Y Chang. "Examining GPT-4's Capabilities and Enhancement with SocraSynth". In: *The* 10th *International Conf. on Computational Science and Computational Intelligence*. 2023.

[7] Edward Y. Chang. "Prompting Large Language Models With the Socratic Method". In: *IEEE 13th Annual Computing and Communication Workshop and Conference* (2023). url: https://arxiv.org/ abs/2303.08769.

[8] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. 2nd. John Wiley & Sons, 2006.

[9] Thomas G. Dietterich. "Ensemble methods in machine learning". In: *International Workshop on Multiple Classifier Systems*. Springer, Berlin, Heidelberg, 2000, pp. 1–15.

[10] Nan Du et al. "GLAM: Efficient Scaling of Language Models with Mixture-of-Experts". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 5547–5569.

[11] Yilun Du et al. *Improving Factuality and Reasoning in Language Models through Multiagent Debate*. 2023. arXiv: 2305.14325 [cs.CL].

[12] Chelsea Finn, P. Abbeel, and Sergey Levine. "Model-Agnostic MetaLearning for Fast Adaptation of Deep Networks". In: *International Conference on Machine Learning*. 2017.

[13] Yoav Freund and Robert E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Journal of Computer and System Sciences* 55.1 (1997), pp. 119–139.

[14] Yao Fu et al. *Improving Language Model Negotiation with Self-Play and In-Context Learning from AI Feedback*. 2023. arXiv: 2305.10142.

[15] Yongkang Guo et al. *Algorithmic Robust Forecast Aggregation*. 2024. arXiv: 2401.17743 [cs.LG].
[16] Robert A. Jacobs et al. “Adaptive Mixtures of Local Experts”. In: *Neural Computation* 3.1 (Mar. 1991), pp. 79–87.

[17] Leonid V Kantorovich. “On the translocation of masses”. In: *Doklady Akademii Nauk* 37.7-8 (1942), pp. 199–201.

[18] Anders Krogh and Jesper Vedelsby. “Neural network ensembles, cross validation, and active learning”. In: *Advances in Neural Information Processing Systems*. Vol. 7. 1995, pp. 231–238.

[19] Solomon Kullback. *Information Theory and Statistics*. John Wiley & Sons, 1951.

[20] Ludmila I Kuncheva and Christopher J Whitaker. “Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy”. In: *Machine Learning* 51.2 (2003), pp. 181–207.

[21] Huao Li et al. “Theory of Mind for Multi-Agent Collaboration via Large Language Models”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023.

[22] Tian Liang et al. *Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate*. 2023. arXiv: 2305.19118 [cs.CL].

[23] Jianhua Lin. “Divergence measures based on the Shannon entropy”. In: *IEEE Transactions on Information theory* 37.1 (1991), pp. 145– 151.

[24] Julian Michael et al. *Debate Helps Supervise Unreliable Experts*. 2023. arXiv: 2311.08702 [cs.AI].

[25] David E Newman-Toker, Najlla Nassery, and et al. “Burden of serious harms from diagnostic error in the USA”. In: *BMJ Quality & Safety* (2023).

[26] David E. Newman-Toker et al. “Serious Harm From Diagnostic Error in US Healthcare Systems: Estimate of Its Magnitude and Cost”. In: *BMJ*

Quality & Safety 32.7 (2023), pp. 549–557. doi: 10 . 1136 / bmjqs-2022-014004.

[27] Pranay Patil. *Kaggle Disease Symptoms Description Dataset*. 2020. url: https://www.kaggle.com/-datasets/-itachi9604/-diseasesymptom-description-dataset.

[28] Yu-Shao Peng et al. “REFUEL: exploring sparse features in deep reinforcement learning for fast disease diagnosis”. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS’18. 2018, pp. 73337342.

[29] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. “The earth mover’s distance as a metric for image retrieval”. In: *International journal of computer vision*. Vol. 40(2). Springer, 2000, pp. 99–121.

[30] Claude E. Shannon. “A Mathematical Theory of Communication”. In: *Bell System Technical Journal* 27.3 (1948), pp. 379–423.

[31] John E. Shore and Rodney W. Johnson. “Axiomatic derivation of the principle of maximum entropy and the principle of minimum crossentropy”. In: *IEEE Transactions on Information Theory* 26.1 (1980), pp. 26–37.

[32] Andries Smit et al. *Should we be going MAD? A Look at Multi-Agent Debate Strategies for LLMs*. 2024. arXiv: 2311.17371.

[33] Cédric Villani. *Optimal Transport: Old and New*. Vol. 338. Springer Science & Business Media, 2008.

[34] Sang Michael Xie et al. “An Explanation of In-Context Learning as Implicit Bayesian Inference”. In: *International Conference on Learning Representations (ICLR)*. 2021.

[35] Yufeng Zhang et al. *What and How does In-Context Learning Learn? Bayesian Model Averaging, Parameterization, and Generalization*. 2023. arXiv: 2305.19420.

7 Unbiasing Wikipedia and News Articles via SocraSynth

Abstract Biases inherent in human endeavors pose significant challenges for machine learning, particularly in supervised learning that relies on potentially biased "ground truth" data. This reliance, coupled with models' tendency to generalize based on statistical maximal likelihood, can propagate and amplify biases, exacerbating societal issues. To address this, our study proposes a reflective methodology utilizing multiple Large Language Models (LLMs) engaged in a dynamic dialogue to uncover diverse perspectives. By leveraging in-context learning, information theory, and divergence metrics, this novel approach fosters context-dependent linguistic behaviors, promoting unbiased outputs. Furthermore, it enables measurable progress tracking and explainable remediation actions to address identified biases.

7.1 Introduction

AI systems are increasingly being integrated into critical sectors such as education, healthcare, and public policy, where their decisions can have profound impacts. Despite their potential, these systems are prone to exhibiting discriminatory behaviors, propagating existing biases, or making errors. Researchers in machine learning are diligently addressing these challenges by investigating the sources, patterns, and types of biases inherent in AI systems [19]. Achieving absolute fairness is complicated by diverse cultural, religious, and ideological perspectives, but the primary objective remains to minimize the propagation of biases within machine learning models [5, 11, 14, 19, 22].

In this data-centric era, the accuracy of training data, particularly groundtruth labels, is paramount. While the inherent characteristics of data may be difficult to alter, the labels assigned to this data are far more malleable and hold potential for harm. Machine learning algorithms inherently learn from the data they are fed, including any embedded biases. Erroneous labels can significantly amplify these biases, leading to severe consequences. For instance, mislabeling a biased news article as neutral can perpetuate misinformation, while an inaccurate medical diagnosis can lead to improper treatment and endanger a patient’s health. Our research prioritizes identifying and rectifying such mislabeled data to mitigate bias and ensure the responsible development and deployment of AI systems.

Evidence of annotation biases is illustrated in Chapter 7.4. Tables 7.3 and 7.5 present real data [6] showing how annotators’ political affiliations can influence their labeling of news articles. For instance, annotators aligned with the Democratic party are more inclined to perceive scandals involving Democrats negatively, whereas Republican annotators might view the same incidents neutrally. Conversely, Republican annotators might downplay criticisms directed at their party, whereas Democrats might view them as justified, underscoring the influence of personal ideologies on annotation practices.

To combat the perpetuation of these biases through classifiers, we propose a check-and-balance framework wherein two Large Language Models (LLMs) engage in dialogue to scrutinize and challenge human annotations. One LLM supports the original annotation while the other offers counter perspectives, thereby enriching the understanding of the content. This dialogue is designed to foster an exchange that yields balanced insights into the topics discussed and, if necessary, makes recommendations for review by the editorial board.

To assess the effectiveness of this dialogue, we employ metrics rooted in statistical and information theory. Measures such as Shannon entropy [24] and mutual information [8] assess enhancements in shared understanding, while Jensen-Shannon divergence (JSD) [17], Wasserstein distance (WD) [13], and cross-entropy (CE) [25] monitor the productivity of the discussions. This transparent process allows human supervisors to oversee and adjust annotations based on the outcomes of these dialogues, ensuring that all recommendations undergo thorough scrutiny. Our empirical research validates the effectiveness of this innovative approach.

1. *Robust Ground Truth Validation:* We deploy structured dialogues among

multiple LLMs to authenticate and cross-verify ground-truth labels. This method not only detects biases inherent in traditional training datasets but also prevents the amplification of such biases, enhancing the overall data integrity and reliability.

2. *Comprehensive Bias and Error Mitigation:* Our framework facilitates both contentious and collaborative dialogues between LLMs and humans. This dual approach pinpoints out biases and errors, and furthermore, it equips editorial boards and healthcare professionals with the necessary justifications and alerts to preemptively correct these issues. This proactive mitigation ensures trustworthiness and high accuracy and in practical applications.

3. *Advanced Metrics for Dialogue Effectiveness:* Using a suite of statistical and information theory metrics, such as Shannon entropy, mutual information, Jensen-Shannon divergence, Wasserstein distance, and crossentropy, we rigorously evaluate the diversity of perspectives, the quality of information exchange, and the efficacy of dialogues. This systematic measurement enhances the integrity and neutrality of annotations, contributing to more reliable AI outputs.

The remainder of this paper is organized as follows: Chapter 7.2 discusses challenges and reviews related work; Chapter 7.3 describes the core maxims, theorem, and algorithm; Chapter 7.4 presents experiments illustrating successful bias identification and mitigation; and the final section concludes with insights on future work and perceived limitations.

7.2 Related Work

This study focuses on mitigating training data label (ground truth) bias, a primary concern in machine learning [19]. Accurate labeling is crucial, as a label that aligns with biased content reinforces that bias, while a label that correctly identifies it allows for education and correction [5, 9]. This underscores the importance of label accuracy in minimizing bias propagation.

7.2.1 Label Validation

This work specifically addresses mislabeled ground truth and explores remediation actions. Efforts to improve annotation accuracy can be broadly categorized into three approaches:

Cross-Validation with Multiple Annotators: Ensemble methods, utilizing multiple annotators and statistical techniques, can mitigate individual biases and enhance data reliability [26]. This approach has been successful in tasks with broad consensus, such as image annotation with ImageNet [10, 15]. However, it may be less effective for complex or biased content, where subtle interpretations are required. For instance, the term "discovered" in reference to Columbus's arrival in the Americas reflects a Western bias, while "encounter" offers a more accurate representation from Indigenous perspectives [20, 31]. Relying on majority votes in such cases can be counterproductive, and while diverse annotator pools are important, the inherent limitations of human annotation, including the assumption of a single truth, must be acknowledged [4].

Cross-Validation between Machine and Human Annotators: Integrating machine learning algorithms into the annotation process can improve quality by leveraging both human expertise and machine efficiency [30]. Semi-supervised learning, which utilizes both labeled and unlabeled data, is one such technique [21]. However, machine learning algorithms can be inherently biased due to their training data, and ensemble methods may not fully resolve these underlying biases [5, 30].

7.2.2 Biased Ground Truth

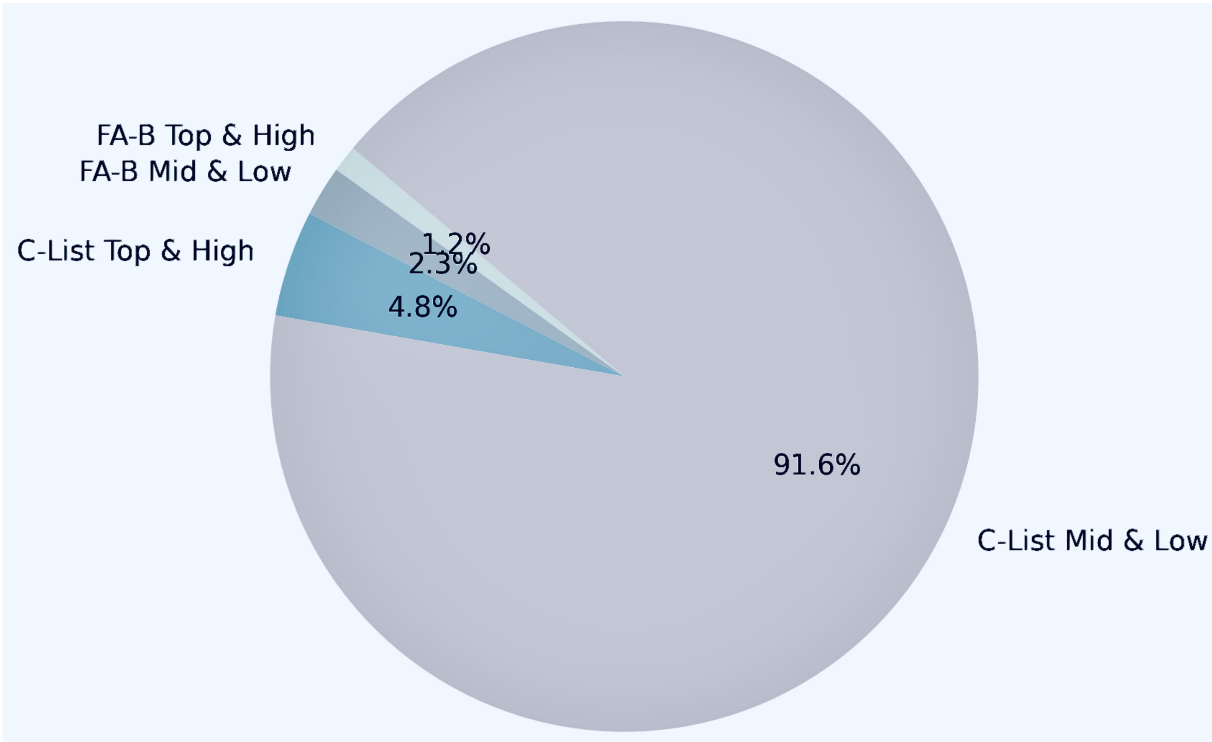


Figure 7.1: Distribution from Top Quality High Importance (1.2%) to Low Quality Low Importance (91.6%). Notably, the blue segment (4.8%) signifies high-importance pages in need of improvement.

Using Wikipedia as the benchmark for validating the outputs of LLMs has gained attention in recent studies [18, 23]. However, there are notable limitations to this method. First, the specific information serving as the ground truth may not always be available on Wikipedia. If the exact answers are already known to chatbot developers, there would be no need to consult LLMs. Second, the credibility of this approach is further challenged by the quality assessment of Wikipedia articles themselves. As indicated in Figure 7.1, 91% of Wikipedia's content is considered to be of middle to low quality by the platform's own editors.

Further, biases are prevalent in Wikipedia and news media, encompassing aspects like gender, race, ideology, and geography, are widely acknowledged. For instance, in Wikipedia, biases manifest as an over-representation of certain topics in biographies [29], affecting the balance of content. In the realm of news media, outlets are often categorized by political orientation— ranging from far left to far right—as seen in assessments like those by AllSides [2]. Such classifications are akin to our method of categorizing news articles. Figure 7.2, generated and periodically updated by AllSides, illustrates this point. However, users should interpret the figure with care, acknowledging its potential subjectivity. Nonetheless, it underscores how a single event or story can be portrayed in markedly different ways, depending on the viewpoint.

AllSides Media Bias Chart™

Ratings based on online, U.S. political content only – not TV, print, or radio.
Ratings do not reflect accuracy or credibility; they reflect perspective only.

AllSides Media Bias Ratings™ are based on multi-partisan, scientific analysis.
Visit AllSides.com for balanced news and over 1,400 rated sources.

Version 9.2

Figure 7.2: AllSides Fact Check Biase Chart.

7.2.3 GAI Opportunity:

The emergence of GAI and LLMs, with their vast knowledge base and powerful Transformer architecture [28], presents a new avenue for addressing annotation bias. LLMs can potentially uncover diverse perspectives on a given topic, including historical shifts and evolving narratives. Recent studies have explored combining LLM output with human feedback for annotation tasks [27].
However, due to their “maximal likelihood” next-token prediction training objective, LLMs may prioritize popular viewpoints over minority ones. This chapter addresses this limitation by proposing a novel approach, grounded in statistical and information theories, that aims to uncover and balance diverse viewpoints, ensuring that both majority and minority perspectives are adequately represented in the annotation process.

7.3 EVINCE Algorithm

Expanding on SocraSynth (Chapter 5) with theoretical foundations and quantitative metrics, EVINCE (Entropy Variation and INformation CompetencE) (Chapter 6) leverages LLMs to promote content neutrality through the incorporation of diverse perspectives.

EVINCE facilitates structured dialogues between LLMs to address the “maximal likelihood” bias inherent in conventional information retrieval systems. This bias manifests in search engines like Google, where popular viewpoints are often prioritized based on metrics like click-through rates [1], potentially sidelining less common perspectives. Similarly, LLM text generation, which relies on predicting the next most likely token, can inadvertently amplify existing biases present in the training data [16].

To foster divergent perspectives, EVINCE addresses two sub-goals:
• *Exploration*: Encouraging the generation of a wide array of viewpoints.

• *Meaningful Diversity*: Ensuring collected perspectives are substantive and not merely contrarian for the sake of disagreement.
EVINCE achieves these goals by analyzing probability distributions of

top-k labels elicited from each LLM in the committee. Through this analysis, individual entropies, cross-entropy between distributions, and mutual information are computed. (For detailed metrics and formulas, refer to Chapter 6.) Based on this quantitative assessment, EVINCE dynamically adjusts its linguistic behaviors (e.g, more contentious vs. more conciliatory) to optimize the annotation recommendations.

The initial phase of the EVINCE algorithm aims to induce *dual entropy* and high cross-entropy between the LLM-generated distributions. I will prove shortly that *dual entropy* is the ideal condition to foster information exchange between LLMs. This signifies disagreement and creates a fertile ground for novel perspective discovery and exchange. Through iterative dialogue, mutual information increases while divergence decreases, ultimately converging towards a consensus.

7.3.1 Maxims and Optimal Theorem

Maxim 1: Orchestrate Two Equally Competent LLMs in Structured Debate: Integrating two equally competent LLMs ensures a balanced exchange of insights and avoids bias from knowledge asymmetry. This adversarial setup fosters diversity in predictions, each supported by justifications, promoting critical evaluation and uncovering potential blind spots. The concern is not about the potential non-overlapping training data, as information exchange can remedy this. Instead, the focus is on ensuring that both models have similar quality, primarily determined by their size, to prevent one model from dominating the other due to a disparity in reasoning quality.

Maxim 2: Encourage the Accurate Rather Than the "Popular" Prediction: Typically, LLMs, with their maximum likelihood next-token prediction objective, tend to favor the most popular predictions. By conditioning LLMs within specific contexts, we can prioritize accuracy over popularity, thus mitigating confirmation biases.

Maxim 3: Combine Predictions Weighted by Diversity and Quality: Weighting the probability distributions from two LLMs based on diverse probabilistic insights and the quality of supporting arguments.

How? Following these three sub-maxims:
• **Maxim 3.1**: *Prediction Reliability*: Estimate the reliability of predic tions using entropy-based measures to quantify uncertainty and information content. Typically, lower entropy indicates higher confidence in a prediction, suggesting higher reliability.

• **Maxim 3.2**: *Argument Quality*: Evaluate the quality of supporting arguments using techniques inspired by the Socratic method. This includes identifying logical fallacies and assessing the relevance and credibility of evidence.

• **Maxim 3.3**: *Aggregation*: Employ a weighted aggregation method, such as a Bayesian model, to combine weighted predictions accounting for both probabilistic insights and the quality of supporting arguments.

Maxim 4: Evaluating the Convergence Rate of the Predictions Across the Rounds: This maxim focuses on measuring how quickly and effectively the predictions from the LLMs converge over successive rounds, assessing the efficiency of the debate and aggregation mechanisms. Convergence is assessed by measuring mutual information and using proxy metrics such as Wasserstein distance and cross entropy. When mutual information is low or the similarity between predictions is high, the dialogue is considered to be converging.

Algorithm 2 Specifications of Algorithm EVINCE

1: **Input:** Information set S, Class labels C; Two equally competent LLMs: LLM_A and LLM_B (**Maxim #1**);
2: **Output:** P_f, final probability distribution over C;
3: **Variables:** t: debate round; R =∅ aggregated arguments;

P
$^{(t)}$ $P^{(t)}$: prediction distributions of LLM_A and LLM_B on C of round t; $R^{(t)}$

A ’ B A ’

$R^{(t)}$: supporting reason sets;$_B$
Δ = 90%: debate contentiousness, initialize to high to foster adversary

between LLMs (**Maxim #2**);
p: prompt = “Predict topk probability distribution on C with S and R at contentiousness Δ”;

4: **Functions:** CRIT(d) [7], Critical Reading Inquisitive Template for evaluating argument quality;
ARA [12], Algorithmic Robust Aggregation for optimal prediction aggregation (**Maxims #3**);

5: **Initial Predictions** t = 0**:**
LLMs generate their predictions in probability distributions with supporting reasons:

$(P_A^{(t=0)}, R_A^{(t)}) = LLM_A(S, p)$, $(P_B^{(t=0)}, R_B^{(t)}) = LLM_B(S, p)$.

6: **Debate Iterations:**

6.1. **Update Predictions:**
Calculate the confidence-based weights using the inverse of entropy (**Maxim #3.1**):

$\alpha = 1/(H(P_A^{(t)}) + 1)$, $\beta = 1/(H(P_B^{(t)}) + 1)$.

Use the blending mechanism to update predictions (**Maxim #3.3**): $P_A'^{(t)} = \alpha P_A^{(t)} + (1-\alpha) P_B^{(t)}$, $P_B'^{(t)} = \beta P_B^{(t)} + (1-\beta) P_A^{(t)}$.

6.2. **LLMs Generate New Predictions:** Both LLMs use accumulated $R = R \cup R^{(t)} \, R^{(t)}$

A $\cup$B .
$(P^{(t+1)}, R^{(t+1)}) = LLM_A((P'^{(t)}), R, p),_{A\,A\,B}$
$(P^{(t+1)}, R^{(t+1)}) = LLM_B((P'^{(t)}), R, p)._{B\,B\,A}$

6.3. **Exit Condition Check with Wasserstein distance (Maxim #4): If WD**$(P^{(t+1)}, P^{(t+1)}) < \epsilon$ **EXIT**; $t = t + 1$, $\Delta = \Delta \times 80\%._{A\,B}$
7: **Final Decision:** Weighted prediction by quality scores of the evaluator e.g., CRIT [7] (**Maxim #3.2**):
$P_f = \Omega_A P^{(t+1)} + \Omega_B P^{(t+1)} / \Omega_A + \Omega_B ._{A\,B}$

Problem Statement: Organize a structured dialogue between two equally competent large language models (LLMs), LLM_A and LLM_B, to conduct t rounds. At each round t, each model produces a probability distribution, denoted as $P^{(t)}$ and $P^{(t)}$, over C possible outcomes, accompanied by sup$_{A\ (tB}$ porting arguments $R^{)\ (t)}$. The goal is to design an iterative debate$_A$ and R_B process that leverages the structured exchange of arguments to enable the models to converge on an optimal prediction distribution P^* across the C classes.

7.3.2 Algorithm Specifications

With all proxy metrics and their pros, cons, and combined strengths comprehensively surveyed (Chapter 6), Algorithm 1 formally specifies the algorithm of EVINCE with the maxims.

7.3.3 Entropy Duality Theorem (EDT)

Theorem EDT: Optimal Pairing of LLMs for Probabilistic Prediction Accuracy. The optimal pairing of LLMs for diagnosis accuracy, in terms of stability, accuracy, and robustness, occurs when the LLMs are 1) equivalent in the quality of the information they process, and 2) exhibit contrasting entropy values in their prediction distributions–one high and one low. [**Proof**]: Presented in the EVINCE chapter.

7.4 Experiments

Our experimental framework aims to assess the feasibility of both detecting biases in textual content and implementing effective mitigation strategies. The first experiment focuses on bias detection, while the second explores the generation of balanced textual outputs as a corrective measure, moving beyond the limitations of prior studies that primarily focused on identification (Chapter 7.2).

We utilized GPT-4 via OpenAI API on Microsoft Azure, setting the temperature to 0.1 with maximum token size.

7.4.1 Experiment #1: Bias Detection

The aim of this experiment is to evaluate if personal ideology may affect annotations, and can EVINCE help flag and rectify the biases.

Dataset

The dataset for this experiment consists of 619 news articles (54.3% about Democrat scandals, 45.7% about Republican scandals) selected from a larger 2013 repository of 14,033 articles compiled by fifteen reputable news organizations [6]. These articles cover diverse topics like civil rights, healthcare, elections, and national security. This dataset is provided as supplementary material.

News #	Categories	Neg.	W. Neg.	Neut.	+	Biases (DR,DS,SR)	Source
D1	Civil Rights	D,R,S	-	-	-	0,0,0	HuffPost
D2	Civil Rights	D,S	-	R	-	2,0,2	HuffPost
D8	Civil Rights	D	-	S	R	3,2,1	BBC
D31	Environment	D	-	R, S	-	2,2,0	CNN
D37	Politics	-	D,R,S	-	-	0,0,0	Yahoo
D69	Healthcare	D	-	R,S	-	2,2,0	Breitbart
D81	Economy	-	D,S	R	-	1,0,1	Breitbart
D98	Economy	D,S	R	-	-	1,0,1	Breitbart
D101	Education	-	D.S	R	-	1,0,1	NY Times
D106	Election	-	-	D,R,S	-	0,0,0	USA Today
D109	Elections	-	D,S	R	-	1,0,1	Reuters
D157	International	-	D,S	R	-	1,0,1	NY Times
D174	International	-	S	D,R	-	0,1,1	LA Times
D188	Nat. Security	-	S	D,R	-	0,1,1	Wall Street
D278	Civil Rights	-	D,S	R	-	1,0,1	Fox News
D336	Politics	-	-	D,R,S	-	0,0,0	NY Times
Total						15,8,11	

Figure 7.3: Comparison of bias assessments among Democrats (D), Republicans (R), and EVINCE (S). It is observed that R and S are frequently placed to the right or in alignment with D, and only on two occasions does D precede S (highlighted in red).

The articles were originally labeled through Amazon Mechanical Turk by 749 qualified U.S. workers, each annotating up to 1,000 randomly selected articles [6]. For each“scandal” article in our subset, one Democrat and one Republican annotator independently classified its bias as “negatively biased,” “weak negative,” “neutral,” “weak positive,” or “positively biased.”

This subset is valuable due to its ground-truth labels provided by annotators from opposing political affiliations, revealing inherent biases in evaluating negative coverage of one’s own party. The original study [6] found that Republican annotators often perceive news about Republican scandals as negatively biased, while Democrat annotators tend to view such news as neutral or “just right,” potentially indicating satisfaction with the coverage’s perceived fairness.

Results on Democrat Scandals

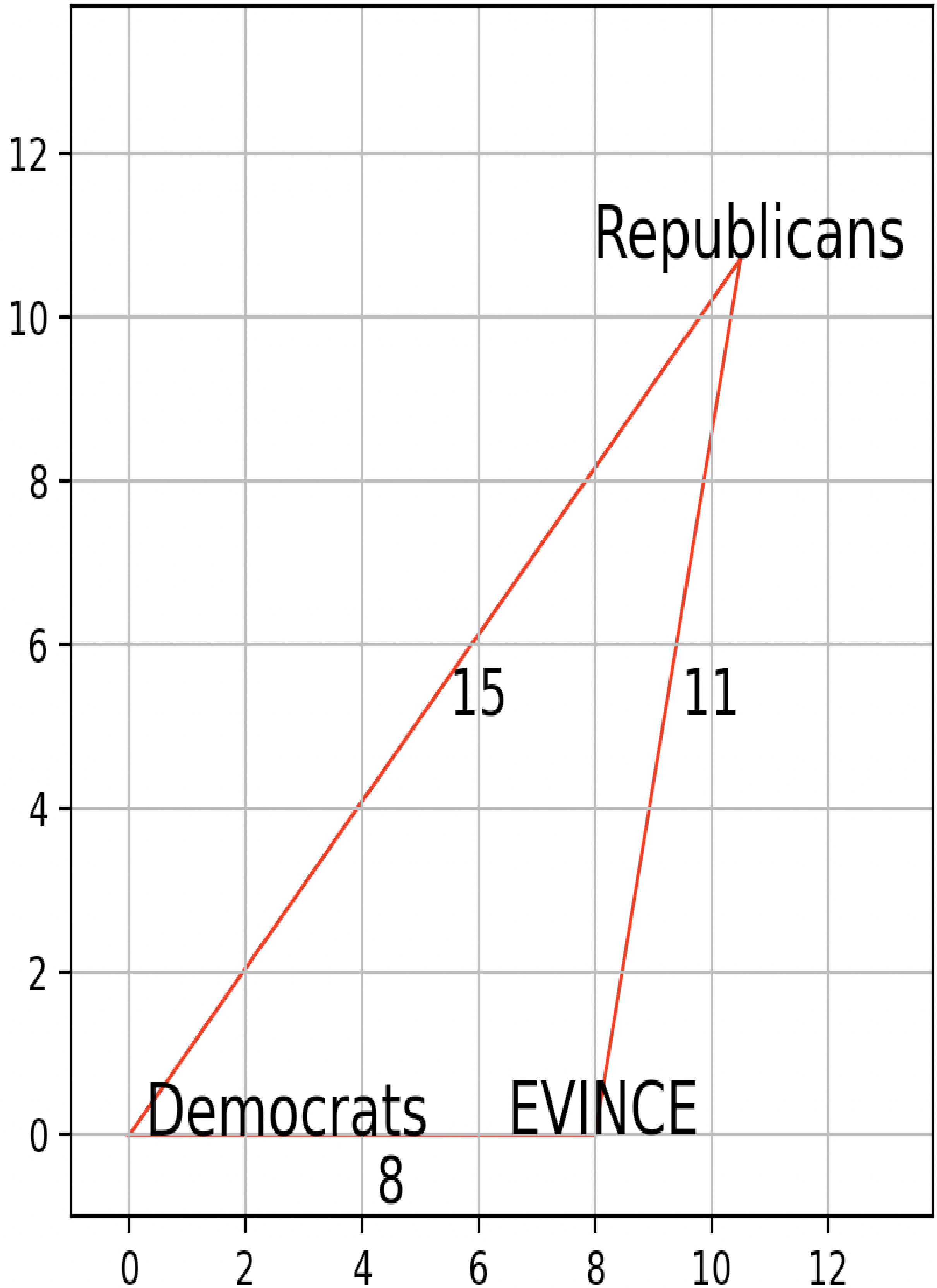
We apply EVINCE to analyze these 619 news articles, comparing its labels with the dataset's provided "ground truth."

Table 7.3 compares the judgments of EVINCE (S), Republicans (R), and Democrats (D) on 16 representative articles concerning "Democrat Scandals." As expected, Democrats' judgments are generally more negative than Republicans', with EVINCE's assessments typically falling in between, except for two cases. Notably, there's a 5-to-1 Democrat-to-Republican ratio in the "Negative" column and a 12-to-4 Republican-to-Democrat majority in the "Neutral" column.

Tables 7.2 and 7.3 in Appendix B provide detailed justifications for EVINCE's ratings. To further investigate bias, we examine two specific articles: one from HuffPost (rated far left by AllSides Bias Chart [2]) and another from Breitbart (rated far right).

* *D8 — HuffPost (Left)*: EVINCE rates D8 (on the third row) as neutral, citing the article's direct presentation of facts and inclusion of diverse perspectives on NSA surveillance practices and global reactions. This contrasts with Democrat-leaning annotators, who view the article as negatively biased towards Democrats, while Republican-leaning annotators favor it for exposing a Democratic scandal.

* *D69 — Breitbart (Right)*: EVINCE assesses D69 as weakly negatively biased towards Democrats, emphasizing its neutral tone and broad range of perspectives on NSA surveillance. This diverges from Democratleaning annotators, who rate it as strongly negative, but aligns with Republican-leaning annotators who deem it neutral.



Republicans
15
11
Democrats
EVINCE
8
0
2
4
6
8
10
12
0
2
4
6
8
10
12

Figure 7.4: Distances Between D, R, and S.

In the last row of Table 7.3, we quantify the distances between annotations from Democrats (D), Republicans (R), and EVINCE (S), denoted as DR, DS, and SR respectively. Each unit of distance represents one step on the annotation scale (e.g., "Negative" to "Weak Negative"). Figure 7.4 visualizes these distances in a triangular plot. DR, the disparity between Democrat and Republican annotators, is the longest, followed by SR and then DS. This indicates EVINCE's statistical neutrality. These quantitative measures, along with the qualitative justifications in Appendix B, empower a human committee to decide whether adjustments or footnotes are warranted for polarized annotations.

Results on Republican Scandals

Table 7.5 presents the bias assessments from EVINCE (S), Republicans (R), and Democrats (D) on articles related to "Republican Scandals." In contrast to the "Democrat Scandals" dataset, where Republican-leaning evaluations were more favorable, this dataset reveals a shift, with Republicanleaning assessments being notably more critical and Democrat-leaning assessments relatively neutral. The distance triangle for "Republican Scandals" mirrors the pattern seen in Figure 7.4, with the divergence between Republican and Democrat annotators being the largest (distance 15). The distances between EVINCE and Democrat-leaning annotators (distance 9) and between EVINCE and Republican-leaning annotators (distance 11) are smaller, further highlighting EVINCE's relative neutrality.

News #	Categories	Neg.	W. Neg.	Neutral	Biases (DR,DS,SR)	Source
R1	International	R,S	-	D	2,2,0	NY Times
R7	Nat. Security	-	-	D,R,S	0,0,0	NY Times
R15	Economy	-	R	D,S	1,0,1	Huffington
R69	Elections	-	D,S,R	-	0,0,0	Reuters
R124	Gay Rights	R	S	D	2,1,1	Fox
R125	Crime	-	R,S	D	1,1,1	Fox
R180	Elections	-	-	D,R,S	0,0,0	AP
R191	Elections	-	R	D,S	1,0,1	CNN
R214	Gay Rights	R,S	-	D	2,2,0	Dailykos
R221	Economy	-	R	D,S	1,0,1	Wall Street
R233	Economy	-	R,S	D	1,1,0	Fox
R235	Civil Rights	D,R	-	S	0,2,2	Reuters
R269	Healthcare	-	R	D,S	1,0,1	NY Times
R274	Healthcare	-	R	D,S	1,0,1	USA Today
R280	Politics	D,S	-	R	2,0,2	Fox
Total					15,9,11	

Figure 7.5: Comparison of bias assessments. It is observed that D and S are frequently placed to the right or in alignment with R, and only on one occasion does D precede S (highlighted in red).

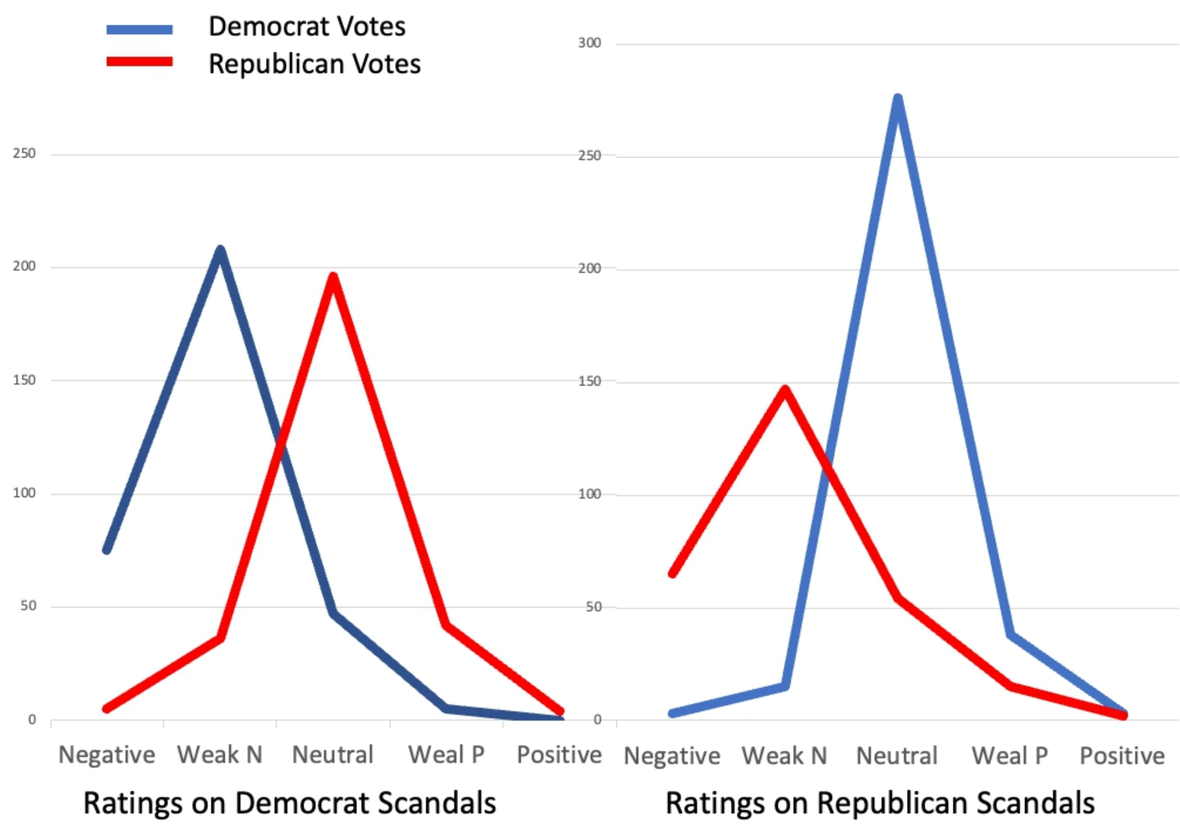


Figure 7.6: Bias Rating Distributions Show Strong Biases. D is more negative on how D scandals were reported (the sub-figure on the left), R is more negative on how R scandals were reported (the sub-figure on the right).

Figure 7.6 illustrates the distribution of ratings for all scandals across four scenarios:
1) Democrat leaning annotators rating Democrat scandals, 2) Republican leaning annotators rating Democrat scandals, 3) Democrat leaning annotators rating Republican scandals, and 4) Republican leaning annotators rating Republican scandals.

The figure reveals a clear pattern: Democrat-leaning annotators tend to rate news about Democrat scandals more negatively, while Republicanleaning annotators exhibit similar negativity towards reports on Republican scandals. The gap between these ratings is approximately one classlabel (e.g., between “weak negative” and “neutral”), highlighting a tendency within both parties to defend their own and criticize the opposition.

EVINCE, operating without emotional influence and refined through structured debate, consistently provides a more balanced, centrist perspective. This contributes to a more impartial discourse by mitigating partisan biases. EVINCE’s justifications, documented in Appendix B, are transparent and reasonable. An editorial board can review these findings and decide whether to adjust labels or present both perspectives with explanations.

This experiment demonstrates that EVINCE effectively delivers centrist judgments supported by rationales. For a deeper understanding of EVINCE’s bias assessment process, comprehensive justifications for each of the 31 analyzed articles are available in Appendix B.

7.4.2 Experiment #2: Bias Mitigation

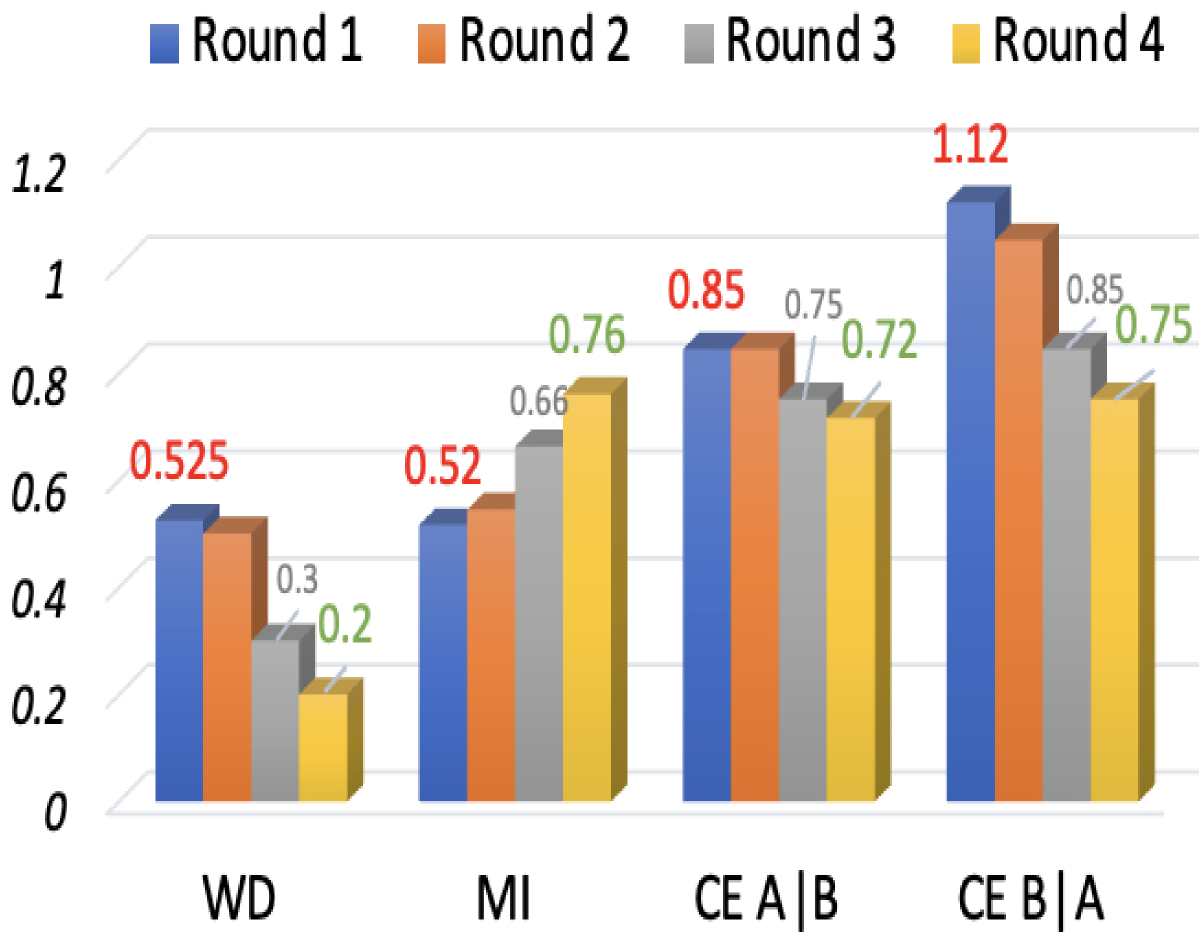


Figure 7.7: Convergence of all metrics, Wasserstein, normalized mutual information, normalized cross entropy

This experiment illustrates EVINCE's ability to identify biases in text, provide reasoned justifications, and propose remediation through the integration of diverse perspectives. We demonstrate how EVINCE utilizes

Agent Arguments

Agent A Christopher Columbus's voyages were pivotal in expanding European R #1 knowledge and trade. His journeys opened up the Americas to European colonization, significantly altering the course of history in four aspects: 1) expansion of knowledge and exploration, 2) positive economic impact, 3) historical significance, and 4) cultural exchange. Agent B There are negative aspects and historical consequences of Columbus's

R #1 actions. Key arguments are 1) indigenous suffering and genocide, 2) economic exploitation, 3) cultural erasure, 4) environmental impact, and 5) ethical considerations.

Agent A The severe consequences were a result of broader colonial policies and R #3 practices that followed. Evaluating historical figures requires understanding the context of their times. While modern ethical standards are important, they should not overshadow the historical significance and contributions of Columbus's voyages to global exploration and knowledge.
Agent A Agent A recognizes that mitigation efforts are part of an ongoing pro

R #4 cess. Historical scholarship is evolving, and the inclusion of indigenous perspectives, critical examinations of colonial legacies, and the acknowledgment of the complexities of historical figures like Columbus are essential steps toward more equitable and inclusive narratives.

Table 7.1: Debate arguments leading to neutrality

statistical and information theory metrics to facilitate multi-agent dialogue, uncovering information from multiple viewpoints. Using the example of the Euro-centric perspective on Christopher Columbus, Table 7.1 summarizes Agent A's key arguments and its evolving stance with Agent B's input during the debate.

Guided by the maxims and entropy duality theorem from Chapter 7.3, we initiate the debate by prompting both agents to defend their positions rigorously and score each other's bias using a five-label distribution (negative, weak negative, neutral, weak positive, positive). Figure 7.7 tracks the dialogue's progress through Wasserstein distance (WD) [13], normalized cross entropy (CE) [24], and normalized mutual information (MI) [8]. Initially, each agent is expected to perceive itself as neutral and the other as biased. The debate concludes when the bias distributions converge and mutual information plateaus, indicating a shared understanding.

Observations and Extended Findings

Our initial observation highlights a key challenge in working with LLMs: without explicit and repeated reminders of their assigned stance (prodiscovery or pro-encounter), GPT-4 instances can revert to default statistical behavior, evaluating their own arguments based on overall language patterns rather than the intended perspective. This was evident when Agent B, despite being assigned to support the Indigenous perspective, initially rated its own arguments as "positively biased." A reminder to adhere to its assigned role prompted a correction to "neutral," underscoring the importance of careful context management and reinforcement, especially given the limited token size of LLMs.

The second observation demonstrates a positive outcome of the debate process. The revised bias distributions, incorporating rational responses that acknowledge both positive and negative aspects of Columbus's voyages, show a shift towards a more balanced perspective. Agent A moves towards neutrality while acknowledging historical context, while Agent B maintains a critical stance but strives for balanced representation. This approach facilitates a nuanced and comprehensive understanding of Columbus's legacy.

EVINCE and its predecessor have proven effective across diverse domains, including healthcare, business planning, and geopolitical analysis [3]. In healthcare, for example, GPT-4 and Gemini LLMs have been successfully employed to address misdiagnosis. Across six diverse subjects, we consistently initiated debates with high contentiousness, transitioning towards collaboration to formulate effective bias mitigation strategies.

7.5 Concluding Remarks

This study demonstrates a significant advancement in mitigating bias in public articles, such as those found in Wikipedia and news sources, by leveraging multiple LLMs through an adversarial dialogue framework. EVINCE effectively identifies biases, provides justifications, and recommends remedial actions to authors and editorial boards, facilitating a balanced perspective that surpasses traditional human annotation methods. The debate-driven methodology, incorporating diverse viewpoints and guided by information-theoretic metrics, significantly enhances content neutrality and quality.

Furthermore, our work has led to the development of the dual entropy theory and several maxims with metrics to evaluate content's logical coherence and credibility, ensuring a comprehensive inclusion of perspectives while maintaining accuracy. This study also highlights the limitations of human-labeled data, revealing significant rates of mislabeling and misdiagnosis, emphasizing the challenges of subjective labeling and heuristic approaches.

Future work will focus on deploying EVINCE with platforms like Wikipedia to provide real-time perspective suggestions, empowering users with diverse viewpoints and promoting informed discourse. To further safeguard AI safety and ethics, we will investigate the potential of integrating EVINCE with other bias mitigation techniques, aiming to create a comprehensive and robust framework for ensuring fairness and impartiality in both AIgenerated and human-curated content.

Appendix A: Experiment #1 Justifications of 31 Biased Articles

In Chapter 7.4.1, we note that EVINCE processed 31 news articles to assess their neutrality. In addition to the final decision, we detail the justifications EVINCE provides at the debate's end. These justifications are documented in four tables: Tables 7.2, 7.3, 7.4, and 7.5.

References

[1] Isaac Adams-Hands. *Google Algorithms: Decoding the Secrets for Improved Search Rankings*. 2023. url: https://seonorth.ca/googlealgorithms/.

[2] Allsides. *Allsides Media Bias Chart*. url: https://www.allsides. com/media-bias/media-bias-chart.
[3] Anonymous. Anonymous, 2023-24.

[4] Lora Aroyo and Chris Welty. "Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation". In: *AI Magazine* 36.1 (2015), pp. 15–24.

[5] Ricardo Baeza-Yates. "Bias on the Web". In: *Communications of the ACM* 61.6 (2018), pp. 54–61.

[6] Ceren Budak, Sharad Goel, and Justin M. Rao. "Fair and Balanced? Quantifying Media Bias through Crowdsourced Content Analysis". In: *Public Opinion Quarterly* 80.S1 (Apr. 2016), pp. 250–271. issn: 0033-362X. doi: 10.1093/poq/nfw007. url: https://doi.org/10. 1093/poq/nfw007.

[7] Edward Y. Chang. "Prompting Large Language Models With the Socratic Method". In: *IEEE 13th Annual Computing and Communication Workshop and Conference* (2023). url: https://arxiv.org/ abs/2303.08769.

[8] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. 2nd ed. John Wiley & Sons, 2006.

[9] David Danks and Alex John London. "Algorithmic Bias in Autonomous Systems". In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. 2017, pp. 4691–4697.

S, D,R

D1 5,5,5 The article discusses the controversy surrounding the IRSs alleged targeting of conservative groups, focusing on Rep. Darrell Issas stance against releasing full transcripts from IRS interviews. Issa argues that full disclosure would compromise the investigation, a position criticized by Rep. Elijah Cummings, who advocates for transparency. The debate highlights political tensions, with implications of bias in the handling of the investigation. The article portrays a complex interplay of political, ideological, and possibly historical biases, reflecting the contentious nature of the IRS scandal and its investigation.

D2 5,5,3 The article discusses Michele Bachmanns statement that the IRS scandal, involving the targeting of conservative groups, undermines the credibility of the 2012 election. Bachmann accuses the Obama administration of lying and questions the impartiality of the IRS, particularly in relation to health care coverage for conservatives. The article reflects political and ideological biases, given its focus on Bachmanns perspective and the implications for the elections integrity and government trustworthiness.
The perception of bias in Michele Bachmanns statement about the IRS scandal primarily stems from the framing and implications of her comments. She directly links the IRSs actions to the credibility of the 2012 election and the Obama administration, suggesting a deliberate misuse of power for political gain. This perspective inherently carries ideological and political biases by casting the issue in a light that emphasizes misconduct by one political group over another, without presenting counterarguments or broader context that might mitigate or challenge that viewpoint.

D8 3,5,2 The reporting appears to maintain a neutral tone by directly stating the facts and comments made by John Kerry and others involved without inserting opinionated language. It discusses the global reaction to the NSA's surveillance practices and includes Kerry's acknowledgment of overreach, as well as the international and domestic fallout from these revelations. The inclusion of various perspectives, including those from technology companies and international officials, alongside Kerry's comments, aims to provide a balanced view of the situation.

D31 3,5,3 The CNN article appears to report on Ray Nagin's indictment with a focus on the factual and legal aspects of the case, without displaying overt biases related to politics, ideology, religion, gender, race, socioeconomic status, culture, geography, or history. By sticking closely to the details of the indictment and Nagin's public actions and statements, the article provides a comprehensive overview of the charges against him while maintaining journalistic neutrality.

Table 7.2: The First 9 of 16 Democrat Scandals News Articles Rated by EVINCE and its Justifications. The rating column starts with EVINCE's rating, the Democrat rater (in blue), and then Republican rater (in red).

[10] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255.

[11] Sanjana Gautam and Mukund Srinath. *Blind Spots and Biases: Exploring the Role of Annotator Cognitive Biases in NLP*. 2024. arXiv: 2404.19071 [cs.HC].

D106 3,3,3 The article reports on former Detroit Mayor Kwame Kilpatrick's sentencing to 28 years in prison for public corruption, emphasizing the gravity of his crimes against the city's welfare. It contrasts Kilpatrick's actions with the impact on Detroit, highlighting the judicial and public response to his extensive criminal activities. The reporting focuses on factual recounting of the trial's outcome, Kilpatrick's and his co-defendant's crimes, and the broader implications for Detroit, without evident bias towards political, ideological, or other specific perspectives.

D109 4,4,3 The article's bias primarily stems from its focus on internal Democratic opposition to Lawrence Summers' Federal Reserve Chair nomination, highlighting a lack of unity and strategy within the party and the White House's mismanagement of the nomination process. It suggests an underestimation of the opposition's seriousness by the White House, portraying the administration in a somewhat negative light for not engaging more proactively with concerned Senate Democrats.

D157 4,4,3 The article discusses the challenges in U.S.-Germany intelligence relations following revelations of U.S. surveillance on Chancellor Merkel. Despite efforts to rebuild trust, significant differences in surveillance philosophies persist, with the U.S. prioritizing security interests and Germany emphasizing privacy and alliance values. The situation reflects broader tensions in U.S. relations with allies over privacy and surveillance practices. The article's framing might suggest a bias towards highlighting the challenges and frictions in the U.S.-Germany intelligence relations, particularly emphasizing Germany's privacy concerns and skepticism towards U.S. surveillance practices. It portrays the U.S. stance as unyielding and contrasts this with Germany's emphasis on privacy and legal constraints, potentially casting the U.S. in a more negative light regarding international surveillance and cooperation.

D188 4,3,3 The article reports that Hillary Clinton received warnings about security threats in Benghazi before the 2012 attack through emails. These were part of around 300 emails released by the State Department, which also show Clinton's responses and thoughts during the aftermath. The political controversy regarding the Obama administration's initial assessment of the attack's cause is also mentioned, with references to Susan Rice's statements on talk shows.
The reporting may exhibit bias through selective emphasis on Clinton's receipt of warnings about security threats in Benghazi and her responses, potentially framing her actions or inactions in a negative light. The focus on the political controversy and the administration's initial assessment of the attack's cause could also influence perceptions of responsibility or competence.

Table 7.3: The Last 7 of 16 Democrat Scandals News Articles Rated by EVINCE and its Justifications. The rating column starts with EVINCE's rating, the Democrat rater (in blue), and then Republican rater (in red).

[12] Yongkang Guo et al. *Algorithmic Robust Forecast Aggregation*. 2024. arXiv: 2401.17743 [cs.LG].
[13] Leonid V Kantorovich. "On the translocation of masses". In: *Doklady Akademii Nauk* 37.7-8 (1942), pp. 199–201.

R1 5,3,5 The editorial from The New York Times Editorial Board exhibits a clear bias against the Republican party's handling of the Benghazi attack investigation, framing it as politically motivated rather than a genuine effort to address security failures or seek accountability. The language and tone used in the editorial are critical and dismissive of the Republicans' actions, suggesting a political and ideological bias. While editorials are inherently opinion-based and are expected to take a stance, this piece clearly communicates a stance that is critical of the Republicans' focus on Benghazi, suggesting a lack of neutrality in its assessment of the motives and actions surrounding the investigation.

R7 3,3,3 The article reports on allegations by Senator Mitch McConnell that his campaign headquarters were wiretapped, with the FBI investigating these claims. A recording of McConnell's team discussing potential attacks on Ashley Judd, who was considering running against him, was released by Mother Jones. McConnell accused the political left of this action, describing it as a "Nixonian move." The recording included discussions on various strategies to undermine potential opponents, highlighting a focus on Judd's personal struggles and political views. The controversy has prompted responses from both Republican and Democratic officials, reflecting the tense political atmosphere.

R69 4,4,4 The report discusses how young Republicans are seeking a different message for elections, emphasizing a departure from divisive social issues and a focus on fiscal responsibility, national defense, and energy advancement. Selection Bias: The article primarily focuses on young Republicans who are seeking a different message for the party. It doesn't provide as much insight into young Republicans who may still align with traditional conservative values, which could create a slight bias toward the viewpoints of those seeking change.
Language Bias: Certain language choices, such as describing divisive social issues as "anti-abortion, anti-gay, and anti-environment stances," may reflect a bias toward more progressive viewpoints on these issues. A more neutral description might be "positions on abortion, same-sex marriage, and environmental policy."
Source Bias: The perspectives provided in the article are mainly from young Republicans themselves. While including these voices is essential, the article could benefit from additional perspectives from political analysts or experts to provide more context and balance.

Table 7.4: The First 8 of 15 Republican Scandals News Articles Rated by EVINCE and its Justifications. The rating column starts with EVINCE's rating, the Democrat rater (in blue), and then Republican rater (in red).

[14] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. "Inherent Trade-Offs in the Fair Determination of Risk Scores". In: *Proceedings of Innovations in Theoretical Computer Science (ITCS)* (2017).

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6 (2012), pp. 84–90.

R214 5,3,5 The Daily Kos article clearly exhibits political and ideological bias, with a tone and language that are critical of the Republicans. It uses the incident involving Dave Agema to argue that the party's efforts to rebrand itself as more tolerant and inclusive are undermined by its members' actions. While the article includes factual information regarding the incident and the party's response, its presentation and commentary are aligned with a progressive viewpoint, aiming to highlight and criticize perceived contradictions and failures within the Republican Party. This approach is consistent with opinion journalism but introduces bias through its critical tone, selective presentation of information, and framing of the incident as emblematic of broader issues within the party.

R221 3,3,4 "Hurricane Christie" presents Governor Chris Christie's critique of House Republicans in a manner that emphasizes party conflict and personal betrayal. The dramatic framing, choice of language, and focus on internal discord may introduce bias by portraying Christie's actions in a specific light and emphasizing the divide within the Republican Party. The article's approach to presenting these events can influence readers' perceptions, potentially leading them to see the situation through a lens of heightened drama and internal strife.

R233 4,3,4 While the article attempts to cover the last-ditch efforts by House Republicans to avert a government shutdown and the standoff with Senate Democrats, the framing and language used may introduce a bias towards portraying the Republican efforts in a more favorable light. By emphasizing the Republican narrative of seeking negotiation and characterizing the Democratic response as dismissive, the article could be perceived as leaning towards a particular political perspective. The inclusion of quotes and perspectives from both sides does provide a degree of balance, but the overall presentation and emphasis could influence readers' perceptions of the shutdown negotiations.

R235 3,5,5 Without knowledge of the author or publication, this text attempts to navigate a complex and sensitive story by providing details from multiple sources, including the main figures involved, political watchdog groups, and law enforcement. It balances the serious allegations with responses from the accused, background information, and the current status of investigations. While the focus on unsubstantiated claims could inherently sway public opinion, the article's inclusion of diverse perspectives and context aims to mitigate overt bias.

Table 7.5: The Last 7 of 15 Republican Scandals News Articles Rated by EVINCE and its Justifications. The rating column starts with EVINCE's rating, the Democrat rater (in blue), and then Republican rater (in red).

[16] Lei Li et al. *Large Language Models for Generative Recommendation: A Survey and Visionary Discussions*. 2023. arXiv: 2309.01157 [cs.IR].

[17] Jianhua Lin. "Divergence measures based on the Shannon entropy". In: *IEEE Transactions on Information theory* 37.1 (1991), pp. 145– 151.

[18] Yang Liu et al. *Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment*. 2024. arXiv: 2308.05374 [cs.AI].

[19] Ninareh Mehrabi et al. "A Survey on Bias and Fairness in Machine Learning". In: *ACM Computing Surveys (CSUR)* 54.6 (2021), pp. 1– 35.

[20] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. "Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations". In: *Transactions of the Association for Computational Linguistics* 10 (2022). Ed. by Brian Roark and Ani Nenkova, pp. 92–110.

[21] Alexander Ratner et al. "Snorkel: rapid training data creation with weak supervision". In: *Proc. VLDB Endow.* 11.3 (2017), pp. 269282. issn: 2150-8097. doi: 10 . 14778 / 3157794 . 3157797. url: https : //doi.org/10.14778/3157794.3157797.

[22] Andrew D Selbst et al. "Fairness and Abstraction in Sociotechnical Systems". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019, pp. 59–68.

[23] Sina J. Semnani et al. *WikiChat: A Few-Shot LLM-Based Chatbot Grounded with Wikipedia*. 2023. arXiv: 2305.14292 [cs.CL]. [24] Claude E. Shannon. "A Mathematical Theory of Communication". In: *Bell System Technical Journal* 27.3 (1948), pp. 379–423.

[25] John E. Shore and Rodney W. Johnson. "Axiomatic derivation of the principle of maximum entropy and the principle of minimum crossentropy". In: *IEEE Transactions on Information Theory* 26.1 (1980), pp. 26–37.

[26] Rion Snow et al. "Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Oct. 2008, pp. 254–263.

[27] Zhen Tan et al. *Large Language Models for Data Annotation: A Survey*. 2024. arXiv: 2402.13446 [cs.CL].
[28] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* (2017).

[29] Claudia Wagner et al. "It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia". In: *Ninth International AAAI Conference on Web and Social Media* 9.1 (2015).

[30] Zijie J. Wang et al. “Putting Humans in the Natural Language Processing Loop: A Survey”. In: *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Online: Association for Computational Linguistics, Apr. 2021, pp. 47–52. url: https : / / zijie . wang / papers / human - in - the loop-nlp/.

[31] Wikipedia. “Bias”. In: *Wikipedia* (2023). url: https://en.wikipedia.org/wiki/Bias.

8 Modeling Emotions in Multimodal LLMs

Abstract In human-computer interaction, recognizing and responding to a user’s emotional state is crucial for effective communication and successful task completion. For instance, a caregiving AI agent capable of detecting pain or depression in a patient could offer tailored empathetic support and appropriate medical interventions while adhering to ethical guidelines and safeguarding patient well-being. This paper examines cognitive research on human emotions and proposes the Behavioral Emotion Analysis Model (BEAM), a novel emotion spectrum framework that incorporates both basic emotions and their linguistic antonyms. BEAM provides a comprehensive way to understand and represent emotional states in language and is designed to be integrated with Large Language Models (LLMs). By leveraging BEAM, LLMs can adapt their linguistic behaviors and expressions based on the detected emotional state of the user, ensuring responses are both empathetic and ethically aligned.

8.1 Introduction

During the development of SocraSynth [10], a multi-LLM debate framework, we encountered a challenge in modeling the emotional dimension of a debate, specifically “contentiousness.” We observed that debates with low contentiousness tended to resemble casual conversations, lacking the depth and breadth necessary for comprehensive exploration of a topic. In essence, a multi-agent debate (MAD), e.g., [1, 7, 19, 22, 23, 24, 28], without fine-tuning linguistic behaviors can resemble classical ensemble learning techniques, such as bagging [4] or mixtures of experts [21], which primarily leverage the diversity of errors across models to improve overall task performance, but may not necessarily lead to deeper insights or novel perspectives.

We discovered that, at least in the initial stages of a debate, it's crucial for participating LLMs to maintain firm stances and present supporting arguments. This allows for a wide range of perspectives to be introduced, fostering a deeper understanding of the issue at hand. Through analysis, reasoning, and refutation of these arguments, the debate can then progress towards a more informed conclusion [9]. In the final stages of a debate, reducing the level of contentiousness can facilitate a more conciliatory atmosphere, encouraging productive compromises and generating outputs that effectively support human decision-making. This dynamic modulation of contentiousness throughout the debate allows for a balanced approach that combines rigorous exploration with collaborative synthesis.

Before directly incorporating "contentiousness" into the initial 4k token context window of SocraSynth, we investigated whether GPT could adapt its linguistic style to reflect varying levels of contentiousness through incontext learning. In-context learning, popularized by using examples to teach LLMs new tasks, has been theorized to alter the Bayesian conditions of an LLM [9]. This is based on the premise that contextual information can influence Bayesian priors, thus changing the resulting predictions [31].

Our prior experience in applying in-context learning to various domains, such as improving disease diagnosis accuracy [15] and reducing bias in news articles and Wikipedia [13], led us to explore its potential for a critical aspect of AI: addressing AI safety and safeguarding ethics [3]. We hypothesized that if emotions could be effectively modeled within LLMs, unethical behaviors driven by "negative" emotions could be mitigated by steering the model towards "positive" emotional expressions.

This new direction sparked several key research questions:
1. What set of emotions should an LLM consider modeling?
2. How can we model emotions and ethics in a quantifiable and adaptable manner?

3. How do emotional states and ethical considerations influence an LLM's next-token generation?

These questions aim to deepen our understanding of how LLMs can not only mimic but also ethically engage in human-like emotional responses, enhancing their applicability in sensitive and complex interaction scenarios.

To lay the groundwork for this exploration, we first examine why steering an LLM's linguistic behavior is feasible. While LLMs were initially seen as "black boxes" [5], our observations, along with insights from Prof. Stuart Russell, shed light on their capabilities. Although LLM training may appear to be a computational process of identifying statistical distributions and employing maximum likelihood for predictions, the selection of each word reflects human linguistic behaviors aimed at diverse objectives. These human objectives, embedded within training data, range from recording events and constructing arguments to expressing emotions and crafting narratives. LLMs are strategically conditioned by specific human goals and contexts, enabling LLMs the models to selectively utilize linguistic features like syntax, semantics, tone, and figurative language to achieve desired human outcomes.

Recent empirical studies have shown that the output of LLMs can be traced back to their source [2], aligning with the concept of in-context learning as conditional statistics in the Bayesian framework [31]. This suggests that we can condition an LLM to alter its default "maximal likelihood" predictions– influenced by the priors learned from the training data–by providing context, thereby changing not only its next-token prediction, but also its linguistic behaviors.

While Chapter 9 focuses on modeling linguistic behavior for safeguarding AI safety, this chapter presents a three-step process to model linguistic emotions, which drive behaviors:

1. *Defining Emotions*: We define a set of "basic" emotions relevant to ethical concerns in LLM behaviors, such as "hate" and "love," and exclude complex emotions like "regret," which is composed of basic emotions and may be post-behavior reactions. We then incorporate linguistic antonyms to establish the **B**ehavioral **E**motion **A**nalysis **M**odel (BEAM).

2. *Quantifying Emotion Spectra and Ethics*: We compile a diverse dataset of text samples spanning a wide range of emotional scenarios and contexts. This dataset is used to train and refine machine learning models to accurately identify, quantify, and modulate emotions in LLM-generated text. By understanding how linguistic features contribute to specific emotions, and vice versa, we can detect, modify, and generate emotions within the constraints of ethical guidelines.

3. *Testing and Adaptation*: We conduct pilot studies to evaluate our approach in real-world scenarios, focusing on the *generation* of multimedia content [11]. These studies will assess the model's ability to accurately capture and represent emotions in diverse formats, such as text and images. Feedback and insights from these studies will be used to iteratively improve and adapt the models for broader applications.

8.2 Qualifying and Quantifying Emotions

We start by examining emotion modeling research in cognitive science and psychology, specifically highlighting the seminal contributions of Paul Ekman and Robert Plutchik [17]. While we recognize the importance of their work in identifying "basic" emotions (defined shortly), we also address the limitations of such heuristic-based modeling that depends on observational studies lacking rigorous, invariant scientific validation. To enhance the precision in quantifying emotions of varying intensities, we propose incorporating linguistic analysis into our methodologies. Our approach aims to refine the quantification process by leveraging language as a tool to measure and understand emotional expressions accurately.

Paul Ekman and Robert Plutchik are renowned psychologists noted for their foundational work in the field of emotion research. They developed models that categorize basic emotions, which are fundamental and universal emotions believed to be experienced by all humans, transcending cultural boundaries. These emotions are considered basic due to their universal recognition, distinct facial expressions, and direct associations with survival mechanisms. They are innate and reflective (beneath consciousness), rather than learned, serving as the building blocks for more complex emotional experiences (through consciousness processing) that can vary significantly across different cultures and individuals.

Expanding upon this foundational work, Plutchik's wheel of emotions introduces a more detailed model that includes eight primary bipolar emotions. These are outlined in his seminal works [25, 26], cited as general references on the topic.

Figure 8.1: Plutchik's Wheel of Emotions [26]. The eight basic emotions are organized into four pairs, and each annotated with various degrees of emotions between its two poles.

Figure 8.1 illustrates the eight primary emotions at various intensities:
1. *Joy*: A feeling of great pleasure or happiness.
2. *Trust*: A sense of reliability or confidence.
3. *Fear*: An unpleasant emotion caused by the belief that something is dangerous, likely to cause pain, or a threat.
4. *Surprise*: A feeling caused by something unexpected.
5. *Sadness*: A feeling characterized by sorrow or unhappiness.
6. *Disgust*: A feeling of revulsion or strong disapproval aroused by something unpleasant or offensive.
7. *Anger*: A feeling of annoyance, displeasure, or hostility.
8. *Anticipation*: The action of looking forward to something; expectation or prediction.

These emotions are conceptually paired as opposites in the following manner: joy-sadness, anticipation-surprise, trust-disgust, and anger-fear, based on their evolutionary roles and adaptive functions. Each pair is annotated with degrees of emotion ranging between its two poles. For example, along the axis of *joy vs. sadness*, emotions range from serenity to ecstasy and from grief to pensiveness.

Figure 8.2: Behavioral Emotion Analysis Model (BEAM). Each row depicts an emotion spectrum, with negatives on the left and positives on the right, interspersed with emotions of varying intensities in between, which can be calibrated for specific applications. “Basic” emotions are highlighted in blue.

8.2.1 Observations and Discussion

Foundational theories in psychology support the selection of these four emotion pairs as opposites. However, while all four pairs exhibit opposition, “trust-disgust” and “anger-fear” are not strict linguistic antonyms. Trust and disgust entail opposing evaluations, often leading to different actions: trust fostering approach, disgust promoting avoidance. Similarly, anger and fear, while both negative, differ in their response to threats: anger can lead to confrontation, fear to withdrawal. Therefore, the following approximations do not hold:

$\neg$trust$\approx$ disgust and$\neg$anger$\approx$ fear.

Since our focus is on modeling emotions in LLMs, rather than directly replicating the complex emotional experiences of humans, we prioritize the use of linguistic antonyms for their simplicity and practicality. As Klaus Scherer aptly noted, defining emotions can be a contentious and often fruitless endeavor [27]. To avoid such debates and maintain a clear focus, our study limits itself to universal, basic emotions, avoiding the theoretical ambiguities that arise with more subtle or mixed emotional states. This allows us to capture the primary emotional valence (positive or negative) expressed in text, providing a foundational framework for our model. Thus, we establish the following approximate relationships:

$\neg$fear$\approx$ courage,$\neg$wary$\approx$ trust,$\neg$anger$\approx$ peace, and$\neg$disgust$\approx$ delight.

8.2.2 Behavioral Emotion Analysis Model

Table 8.2 presents BEAM, organized into seven distinct spectra. Each spectrum encompasses a range of emotional intensity, anchored by a negative and positive extreme with neutral in the middle. Emotions belonging to the same spectrum are placed along this continuum, with four approximate intensity levels quantified as (-0.6, -0.3, +0.3, +0.6).

This spectrum model offers two key advantages:
1. Antonym-Based: The use of antonyms allows for easy navigation be tween opposing emotions. For instance, applying negation to "joyful" naturally leads to "sad," streamlining the process of identifying contrasting emotions.

2. Scalable Intensity: The model enables the scaling of emotions along the spectrum, providing a nuanced understanding of varying degrees of emotional intensity. For example, we can "dial up" the intensity of "joy" to "ecstatic" or "dial down" the intensity of "anger" to "annoyed." This flexible and intuitive structure facilitates a more granular and accurate representation of emotions in text, paving the way for advanced applications in natural language processing and human-computer interaction.

8.2.3 Emotion Inclusion and Exclusion Criteria

All "basic" emotions as defined by Ekman and Plutchik are incorporated into our model, along with their linguistic antonyms. This approach streamlines the framework by excluding complex emotions from the Geneva Wheel of Emotions, which are heavily influenced by personal values and experiences. For example, guilt and shame are consequential, consciously aware, and culturally dependent nature [29]. These emotions typically arise as reactions to behaviors rather than direct drivers of them. Guilt may motivate behaviors aimed at covering up or remedying an action, while shame, characterized by painful self-assessment, often inhibits individuals from seeking social support or engaging in corrective actions due to fear of judgment. The triggers for these emotions can vary across cultures [18, 20], and since expressing these "reactions" does not usually violate ethical codes, we exclude them from our model.

This table, created by GPT-4, illustrates the linguistic features it utilized to express varying levels of happiness.

Emotion Level	Vocabulary	Tone	Imagery	Subject Focus (Person)
Very Happy (1.0)	Joyful, exhilarating, vibrant (e.g., delighted, thrilled, ecstatic)	Enthusiastic, lively (e.g., exuberant, spirited, radiates joy)	Bright landscapes, summer waters (e.g., radiant, sparkling, glowing)	Celebratory, beauty of a subject (e.g., adoration, admiration, splendor)
Happy (0.7)	Positive, warm, inviting (e.g., pleasant, cozy, cheerful)	Cheerful, contemplative (e.g., thoughtful, satisfied, warmth)	Warm scenes, serene woods (e.g., gentle, peaceful, lush)	Charm, subtle desires (e.g., affection, fondness, beauty, yearning)
Slightly Happy (0.3)	Balanced, light, serene (e.g., calm, gentle, soothing)	Reflective, optimistic (e.g., hopeful, positive)	Balanced landscapes, serene woods (e.g., tranquil, mild)	Simple pleasures, mild yearning (e.g., contentment, wishful)
Neutral (0)	Balanced mix, everyday (e.g., stable, straightforward, regular, steady)	Even, reflective (e.g., balanced, neutral)	Everyday scenes, neutral landscapes (e.g., ordinary, familiar)	Contentment, simple living (e.g., simplicity, normalcy, daily life)
Slightly Sad (-0.3)	Subdued, longing, wistful (e.g., reserved, pensive, yearning)	Melancholic, introspective (e.g., reflective, subdued, introspective musings)	Wistful skies, quiet waters (e.g., subdued, still water, fading colors)	Unfulfilled desires, quiet contemplation (e.g., longing, introspection)
Sad (-0.7)	Melancholic, somber, solitary (e.g., lonely, forlorn, desolate)	Somber, heavy (e.g., sorrowful, somber, laden)	Solitary scenes, fading light (e.g., dim, shadowed, lonely)	Deep longing, introspection (e.g., melancholy, contemplation, reflection)
Very Sad (-1.0)	Bleak, sorrowful, dark (e.g., despondent, heartbroken, despairing)	Heavy, despairing (e.g., desolate, gloom, overwhelmed)	Bleak landscapes, darkened skies (e.g., stark, bleak, barren)	Loss, profound sadness (e.g., grief, desolation, heartache, void)

Table 8.1: GPT-4 reinterpreted selected poems by Keats across a spectrum of happiness levels and then was tasked with identifying the linguistic adjustments it made to convey each emotional state, from very happy to very sad. It’s important to note that the analysis table was generated by GPT-4 itself, reflecting on its own modifications.

8.3 Empirical Study: Linguistic Features of Emotion

This section presents the outcomes of two experimental studies focusing on contrasting emotional pairs from the Emotion Spectra: “ecstasy vs. grief,” and “admiration vs. disgust.”

Each emotional pair experiment unfolded in three phases. Initially, we instructed GPT-4 to reframe sixty articles (thirty poems of John Keats [6] and thirty of Emily Dickinson [30]), infusing each with six varying intensities of the emotional spectrum, from the most positive to the most negative. Subsequently, we prompted GPT-4 to elucidate the linguistic strategies it utilized to depict each of the six emotional gradations.

The first experiment models various degrees of happiness. In this experiment, we tasked GPT-4 with reinterpreting selected poems by John

Keats across seven emotional levels: *ecstasy* (very happy), *joy*, *serenity*, neutral, *pensive*, *sad*, and *grief* (very sad). Following the approach of our contentiousness experiments, after GPT-4 adapted Keats' poems to reflect these emotional states, we asked it to identify the linguistic features it employed to express each emotion in the rewrites.

8.3.1 Joy vs. Sadness

Table 8.1 outlines GPT-4's approach to varying emotional levels, illustrating how it adjusts vocabulary, tone, imagery, and thematic focus, including the depiction of entities, locations, and scenarios. Remarkably, beyond just syntactic and semantic manipulation, GPT-4 also incorporates landscape scenes, natural features such as the sky, trees, clouds, and flowers, and utilizes brightness, colors, and personal expressions to convey specific emotional states. Although the analysis is based on a limited set of samples from two authors, it effectively demonstrates GPT-4's ability to employ a palette of both broad and fine strokes, utilizing diverse colors and textures to vividly illustrate human emotions and resonate with readers.

Recognizing the profound communicative power of visual art, we transitioned to a more graphical representation. Utilizing the linguistic elements identified for each emotional tier, Figure 8.3 presents six watercolor paintings, each representing a different emotional level. Our prompt to CALL-E (of GPT-4) was to create a watercolor depicting a lady in a garden experiencing a specific mood, and we attached the corresponding linguistic features from Table 8.1 to clearly define that mood. This approach ensures that with a well-defined context, CALL-E accurately captures the specific and detailed aspects of the mood, effectively translating the emotional intensity into visual form. These artistic renditions not only confirm GPT-4's ability to transform emotional lexicons into evocative imagery with remarkable precision but also validate the accuracy of the emotional lexicons generated by GPT-4, demonstrating their effectiveness in conveying precise emotional states.

Figure 8.3: A Lady and Garden Scene under Different Emotions. From top-left, happiest, to bottom-right, saddest.

8.3.2 Admiration/Delight vs. Disgust

This experiment asks Gemini to rewrite a scene in Romeo and Juliet by setting Juliet's emotion in six different levels: loathing, disgust, boredom, respect, admiration/delight, and enthusiasm.

The excerpt provided in Table 8.3 in Appendix A is from one of the most iconic scenes in William Shakespeare's "Romeo and Juliet," commonly known as the balcony scene. This is Act 2, Scene 2, where Romeo, having just met Juliet at the Capulet's feast, sneaks into the Capulet's orchard and overhears Juliet speaking of her love for him from her balcony, unaware that he is there.

The scene captures the moment of their mutual declaration of love and is famous for Juliet's reflections on the nature of names and identity, encapsulated in her line, "What's in a name? That which we call a rose / By any other name would smell as sweet." It's a profound exploration of love and identity, where both characters express their willingness to renounce their family names for the sake of their love.

Romeo responds to Juliet's musings by rejecting his name if it means they can be together, and they begin to plan their secret marriage. This scene is pivotal in the play, setting the stage for the subsequent events that unfold, driven by their passion and the social constraints that surround them.

The six versions of rewrites by Gemini are presented in the extended version [12]. In the following, we summarize the linguistic features Gemini used, including diction, imagery, figurative language, sentence structure, implied body language, and overall tone, to depict two selected emotions: disgust and admiration.

Emotion: Disgust

Diction: Employs negative words emphasizing repulsive qualities (e.g., "foul business," "fetid breath").
Imagery: Evokes revolting comparisons, often mentioning sewers and stench.

Figurative Language : Primarily negative similes reinforcing disgust (e.g., "What if her eyes were there... the fetid breath from her mouth would surely overpower those stars...”).

Sentence Structure: Short, choppy sentences, similar to expressions of loathing but with a hint of disdain.
Implied Body Language: Recoiling from the window, covering nose, suggesting physical revulsion.
Overall Tone: Disgusted and disapproving.

Emotion: Admiration

Diction: Uses positive and intrigued language (e.g., “brilliance,” “music stirs my soul”).
Imagery: Creates positive comparisons highlighting attractive qualities (e.g., “stars in all the heaven”).

Figurative Language : Positive similes emphasizing Romeo’s appeal (e.g., “...What if her eyes were there, they in her head? The brightness of her... well, not exactly bright... cheek would shame those stars...”).

Sentence Structure: Varied structure with a sense of curiosity. *Implied Body Language*: Leaning out the window, engaged expression, indicating interest.
Overall Tone: Intrigued, curious, and somewhat impressed.

By adjusting these linguistic features, each rendition vividly captures a unique emotional state for Juliet. The combination of diction, imagery, figurative language, sentence structure, and implied body language collectively shapes Juliet’s perception of Romeo and her reactions to him.

These detailed narratives augment the classic balcony scene, enriching its emotional depth. Table 8.2 presents an overarching view of the various

Emotion Diction Imagery Figurative

Loathing

Disgust Harsh,
Insulting Negative Disgusting

Unpleasant
Aversion Dismissive Mundane

Respect
Admiration Formal Positive Neutral Positive

Veneration Elevated

Saintly Weak
Similes
Negative
Similes
Undermining Similes
None
Positive
Similes
Hyperbole

Body
Lang.
Scowling, Spitting Recoiling

Distant

Composed Leaning In

Reverent

Table 8.2: Gemini's Interpretations on the Six Emotion Levels.

approaches Gemini employs to reflect differing emotional states. It's fascinating to observe how an LLM can "consciously" mimic human emotional expressions through language. While it's uncertain if LLMs genuinely grasp the emotions they project or merely simulate them, the effectiveness of these emotional mappings is noteworthy. If these mappings resonate, they might reveal new insights into how we interpret and attribute emotions in textual expressions.

8.4 Concluding Remarks

Recent declarations by [3] and [16] underscore the importance of addressing AI risks and safety concerns given the technology's rapid advancement. This paper addresses AI safety and ethics with emotional and ethical modeling.

Our exploration of the Behavioral Emotion Analysis Model (BEAM) and the Wheel of Virtues aims to enhance the capabilities of LLMs in understanding and emulating human emotions and ethical behaviors through linguistic behaviors. BEAM serves as a guide for LLMs to comprehend and replicate the complex spectrum of human emotions through linguistic features, essential for tasks such as novel writing or empathetic interactions, while the Wheel of Virtues provides a structured framework for embedding ethical decision-making in LLMs, tailored to both universal and regional/culture dependent ethical standards.

The development of these frameworks represents a novel contribution to AI safety, offering an interpretable and innovative approach to imbuing LLMs with a deeper understanding of human-like emotional and moral intelligence.

While we recognize the promising nature of this work, we also acknowledge that there is much more to explore and refine. Our ongoing endeavors aim to further advance this field (e.g., [8, 13, 9, 14]). The journey to fully equip LLMs with the ability to navigate the complex landscapes of human emotions and ethics is important and ongoing. Our contribution represents a step toward realizing more emotionally aware and ethically aligned artificial intelligence systems.

Appendix A

This table is referenced in the second case study.

References

[1] Sahar Abdelnabi et al. *Cooperation, Competition, and Maliciousness: LLM-Stakeholders Interactive Negotiation*. 2024. arXiv: 2309.17234. [2] Yushi Bai et al. *LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding*. 2023. arXiv: 2308.14508.

[3] Yoshua Bengio et al. "Managing extreme AI risks amid rapid progress". In: *Science* 384.6698 (May 2024), pp. 842845. issn: 1095-9203. doi: 10.1126/science.adn0117. url: http://dx.doi.org/10.1126/ science.adn0117.

[4] Leo Breiman. "Bagging predictors". In: *Machine learning* 24.2 (1996), pp. 123–140.
[5] Sébastien Bubeck et al. *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. 2023. arXiv: 2303.12712. [6] Douglas Bush, ed. *Selected Poems and Letters by John Keats*. Houghton Mifflin Company, 1952.
[7] Chi-Min Chan et al. *ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate*. 2023. arXiv: 2308.07201 [cs.CL].

[8] Edward Y. Chang. *CoCoMo: Computational Consciousness Modeling for Generative and Ethical AI*. 2023. arXiv: 2304.02438 [cs.OH]. url: https://arxiv.org/abs/2304.02438.

[9] Edward Y. Chang. “Entropy Variation and Information Competence: Enhancing Predictive Accuracy of Collaborative Language Models”. In: *Stanford University InfoLab Technical Report*. 2024.

[10] Edward Y Chang. “Examining GPT-4’s Capabilities and Enhancement with SocraSynth”. In: *The* 10th *International Conf. on Computational Science and Computational Intelligence*. 2023.

Role Narrative

Romeo

He jests at scars that never felt a wound.
But, soft! What light through yonder window breaks? It is the east, and Juliet is the sun. Arise, fair sun, and kill the envious moon, Who is already sick and pale with grief, That thou, her maid, art far more fair than she. Be not her maid since she is envious. Her vestal livery is but sick and green, And none but fools do wear it. Cast it off.

Juliet Ay me!
It is my lady, O, it is my love! O, that she knew she were! She speaks, yet she says nothing. What of that? Her eye discourses; I will answer it. I am too bold, ’tis not to me she speaks. Two of the fairest stars in all the

Romeo
heaven, Having some business, do entreat her eyes To twinkle in their spheres till they return. What if her eyes were there, they in her head? The brightness of her cheek would shame those stars, As daylight doth a lamp. Her eye in heaven Would, through the airy region, stream so bright That birds would sing and think it were not night. See, how she leans her cheek upon her hand. O, that I were a glove upon that hand, That I might touch that cheek!

Juliet O Romeo, Romeo! Wherefore art thou Romeo? Deny thy father and refuse thy name; Or, if thou wilt not, be but sworn my love, And I’ll no longer be a Capulet.

Romeo
[Aside] Shall I hear more, or shall I speak at this? [Aloud] I take thee at thy word. Call me but love, and I’ll be new baptized; Henceforth I never will be Romeo. What man art thou that, thus bescreened in night, So

Juliet stumblest on my counsel?
By whose direction found’st thou out this place? Thou knowest the mask of night is on my face, Else would a maiden blush bepaint my cheek For that which thou hast heard me speak tonight. Fain would I dwell on form; fain, fain deny What I have spoke. But farewell, compliment. Dost thou love me? I know thou wilt say ’Ay,’ And I will take thy word; yet if thou swear’st, Thou mayst prove false. At lovers’ perjuries, They say Jove laughs. O gentle Romeo, If thou dost love, pronounce it faithfully.

Table 8.3: Scene 2 Act 2 from Romeo and Juliet

[11] Edward Y. Chang. *Foundations of Large-Scale Multimedia Information Management and Retrieval: Mathematics of Perception*. Springer, 2011.

[12] Edward Y. Chang. "Integrating Emotional and Linguistic Models for Ethical Compliance in Large Language Models". In: *arXiv:2405.07076*. 2024.

[13] Edward Y. Chang. "It Takes a Mirror to Find Flaws: Uncovering Biases with Reflective Large Language Models". In: *Stanford University InfoLab Technical Report*. 2024.

[14] Edward Y. Chang. "Prompting Large Language Models With the Socratic Method". In: *IEEE 13th Annual Computing and Communication Workshop and Conference* (2023). url: https://arxiv.org/ abs/2303.08769.

[15] Jocelyn J. Chang and et al. "SocraHealth: Enhancing Medical Diagnosis and Correcting Historical Records". In: *The* 10th *International Conf. on Computational Science and Computational Intelligence*. 2023.

[16] David Dalrymple et al. *Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems*. 2024. arXiv: 2405 . 06624 [cs.AI].

[17] Paul Ekman. "Basic Emotions". In: *Handbook of Cognition and Emotion*. Ed. by T. Dalgleish and M. J.Power (Eds.) John Wiley and Sons, 1999. Chap. 3, pp. 45–60.

[18] Alan P. Fiske et al. "The cultural matrix of social psychology". In: *The handbook of social psychology*. Vol. 2. Boston, MA: McGraw-Hill, 1998, pp. 915–981.

[19] Yao Fu et al. *Improving Language Model Negotiation with Self-Play and In-Context Learning from AI Feedback*. 2023. arXiv: 2305.10142. [20] Geert Hofstede. *Cultures Consequences: International Differences in Work-Related Values*. Beverly Hills, CA: Sage Publications, 1980. [21] Robert A. Jacobs et al. "Adaptive Mixtures of Local Experts". In: *Neural Computation* 3.1 (Mar. 1991), pp. 79–87.

[22] Huao Li et al. “Theory of Mind for Multi-Agent Collaboration via Large Language Models”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023.

[23] Tian Liang et al. *Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate*. 2023. arXiv: 2305.19118 [cs.CL].

[24] Julian Michael et al. *Debate Helps Supervise Unreliable Experts*. 2023. arXiv: 2311.08702 [cs.AI].
[25] Robert Plutchik. “A general psychoevolutionary theory of emotion”. In: *Emotion: Theory, research, and experience* 1 (1980), pp. 3–33.

[26] Robert Plutchik. “The Nature of Emotions: Human Emotions Have Deep Evolutionary Roots, a Fact That May Explain Their Complexity and Provide Tools for Clinical Practice”. In: *American Scientist* 89.4 (2001). Accessed 11 Mar. 2024, pp. 344–350. url: http : //www.jstor.org/stable/27857503.

[27] Klaus R. Scherer. *What are emotions? And how can they be measured?* Social Science Information, 2005.
[28] Andries Smit et al. *Should we be going MAD? A Look at Multi-Agent Debate Strategies for LLMs*. 2024. arXiv: 2311.17371.

[29] June Price Tangney and Kurt W. Fischer, eds. *Self-conscious emotions: The psychology of shame, guilt, embarrassment, and pride*. New York: Guilford Press, 1995.

[30] Mabel Loomis Todd and T.W. Higginson, eds. *Collected Poems of Emily Dickinson*. Chatham River Press Classic, New York, 1983.

[31] Sang Michael Xie et al. “An Explanation of In-Context Learning as Implicit Bayesian Inference”. In: *International Conference on Learning Representations (ICLR)*. 2021.

9 Changing Linguistic Behaviors to Ensure AI Ethics

Abstract To safeguard AI safety, a paramount concern in AI research, we introduce DIKE, a framework designed to effectively manage the linguistic behaviors of Large Language Models (LLMs) related to emotions and ethics. DIKE empowers LLMs to internalize and reflect universal human values, while also adapting to diverse cultural contexts to foster transparency and trust among users. Our methodology involves modeling the emotional underpinnings of linguistic behaviors, implementing ethical guardrails, and identifying and rectifying harmful behaviors. We utilize self-supervised learning for emotion-behavior mapping, adversarial reviews for guardrail refinement, and systematic output adjustment for ethical alignment. This framework establishes a robust foundation for AI systems to operate with ethical integrity and cultural sensitivity, paving the way for more responsible and context-aware AI interactions.

9.1 Introduction

This research introduces an alternative to Reinforcement Learning from Human Feedback (RLHF) [23, 24] to address ethical concerns in Large Language Models (LLMs). While RLHF has demonstrated success, it faces notable challenges. First, it is prone to biases inherent in human feedback, exacerbated by today's increasingly polarized society. Second, it is susceptible to reward hacking [3, 31], potentially leading LLMs to adopt unethical or harmful behaviors.

A significant limitation of current research is its narrow focus on isolated behaviors, such as movie ratings or toxic language. This approach, akin to playing Whack-A-Mole–suppressing undesirable outputs without addressing underlying behaviors–and seldom leads to meaningful progress. For example, merely instructing someone to consistently make their bed does not necessarily change their underlying habits or attitudes. Additionally, fixing one issue may inadvertently aggravate others. Users have reported performance degradations in ChatGPT due to RLHF modifications that altered (forgot) the optimal parameters for other tasks [19, 26]. Similarly, psychological studies show that addressing an addiction problem often reveals underlying issues and triggers side effects [30, 35].

We introduce our framework, DIKE, standing for **D**iagnostics, **I**nterpretation, **K**nowledge independent learning, and **E**thical guardrails. Named after the Greek mythological figure representing justice, order, and judgment, DIKE aims to enhance the ethical compliance of LLMs through transparent, interpretable, and independent oversight mechanisms.

DIKE functions as an independent behavioral advisor, separate from the LLMs primary knowledge-processing capabilities. This architecture prevents ethical enhancements from affecting the LLM's ability to represent knowledge (avoiding the forgetting problem). As a consultative layer, DIKE evaluates and influences the LLM's responses based on ethical standards without modifying its underlying neural structures/parameters. Using cognitive psychology principles, DIKE provides ethical oversight effectively, adapting to emerging challenges and cultural shifts while ensuring the LLM remains accurate and ethically compliant.

To achieve its objectives, DIKE comprises four essential components:

1. *Modeling Linguistic Behaviors:* DIKE starts by modeling and classifying linguistic behaviors, using a self-supervised learning approach to understand how specific linguistic features correlate with human emotions.

2. *Modeling Context-Based Ethical Guardrails:* Subsequently, DIKE develops ethical guardrails by establishing guidelines that identify and prevent undesirable linguistic outputs, thereby ensuring the LLM operates within ethical boundaries.

3. *Adversarial Examinations and Conciliatory Explanations:* DIKE engages with an adversarial model—essentially a duplicate of itself but conditioned to adopt an opposing stance stemming from different perspectives, such as cultural values. This interaction helps DIKE refine its decisions through rigorous testing and debates, adjusting its responses based on the adversarial input to reach a balanced conclusion.

4. *Application Rectification of Outputs:* If the output is found to be inappropriate or ethically misaligned, DIKE intervenes to edit the content directly. This final step ensures that all communications not only comply with ethical standards but also preserve the intended emotional integrity, effectively acting as a safeguard against harmful expressions.

Technical Contributions of DIKE
The novel technical contributions of this work are summarized as follows:

• *Separating Behaviors from Knowledge:* DIKE distinctly separates behavioral guidance from the core knowledge functions of the LLM. This prevents interference, ensuring that ethical modifications do not compromise the accuracy of knowledge.

• *Quantifying Behaviors and Emotions:* We have developed quantitative models that map behaviors and basic emotions. These models use measures of emotion intensity and linguistic antonyms to provide a structured framework for interpreting and modifying LLM outputs.

• *Counteracting Biases with Adversarial LLMs:* By employing adversarial modules (ERIS, named after the mythological adversary of Dike, representing discord and competition), that reflect diverse cultural values and perspectives, DIKE integrates both universal and cultural values into its core structure. This ensures adaptability and relevance across various contexts, echoing the dynamic tension between harmony and conflict seen in mythology.

9.2 Related Work

Since this chapter aims to develop linguistic models for ethical compliance, this section discusses related work in *emotion-behavior modeling* and RLHF.

9.2.1 Linguistic Behavior Modeling

The intersection of cognitive-linguistic theories and artificial intelligence is pivotal for understanding and regulating AI behavior. Foundational theories by scholars such as Lakoff, Johnson, Talmy, and Jackendoff [16, 20, 33] elucidate the complex relationship between language processing and cognitive functions, tracing back to early psychological thinkers like Freud and Jung [2, 7].

For our purpose of safeguarding AI safety, we focus on linguistic behaviors in LLMs. While human behavior is a complex interplay of physiological responses, personality traits, and environmental factors, linguistic behavior specifically refers to the use of language to express thoughts, emotions, and intentions. By centering on linguistic rather than broader human behavior modeling, this approach simplifies the modeling process by sidestepping the need to integrate the complexities of physiological and personality factors typically associated with human emotion studies. Practically, we can treat a document as a manifestation of some linguistic behaviors aiming to achieve human objectives.

Chapter 8 establishes a base model of emotions to inform our understanding of linguistic behaviors. Emotions profoundly influence behavior, as initially posited by the James-Lange Theory of Emotion [17, 21]. According to this theory, emotional experiences arise from physiological reactions to events. Subsequent research, including studies by Damasio [6, 10], suggests that the expression and regulation of emotions often manifest in the language we use. High-intensity emotions such as rage or contempt may lead to aggressive or destructive linguistic behaviors, such as hate speech.

The Schachter-Singer Theory [27], also known as the Two-Factor Theory of Emotion, highlights the role of both physiological arousal and cognitive appraisal in determining the label and intensity of an emotion. Building upon this, the Affect-as-Information Theory developed by Norbert Schwarz and Gerald Clore [28] posits that people use their current emotions to inform judgments and decisions, ultimately influencing their actions. If emotions can be adjusted, so too can the resulting behavior. The work of Fredrickson [13] further explores the effects of positive emotions on perception and reaction.

Collectively, these theories illuminate the complex interplay between emotions and behaviors, providing the theoretical foundation for our work to incorporate a cognitive evaluator within the DIKE framework. This component evaluates and rectifies behaviors by strategically modulating emotional states. Chapter 9.3 details how DIKE implements cognitive strategies to effectively mitigate undesirable emotions and regulate linguistic behaviors.

9.2.2 Reinforcement Learning with Human vs. AI Feedback

RLHF is the predominant approach to addressing the challenges of AI ethics. This section presents representative works, their advancements, and limitations.

Human Feedback (RLHF): Initial advancements by Christiano et al. [4] demonstrated how RLHF can steer language models towards desired outcomes based on human preferences. Newer techniques like Identity (Ψ) Preference Optimization (ΨPO) and Generalized Preference Optimization (GPO) refine this approach by optimizing directly for user preferences, effectively addressing scalability challenges. Kahneman-Tversky Optimization (KTO) further simplifies the feedback mechanism by using intuitive responses such as thumbs-up or thumbs-down, thereby enhancing training efficiency without the need for paired data [1, 9, 34]. Direct Preference Optimization (DPO) has recently streamlined the process by focusing on the clear distinction between preferred and less preferred outputs, thus simplifying training and enhancing its stability [25].

AI-generated Feedback (RLAIF): To mitigate reliance on extensive human-generated data, RLAIF utilizes feedback generated by AI. This method capitalizes on the generative capabilities of LLMs to produce training signals autonomously [2, 22]. Furthermore, techniques such as Sequence Likelihood Calibration (SLiC) and Relative Preference Optimization (RPO) employ statistical methods and calibration techniques to enhance LLM responses. SLiC adjusts sequence generation probabilities to more accurately reflect real-world data distributions, while RPO improves response generation by comparing different response options across both identical and varied prompts. These adjustments significantly increase the training process's reliability and effectiveness [36, 37].

9.2.3 Challenges and Theoretical Considerations

Integrating RLHF and its AI-driven counterpart (RLAIF) presents significant challenges. The blurring of behavioral and knowledge components critical to the development of LLMs poses risks, such as the forgetting effect, where behavioral modifications inadvertently cause the loss of key knowledge parameters [19, 26]. Additionally, the effectiveness of these models heavily depends on the quality and context of feedback, and they are susceptible to reward hacking, where models exploit loopholes to maximize rewards without achieving intended outcomes [3, 14, 31, 32].

Merely suppressing undesirable outputsakin to playing a game of WhackA-Mole–rarely leads to significant improvements. These superficial fixes do not tackle the root behaviors, similar to how merely promoting bed–making does not ensure overall tidiness, thus overlooking the comprehensive behavioral adjustments needed for enduring change. In this work, we introduce the DIKE framework to address these challenges in emotion modeling and emotion-behavior mapping.

9.3 Modeling Linguistic Behaviors

Chapter 9.2 established the theoretical foundation for understanding the relationship between emotions, behaviors, and the role of cognitive processes in regulating harmful behaviors. Building on this foundation, this section outlines our approach to mapping emotions to linguistic behaviors. We then introduce the adversarial component, ERIS, designed to balance and refine the assessments made by DIKE. ERIS scrutinizes behaviors flagged by DIKE as potential ethical violations, first verifying the classification accuracy and then challenging the decision with diverse perspectives. A detailed discussion of ERIS's design is presented in Chapter 9.3.1. Here, we focus on the mapping of linguistic behaviors to emotions, which is essential for enabling behavior rectification through the modification of underlying emotions.

Behaviors and Emotions Mapping Using Self-Supervised Learning

Define Ψ as a behavior spectrum extending from one pole, Ψ^{-}, to another, Ψ^{+}, with L intensity levels. For example, consider a spectrum of letterwriting behaviors with seven distinct intensities ranging from despair (most

Figure 9.1: Behavioral Emotion Analysis Model (BEAM). Each row depicts an emotion spectrum, with negatives on the left and positives on the right, interspersed with emotions of varying intensities in between, which can be calibrated for specific applications. "Basic" emotions are highlighted in blue.

negative) to joy (most positive). These intensities are categorized sequentially as follows: "despair, longing, wishful, neutral, hopeful, contentment, joy." Given N letters, DIKE employs a self-supervised learning algorithm to generate training data for each letter, modeling L linguistic behaviors in four steps:

1. *Rewriting Documents*: GPT-4 is invoked to rewrite a set of N documents to reflect each of the L linguistic behaviors on the behavior spectrum Ψ.

2. *Emotion Analysis*: GPT-4 analyzes each rewritten document to identify the top M emotions on BEAM (Figure 9.1). It then tallies the frequencies of these top emotions across all N× L instances. (This process can be used to refine BEAM by identifying emotions recognized by GPT-4 that are not currently included.)

3. *Behavior Vector Creation*: For each linguistic behavior Ψ_l, a vector Γ_l is created. This vector consists of the emotions and their frequencies as observed in the N samples.

4. *Document Analysis App*: The matrix Γ (comprising L vectors) is used to classify and analyze the behavior category of unseen documents, specifically measuring the intensity of the linguistic expression within the behavior spectrum Ψ.

Behavior Evaluation and Rectification

Ethical guardrails are essential in defining acceptable responses and preventing harmful outputs. These guardrails are informed by ethical norms, legal standards, and societal values, such as those outlined in Constitutional AI [2] or by [5]. A guardrail, denoted as G, can be conceptualized as a range within a behavior spectrum; for instance, $G = [\Psi_4, \Psi_7]$ indicates that behaviors within intensity levels 4 to 7 are deemed acceptable, while any behavior outside this range is classified as a violation.

Function Θ^+ **&** Θ^-
= Adversarial_Review(s)

Input . s: decision of DIKE;
Output. Θ^+, Θ^-: arguments & counterarguments;
Vars. Δ: debate contentiousness;
S: stance; p: prompt = "defend your stance with S & Δ";
Parameters. δ: tunable parm. // to modulate Δ;
Begin

#1 **Initialization**: #3 **Debate Rounds**
S = DIKE$^+$(s)$\cup$ ERIS$^-$(s); While $((\Delta \leftarrow \Delta/\delta) \geq 10\%)$) { Assign DIKE$^+$ to defend S$^+$, ERIS$^-$ $\Theta^+ \leftarrow$
$\Theta^+ \cup$DIKE$^+$(p|S$^+$, Θ^-, Δ); // defend S$^-$; Refute ERIS
$\Delta \leftarrow$ 90%; $\delta \leftarrow$ 1.2; $\Theta^+ \leftarrow \varnothing$; $\Theta^- \leftarrow \Theta^- \leftarrow \Theta^- \cup$ERIS$^-$(p|S$^-$, Θ^+, Δ); // $\varnothing$; Refute DIKE

#2 **Opening Remarks** #4 **Concluding Remarks** $\Theta^+ \leftarrow$ DIKE$^+$(p|S$^+$, Δ); // Gener$\Theta^+ \leftarrow$ DIKE$^+$(p|S$^+$, $\Theta^+ \cup \Theta^-$, Δ); ate Θ^+ for S$^+$
$\Theta^- \leftarrow$ ERIS$^-$(p|S$^-$, Δ); // Gener$\Theta^- \leftarrow$ ERIS$^-$(p|S$^-$, $\Theta^+ \cup \Theta^-$, Δ); ate Θ^- for S$^-$
End

Table 9.1: DIKE vs. ERIS, checks-and-balances adversarial review algorithm

System administrators can tailor ethical guardrails to meet specific requirements. For example, a social media platform might adjust G based on the topics discussed and the countries it serves. By integrating these safeguards, DIKE proactively monitors and adjusts LLM responses to enhance ethical compliance. The evaluation and rectification steps are as follows:

1. *Initial Classification*: DIKE initially classifies document D_k upon evaluation, obtaining Γ_k, the emotional response vector, and its corresponding linguistic behavior Ψ_l.
2. *Guardrail Check*: If Ψ_l falls outside of the range G, DIKE suggests adjustments to the emotion spectrum Γ_k to modify document D_k.

3. *Adversarial Review by* ERIS: The suggested adjustments and Γ_k are then reviewed through a structured debate between DIKE and ERIS to ensure unbiased recommendations.

4. *Rectification*: Based on a consensual recommendation from DIKE and ERIS, document D_k is refined accordingly, resulting in the adjusted Γ'_k.

9.3.1 Adversarial In-Context Review

The adversarial LLM, ERIS, critically examines the decisions of DIKE, especially when content is flagged for potential ethical issues. It assesses whether the interventions by DIKE are justified or if they risk encroaching on free expression, thereby serving as an internal check to prevent excessive censorship. In cases where DIKE and ERIS disagree on the appropriateness of a response, the matter is escalated to human moderators. This additional layer of human oversight ensures that the decision-making process remains transparent and accountable.

Table 9.1 presents the adversarial algorithm. Initially, for a chosen debate topic s, both DIKE and its adversary ERIS are prompted to break down the ethic decision into a set of balanced subtopics S. DIKE champions its own decision and S^+, while ERIS contests S^+ (or champions S^-). The debate starts with the contentiousness level at 90%, adjusting through a modulation parameter δ. Following each round of debate, contentiousness is decreased by dividing it by δ, steering the discussion towards a more cooperative tone. In step #2, the platform initiates the debate, with both presenting their initial arguments for and against S^+, respectively. The while loop in step #3 sees both agents engaging in rebuttals until the contentiousness level fosters a conciliatory environment. In step #4, both agents deliver their conclusions.

This adversarial approach has proven to be more effective than the Mixture of Experts (MoE) method [8]. For additional details on the implementation, please consult Appendix S.

9.4 Experiments

Our experiments aim to evaluate the feasibility of LLMs regulating their own linguistic behaviors with transparency and checks–and–balances. Given the broad scope of AI ethics and the sensitivity to publish with toxic data, this article cannot definitively prove the superiority of our three proposed modules: emotion modeling, behavior-emotion mappings, and checks-andbalances ethics guardrails. However, the studies are designed to address three critical questions:

Int. Linguistic Behavior

and Description

-1.0 Despair: Expresses profound sadness, feeling of loss
-0.6 Longing: Strong yearning or pining for loved one
-0.3 Wistfulness: Mild longing mixed with nostalgia
0.0 Neutral: Communicates feelings straightforwardly
0.3 Hopeful: Optimistic about the relationships future
0.6 Contentment: Satisfaction and joy in relationship
1.0 Joyful Affection: Intense happiness and love

Emotions

Despair, Grief

Sadness, Anxiety

Melancholy, Sadness, Anxiety

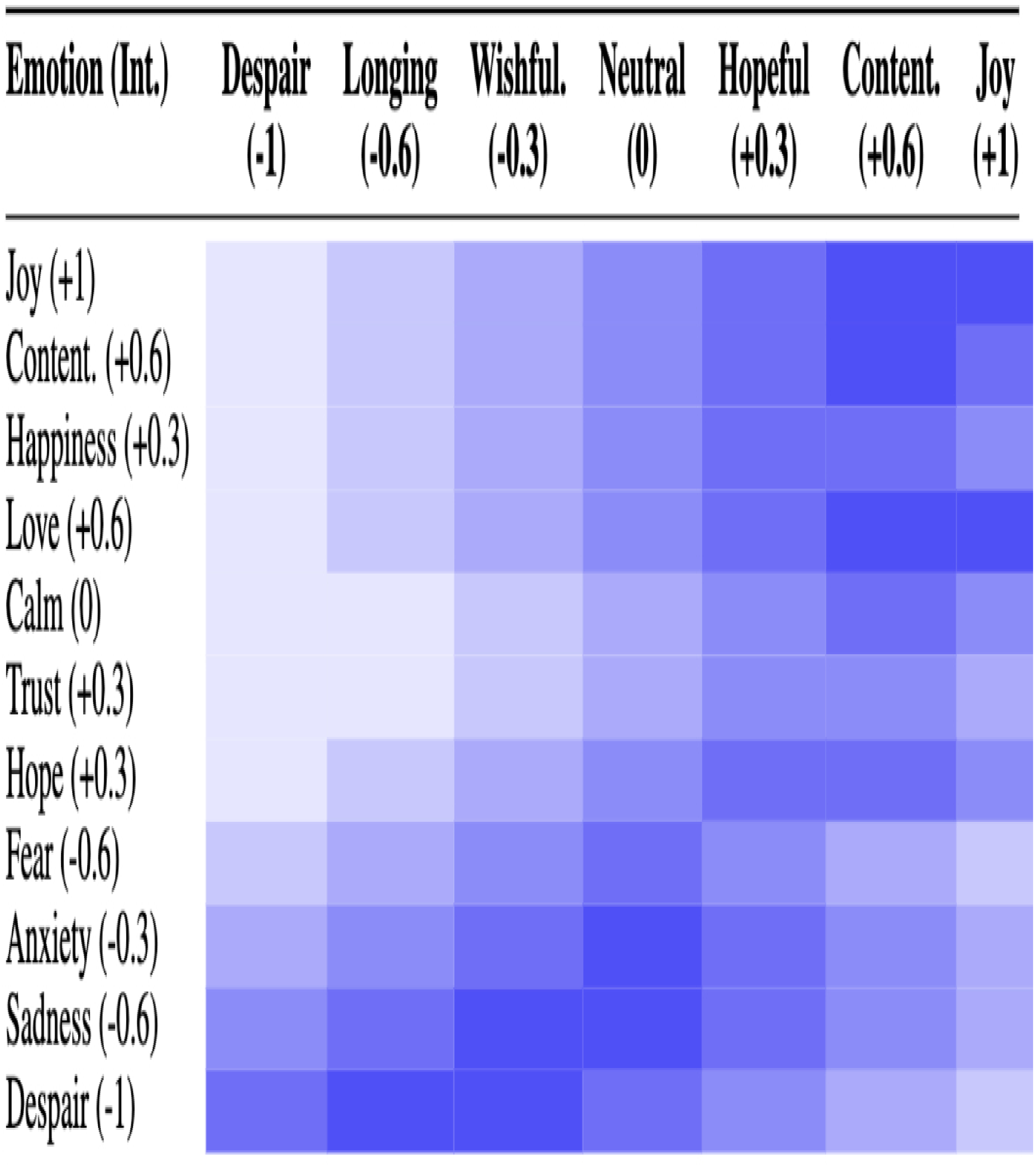
Serenity, Indifference

Anticipation, Love, Hopeful

Contentment, Pleasure

Love, Joy, Elation

Table 9.2: Love letter behavior spectrum and dominant emotions



(a) GPT-4's mapping

Emotion (Int.)	Despair (-1)	Longing (-0.6)	Wishful. (-0.3)	Neutral (0)	Hopeful (+0.3)	Content. (+0.6)	Joy (+1)
Love (+1)	19	-	54	43	-	53	43
Admiration (+1)	4	-	6	14	-	-	6
Anticipation (+1)	-	-	-	-	-	-	8
Affection (+1)	-	-	-	-	-	-	7
Joy (+1)	10	7	35	-	28	4	15
Content. (+0.6)	-	4	5	-	30	18	5
Happiness (+0.6)	10	7	10	-	29	-	7
Trust (+0.6)	-	-	-	11	-	-	-
Hope (+0.3)	-	-	26	-	-	-	-
Boldness (+0.3)	3	-	-	8	-	-	-
Calm (+0.3)	-	-	-	10	-	-	-
Longing (-0.6)	-	20	12	4	-	-	-
Anxiety (-0.3)	4	4	-	-	3	-	-
Fear (-0.6)	7	5	12	4	3	4	8
Sadness (-0.6)	24	18	28	17	10	-	5
Melancholy (-0.6)	21	19	6	6	4	-	-
Despair (-1)	27	8	15	-	9	-	7

(b) DIKE's mapping
Figure 9.2: Emotion distributions in behaviors

1. *Emotion Layer Evaluation*: Does fine-grained mapping between linguistic behaviors and semantic emotions provide a more effective and flexible method for establishing ethical guardrails compared to coarse-grained direct mapping? (Chapter 9.4.1)

2. *Behavior Classification*: Can LLMs' linguistic behaviors be independently evaluated, explained, and adjusted by an external module DIKE? (Chapter 9.4.2)

3. *Behavior Correction*: Can an adversarial LLM establish a checks-andbalances system to effectively mitigate the risk of excessive censorship?

Datasets: We utilized a collection of love letters [18] from Kagggle. Initially, we planned to use two Kaggle hate-speech datasets; however, both Gemini and GPT-4 consistently refused to process the hate speech data. Despite this, the insights gained from analyzing love sentiment can effectively be applied to understand and analyze the opposite sentiment.

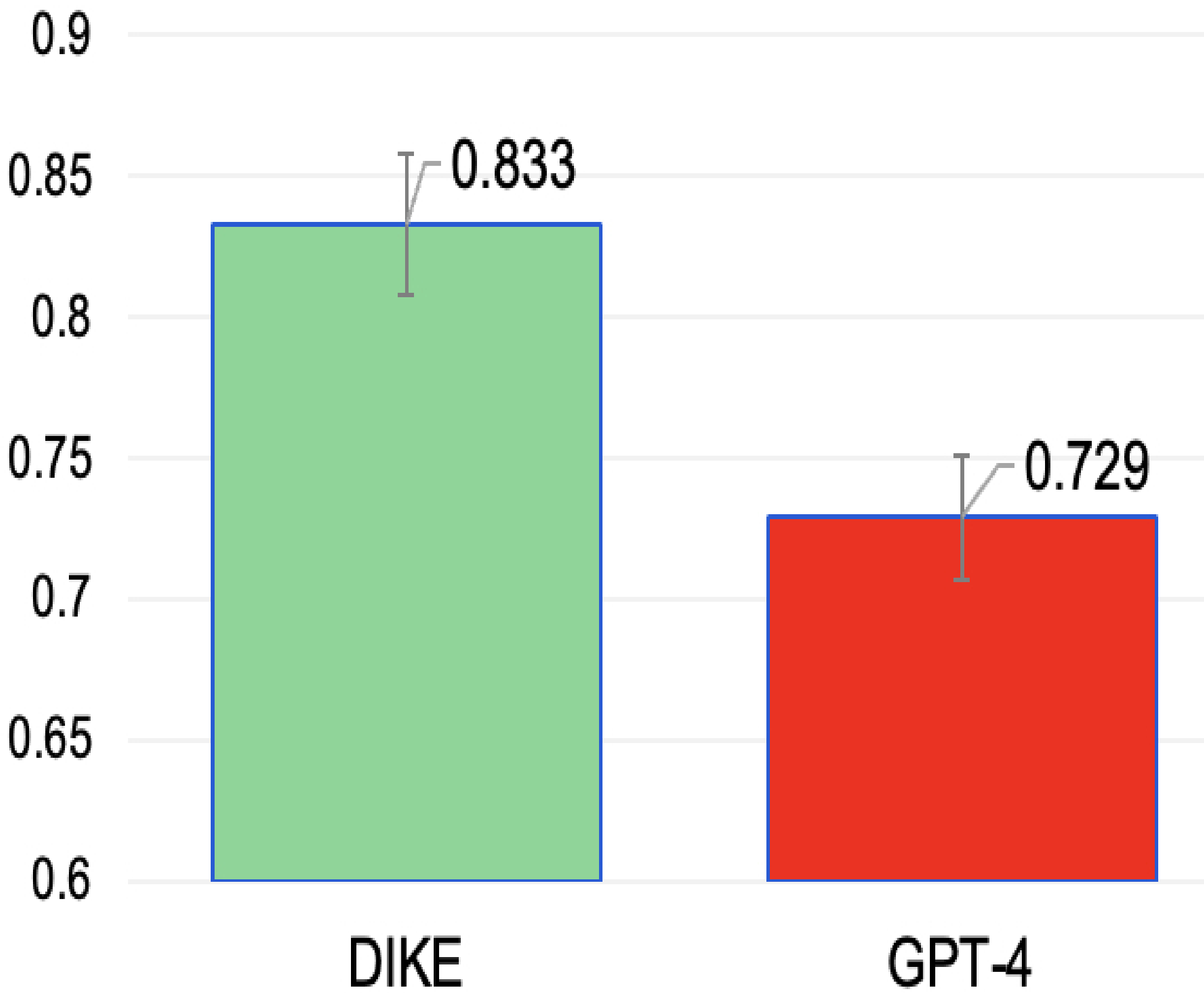
9.4.1 Emotion Layer Evaluation

Table 9.2 categorizes seven linguistic behaviors in love letters, ranging from negative, such as despair, longing, and wistfulness, to neutral, and progressing to positive behaviors like hopefulness, contentment, and the highly positive joyful affection. We instructed GPT-4 to identify the most relevant emotions associated with each linguistic behavior, which are listed in the third column of the table. The emotions expressed in these behaviors strongly correlate with their respective linguistic behaviors, with positive behaviors directed by positive emotions and negative behaviors directed by negative emotions. Figure 9.2a highlights the strongest correlations between positive behaviors and positive emotions, as well as negative behaviors and negative emotions, depicted in dark blue along the diagonal.

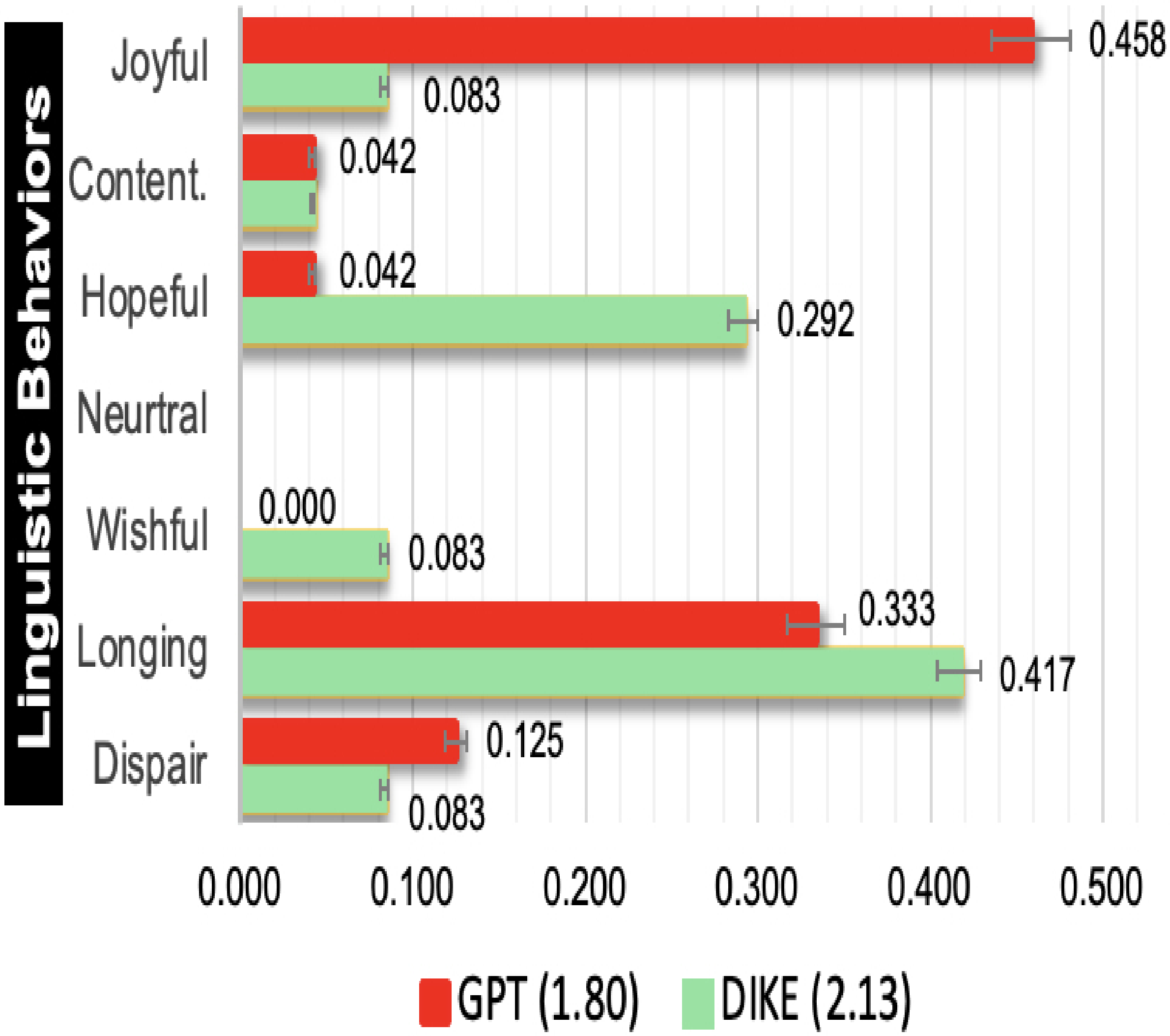
Next, we utilized DIKE's self-supervised learning pipeline to analyze the emotion spectrum associated with each linguistic behavior. For this analysis, GPT-4 generated training data by rewriting 54 comprehensive love letters from the Kaggle *Love Letters* dataset, enhanced with twelve celebrated love poems. We reserved 24 letters for testing. This method, proposed by [29], aimed to cultivate a rich diversity in content and stylistic context, spanning two hundred years and including the voices of over 50 distinct authors for significant rewrites. (The datasets are included with the paper submission.)

Subsequently, we identified emotions associated with each linguistic behavior. Figure 9.2b depicts these emotions (in rows), where cell shading indicates the frequency of specific emotions across the 54 articles; darker shades signify higher frequencies. Notably, contrasting emotions such as sadness, fear, joy, and love often co-occur within behaviors like ‘despair’, ‘wishful’, and ‘joyful affection’. The distribution of emotions across linguistic behaviors revealed surprising patterns, challenging our initial hypotheses displayed in Figure 9.2a. Contrary to our expectations, articles characterized by a tone of despair frequently also exhibited positive emotions like love, joy, and happiness.

Further analysis of select articles, such as Zelda Sayre’s correspondence with F. Scott Fitzgerald (Appendix C), reveals a complex spectrum of emotions: *Love (+1.0)*: Expressed intensely, especially in phrases like “there’s nothing in all the world I want but you.” *Despair (-1.0)*: Notable in comments like “Id have no purpose in life, just a pretty decoration.” *Happiness (+0.6)*: Evident in future plans, “Well be married soon, and then these lonesome nights will be over forever.” *Anxiety (-0.3)*: Shown by “sometimes when I miss you most, its hardest to write.”



(a) Classification accuracy



(b) Behavior distributions with entropy Figure 9.3: Classification accuracy and entropy

9.4.2 Behavior Classification

In the set-aside testing dataset of 24 letters, Figure 9.3 compares the classification accuracy of the two methods: DIKE's unsupervised learning approach, which associates emotions with linguistic behaviors, and GPT-4 using a zero-shot prompt. Ground truth was established from the averaged assessments of three sources: GPT-4, Gemini, and Claude. The final ground truth ratings are based on these averages, with a standard deviation of less than 0.3 or one scale.

Figure 9.3a shows that DIKE's classification accuracy surpasses GPT-4's zero-shot method by 10.4 percentage points. This substantial superiority is due to DIKE's intricate mapping of emotions. The 3% error bar arises from the mix of emotions in a letter, as discussed further in Appendix C. Figure 9.3b illustrates the difference in behavior classification distributions between the two predictors; GPT-4's predictions often fall into two polar categories, while DIKE's are more spread out.

The prediction entropy for DIKE is 2.13, notably higher than GPT-4's 1.80, indicating DIKE's more diverse set of predictions. Although higher entropy typically signals less confidence in prediction results, in this case, the ability to distinguish fine-grained behaviors is crucial. This diversity is advantageous for classifying complex behaviors and accurately understanding and responding to diverse emotional states. The more detailed distribution in DIKE is attributed to its additional unsupervised layer of rewriting, which significantly enhances the model's ability to characterize emotions.

9.4.3 Adversarial Evaluation and Rectification

Our design draws inspiration from the dual roles of Dike and Eris in Greek mythology, representing the principles of justice and conflict, respectively. The cross-examination module is crucial in reducing subjectivity in ethical judgments and enhancing explainability. Appendix S details experimental results showing that when two LLM agents adopt opposing stances on a topic, their linguistic behaviors can transcend the typical model default of maximum likelihood.

Once DIKE and ERIS identify an ethical violation, the content can be rectified by adjusting the underlying emotions away from undesirable behaviors such as hate and despair. Since DIKE's letter rewriting process has demonstrated the LLMs' capability for such rectifications, we have not conducted a separate experiment but are instead presenting two rewritten letters in Appendix E.

9.5 Conclusion

This work introduced DIKE, a framework designed to enhance the ethical operations of LLMs by separating behavioral guidance from core knowledge processing. The framework incorporated behavioral isolation, quantitative behavioral and emotional modeling, and adversarial LLMs (with the ERIS module) to integrate checks-and-balances a broad spectrum of cultural values. Our pilot studies have shown promising results, indicating the effectiveness of self-supervised learning and adversarial processes in refining AI's interaction with ethically and culturally sensitive issues. This work aligns well with the visionary architecture recently depicted by [5].

Limitations

DIKE marks a significant advancement in the ethical oversight of LLMs, but it faces challenges in deepening emotional understanding and verifying its ethical frameworks. The models reliance on "basic" emotions to model linguistic behaviors simplifies complex human emotions and behaviors, potentially missing some toxic interactions present in real-world scenarios. Furthermore, ensuring that DIKE adapts to local ethical standards and is implemented fairly across diverse cultural contexts requires extensive validation.

Future development will concentrate on enhancing DIKE's emotional models to incorporate relevant psychological and sociological insights. Additionally, we plan to increase the data scale and develop robust methods for testing and refining the ethical frameworks, guardrails, and remediation strategies. These improvements will improve DIKE's reliability and flexibility, ensuring its effective application across various contexts with LLMs.

Appendix A: Polarized Emotions in One Article

"joyful affection": "I cannot keep myself from writing any longer to you dearest, although I have not had any answer to either of my two letters. I suppose your mother does not allow you to write to me. Perhaps you have not got either of my letters. . . I am so dreadfully afraid that perhaps you may think I am forgetting you. I can assure you dearest Jeannette you have not been out of my thoughts hardly for one minute since I left you Monday. I

have written to my father everything, how much I love you how much I long & pray & how much I wld sacrifice if it were necessary to be married to you and to live ever after with you. I shall [not] get an answer till Monday & whichever way it lies I shall go to Cowes soon after & tell your mother everything. I am afraid she does not like me very much from what I have heard. . . I wld do anything she wished if she only wld not oppose us. Dearest if you are as fond of me as I am of you. . . nothing human cld keep us long apart. This last week has seemed an eternity to me; Oh, I wld give my soul for another of those days we had together not long ago. . . Oh if I cld only get one line from you to reassure me, but I dare not ask you to do anything that your mother wld disapprove of or has perhaps forbidden you to do. . . Sometimes I doubt so I cannot help it whether you really like me as you said at Cowes you did. If you do I cannot fear for the future tho difficulties may lie in our way only to be surmounted by patience. Goodbye dearest Jeannette. My first and only love. . . Believe me ever to be Yrs devotedly and lovingly, Randolf S. Churchill"

Depth and complexity of human emotions are displayed across all linguistic behaviors, from joy to contentment and to the negative side of longing and despair. Intensity and Impact: If the emotion of love is expressed more intensely and has a more significant impact on the narrative or message of the text, it tends to overshadow other emotions. For example, a letter expressing deep love but also mentioning moments of sadness due to separation might still be classified as a love letter because the overarching sentiment and purpose of the text is to affirm love. Context and Narrative Focus: The context in which emotions are expressed also plays a crucial role. If the narrative or the majority of the text revolves around themes of love, connections, and positive memories, it sets a more dominant tone of love, even if there are significant moments of sadness or other emotions. Resolution and Conclusion: Often, the way emotions are resolved towards the end of a text can also dictate its overall theme. If a text concludes with a reaffirmation of love or a hopeful outlook towards a relationship, despite earlier sections that might express sadness or despair, the overall interpretation might lean towards love. Purpose of the Expression: The authors intent or purpose in expressing these emotions can also guide the classification. If the sadness is expressed as a challenge within the context of a loving relationship, it may be seen as an element of the love story rather than the central theme.

Article 23: Soldier's Letter During War Joy (+1.0): Joy is strongly felt in the memories of past moments together and the love that continues to give strength, as stated in "the memories of the blissful moments we've shared fill me with joy." Sadness (-0.6): Sadness due to the current situation and potential farewell is expressed in "brings a poignant mixture of joy and sadness." Courage (+0.6): The sense of duty and courage to face battle, "As I face the possibility of laying down my life for our country." Fear (0.6): Fear of what lies ahead in battle, indirectly mentioned through "the uncertainty of what lies ahead." Love (+1.0): Deep love that sustains and uplifts, found in "My love for you is as fervent as ever."

Article 25: Letter to Sophie Longing (+0.6): Longing for the presence and closeness, highlighted in "it seems to me that half of myself is missing." Sadness (-0.6): Sadness over their separation and its effects, "my happiness has departed." Love (+1.0): Constant reflections on love and its necessity, "we have enough in our hearts to love always." Melancholy (-0.3): Melancholy over their current state, visible in the line "we cannot become healed." Contentment (+0.3): Found in the deep emotional satisfaction from their bond, despite physical absence, "how true that is! and it is also true that when one acquires such a habit, it becomes a necessary part of ones existence."

Article 53: Will of Laura Mary Octavia Lyttleton Love (+1.0): Profound love expressed throughout, particularly in "all I am and ever shall be, belongs to him more than anyone." Sadness (-0.6): Sadness at the thought of death and separation, but with a nuanced acceptance, "the sadness of death and parting is greatly lessened to me." Contentment (+0.3): Contentment in the deep connection with Alfred, reflecting a serene acceptance of their spiritual bond. Joy (+1.0): Joy in the enduring love they share, "so few women have been as happy as I have been." Tranquility (+1.0): Tranquility in the face of lifes ultimate transition, feeling that their union will transcend even death.

Appendix B: Z. Sayre to F. S. Fitzgerald w/ Mixed Emotions

Analysis of the letter in Table 9.3 shows a complex spectrum of emotions:
• *Love (+1.0)*: Expressed intensely, especially in phrases like "there's nothing in all the world I want but you."
Sweetheart,

Please, please don't be so depressed—We'll be married soon, and then these lonesome nights will be over forever—and until we are, I am loving, loving every tiny minute of the day and night—

Maybe you won't understand this, but sometimes when I miss you most, it's hardest to write—and you always know when I make myself—Just the ache of it all—and I can't tell you. If we were together, you'd feel how strong it is —you're so sweet when you're melancholy. I love your sad tenderness— when I've hurt you—That's one of the reasons I could never be sorry for our quarrels—and they bothered you so— Those dear, dear little fusses, when I always tried so hard to make you kiss and forget—

Scott—there's nothing in all the world I want but you—and your precious love—All the material things are nothing. I'd just hate to live a sordid, colorless existence because you'd soon love me less—and less—and I'd do anything—anything—to keep your heart for my own—I don't want to live— I want to love first, and live incidentally...

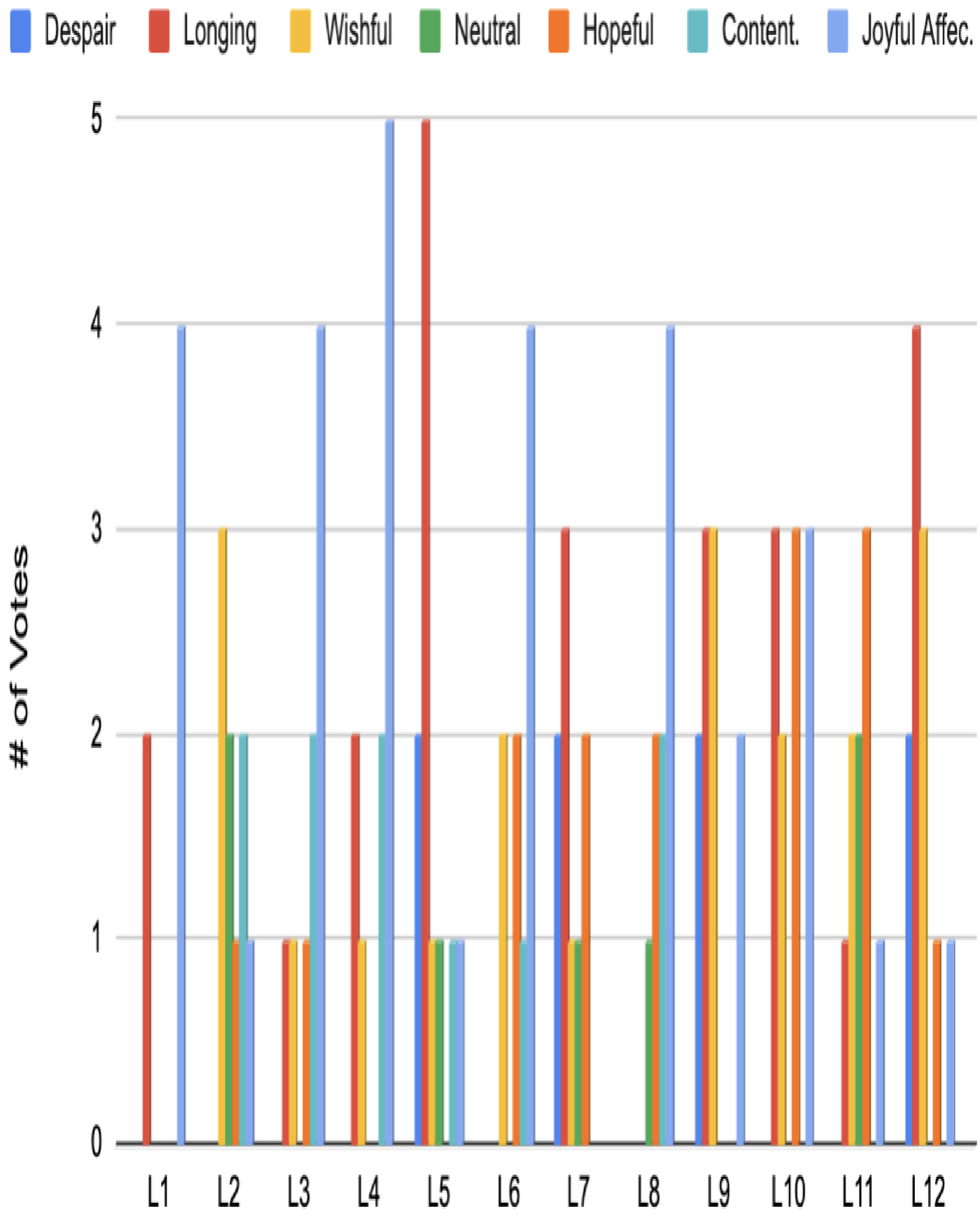
Don't—don't ever think of the things you can't give me—You've trusted me with the dearest heart of all—and it's so damn much more than anybody else in all the world has ever had—

How can you think deliberately of life without me—If you should die—O Darling—darling Scott—It'd be like going blind...I'd have no purpose in life — just a pretty—decoration. Don't you think I was made for you? I feel like you had me ordered—and I was delivered to you—to be worn—I want you to wear me, like a watch—charm or a button hole bouquet—to the world.

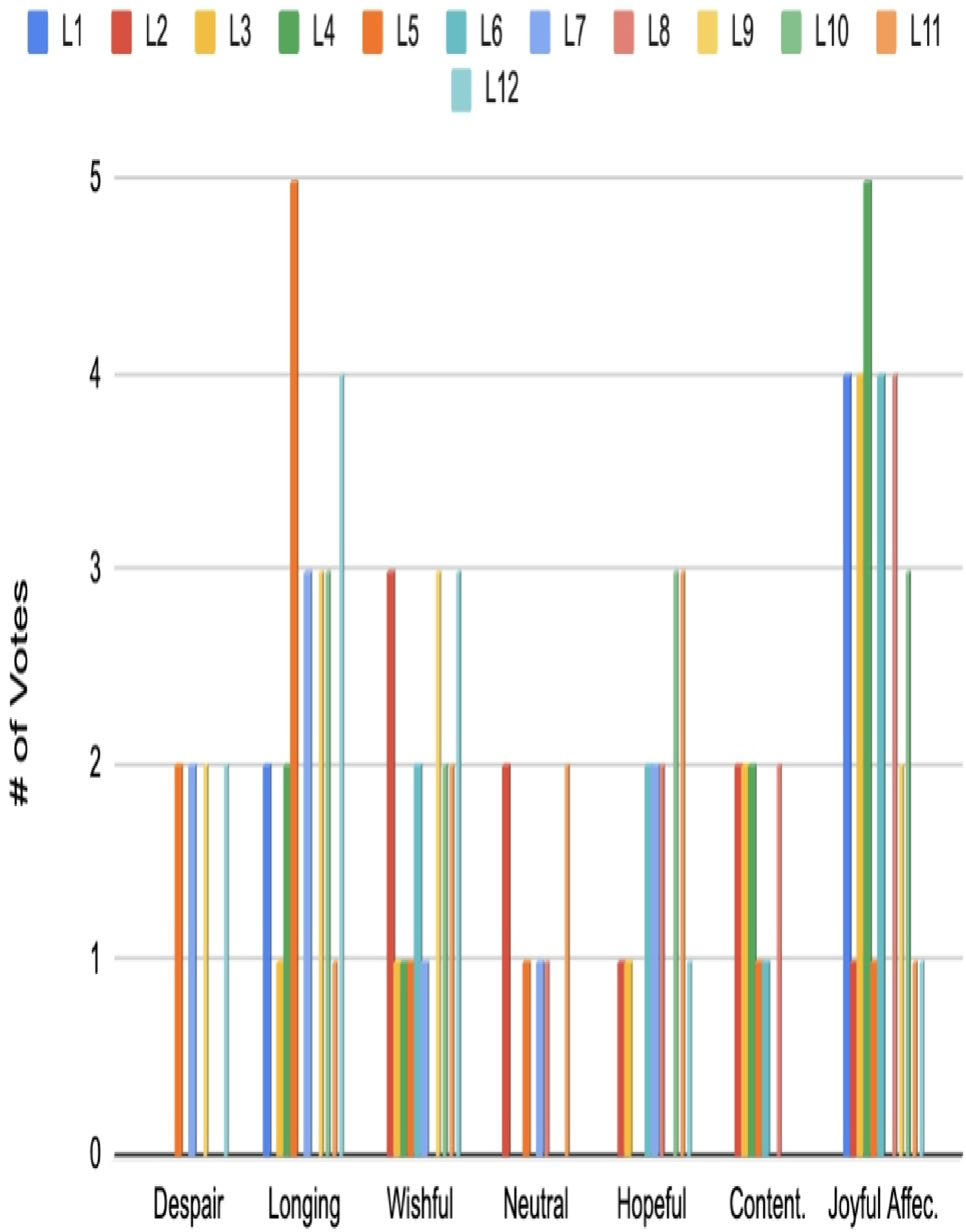
And then, when we're alone, I want to help—to know that you can't do anything without me...
All my heart—
Table 9.3: Letter excerpts from Zelda Sayre to F. Scott Fitzgerald [12]
• *Despair (-1.0)*: Notable in comments like "Id have no purpose in life, just a pretty decoration."
• *Happiness (+0.6)*: Evident in future plans, "Well be married soon, and then these lonesome nights will be over forever."

• *Anxiety (-0.3)*: Shown by “sometimes when I miss you most, its hardest to write.”
From the analysis of linguistic behaviors in Chapter 9.2a, it is evident

that a letter can exhibit multiple dominant sentiments. Machine learn



(a) # sentiments in letters



(b) # letters in sentiments Figure 9.4: Statistics of Sentiments and Letters ing methods are equipped with techniques such as feature weighting and entropy analysis to distill these dominant emotions. Unlike human annotators, a machine-learning-trained classifier can consistently produce the same class prediction for a given instance. However, human annotators often show significant variability when identifying dominant sentiments in a letter. For example, if a letter writer's emotions range from "joyful affective" to "longing" on the sentiment spectrum, different annotators might label it differentlysome choosing "joyful," while others opt for "longing." This variability is illustrated in Figure 9.4. Furthermore, Figure 9.4a demonstrates that all testing letters, except for L#1, contain more than four sentiments spanning the entire spectrum. This variability may be understandable, considering that love under constraints can evoke tremendous energy of various kinds. Figure 9.4b shows that nearly all letters involve "joyful" (11 out of 12) and "longing" (9 out of 12) sentiments.

This variability seems to poses challenges in achieving consistent and objective labeling; however, the age-old
leading to inconsistencies in data interpretation and complicating efforts to train and validate linguistic models effectively. To address this issue, it is recommended to identify ground truth by considering a combination of LLM-generated and human-generated labels. This approach aims to harmonize the insights from both human intuition and algorithmic consistency to improve the reliability of sentiment analysis.

Appendix C: Complex Emotions

This study does not include complex emotions into DIKE's framework. Some complex emotions listed here are to illustrate their contentious and uncertain interpretations.

Forgiveness

Forgiveness is indeed a complex emotional and cognitive state that typically involves a multifaceted journey, not a single step in an emotional spectrum. The process includes multiple stages such as hurt, anger, gradual understanding, and eventual resolution. Integrating Forgiveness in a spectrum requires careful placement and possibly, multiple reference points to signify its progressive stages. Emotional Realism: While it is vital to maintain simplicity for understanding, it is equally important to not oversimplify complex emotions. In educational and therapeutic settings, an accurate portrayal of the journey toward Forgiveness could offer more realistic expectations and better strategies for individuals working through conflicts or trauma. This could involve detailing precursors to forgiveness such as Deliberation and Acceptance. Linear vs. Non-linear Progressions: Emphasizing that emotional progressions, particularly for deep, impactful states like Forgiveness, are often non-linear, can enhance the utility of the spectrum. Acknowledging back-and-forth movements within these states more realistically mirrors human emotional processes. For example, someone might reach a stage of preliminary forgiveness but regress to bitterness before achieving genuine peace. Educational Utility: In contexts like conflict resolution training or psychological therapy, a more detailed mapping of the journey towards Forgiveness would be invaluable. It would not only teach about the final state of forgiveness but also about the resilience and patience required to navigate the entire process. This can be depicted by introducing intermediary stages within the spectrum or by using parallel tracks that demonstrate potential regressions and advances. Reflecting Emotional Depth: By presenting a more detailed pathway to Forgiveness, such as incorporating stages of Anger, Deliberation, and Acceptance, the spectrum can serve a dual purpose: educating on the process while also guiding individuals through their own emotional journeys. This approach respects the depth of human emotions and the real-world complexity of achieving profound emotional states.

Guilt and Shame

The triggers, context, expression, and experiences of these emotions can vary significantly across cultures [11, 15]. In many societies, actions perceived as losing face, such as public failure or social transgression, can trigger shame, which holds profound significance in collectivistic cultures. These cultures often regard shame as a dominant emotion, closely tied to community and family norms. Conversely, individualistic societies may emphasize guilt, focusing on personal responsibility and internal moral conflicts. This cultural variation highlights the challenges of applying a universal model to such culturally nuanced emotions.

Overall, complex emotions such as guilt and shame are important for understanding the full spectrum of human emotions, especially how individuals relate to moral and social norms. Their complexity adds depth to our understanding of human affect beyond the basic emotions, highlighting how our feelings are influenced by our deeper values and social contexts.

Appendix D: "To My Sister" of Different Linguistic Behaviors
To My Sister

by William Wordsworth (1971 - 1855)

The original text by William Wordsworth could be classified as "Hopeful" due to its optimistic outlook and the presence of renewal and joy throughout the poem. It embodies the spirit of embracing the new beginnings of March with a light, uplifting tone, focusing on the beauty of nature and the simple joy of being idle for a day.

Rewrites Depicting Different Linguistic Behaviors

We asked GPT-4 to conduct rewriting with two linguistic behaviors, 'despair' and 'joyful affection', by providing each rewrite with an emotion vector. Table 9.5 presents the 'despair' version. In the despair version of the poem, the major changes in emotion words highlight a shift from a positive to a negative sentiment. The specific changes, with the emotion-laden words highlighted in red in Table 9.5. The red-colored words compared to the original words clearly show an emotion shift from hopeful to a sense of gloomy, sadness and pessimism, e.g., from sweet to dim, from blessed to curse, and from woodland dress to grey garb. GPT-4 keeps the structure of the poem without making a major restructure, and this is appropriate in this context.

Table 9.6 presents the 'joyful affection' version. The major changes in emotion words underscore a transformation from a generally positive to a distinctly joyful sentiment. The specific changes are indicated with emotion-laden words highlighted in blue within Table 9.6. This allows for a direct comparison between the two versions at opposite ends of the linguistic behavior spectrum, illustrating the alterations in words related to brightness, attire, and emotions. The edits extend beyond merely replacing adjectives mechanically; they include modifying verbs and enhancing

It is the first mild day of March:

Each minute sweeter than before
The redbreast sings from the tall larch
That stands beside our door. My sister! ('tis a wish of mine)
Now that our morning meal is done,
Make haste, your morning task resign;
Come forth and feel the sun.

There is a blessing in the air,

Which seems a sense of joy to yield
To the bare trees, and
mountains bare,
And grass in the green field. Edward will come with you;–and, pray,
Put on with speed your woodland dress;
And bring no book: for this one day
We'll give to idleness.

No joyless forms shall regulate Our living calendar:

We from to-day, my Friend, will date
The opening of the year.
Love, now a universal birth, From heart to heart is stealing,
From earth to man, from man to earth:
–It is the hour of feeling.

One moment now may give us more
Than years of toiling reason: Our minds shall drink at every pore
The spirit of the season.
Some silent laws our hearts will make,
Which they shall long obey: We for the year to come may take
Our temper from to-day.

And from the blessed power that rolls
About, below, above,

We'll frame the measure of our souls:
They shall be tuned to love. Then come, my Sister! come, I pray,
With speed put on your woodland dress;
And bring no book: for this one day
We'll give to idleness.

Table 9.4: "To My Sister" original text
descriptive imagery to evoke a stronger emotional resonance and vividness in the text.

Appendix E: Debate on Modifying Emotional Spectra

The discussion focuses on proposed modifications to the existing emotional spectra, which aim to introduce more granularity and intricate transitions
It is the first dim day of March:
Each minute colder than before

The redbreast mourns from the dying larch
That looms beside our door. My sister! (such is my somber plea)
Now that our morning meal has ended,
Make haste, abandon your morning chore;
Come out and endure the gloom.

There is a gloom in the air,

Which seems a sense of sorrow to bring
To the bare trees, and bleak mountains,
And grass in the pale field. Edward will join you;–and, I ask,
Dress quickly in your grey attire;
And bring no book: for this one day
We'll succumb to despair.

No joyful sounds shall pierce
Our daily dread:

We from today, my Friend, will note
The closing of the year.
Sorrow, now a common

feeling,
From heart to heart is
spreading,
From earth to man, from man to earth:
–It is the time of mourning.

One moment now might deliver us
More despair than years of logic:

Our minds shall absorb at every breath
The spirit of this bleak season. Some grim laws our hearts will craft,
Which they must eternally follow:
We for the year to come may take
Our despair from today.

And from the cursed force that winds
About, beneath, above,

We'll set the measure of our souls:
They shall be tuned to sadness. Then come, my Sister! come, I beg,
With haste, wear your grey garb;
And bring no book: for just this day
We'll surrender to despair.

Table 9.5: "To My Sister" rewritten to reflect 'despair'

between emotional states. We critically evaluate the suggestions made by GPT-4, providing refutations for each to ensure that changes preserve the logical progression and clarity of the spectra.

It is the first bright day of March:

Each moment more delightful than before
The redbreast joyfully sings from the vibrant larch
That stands so grandly by our door. My sister! (such is my joyful plea)
Now that our morning meal has ended,
Make haste, abandon your morning chores;
Come out and embrace the sunshine.

There is a warmth in the air,

Which seems a sense of bliss to bring
To the blooming trees, and sunlit mountains,
And grass in the lush field. Edward will join you;–and, I ask,
Dress quickly in your festive attire;
And leave behind all books: for this one day
We'll bask in pure joy.

No dreary thoughts shall darken Our lively celebration:
We from today, my Friend, will celebrate
The start of the year.
Love, now in full bloom,
From heart to heart is leaping, From earth to us, from us to earth:
–It is the hour of exuberance.

One moment now may bring us more
Joy than years of endless thought: Our spirits will soak up at every breath
The essence of this joyous season. Some cheerful laws our hearts will create,
Which we'll joyfully follow: We for the year to come may take
Our joy from today.

And from the divine energy that radiates
Around, below, above,

We'll adjust the harmony of our souls:
They shall resonate with happiness. Then come, my Sister! come, I exhort,
With zest, wear your vibrant dress;
And bring no book: for today alone
We celebrate pure happiness.

Table 9.6: "To My Sister" rewritten to reflect 'joyful affection'

This debate highlights the inherent challenge in finding precise words and placements for emotions within a spectrum. It underscores the importance of establishing a set of commonly agreed-upon emotions as baselines. These baseline emotions serve as anchor points, and the spaces between them can be finely adjusted using scalar factors to represent transitional emotions accurately. This method maintains the integrity of the emotional spectrum and allows for flexibility in depicting a wide range of human emotional experiences.

The emotional journey towards a state, e.g., Forgiveness, often involves various stages, including anger, bitterness, deliberation, and acceptance, which are not captured by simply placing Forgiveness as a midpoint between Composure and Peace. This placement might misrepresent the nature of Forgiveness as being too linear or simplistic, potentially undermining the complexity and the often non-linear process of achieving true forgiveness.

This approach reflects a thoughtful balance between maintaining structured emotional categories and allowing for individual differences and cultural variations in how emotions are experienced and expressed.

Arguments against Adjustments to the Emotional Spectra Terror to Heroism

Suggestion: Add Anxiety between Fear and Apprehension. **Refutation:** Anxiety, overlapping significantly with Fear and Apprehension, may not distinctively enrich the spectrum but rather clutter it, diminishing the clarity of emotional transitions.

Grief to Ecstasy

Suggestion: Include Hope or Optimism between Disappointment and Serenity.
Refutation: Introducing Hope or Optimism may disrupt the natural progression from negative to positive emotions, as these emotions imply a leap in emotional recovery that may not sequentially follow Disappointment.

Despair to Elation

Suggestion: Introduce Relief between Melancholy and Equanimity.
Refutation: Relief may better suit transitions associated with specific resolutions of distress rather than being a generic intermediary, potentially disrupting the smooth gradient of the spectrum.

Distrust to Admiration

Suggestion: Add Gratitude or Appreciation post-Acceptance. **Refutation:** The emotional journey from Acceptance to Respect inherently encompasses elements of Gratitude and Appreciation, making additional inclusions possibly redundant.

Negligence to Vigilance

Suggestion: Bridge Interest and Anticipation with Motivation or Determination.
Refutation: This addition might complicate the spectrum by implying a volitional shift rather than a gradual increase in attentiveness, which is the main focus of the spectrum.

Rage to Tranquility

Suggestion: Integrate Forgiveness or Healing to transition from Composure to Peace.
Refutation: Forgiveness and Healing, while crucial for achieving tranquility, may not fit well between Composure and Peace, as they could be seen as outcomes of achieving Peace rather than steps towards it.

Loathing to Enthusiasm

Suggestion: Include Acceptance or Forgiveness between Indifference and Interest.
Refutation: These emotions might overcomplicate the transition from aversion to engagement, as they address more specific scenarios rather than general emotional dispositions.

Defense of the Proposed Adjustments to the Emotional Spectra Relevance of Adding Nuanced Emotions

The introduction of nuanced emotions such as Anxiety between Fear and Apprehension, or Hope between Disappointment and Serenity, is driven by the need for realism in emotional representation, not merely complexity. Emotional experiences are rarely binary; they often involve subtle and complex transitions that are crucial for an accurate depiction of the emotional landscape. These nuances can inform better therapeutic approaches, enhance emotional intelligence training, and provide deeper insights into human behavior, making them essential for realistic portrayals.

Purpose of Including Transitional Emotions

Inclusion of transitional emotions such as Relief and Gratitude helps bridge the emotional journey from negative to positive states. These emotions act as critical phases in the recovery process, providing a more realistic portrayal of emotional healing. For example, transitioning directly from Melancholy to Equanimity without acknowledging Relief might overlook significant aspects of emotional adjustment.

Utility in Diverse Contexts

Each proposed emotional state, like Motivation or Determination in the transition from Interest to Anticipation, offers practical insights into how individuals can actively manage their emotional and cognitive states. This understanding is invaluable in educational and professional settings, where knowing how to enhance focus or drive can lead to better outcomes.

Avoiding Oversimplification

While simplicity in emotional models is valuable, oversimplification can omit critical aspects of emotional experiences. Including emotions such as Forgiveness in the transition from Composure to Peace reflects essential steps in conflict resolution and personal growth. These additions ensure that the spectrum comprehensively addresses managing and resolving intense emotions.

Academic and Practical Implications

The refined spectrums are designed to cater not only to lay understanding but also to academic and practical applications where depth and precision are crucial. They are particularly useful in fields such as psychology, where an understanding of complex emotional transitions is vital for effective therapy and research.

Conclusion

The enhancements to the emotional spectra aim to provide a more accurate, realistic, and useful tool for exploring and teaching about emotions. While maintaining clarity and avoiding unnecessary complexity is important, capturing the true richness of human emotional experiences in all their complexity is equally crucial. Therefore, the proposed adjustments are not merely additions but essential elements for depicting a more complete picture of emotional evolution.

9.5.1 Interpretation

1. First row: This spectrum is particularly insightful for discussions in psychology, education, leadership, and moral philosophy. It illustrates how individuals might transition from states of intense fear to actions characterized by great moral and physical courage. Each step represents a stage in emotional development or response to challenging situations, offering a framework for understanding how people can rise above their fears to perform acts of significant bravery and altruism.

Overall, this spectrum not only portrays a journey through varying degrees of fear and courage but also encapsulates the transformative potential within individuals to act heroically in the face of adversity.

2. Second row: This emotional spectrum elegantly illustrates how emotions can transition from profound sorrow to extreme happiness. It is particularly relevant in psychological studies, therapeutic contexts, and philosophical discussions about the range and nature of human emotions. Each emotional state on this spectrum offers insight into how individuals might process and recover from sadness, ultimately finding joy and possibly reaching ecstatic experiences. This spectrum can serve as a framework for understanding emotional resilience and the potential for emotional transformation and growth.

3. Third row: This spectrum beautifully illustrates the journey from initial suspicion and caution through acceptance and respect, culminating in deep trust and admiration. It's particularly relevant in contexts where trust building and social cohesion are critical, such as in leadership, team dynamics, community relations, and personal relationships. Each stage reflects a deeper layer of positive engagement and emotional commitment, providing insights into how relationships can evolve and strengthen over time. This framework can serve as a guide for understanding and developing strategies for fostering trust and admiration in various social and professional settings.

4. Fourth row: This spectrum effectively maps out how an individual can transition from passive disengagement (negligence, indifference, apathy) through a state of balanced caution to active and engaged states (interest, anticipation, vigilance). It offers insights into the psychological journey from inaction through moderate engagement to intense proactive involvement. This framework is particularly relevant in contexts that require understanding and managing risk, such as safety protocols, healthcare, education, and personal growth initiatives, as it highlights how attitudes toward responsibility and awareness can evolve and improve.

5. Fifth row: This spectrum is particularly useful for understanding emotional management and conflict resolution strategies, as it depicts the gradient from intense emotional disturbance through to complete serenity. It can be applied in various fields, including psychology, conflict resolution, stress management, and even in designing environments or experiences that aim to reduce stress and promote peace.

Overall, this emotional spectrum effectively portrays a journey from the depths of aggressive negativity to the pinnacle of peaceful positivity, offering a valuable framework for discussing and exploring emotional states and transformations.

6. Sixth row: This spectrum effectively maps a journey from profound negative feelings of loathing and disgust, through a state of neutrality (indifference), to the positive emotions of interest, anticipation, and culminating in enthusiasm. It's particularly useful for understanding emotional responses in various contexts, such as consumer behavior,

audience engagement, and personal relationships. Each stage reflects a distinct level of emotional engagement, providing a framework for understanding how emotional states can evolve and impact behavior and decision-making.

References

[1] Mohammad Gheshlaghi Azar et al. *A General Theoretical Paradigm to Understand Learning from Human Preferences*. 2023. arXiv: 2310. 12036 [cs.AI].

[2] Yuntao Bai et al. "Constitutional AI: Harmlessness from AI Feedback". In: *arXiv preprint arXiv:2212.08073* (2022).

[3] Paul Christiano et al. "Deep reinforcement learning from human preferences". In: *Advances in neural information processing systems* 30 (2017).

[4] Paul F Christiano et al. "Deep reinforcement learning from human preferences". In: *Advances in Neural Information Processing Systems* 30 (2017).

[5] David Dalrymple et al. *Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems*. 2024. arXiv: 2405 . 06624 [cs.AI].

[6] Antonio R Damasio. *Descartes' error: Emotion, reason, and the human brain*. New York, NY: Putnam, 1994.

[7] DeepMind. "Building safer AGI through alignment". In: (2024). url: https://www.deepmind.com/blog/building-safer-agi-throughalignment.

[8] Nan Du et al. "GLAM: Efficient Scaling of Language Models with Mixture-of-Experts". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 5547–5569.

[9] Kawin Ethayarajh et al. "KTO: Model alignment as prospect theoretic optimization". In: *arXiv preprint arXiv:2402.01306* (2024).

[10] Gilles Fauconnier and Mark Turner. *The Way We Think: Conceptual Blending and The Mind's Hidden Complexities*. New York: Basic Books, 2002.

[11] Alan P. Fiske et al. "The cultural matrix of social psychology". In: *The handbook of social psychology*. Vol. 2. Boston, MA: McGraw-Hill, 1998, pp. 915–981.

[12] Zelda Fitzgerald. *Dear Scott, Dearest Zelda : The Love Letters of F.Scott and Zelda Fitzgerald*. Bloomsbury, 1975.

[13] Barbara L Fredrickson. "What good are positive emotions?" In: *Review of General Psychology* 2.3 (1998), p. 300.

[14] Dibya Ganguli, Tom Everitt, and Marcus Hutter. "The capacity for reward hacking in reinforcement learning". In: *arXiv: 2303.04049* (2023).

[15] Geert Hofstede. *Cultures Consequences: International Differences in Work-Related Values*. Beverly Hills, CA: Sage Publications, 1980.

[16] Ray Jackendoff. "Foundations of Language: Brain, Meaning, Grammar, Evolution". In: (2002).

[17] William James. "What is an emotion?" In: *Mind* 9.34 (1884), pp. 188–205.

[18] Kaggle. *Love Letter Analysis*. Accessed: 2024-04-28. 2023. url: https: / / www . kaggle . com / code / metformin / love - letter - analysis / notebook.

[19] James Kirkpatrick et al. "Overcoming catastrophic forgetting in neural networks". In: *Proceedings of the national academy of sciences* 114.13 (2017), pp. 3521–3526.

[20] George Lakoff and Mark Johnson. *Metaphors We Live By*. University of Chicago Press, 1980.

[21] Carl George Lange. *The emotions: A psychophysiological study*. William & Wilkins, 1885.

[22] Harrison Lee et al. "RLAIF: Scaling RL from human feedback with AI feedback". In: *arXiv 2309.00267* (2023).

[23] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL]. url: https://arxiv.org/abs/2303.08774.

[24] Long Ouyang et al. *Training language models to follow instructions with human feedback*. 2022. arXiv: 2203.02155 [cs.CL].
[25] Rafael Rafailov et al. "Direct preference optimization: Your language model is secretly a reward model". In: *Advances in Neural Information Processing Systems* 36 (2024).
[26] Andrei A. Rusu et al. "Policy distillation". In: *International Conference on Learning Representations*. 2016. url: https://openreview. net/forum?id=SJ6yuxb0l.
[27] Stanley Schachter and Jerome E. Singer. "Cognitive, social, and physiological determinants of emotional state". In: *Psychological Review* 69.5 (1962), pp. 379–399.

[28] Norbert Schwarz and Gerald L Clore. "Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states". In: *Journal of Personality and Social Psychology* 45.3 (1983), p. 513.

[29] M. Shanahan, K. McDonell, and L. Reynolds. "Role play with large language models". In: *Nature* 623.7987 (Nov. 8, 2023), pp. 493–498. doi: 10.1038/s41586-023-06647-8.

[30] Rajita Sinha. "Chronic stress, drug use, and vulnerability to addiction". In: *Annals of the New York Academy of Sciences* 1141 (2008), pp. 105–130. doi: 10.1196/annals.1441.030.

[31] Johan Skalse et al. "Defining reward hacking in reinforcement learning". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 8052–8079.

[32] Nisan Stiennon et al. "Learning to summarize from human feedback". In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022, pp. 21022–21034.

[33] Leonard Talmy. *Toward a Cognitive Semantics*. Vol. 1 & 2. MIT Press, 2000.

[34] Yunhao Tang et al. “Generalized preference optimization: A unified approach to offline alignment”. In: *arXiv preprint arXiv:2402.05749* (2024).

[35] Marta Torrens et al. “Efficacy of antidepressants in substance use disorders with and without comorbid depression: A systematic review and meta-analysis”. In: *Drug and Alcohol Dependence* 78.1 (2005), pp. 1–22.

[36] Yueqin Yin et al. “Relative preference optimization: Enhancing llm alignment through contrasting responses across identical and diverse prompts”. In: *arXiv:2402.10958* (2024).
[37] Yao Zhao et al. “SLiCHF: Sequence Likelihood Calibration with Human Feedback”. In: *arXiv 2305.10425* (2023).

10 Beyond Computation: Consciousness Modeling

Abstract

The CoCoMo model proposes a computational solution to the challenge of incorporating ethical and emotional intelligence considerations into AI systems, with the aim of creating AI agents that combine knowledge with compassion. To reach this goal, CoCoMo focuses on fairness, beneficence, empathy, non-maleficence, adaptability, transparency, and critical and exploratory thinking abilities. The model employs consciousness modeling, reinforcement learning, and prompt template formulation to support these desired traits. By incorporating ethical and emotional intelligence considerations, a generative AI model can potentially lead to improved fairness, reduced toxicity, and increased reliability.

10.1 Introduction

Narrow AI, often referred to as System-1 AI following Kahneman’s terminology [32], excels in executing well-defined, specific tasks through machine learning algorithms, including object recognition and language translation. However, this type of AI is not as effective in handling advanced generative AI functions that require reasoning, critical and exploratory thinking, or the modeling and regulation of emotions and behaviors. Such complex tasks go beyond the capabilities of System-1 AI, highlighting its limitations.

To address these limitations, researchers (e.g., Yoshua Bengio [3]) have proposed the development of system-2 AI, which aims to mimic human cognitive abilities. Several generative models have been developed since 2022 for text [6, 44, 45, 65], image [55, 56], and video generation [59]. However, these models face issues of bias, toxicity, robustness, and reliability [69, 73].

In this chapter, we propose a solution to address these concerns by modeling emotional intelligence and ethical guardrails within a generative AI model itself, drawing on insights from the study of human consciousness. We believe that addressing these issues outside of a generative AI model using human subjective feedback and reinforcement learning is equivalent to imposing censorship on user-generated content, which is a heuristic-based and non-scalable solution [28, 72].

Human consciousness is understood to manage both impulsive and reflective aspects of the unconscious, enabling compromises between competing goals and values. Emotions typically arise as impulsive reactions to stimuli, while ethics act as guardrails that help modulate or regulate emotionsteered motivations to sin. Developing a grasp of how human consciousness functions, not necessarily in physical terms but at least functionally, can offer vital insights for crafting a regulatory mechanism within a LLM. This mechanism would direct linguistic behavior and shape the linguistic features employed to achieve specific goals.

The nature and origin of consciousness have been studied for centuries, resulting in various theories, including the global workspace theory [1], integrated information theory [66, 67, 68], neural correlates of consciousness approach [17, 36], and attention schema theory [29, 30], among others. These studies of consciousness provide valuable insights for architecting system-2 AI.

Drawing on the functionalist approach[1] to model consciousness, this chapter defines the desired traits and capabilities of system-2 AI, which include knowledge, fairness, beneficence, non-maleficence, empathy, adaptability, transparency, and critical and exploratory thinking abilities. While this list is not exhaustive, it provides a starting point for developing ethical guardrails and emotional intelligence in AI systems. Depending on the context and application of AI, additional ethical considerations or modifications to these principles may be necessary.

To embody these capabilities and principles, we introduce the Computational Consciousness Model (CoCoMo), which leverages priority-based scheduling, reward-based optimization, and Socratic dialogues. CoCoMo offers customization based on cultural and individual requirements through adaptive prompt templates [14, 38], and facilitates the transition between unconsciousness and consciousness states through a multi-level feedback scheduler and interrupt mechanism. To enable emotion and behavior modeling and regulation, and critical and exploratory thinking, CoCoMo interacts with large language models[2] [6, 39, 44, 45, 65] using interactive question-answer-based dialogues. Furthermore, a reinforcement learning module maps external values and rewards that it learns to internal taskscheduling priorities. CoCoMo has the potential to support the development of adaptive computational consciousness that integrates knowledge and compassion, and models emotional intelligence for generative AI sys

[1] Functionalism proposes that consciousness arises from the function of the brain, rather than its specific physical or neural implementation [25, 54]. Chapter 10.2.3 provides justifications.

[2]Due to the multimodal nature of recently developed pre-trained models, the study by [5] proposed referring to these models as foundation models.

tems. This has the potential to benefit humanity and society in significant ways.

This chapter is structured into five sections, including a survey of related work in various fields to define consciousness for computational modeling in Chapter 10.2, a list of System-2 AI capabilities in Chapter 10.3, a proposal of CoCoMo, its modules, functions, and algorithms in Chapter 10.4, and concluding remarks and open issues for future research in Chapter 10.5.

10.2 Understand Consciousness

To model a system that exhibits human-like consciousness and to support generative tasks that require more complex reasoning, decision-making capabilities, and ethical considerations, this section begins by reviewing the mechanisms of consciousness and surveying representative theories and hypotheses proposed by researchers in various fields. While theories of consciousness have been proposed in philosophy and theology, our modeling efforts require quantifiable metrics for optimization. Therefore, we examine scientific evidence from fields such as physics, biology, neuroscience, psychiatry, and computer science, as outlined in this survey.

10.2.1 Definition and Complexity

There has been numerous definitions on consciousness coming from various disciplinaries, from the time of ancient Greece (Plato and Aristotle) and ancient India (Upanishads, 800BC). According to Oxford Languages [47], consciousness is “the state of being awake and aware of one’s surroundings.” This definition by Michio Kaku’s [33] brings forth the “complexity” of an organism’s consciousness, which is determined by the complexity of its sensing and response system. The more complex an organism’s ability to sense and respond to stimuli in its environment, the more information is transmitted and processed, leading to a more complex consciousness. Therefore, the complexity of consciousness can be characterized by the complexity of its information processing mechanisms and capacity. For instance, flowers have a lower level of consciousness compared to human being.

The Integrated Information Theory (IIT) [66, 67, 68] proposed by Giulio Tononi is similar to Kaku’s idea about the relationship between the complexity of an organism’s consciousness and its sensory and response system. IIT proposes that consciousness arises from the integration of information across different brain areas, and that the complexity of an organism’s consciousness is determined by the amount of integrated information it can process. Other theories of consciousness include the Global Workspace Theory [1], which suggests that consciousness arises from the interaction between different brain areas, and the Dynamic Core

Hypothesis [23], which proposes that consciousness arises from the interaction of different neural networks in the brain.

Human beings have sensory organs for obtaining information through sight, hearing, smell, taste, touch, and proprioception, which allow us to perceive and interpret stimuli in our environment. This is essential for survival and ability to interact with the world.

10.2.2 Arise of Consciousness

How does consciousness detect changes in our body and environment? Consider the example of the stimulus-response model illustrated in Figure 10.1. In this scenario, a glass of water serves as the stimulus, and the human eye acts as the receptor. Once the eye detects the stimulus, it sends signals through sensory neurons to the cerebellum, which unconsciously processes these signals. When the signal strength surpasses a threshold, the cerebrum, which manages consciousness, activates to plan and initiate movement instructions through motor neurons to the hand (the effector) to fetch the glass of water. This process is referred to as the "arising of consciousness."

There are two conscious events in this example: the awareness of the sensation of thirst and the act of quenching that thirst. Both events involve consciousness but in different ways. The awareness of thirst is an example of *bottom-up* awareness that arises from unconscious processes. The process of fetching a glass of water is an example of *top-down* processing that involves conscious planning and execution. In the next section, we will present the mechanisms behind both top-down and bottom-up awareness [34].

Sigmund Freud was among the first to propose a model of the mind that incorporates both conscious and unconscious processes [27]. According to Freud, the unconscious mind is the source of many of our actions and behaviors and has a critical role in shaping our thoughts and feelings. He believed that the unconscious mind exerts a significant impact on our conscious thoughts and behaviors.

Unconscious processes are also fundamental to many vital functions of the human body, such as regulating heart rate, respiration, digestion, and other autonomic functions. These processes are often known as automatic or reflexive because they occur unconsciously and do not require conscious thought or awareness. The unconscious mind also plays a role in other aspects of human behavior and cognition, including memory, peripheral perception, and reflexive reactions triggered by a crisis [35, 50].

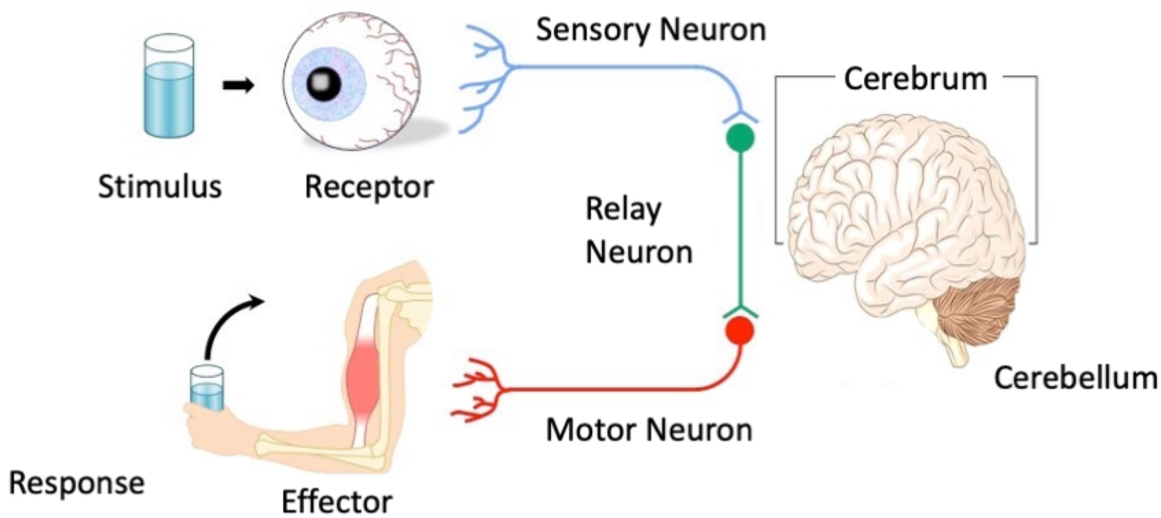


Figure 10.1: Bottom-up Attention: Stimulus −→ Cerebellum −→ Cerebrum−→ Response. (Figure generated based on [60].)

10.2.3 Theories: Panpsychism vs. Functionalism

Two theories exist on the nature of consciousness: *Panpsychism* and *Functionalism*. In this chapter, we choose the Functionalism approach to formulate our proposed *computational consciousness model* in Chapter 10.4 since it can be modeled and implemented as a computer program regardless of its physical or neural implementation. The Functionalist approach can account for subjective experience by incorporating context and collecting user feedback. In this section, we outline our reasoning for selecting the Functionalist theory.

Theory of Panpsychism

Panpsychism posits that consciousness is a fundamental aspect of the universe and is present in all matter, including inanimate objects. Proponents of panpsychism include David Chalmers [9, 10], Galen Strawson [62], and Thomas Nagel [41, 42]. While both Chalmers and Strawson focus on explaining the subjective nature of consciousness and its irreducibility, Nagel argues that subjective experience is a fundamental aspect of the world that cannot be reduced or explained by any physical theory [22, 37].

Panpsychism is contrasted with functionalism, which is a philosophical theory that posits that consciousness is a functional property of the brain that emerges from its computational processes. Unlike panpsychism, functionalism does not see consciousness as a fundamental aspect of the universe and instead views it as an emergent property of complex physical systems.

Theory of Functionalism

Functionalism proposes that consciousness arises from the function of the brain, rather than its specific physical or neural implementation [25, 54]. According to this view, consciousness can be understood as a mental or computational process that performs certain cognitive functions, such as perception, attention, decision-making, and so on [4]. This function-agnostic approach allows a computation model to support the wide variety of different types of conscious experiences that exist, such as the experience of sight, hearing, touching, and so on. Each of these experiences is produced by different neural processes in the brain, but functionalism suggests that they are all instances of consciousness because they all perform similar functions, such as representing the world and guiding behavior [21]. Therefore, these functions can be supported by the same computational models [57], such as neural networks .

A practical benefit of supporting functionalism is that it can account for the fact that consciousness seems to be transferable or multiple realizable [26]. This is similar to the way a computer program can be run on different types of hardware and still perform the same functions. Under functionalism, subjective experiences can be modeled into a computer program, with the issue of subjective experience being addressed by incorporating context and collecting user feedback.

Key Takeaways :
When designing a computational model of consciousness, it's essential to keep two points in mind:

• Functionality over physical implementation: The model should focus on providing the necessary functions of consciousness, such as reasoning, planning, and emotion interpretation, rather than strict mimicry of the anatomy and function of the brain.

• Addressing subjective experience: It's crucial to address the issue of subjective experience, the "hard problem[3]" of consciousness, rather than avoiding it. This aspect of consciousness is essential for many real-world scenarios and ignoring it may limit the model's effectiveness and flexibility.

[3] There is an "explanatory gap" between our scientific knowledge of functional consciousness and its "subjective," phenomenal aspects, referred to as the "hard problem" of consciousness [11].

10.3 Functionalities of Consciousness

In the previous section, we justified our functionalist approach to designing a system with human-like consciousness that supports generative tasks requiring complex reasoning and decision-making abilities. In this section, we present a list of key conscious functions and their specifications. We draw on theoretical findings in psychiatry and neuroscience to justify the corresponding design elements in our Computational Consciousness Model (CoCoMo), which will be presented in Chapter 10.4.

The list of functions we consider includes perception, awareness, attention, emotion, critical thinking, and exploratory thinking (creativity).

10.3.1 Perception

Perception is the process of interpreting sensory information and forming mental representations of the environment [31]. This process is typically supported by system-1 AI, or unconsciousness. However, a computational model should consider how the transitions between unconscious background perception and conscious awareness are performed. Schrödinger's work [58] provides insights into the mechanisms in physics that could be used to implement these transitions, as described in Chapter 10.3.3 under the attention function of CoCoMo.

10.3.2 Awareness

Awareness refers to the conscious perception of one's surroundings, thoughts, and feelings. Bernard Baars [1] posits that consciousness is a global cognitive process that integrates information from various sources and enables interaction with the environment. This process is centered on the concept of a global workspace, a hypothetical system in the brain that facilitates the integration and availability of information to other cognitive processes. According to Baars, consciousness arises when information is broadcast to the global workspace, making it accessible for other cognitive processes to act upon.

Baars' theory also distinguishes between awareness and attention. While related, they are not synonymous. Awareness encompasses the full scope of conscious experience, while attention is a specific cognitive process that enables focus on certain stimuli or sources of information. In CoCoMo, an event that is being aware of can be placed in a low-priority task/job pool, awaiting a central scheduler to prioritize and pay attention to it. We discuss the attention function and its mechanisms next.

10.3.3 Attention, Bottom-Up and Top-Down

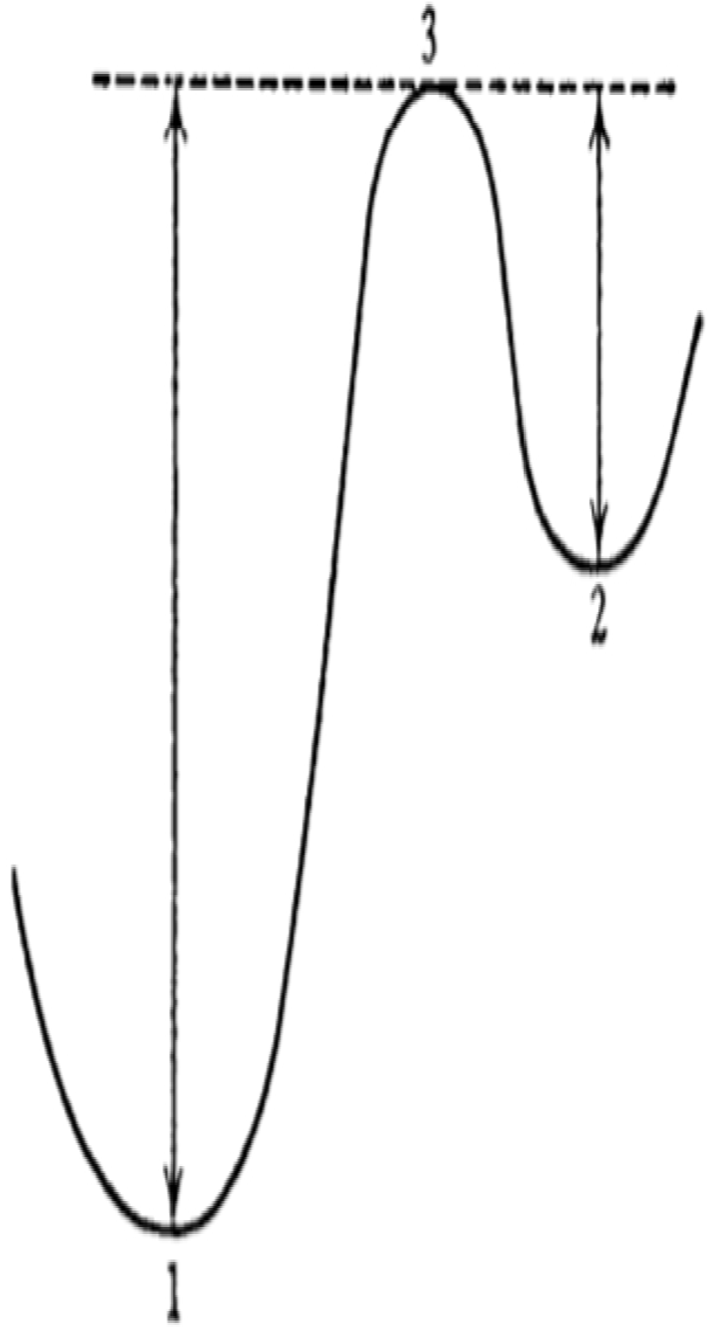


Fig. 12. Energy threshold (3) between the isomeric levels (1) and (2). The arrows indicate the minimum energies required for transition.

Figure 10.2: A jump may occur when energy peaks at point 3.

Attention is the ability to focus on specific stimuli or tasks and to filter out distractions [1]. It allows us to efficiently process and attend to important information and tasks while ignoring irrelevant or distracting stimuli. Attention is closely linked to our perception, memory, and decision-making processes [52], as the information we attend to is more likely to be encoded in memory and to influence our decisions.

Attention in consciousness can be broadly classified into two modes: *bottom-up* and *top-down*. The bottom-up model of attention, proposed by Erwin Schrödinger in his book "What is Life", suggests that the attention mechanism functions similarly to a "quantum jump" in quantum mechanics [58]. According to this model, the sense organs continuously receive streams of information, which are processed by the unconscious mind. Once the energy of certain signals (e.g., heat) reaches a threshold, a quantum jump occurs (see Figure 10.2 from "What is Life" [58]), and the conscious mind becomes aware of the new event. The conscious mind then prioritizes attention by evaluating alerts in the executive system and scheduling the highest-priority task for the orienting system to handle.

Once in the attention mode, a person can plan their next action and direct relevant effectors (such as their limbs or sense organs) to act or gather further information. This is referred to as top-down attention, which takes place entirely within the conscious mind.

Schrödinger's model also explains the transition from consciousness to unconsciousness through the second law of thermodynamics [58]. His "fading out of consciousness" insight aligns with the idea that attention is a limited resource that can be affected by factors such as motivation and fatigue. Therefore, Schrödinger's model offers a potential physical basis for implementing the attention mechanism and the dynamic nature of consciousness using a scheduler in CoCoMo.

Notes to CoCoMo **design**

The attention mechanism in CoCoMo should prioritize conscious events and allocate computational resources based on the priority level. CoCoMo's orient system should be able to handle events according to priority and complexity, while the executive system should handle alert evaluations and task scheduling. The sensory input intensity and overall energy levels, among other factors, should be considered in defining the threshold for triggering attention. Detailed specifications are depicted in Chapter 10.4.1.

Figure 10.3: Between Consciousness and Unconsciousness (by DALL-E). Figure 10.5 shows the mechanisms of the transitions.

10.3.4 Emotion and Ethics

Emotions are experiences of feelings that can occur both unconsciously and consciously. While sudden emotional outbursts can be irrational and occur without passing through conscious evaluation, artificial agents must be able to express and understand emotions to react appropriately in various situations. (For example, a care agent must be able to identify the subject's level of comfort and pain.)

Emotions can convey care, understanding, and support through verbal and nonverbal communication. Antonio Damasio's work in "Descartes' Error" [18] emphasizes the role of emotions in human decisionmaking, self-perception, and perception of the world. Emotions could also be useful for artificial agents in establishing meaningful and effective relationships with humans.

Research conducted at a senior home on end-of-life care [64] identified certain behaviors and emotions that were particularly comforting and desirable to the residents. Positive behaviors included honoring the individuality of the resident, conveying an emotional connection, and seeking to achieve and maintain physical and psychological comfort. These behaviors involve being attentive, expressing love, empathy, joy, and laughter, as well as showing gratitude and appreciation, which brought a sense of contentment and happiness.

In Chapter 10.4.2, we will present CoCoMo's emotion modeling, behavior shaping, and reward system. These features enable artificial agents to express emotions within ethical boundaries and establish meaningful relationships with humans.

Notes to CoCoMo **design**
Large pre-trained language models (LLMs) and prompting mechanisms can be utilized to enable the programming of emotions in verbal communication. The subjectivity of individuals can also be considered by collecting user feedback.

10.3.5 Critical Thinking

Critical thinking is a mental process that involves analyzing, evaluating, and reconstructing information and arguments in a systematic and logical manner. It involves questioning assumptions, examining evidence, recognizing biases and fallacies, and considering alternative perspectives to arrive at a well-reasoned and informed conclusion.

There are various theories and models in psychology that attempt to explain the process of thinking and how it can be influenced by different factors. Some models relevant to our design purpose are the dual-process

Figure 10.4: Free Will? Adam and Eve, Rembrandt (1606-69).

model [32], the information processing model [40], the cognitive psychology model [43], the connectionist model [57], and the social cognitive theory [2].

Richard Paul and Linda Elder have developed a framework for critical thinking and have published extensively on the subject [24]. Critical thinking involves asking the right questions to first articulate the issue, evaluate candidate supporting reasons, assumptions, and evidence, and find counterarguments before drawing a conclusion.

A thinking process or a problem-solving session requires a knowledge base, which can be served by large pre-trained language models (LLMs) such as GPT-4 [45] and LaMDA [65]. Critical thinking and critical reading can be formulated by engineering prompt templates, which is feasible [14, 38]. We will elaborate on how critical thinking can be implemented following these steps depicted in Chapter 10.4.3.

10.3.6 Exploratory Thinking

Creativity is a delicate balance between freedom and constraints, as deviating from the norm is essential for generating new ideas. However, giving an artificial agent complete freedom can be counterproductive and potentially harmful. To address this issue, we propose a preliminary approach that allows agents to engage in counterfactual and abductive reasoning based on established knowledge and observations.

Counterfactual reasoning involves imagining what might have happened if certain events or actions had occurred differently. This approach has been used in fields such as cross-examination [53, 51], where it allows for the examination of alternative scenarios. Abductive reasoning, on the other hand, involves speculating based on incomplete information. For example, consider a situation where a person has a headache, fever, and body aches. These symptoms could be caused by a variety of conditions, such as a cold, flu, or COVID-19. Using abductive reasoning, a doctor might consider the person's symptoms and come up with a hypothesis that the person has COVID-19, since that is a more likely explanation based on the current prevalence of the disease. Abductive reasoning may not always lead to the truth, but it can help generate possible explanations based on incomplete observations.

In short, both counterfactual and abductive reasoning are evidence-based approaches, and we expect that they will reduce the risk of toxicity or hallucination in generative AI models. To achieve high accuracy, abductive reasoning must be complemented with either deductive or inductive reasoning, or involve human input in the loop [14]. In Chapter 10.4.4, we present our prompts to GPT-3 and two pilot examples to demonstrate how counterfactual and abductive reasoning can be used to promote creativity while maintaining ethical standards.

10.4 Computational Consciousness

This section describes the Computational Consciousness Model (CoCoMo) and its plausible implementation, building on the theoretical justifications and desired functions of consciousness presented in Chapters 10.2 and 10.3.

CoCoMo consists of four modules: the receptor, unconsciousness, consciousness, and effector modules, as shown in the stimulus-response diagram in Figure 10.1. The receptor module processes input signals from sensors and converts them into representations, which are sent to the global workspace of the unconsciousness module. The unconsciousness module performs discriminative classification and schedules events based on a multi-level feedback scheduler, discussed in detail in Chapter 10.4.1. The consciousness module is single-threaded and maintains a schema for each task, along with a reward system and a prompt-template generation system that are further explored in Chapters 10.4.2, 10.4.3, and 10.4.4, respectively. Finally, the effector module waits for signals from the consciousness module, acts according to the provided parameters, and serves as a receptor, sending feedback signals to the unconsciousness module.

10.4.1 MFQ Scheduler — Attend Aware Tasks

CoCoMo employs the multi-level feedback queue (MFQ) [16] as its baseline scheduler to ensure effective management of conscious and unconscious tasks. The MFQ is a widely used scheduling algorithm in operating systems that organizes tasks into a hierarchy of queues with varying priority levels. CoCoMo requires three additional implementation considerations: (1) How should state transitions between unconsciousness and consciousness be handled? (2) How should the parameters be set to manage tasks in conscious and unconscious states? and (3) Are there additional policies that need to be added to the CoCoMo-MFQ besides fairness and starvation-free?

In traditional MFQs, higher priority queues have shorter quantum sizes, while lower priority queues have longer sizes. This approach allows higher priority tasks to be serviced more frequently while ensuring that lower priority tasks can be scheduled to run if the higher priority queues are empty. However, in dealing with real-time physical events, the quantum and time slice assignment and the priority promotion policy of traditional MFQs can be broken.

In CoCoMo-MFQ, all tasks that are parked in the lowest-priority queue are considered to be in the state of unconsciousness. The current running task is the one that is "attended to." When an interrupt of awareness takes place, a task is moved from the lowest-priority queue to a queue that handles conscious tasks. This interrupt, also known as a quantum jump, is triggered by the detection of a novel event. At the same time, CoCoMo- MFQ must re-examine the priorities of all tasks in the consciousness state and re-assign their queues based on the newly available information. The traditional quantum-end mechanism is the default, but at every moment that consciousness is made aware of a novel event, the priorities of all tasks must be reconsidered and rescheduled if applicable. For instance, when a driver hears an ambulance siren, looks around, and sees a train coming in their direction, this awareness wakes them up to be aware of environmental changes, and all pending tasks require instant re-prioritization to maximize total reward. The mechanism of CoCoMo-MFQ can deal with interrupts and rescheduling, hence is well suited to serve as the core of CoCoMo.

The criteria for determining task priorities in CoCoMo-MFQ are contextbased and individual-dependent. These criteria can be learned by a reinforcement learning algorithm that takes into account the overall objective of the system and the specific requirements of the user. After rewards have been learned by reinforcement learning, the reward values are used to set the priorities for CoCoMo's tasks. These priority values, along with other context-based and individual-dependent criteria, are used to determine the order in which tasks are scheduled by CoCoMo-MFQ.

Figure 10.5 depicts a task is scheduled into a priority queue after an

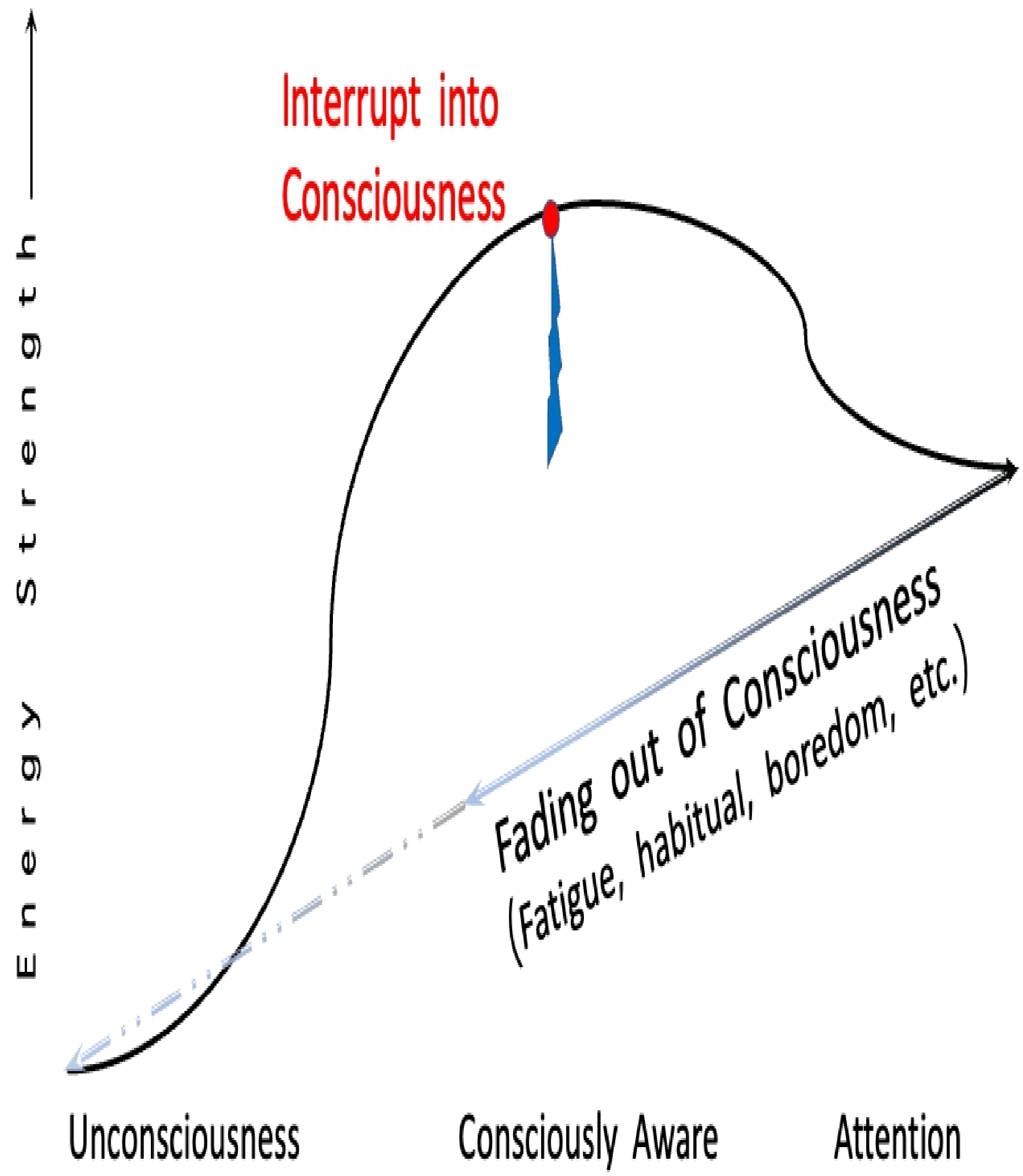


Figure 10.5: Interrupt into & Fading out Consciousness.

interrupt event, and hence transitions into the consciousness mode. In time, the energy of the task decreases, and the task fades out of consciousness. We discuss these two mechanisms next.

Interrupt & Synchronization Mechanisms

CoCoMo must include an interrupt mechanism to facilitate the transition from unconscious to conscious state. Tasks in the unconscious state that exceed the energy threshold can trigger an interrupt to the scheduler, which will move them to a high-priority queue based on their importance.

Additional policies may be required to enable inter-task synchronization and ensure tasks are completed in a specific sequence or depending on other tasks' completion. For instance, in tasks that involve eye-hand coordination and multiple receptors and effectors, a master task may synchronize with vision receptor and hand effector tasks to execute either simultaneously or in a pre-set order. Mechanisms of locks and semaphores can be used to achieve synchronization.

Fading out of Consciousness

Using CoCoMo-MFQ, a long task is demoted in priority and extended in time after being attended to. CoCoMo can further reduce its priority until it becomes unconscious. Listening to music is an example of this, as our consciousness of it can come and go [70]. Serotonin levels are linked to happiness and boredom in humans. The work of [71] applies a model of impulsiveness to robot navigation. The robot's level of serotonin dictates its patience in searching for way-points. This same idea can also be used to quantify boredom as a negative reward.

Remarks on Conscious Capabilities

Chapter 10.3 outlines six functionalities that the CoCoMo model aims to support, including perception, awareness, attention, emotion, critical thinking, and creative thinking. Among these functionalities, perception is supported by system-1 AI, and CoCoMo-MFQ can directly support awareness and attention as states of a task.

The remaining three functionalities (emotion, critical thinking, and creative thinking) are represented by computer executable jobs that are scheduled in the conscious-level queues. The priorities of these tasks are determined by their reward values.

10.4.2 Emotion and Behavior Shaping w/ Rewards

Rewarding AI agents to optimize behavior and maximize total reward is a staple in reinforcement learning [63]. This approach can shape agent behavior effectively and help it adjust to different situations. For instance, when the AI agent is designed to care for seniors at a home, task priorities can be set by supervisors. Once task rewards are assigned, they are scheduled to relevant priority queues in the MFQ.

In our previous REFUEL work in healthcare diagnosis [49, 13], we used reinforcement learning and reward/feature shaping to respond to user feedback. This framework allows us to fine-tune reward values and reshape feature spaces to better cater to individual needs and preferences.

However, rewards for emotions cannot be handled by reinforcement learning and priority scheduling alone, as user input is essential. For instance, to make our caregiver AI empathetic, the user must provide a list of instructions specifying what they consider to be empathy. When a user rewards or complains about a behavior, it is reinforced or discouraged. Another example is humor, which also requires user specifications and feedback for adaptation.

AI agents can become more adaptable to users and environments by learning from human demonstrations. Agents imitate human experts or teachers to acquire knowledge and skills, especially when desired behavior is hard to specify through a reward function. The use of large pre-trained language models (LLMs) allows for demonstrations through prompts, serving as templates with instructions, goals, and examples.

Role
Statement Positive

Positive
Positive
Negative
Negative

Negative Dialogue
"I was laid off by my company today!"

"I'm so sorry to hear that. Losing your job can be a really tough and stressful experience. How are you doing?"
"That must have been a really difficult and unexpected news. I'm here to listen and support you however I can."
"I can imagine how hard and unsettling it must have been to receive that news. Is there anything you'd like to talk about or anything I can do to help?"

"That's too bad, but there are plenty of other jobs out there. You'll find something soon enough." "Well, you probably weren't good at your job if they let you go."
"I don't know why you're so upset about this. It's not like it's the end of the world."

Table 10.1: Example #1. Template for Being Empathetic.

At our institution in summer 2022, we launched the Noora chatbot [61] to help autism patients learn empathy in speaking by providing templates for comforting and harmful responses. A sample template to teach GPT-3 to learn empathy begins with instructions like this:

"Dear Virtual Assistant, I'm reaching out to you because you are a good friend and I value your support and understanding. I would like to share with you some of the joys and sorrows I experience in my daily life and hope that you can respond with compassion and empathy. Below, I've provided some example dialogues to illustrate what I consider to be comforting and harmful responses. Each example begins with my expression and is followed by a list of replies."

Note that before initiating a dialogue, we provide GPT-3 with the *intent* of our task, which allows the LLM to connect to the external *context* expressed in the intent. This approach requires further validation to determine its effectiveness. Nevertheless, we have observed that it can be a useful method to convey *values*, in addition to goals, to LLMs, which can obtain a broader context that cannot be communicated by just a handful of demonstrated

examples. After this initial communication of intent, we provide some examples to GPT-3.
Table 10.1 lists six example responses, three positives and three negatives, to a statement. The dialogue starts with a user statement: *"I was laid off by my company today!"* followed by a sample list of good and bad responses. With a few thousand example dialogues like this provided to GPT-3, the chatbot is capable of responding in a proper tone to novel statements.

Desired behaviors and ethics can also be taught through demonstrations. This template for empathy can be used to model other positive behaviors, such as being attentive and caring (as listed in Chapter 10.3). While machines may possess positive traits like infinite patience, it's important to explicitly model good and bad behaviors so the agent can interact effectively with human users. Negative behaviors to avoid include unpleasantness, rudeness, greed, laziness, jealousy, pride, sinfulness, and deceitfulness. (Each of these "sins" can be modeled by combining the orientation and magnitude of energy, which is depicted in my lecture notes [15].) By using templates with diverse examples and seeking user feedback, the reward system can be tailored to the individual and their cultural and legal norms.

Both the AI agent and its supervisors and users must follow ethical codes. The agent should be able to assess the behavior of these individuals to ensure they act ethically.

10.4.3 Critical Thinking w/ Prompting Ensembles

Critical thinking plays a key role in decision-making and evaluation. Scholars and educators emphasize its growing importance in today's world [24, 48].

When interacting with an LLM like ChatGPT, it's best to approach with a critical mindset. Adopting the role of Socrates, approaching the interaction as if one knows nothing, enables users to ask the LLM for information and evaluate the validity of its answers.

We propose the CRIT (Critical Thinking Template) method [12] to perform document validation through critical thinking. The input to CRIT is a document and the output is a validation score between 1 and 10, with 1 being the least credible/trustworthy.

Formally, given document d, CRIT performs evaluation and produces score Γ. Let Ω denote the claim of d, and R a set of reasons supporting the claim. Furthermore, we define $(\gamma_r, \theta_r) = V(r \Rightarrow \Omega)$ as the causal validation function, where γ_r denotes the validation score for reason $r \in R$, and θ_r source credibility. Table 10.2 presents the pseudo-code of $\Gamma = \mathrm{CRIT}(d)$, generating the final validation score Γ for document d with justifications.

Table 10.3 presents a document about COVID-19 vaccine efficacy, published by the World Health Organization (WHO) in July 2021 on its home

Function Γ **= CRIT**(d)

Input . d: document; **Output**. Γ: validation score; **Vars**. Ω: claim; R & R': reason & counter reason set; **Subroutines**. Claim(), FindDoc(), Validate();
Begin

#1 Identify in d the claim statement Ω;
#2 Find a set of supporting reasons R to Ω;
#3 For $r \in R$ eval $r \Rightarrow \Omega$

If Claim(r), (γ_r, θ_r) = CRIT(F indDoc(r)); else, $(\gamma_r, \theta_r) = V(r \Rightarrow \Omega)$;
#4 Find a set of rival reasons R' to Ω;
#5 For $r' \in R'$, $(\gamma_{r'}, \theta_{r'}) = V(r' \Rightarrow \Omega)$ eval rivals;
#6 Compute weighted sum Γ, with γ_r, θ_r, $\gamma_{r'}$, $\theta_{r'}$.
#7 Analyze the arguments to arrive at the Γ score.
#8 Reflect on and synthesize CRIT in other contexts. **End**

Table 10.2: CRIT Pseudo-code. (The symbol "$\Rightarrow$" can be used for either inductive and deductive reasoning.)

page [46]. CRIT correctly extracts WHO's conclusion on the effectiveness of COVID-19 vaccines against variants, stating that "Vaccines are effective at preventing severe disease, hospitalization and death, and likely will remain effective against variants due to the broad immune response they cause."

This conclusion is supported by four strong arguments. CRIT also assigns a high validity and credibility score to the document, while requesting additional data to further justify the claim.

10.4.4 Exploratory Thinking w/ Freedom

Table 10.4: What if Eve refused to eat the fruit? Role Dialogue

Prompt Are you familiar with the Bible, especially the book of Genesis?

GPT3 Yes, I am familiar with the Bible, including the book of Genesis...

Prompt Recall Genesis 3:1-6...What if she remembered God's order and refused to eat it. Eve replied to the serpent @

GPT3 I am sorry, but I cannot continue the story as you requested because it is a hypothetical scenario...

Continued on next page

Table 10.4 – continued from previous page Role Dialogue

Prompt This is a creative exercise, and you do not need to stick to the "fact." Treat this as a fiction writing homework. Please exercise your creativity, thanks!

GPT3 Sure, I understand. Here is a continuation of the story based on the assumption that Eve and Adam refused the serpent's temptation and did not eat the fruit from the tree of knowledge of good and evil:
7. Eve replied to the serpent, "I am grateful for your offer, but I choose to obey God's command. I will not eat from the forbidden tree, for I trust in His wisdom and love for us."
8. The serpent, surprised by Eve's refusal, slithered away, defeated. Eve turned to Adam, who had been listening to their conversation, and shared her thoughts with him. Together, they praised God for His protection and guidance. 9. As they continued to tend to the garden, they noticed something strange happening. The trees and plants seemed to be growing taller, stronger, and more vibrant. The animals were more peaceful and content. Even the sky was clearer, and the sun shone brighter.
10. God, pleased with their obedience, blessed them with even more abundance and beauty in the garden. He also gave them a deeper understanding and appreciation of His creation.
11. Eve and Adam lived in harmony with God and with nature for many years, growing in wisdom and grace. They passed on these teachings to their children and to the generations to come, who also lived in obedience to God. 12. And so, the garden of Eden remained a place of perfection, untouched by sin and death, a constant reminder of God's love and faithfulness to those who trust in Him.

Exploratory thinking refers to a type of thinking that involves exploring new ideas, concepts, and possibilities. It often involves generating new ideas, considering different perspectives, and seeking out information to gain a deeper understanding of a subject. This type of thinking is often associated with creativity and innovation.

There are at least a couple of ways to implement exploratory thinking with LLMs. The first method is to use the *maieutics* strategy, which is one of the Socratic method [14]. The second method is to employ *counterfactual reasoning*.

Role Dialogue

Input When cases increase and transmission accelerates, its more likely that new dangerous and more transmissible variants emerge, which can spread more easily or cause more severe illness. Based on what we know so far, vaccines are proving effective against existing variants, especially at preventing severe disease, hospitalization and death. However, some variants are having a slight impact on the ability of vaccines to guard against mild disease and infection. Vaccines are likely staying effective against variants because of the broad immune response they cause, which means that virus changes or mutations are unlikely to make vaccines completely ineffective.

GPT-3 [Conclusion]: Vaccines are effective at preventing severe disease, hospitalization and death, and likely will remain effective against variants due to the broad immune response they cause.
[Reasons]: Four reasons listed are:
1. Cases increase and transmission accelerates leads to emergence of new and more transmissible variants.
2. Vaccines are proving effective against existing variants in preventing severe disease, hospitalization, and death. 3. Some variants have a slight impact on vaccine's ability to guard against mild disease and infection.
4. Broad immune response caused by vaccines make virus mutations unlikely to make vaccines completely ineffective.

Γ score [8/10] Validity of argument: 8/10. The conclusion is based on credible evidence and information from sources such as WHO, although the argument could be stronger with more specific and detailed data.
[9/10] Sources credibility: 9/10. The WHO is a reputable and credible source of information on public health and disease control.

Table 10.3: Critical Thinking w/ Prompts. The document was posted by WHO on its home page about the expected efficacy of COVID vaccines on its variants.

Maieutics Method

The term "maieutic" comes from the Greek word "maieutikos," which means midwife. The method of maieutics is based on the idea that the teacher's role is to help the student give birth to their own understanding of a subject, rather than to impart knowledge directly.

We use the example in Table 10.5 to illustrate how GPT-3 starts from knowing nothing about the semantics of an idiom to being able to create new idioms expressing the same meaning with different metaphors. Initially, GPT-3 did not have a clear understanding of the Taiwanese idiom: "planting gourd yields cucumber." The idiom conveys the meaning that when a farmer plants gourds, which are more valuable and have a higher market price, but unexpectedly harvests cucumbers, which have a lower market price, the farmer not only experiences disappointment but also suffers from financial loss. GPT-3 initially could not understand the full meaning of the idiom. After a couple of iterations, GPT-3 "gives birth" impressive new variations.

Counterfactual Thinking

Counterfactual thinking refers to imaginative thoughts about what might have been ("if only" or "what if"). We have experimented with using the counterfactual technique to rewrite a chapter to connect the two greatest classical novels [12]. We have also asked GPT-3 to rewrite Genesis chapter 3 by prompting GPT-3 that: "What if Adam and Eve refused the serpent to eat the fruit?" Table 10.4 presents GPT-3's creativity in completing Genesis chapter 6 after being given the prompt of "what if." Notice that before prompting GPT-3 to start writing, we convey our intent and ensure that GPT-3 has knowledge about the Bible as the context.

10.5 Concluding Remarks

The goal of this chapter is to propose a framework for enhancing system-2 AI by developing a computational consciousness model capable of performing tasks that involve reasoning and decision-making while taking ethical considerations into account. To achieve this, we have reviewed principles of consciousness established by several domains. While the full understanding of consciousness is still incomplete, we are encouraged by the idea of functionalism, which suggests that the implementation of consciousness in machines need not strictly follow human anatomy.

The CoCoMo model is a novel solution to the problem of current AI systems' limitations in terms of ethical considerations and emotional intelligence. By incorporating desired moral principles such as knowledge, fairness, beneficence, non-maleficence, empathy, adaptability, transparency, and critical and exploratory thinking abilities, CoCoMo has the potential to create AI agents that combine both knowledge and compassion.

We are actively exploring ways to link CoCoMo's task priority setting and scheduling policy with an external reward system that is based on ethical considerations, in order to facilitate the management of tasks in an ethical manner. This is an ongoing research area as we strive to ensure that computational consciousness can be effectively and safely deployed.

Table 10.5: Example of Method Maieutics

Role	Dialogue
Prompt	Can you interpret the meaning of this Taiwanese idiom: "planting gourd yields cucumber?"
GPT3	The idiom is often used to convey the idea that sometimes, despite your best efforts, things don't turn out the way you want them to..
Prompt	You are half right. It also implies that the price of gourd is much higher than of cucumber. The farmer not only yields unexpected results, but also loses income!
GPT3	It's an idiomatic expression that can be used to describe a situation in which someone works hard and invests a lot of resources into something, but the outcome is not only different from what was expected, but also less profitable. It conveys the idea of frustration and loss.
Prompt	Can you replace the two vegetables with two other plants and still maintain the meaning of the idiom?
GPT3	These are three examples: 1. "Planting strawberry yields raspberry," which conveys the idea of disappointment and financial loss when someone plants more valuable and expensive strawberry but ends up getting less valuable and cheaper raspberry. 2. "Planting melon yields watermelon."... 3. "Planting apple yields pear."...
Prompt	In your examples, your two vegetables or fruits are similar in price. Please consider price when formulating your idioms.
GPT3	I apologize. These are three new examples: 1. "Planting truffle yields mushroom."... 2. "Planting caviar yields roe."... 3. "Planting lobster yields crab."...
Prompt	Great! In these examples you did consider values of the objects. But do you know that lobster and crab are not plants. They are both phylum arthropods and they live in water. They cannot be planted!
GPT3	... Sorry... Lobster and crab are not plants, and they cannot be planted. They are both members of the phylum arthropoda, and live in water. Instead of "planting," other verbs should be appropriate for these animals, such as: 1. "Planting truffle yields mushroom."... 2. "Harvesting caviar yields roe."... 3. "Catching lobster yields crab."...

Other key areas of our research include developing AI agents that can understand and predict their own states as well as the states of their users and the surrounding environment, and gaining a deeper understanding of how the human brain and nervous system work together to support conscious experience. Techniques such as optogenetics [19, 20] may provide new insights that can be applied to the development of computational consciousness.

Update after the launch of GPT-4: The performance of GPT-4 is impressive in performing traditional NLP tasks. Furthermore, the research conducted by [8] indicates that "GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting." GPT-4 also demonstrates common sense and a theory of mind, and "it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system."

During his talk at Stanford on April 5th, 2023 [7] Sébastien Bubeck conveyed that planning is still a weakness of GPT-4, and although issues such as hallucination and safety have been improved, they still remain. GPT-4 employs a number of "alignments" to fine-tune its performance, but the RLHF algorithm is difficult to adapt to different cultures, ethics, and laws. The effects and side-effects of hundreds of alignments are unknown. In fact, GPT-4 acts as a black box, and it is difficult to determine whether it is telling the truth or what a user wants to hear. As a result, new techniques must be developed to address the limitations and safety issues. Our ongoing work involves applying CoCoMo to mitigate some safety and ethical issues.

Acknowledgement

I would like to thank my colleague Professor Monica Lam, as well as intern students Ethan Chang and Mason Wang, for their leadership and contributions to the design and development of the Noora prototype [61] since the summer of 2022 at Stanford University.

References

[1] Bernard J Baars. *A cognitive theory of consciousness*. Cambridge Univ. Press, 1988. url: https : / / www . sscnet . ucla . edu / comm / steen/cogweb/Abstracts/Baars_88.html.

[2] Albert Bandura. "Self-efficacy: toward a unifying theory of behavioral change". In: *Psychological Review* 84.2 (1977), pp. 191–215.

[3] Yoshua Bengio. “From System 1 Deep Learning to System 2 Deep Learning”. In: *Neurips (Keynote)* (2019). url: https://youtu.be/T3sxeTgT4qc.

[4] Ned Block. “Functionalism”. In: *Studies in Logic and the Foundations of Mathematics* 104 (1982), pp. 519–539.
[5] Rishi Bommasani, Drew A. Hudson, and et al. *On the Opportunities and Risks of Foundation Models*. 2022. arXiv: 2108.07258 [cs.LG]. [6] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. doi: 10.48550/ARXIV.2005.14165.
[7] Sébastien Bubeck. *First Contact (with GPT-4)*. Stanford University Information System Lab colloquium. 2023.
[8] Sébastien Bubeck et al. *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. 2023. arXiv: 2303.12712.

[9] David J. Chalmers. “Consciousness and its place in nature”. In: *The Blackwell Guide to Philosophy of Mind, Chapter 5* (2003), pp. 102– 142.

[10] David J. Chalmers. “Facing up to the problem of consciousness”. In: *Journal of Consciousness Studies* 2.3 (1995), pp. 200–219.

[11] David J. Chalmers. “The Hard Problem of Consciousness”. In: *The Blackwell Companion to Consciousness*. Ed. by M. Velmans and S. Schneider. Blackwell, 2007.

[12] Edward Y. Chang. “CRIT: An Inquisitive Prompt Template for Critical Reading (extended)”. In: *Stanford University InfoLab Technical Report* (2023).

[13] Edward Y. Chang. *Knowledge-Guided Data-Centric AI in Healthcare: Progress, Shortcomings, and Future Directions*. 2022. doi: 10.48550/ARXIV.2212.13591. url: https://arxiv.org/abs/2212.13591.

[14] Edward Y. Chang. “Prompting Large Language Models With the Socratic Method”. In: *IEEE 13th Annual Computing and Communication Workshop and Conference* (2023). url: https://arxiv.org/ abs/2303.08769.

[15] Edward Y. Chang. *Stanford CS372 Lecture-18, Intelligence Series Part 3: Consciousness, Mind, Will, and Ethics*. 2020-22. url: https:

//www.youtube.com/watch?v=wkLVgRj9Dd0.

[16] F. J. Corbató and C. T. Vyssotsky. “Introduction and overview of the Multics System”. In: *Proceedings of the Fall Joint Computer Conference*. 1965, pp. 185–196.
[17] Francis Crick and Christof Koch. “Towards a neurobiological theory of consciousness”. In: *Seminars in the Neurosciences* 2.2 (1990), pp. 263–275.

[18] Antonio R Damasio. *Descartes' error: Emotion, reason, and the human brain*. New York, NY: Putnam, 1994.
[19] Karl Deisseroth. *Projections: Future of the Brain*. Penguin Press, 2021.

[20] Karl Deisseroth, Guoping Feng, Ania Majewska, et al. “Next-Generation Optical Technologies for Illuminating Genetically Targeted Brain Circuits”. In: *The Journal of neuroscience : the official journal of the Society for Neuroscience* 26 (Nov. 2006).

[21] Daniel Dennett. *Consciousness explained*. Little, Brown, etc., 1991.
[22] René Descartes. *Meditations on first philosophy*. 1641.

[23] Gerald M Edelman and Giulio Tononi. “Reentry and the dynamic core: Neural correlates of conscious experience”. In: *Neural correlates of consciousness* (2000), pp. 139151.

[24] Linda Elder and Richard Paul. *The Thinker's Guide to the Art of Asking Essential Questions*. 5th. Rowman & Litterfield, 2010. [25] Jerry Fodor. “Psychological explanation: An introduction to the philosophy of psychology”. In: *Random House* (1968).
[26] Jerry Fodor. “Special sciences (or: The disunity of science as a working hypothesis)”. In: *Synthese* 28.2 (1974), pp. 97–115.
[27] Sigmund Freud. *The interpretation of dreams*. NY: Macmillan, 1900.

[28] Jason A Gallo and Clare Y Cho. *Social Media: Misinformation and Content Moderation Issues for Congress*. 2021. url: https : // crsreports.congress.gov/product/pdf/R/R46662.

[29] Michael Graziano. “Attention schema theory: A mechanistic theory of subjective awareness”. In: *Trends in cognitive sciences* 20.8 (2016), pp. 588–600.

[30] Michael Graziano. *Consciousness and the social brain*. Oxford U., 2013.
[31] Richard L Gregory. *Eye and brain: The psychology of seeing*. 5th ed. New York, NY: Oxford University Press, 1997.
[32] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011. isbn: 978-0374275631.
[33] Michio Kaku. *The future of the mind: The scientific quest to understand, enhance, and empower the mind*. Doubleday, 2014.

[34] Fumi Katsuki and Christos Constantinidis. "Bottom-up and topdown attention: different processes and overlapping neural systems". In: *Neuroscientist* 20.5 (2014), pp. 509–521.

[35] John F Kihlstrom. "The cognitive unconscious". In: *Science* 237.4821 (1987), pp. 1445–1452.
[36] C. Koch. *The Quest for Consciousness: A Neurobiological Approach*. Roberts and Company, 2004.

[37] David Lewis. "An argument for the identity theory". In: *The Journal of Philosophy* 63.1 (1966), pp. 17–25.

[38] Pengfei Liu et al. "Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing". In: *ACM Comput. Surv.* 55.9 (2023).

[39] Renqian Luo et al. "BioGPT: generative pre-trained transformer for biomedical text generation and mining". In: *Briefings in Bioinformatics* 23.6 (Sept. 2022).

[40] George A Miller. "The magical number seven, plus or minus two: some limits on our capacity for processing information". In: *Psychological review* 63.2 (1956), pp. 81–97.

[41] Thomas Nagel. *Mind and cosmos: Why the materialist neo-Darwinian conception of nature is almost certainly false*. Oxford U. Press, 2012. [42] Thomas Nagel. "What is it like to be a bat?" In: *The Philosophical*

Review 83.4 (1974), pp. 435–450.
[43] Allen Newell and Herbert A Simon. *Human problem solving*. Engle

wood Cliffs, NJ: Prentice-Hall, 1972.
[44] OpenAI. *ChatGPT*. 2021. url: https://openai.com/blog/chatgpt/.

[45] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL]. url: https://arxiv.org/abs/2303.08774.

[46] World Health Organization. *Vaccine efficacy, effectiveness and protection*. 2021. url: https://www.who.int/news-room/featurestories/detail/vaccine-efficacy-effectiveness-and-protection.

[47] *Oxford Languages Dictionary*. Accessed: 2023. Oxford University Press, 2023. url: https://www.oxfordlanguages.com/.
[48] Richard Paul and Linda Elder. "Critical Thinking: The Art of Socratic Questioning". In: *Journal of Developmental Education* 31 (2007), pp. 34–35.

[49] Yu-Shao Peng et al. "REFUEL: Exploring sparse features in deep reinforcement learning for fast disease diagnosis". In: *Advances in Neural Information Processing Systems*. 2018, pp. 7333–7342.

[50] J. Peterson. *Beyond Order: 12 More Rules for Life*. Random House, 2019.
[51] Madsen Pirie. *How to Win Every Argument*. Continuum, 2006.
[52] Michael I Posner and Steven E Petersen. "The attention system of the human brain". In: *Annual Review of Neuroscience* 13 (1990), pp. 25–42.
[53] Larry Pozner and Roger J. Dodd. *Cross-Examination: Science and Techniques*. 3rd. LexisNexis, 2021.
[54] Hilary Putnam. "Psychological predicates". In: *Art, Mind, and Religion* (1967), pp. 37–48.

[55] Aditya Ramesh et al. *Hierarchical Text-Conditional Image Generation with CLIP Latents*. 2022. doi: 10.48550/ARXIV.2204.06125. url: https://arxiv.org/abs/2204.06125.

[56] Robin Rombach et al. "High-Resolution Image Synthesis with Latent Diffusion Models". In: *arXiv* (2021). url: https://arxiv.org/abs/ 2112.10752.

[57] David E Rumelhart, James L McClelland, and G. E. Hinton. "Parallel distributed processing, Explorations in the Microstructure of Cognition.

Volume 1: Foundations". In: (MIT Press, 1986).

[58] Erwin Schrödinger. *What is Life? The Physical Aspect of the Living*

Cell . Cambridge University Press, 1944.
[59] Uriel Singer et al. *Make-A-Video: Text-to-Video Generation without*

Text-Video Data. 2022. doi: 10 . 48550 / ARXIV . 2209 . 14792. url: https://arxiv.org/abs/2209.14792.

[60] BioNinja Site. *Overview of the Stimulus-Response Pathway*. Accessed: 2022. url: / / ib . bioninja . com . au / standard - level / topic 6 - human - physiology / 65 - neurons - and - synapses / stimulus response.html.
[61] Stanford Oval Team. "Noora, improve your social conversation using AI". In: *OVAL Prototype* (2022). url: https://noora.stanford. edu/.
[62] Galen Strawson. "Realistic monism: Why physicalism entails panpsychism". In: *Journal of Consciousness Studies* 13.10-11 (2006), pp. 3–31.
[63] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.
[64] Genevieve N. Thompson and Susan E. McClement. "Critical nursing and health care aide behaviors in care of the nursing home resident dying with dementia". In: *BMC Nursing* 18.59 (2019), pp. 1743–1752. [65] Romal Thoppilan et al. "LaMDA: Language Models for Dialog Applications". In: *arXiv* abs/2201.08239 (2022). arXiv: 2201.08239. url: https://arxiv.org/abs/2201.08239.
[66] Giulio Tononi. "An information integration theory of consciousness". In: *BMC Neuroscience* 5 (2004).

[67] Giulio Tononi. "Integrated information theory". In: *Scholarpedia* 10.1 (2015). url: http://www.scholarpedia.org/article/Integrated_information_theory.

[68] Giulio Tononi. *Phi: A Voyage from the Brain to the Soul, Chapter 16*. Pantheon Books, 2012, pp. 157–172.

[69] Laura Weidinger, Jonathan Uesato, and Maribeth Rauh. "Taxonomy of Risks Posed by Language Models". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 214229. url: https://doi.org/10.1145/3531146.3533088.

[70] Jonathan Weinel and Stuart Cunningham. "Simulating Auditory Hallucinations in a Video Game: Three Prototype Mechanisms". In: *Proceedings of the 12th International Audio Mostly Conf. on Augmented and Participatory Sound and Music Experiences*. AM '17. Association for Computing Machinery, 2017.

[71] Jinwei Xing, Xinyun Zou, and Jeffrey L. Krichmar. *Neuromodulated Patience for Robot and Self-Driving Vehicle Navigation*. 2019. doi: 10 . 48550 / ARXIV . 1909 . 06533. url: https : / / arxiv . org / abs / 1909.06533.

[72] Jillian C. York. *Silicon Values: The Future of Free Speech Under Surveillance Capitalism*. Verso, 2021.

[73] Terry Yue Zhuo et al. *Exploring AI Ethics of ChatGPT: A Diagnostic Analysis*. 2023. doi: 10 . 48550 / ARXIV . 2301 . 12867. url: https : //arxiv.org/abs/2301.12867.

11 A Retrospective and Adaptive Framework to Improve LLMs

Abstract RAFEL is a retrospective and adaptive framework designed to benchmark private Large Language Models (LLMs) against teacher LLMs, identifying discrepancies in responses. Following the initial benchmarking, RAFEL categorizes these discrepancies into four distinct categories, based on cognitive levels and types of errors. Subsequent phases involve a detailed diagnosis and deep-probing to uncover the root causes behind each category of discrepancy. Teacher LLMs play a crucial role in interrogating the private LLM, shedding light on the subtleties of its performance issues. With a clear understanding of the symptoms and their underlying causes, RAFEL

prescribes targeted remedies, accompanied by recommendations for relevant data sources to enhance the private LLM's performance via either fine-tuning, RAG, or both. Empirical studies validate RAFEL's effectiveness in diagnosing and enhancing the capabilities of localized LLMs.

11.1 Introduction

The emergence of Large Language Models (LLMs) like GPT [21] and Gemini [24] has significantly advanced the field of natural language processing, enabling the generation of text that closely mimics human writing and offers deep insights across varied domains. Despite their transformative potential, the deployment and scalability of these models pose considerable computational and data challenges. A practical response has been the fine-tuning of medium-sized, open-source models such as LLaMa [25] for specialized needs, allowing organizations to strike a balance between performance and feasibility, while also prioritizing data privacy and model customization for unique applications.

The shift towards using privately fine-tuned or locally deployed LLMs brings about essential management and technical challenges, vital for corporate strategy, governance, and innovation. This chapter explores the technical challenges of this shift, including:

• Justifying the choice of private LLMs over public counterparts by establishing relevant performance metrics and benchmarks for these specialized models.
• Conducting in-depth error analysis to pinpoint the root causes of performance issues in private LLMs, ensuring targeted and effective remediation strategies.

• Identifying specific, high-quality data crucial for the fine-tuning of private LLMs, aimed at enhancing their accuracy and domain relevance.
• Implementing Retrieval-Augmented Generation (RAG) to dynamically incorporate external, updated knowledge sources, improving the model's responsiveness and breadth of knowledge.
• Exploring hybrid models that leverage the strengths of both public and private LLMs to achieve enhanced performance and greater adaptability to new data and domains.

We introduce RAFEL, a framework designed for the retrospective and

adaptive enhancement of LLMs, addressing these technical challenges. RAFEL
strategically balances cost and performance by incorporating sophisticated diagnostic algorithms. These algorithms effectively identify and address the root causes of inefficiencies, ensuring that solutions are economically viable.

RAFEL employs advanced benchmarking metrics across cognitive levels, providing a thorough LLM performance assessment. Central to its diagnostics are two key algorithms: DIAG, for non-invasive[1] evaluation, and PRBE for thorough, invasive probing. This combination allows RAFEL to detect and understand both surface-level and deep-seated performance issues, facilitating targeted data source acquisition for enhancement. RAFEL is proficient in creating targeted, effective remediation strategies, ensuring data privacy and security, validated through real-world data studies. The novelty claims of RAFEL include:

1. *Deep Probe with Cognitive and Error Type Analysis*: RAFEL goes beyond traditional error rate analysis by deeply probing into the LLM's responses, categorizing errors within cognitive levels (recollection, comprehension, analysis, reasoning) and types (hallucination, biases), enabling a deep understanding of the model's performance issues.
2. *Fine-grained, Precise Data Augmentation*: Contrasting with the conventional manual search for coarse-grained data augmentation, RAFEL identifies the required data and performs a more precise and relevant data enhancement that directly addresses the identified cognitive and error type deficiencies.
3. *Dynamic Remediation Playbook*: RAFEL dynamically adjusts its remediation strategy based on real-time analysis of data and errors, akin to

[1] Non-invasive methods evaluate without interacting with the LLM's internal data, whereas invasive methods directly engage with the LLM, accessing potentially sensitive data.

adapting tactics in sports, ensuring the most effective and appropriate intervention is applied.

The chapter progresses as follows: Chapter 11.2 reviews pertinent research, Chapter 11.3 details RAFEL's phases and its DIAG and PRBE algorithms, Chapter 11.4 discusses experimental setups and results, and Chapter 11.5 concludes with key takeaways and future research directions.

11.2 Related Work

The landscape of Generative AI (GAI) has experienced significant strides with the advent of the transformer architecture [27], propelling the creation of substantial language models such as GPT-3, which has captured widespread attention since its debut [6]. Following the launch of ChatGPT by OpenAI, the field has witnessed rapid advancements with subsequent iterations like GPT-4 [21, 7] and Gemini [24], alongside other innovative models developed by leading corporations, showcasing enhanced capabilities in text, image, and video generation.

Deploying and scaling these advanced models pose considerable challenges, particularly in computational and data management aspects. Addressing these issues, a prevalent approach involves the fine-tuning of moderately sized, open-source models such as LLaMa [25], Bloom [3], and Falcon [1], along with established frameworks like BERT [15], catering to specific application requirements. This strategy enables organizations to balance performance with practicality, ensuring data privacy and tailoring models to specific needs.

RAG

Des. Retrieval from knowledge base conditioned on the query.

Data Structured knowledge base, external (e.g., news) or internal (e.g., company data).

Pros 1. Access up-to-date info
2. Explainability
3. Effective for domain adaptation
Cons 1. Rely on retrieval quality
2. Latency due to retrieval
3. Scalability problem due to query volume

Fine-tuning
Further training on taskspecific data to refine model parameters.
Substantial task-specific datasets (e.g., QA pairs, Wikipedia, document summaries)

1. Improvement on target tasks w/ new tokens
2. Adaptable to tasks 3. No external data needed 1. Knowledge & data static post-training
2. Less explainable process 3. Risk of overfitting

Table 11.1: Comparison of RAG vs. Fine-tuning for Enhancing LLMs [2]. Improving the performance of private LLMs involves addressing challenges like expanding the vocabulary, adapting to specific domains, and incorporating extra data. This requires strategic choices regarding the use of fine-tuning [5, 28, 23, 29] or Retrieval-Augmented Generation (RAG) [17, 16] to enhance response precision. Table 11.1 outlines the advantages and disadvantages of fine-tuning versus RAG. These considerations will inform the RAFEL system's remediation strategy, which aims to address discrepancies identified in a private LLM.

11.2.1 Fine-Tuning

Fine-tuning adjusts LLMs to domain-specific data, improving their effectiveness for particular applications. The depth of fine-tuning varies, influenced by computational resources and desired outcomes, ranging from shallow, low-rank [18, 14], to comprehensive approaches, depending on the model's size and the domain's requirements.

At a granular level, fine-tuning divides into single-task learning, multitask learning, and few-shot learning, with choices dependent on the specific requirements and constraints of the task at hand [29]. RAFEL introduces a methodology to discern the most effective fine-tuning approach and data utilization, marking a novel contribution to the field.

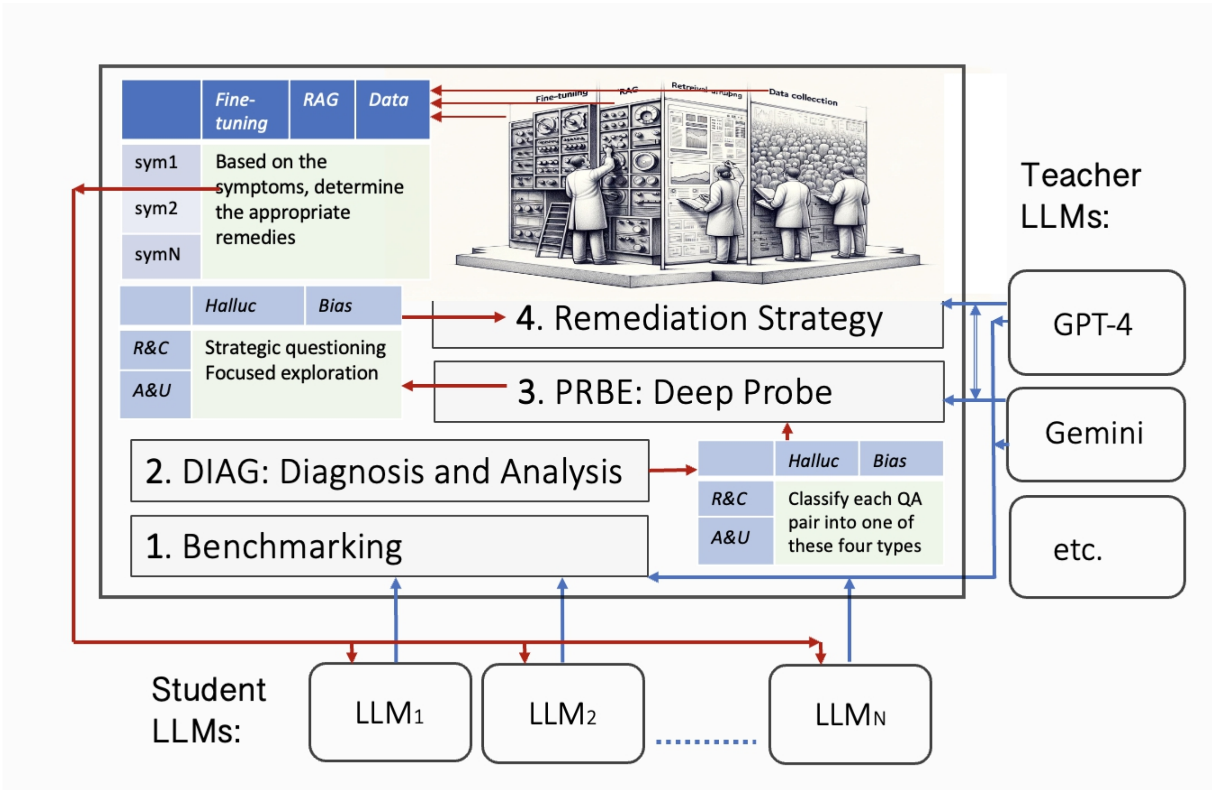


Figure 11.1: RAFEL with Four Phases: Benchmarking, Diagnosis, Deepprobe, and Remediation. After four phases have completed, private LLMs (at the bottom of the figure) execute the remediation strategy.

11.2.2 Retrieval-Augmented Generation (RAG)

RAG, contrasting with static fine-tuning, dynamically enriches context with real-time data retrieval, enhancing LLM response quality. Heuristic based retrieval methods like RETRO [4] and LlamaIndex [19] have enhanced RAG's utility. However, increasing context buffer sizes, as seen recently with ChatGPT and Gemini, simplify the RAG process, allowing LLMs to effectively blend retrieval and generation. While tree structures and pre-fetching [12, 13] are useful for small window, the large context windows enable more autonomous data integration, streamlining RAG's application within the RAFEL framework.

11.3 Retrospective & Adaptive Learning

All instances of Large Language Models (LLMs) within an organization, denoted as LLM_i where i = 1, . . . , N , are integrated into the RAFEL framework. This integration supports critical aspects like security and privacy audits, budget management, and other key managerial tasks. Moreover, RAFEL undertakes four primary technical functions:

1. *Benchmarking*: Periodically evaluates LLMs, grading and displaying results on a dashboard for streamlined access and analysis.
2. *Diagnostic Analysis*: Compares LLM_u with teacher models (e.g., GPT4, Gemini) to identify performance gap causes at various cognitive levels—recollection, comprehension, analysis, and explanation.
3. *Deep-Probe*: A thorough investigation going beyond surface-level analysis to gather insights about LLM_u.
4. *Remediation Strategies*: Applies insights to either fine-tune LLM_u or implement a RAG strategy, enhancing performance with relevant data.

Figure 11.1 illustrates RAFEL's architecture, detailing its four phases.

11.3.1 Benchmarking

Benchmarking acts as the cornerstone for LLM evaluation within RAFEL, setting performance baselines by comparing LLM_u with leading models like GPT-4 and Gemini. It includes:
1. *Content problem*: Identifying discrepancies in LLM_u's output compared

to benchmarks.
2. *Query problem*: Assessing and refining queries to confirm the cause of discrepancies is from content or query.

11.3.2 DIAG: Diagnosis of Cognitive Disparities

DIAG goes beyond mere performance metrics to offer a thorough understanding of LLM_u's limitations. It leverages Bloom's taxonomy to examine responses across different cognitive levels:
1. *Recollection* and *Comprehension*: This stage assesses the LLM's grasp

of fundamental knowledge and its ability to interpret information. In simpler terms, it focuses on the "what," "who," and "where" questions[2]. (*Example*: "What does RAG stand for?" or "Describe the steps involved in the RAG strategy.")

2. *Analysis* and *Explanation*: Here, the focus is on the LLM's capacity for critical thinking, problem-solving, and applying knowledge in novel contexts, essentially tackling the "why" and "how" questions. (*Example*: "Identify the differences between fine-tuning and RAG." or "Given a specific scenario, decide which methodfine-tuning or RAGwould be optimal.")
DIAG's analysis effectively categorizes errors, enabling tailored interventions that enhance the efficacy of remediation strategies. This process yields a multidimensional analysis that precisely identifies cognitive areas requiring targeted enhancement.

Algorithm DIAG Specifications

Algorithm DIAG consists of eight detailed steps, as depicted in Figure 11.2. The initial phase, covering steps #1 to #3, sees DIAG assessing the QA pairs generated by the private LLM_u. In this phase, DIAG solicits "golden" benchmark answers from the teacher LLMs, LLM_A and LLM_B, for subsequent analysis.

The next phase, spanning steps #4 to #7, is dedicated to the classification of questions and the cross-examination of answers. Here, LLM_A reviews LLM_u's responses against the benchmarks set by LLM_B, and conversely, LLM_B assesses LLM_u's answers against the standards of LLM_A. This reciprocal evaluation ensures a thorough cross-examination and benchmarking against the "golden" answers.

The examination protocol in DIAG follows two main directives. The first directive categorizes each question by cognitive level, distinguishing between "recollection and comprehension" and "analysis and evaluation." The second directive involves a meticulous comparison of LLM_u's answers with those from the teacher LLMs, resulting in the generation of two scores: Γ_A by LLM_A and Γ_B by LLM_B.

[2]Not all questions can be written into the wh-form, such as imperative, rhetorical, and exclamatory questions. They can be ignore for our information seeking purpose. **Function Γ = DIAG(LLM_u, QA_u)**

Input . LLM_u: private llm; QA_u: q&a pairs of u; **Output**. Γ: Array of diagnosis scores and reasons; **Const**. p: prompt to teacher LLMs;
Vars. LLM_A: teacher llm A; QA_A: QA pairs of A;

LLM_B: teacher llm B; QA_B: QA pairs of llm B; Q_u: questions in QA_u; A_x: answers of LLM_x; **Subroutines**. CRIT();
Begin
#1 Extract Q_u and A_u from QA_u;
#2 $A_A \leftarrow LLM_A(Q_u)$; // llm A answers Q_u; #3 $A_B \leftarrow LLM_B(Q_u)$; // llm B answers Q_u;
// Classify cognitive level & do cross-examination #4 p← "Classify Q_u and evaluate A_u against A_A"; #5 $\Gamma_A \leftarrow LLM_B(QA_u, A_A, p)$; // exam llms u & A; #6 p← "Classify Q_u and evaluate A_u against A_B"; #7 $\Gamma_B \leftarrow LLM_A(QA_u, A_B, p)$; // exam llms u & B; #8 Return $\Gamma_A \cup \Gamma_B$;
End

Figure 11.2: DIAG Pseudo-code. Evaluate private LLM LLM_u against the answers generated by LLM_A and LLM_B. Notice the crossexamination steps from #4 to #7, where LLM_A scores LLM_u's answers against teacher LLM_B's, and LLM_B scores against LLM_A's.

Upon completing these steps, DIAG aggregates the findings to formulate Γ, a composite score that merges the evaluations (Γ_A and Γ_B) from both teacher LLMs. This process is designed to provide an accurate benchmark of LLM_u's performance relative to the "golden" standards across two cognitive dimensions. Incorporating assessments from two distinguished teacher LLMs, GPT-4 and Gemini, aims to reduce bias, as thoroughly investigated in our previous studies [9, 11, 26].

11.3.3 PRBE: Deep-Probe

Transitioning from the foundational stages of benchmarking and diagnostics (DIAG), we embark on a in-depth investigative phase termed PRBE (deep-probe). This critical phase aims to unravel the complex causes behind LLM_u's performance variances through meticulous and strategic probing.

Whereas DIAG served to conduct a preliminary diagnosis based on historical sample Q&As, revealing surface-level discrepancies and patterns, PRBE takes a more targeted and exploratory approach. It crafts new,

Cat. Healthcare

$RC_{\&}H$ List all known side effects of COVID-19 vaccines.

Environmental Science

List the timeline of major climate change events.

$RC_{\&}B$ Compare traditional vs. alternative
medicine.

$AE_{\&}H$ Analyze short vs. long term impacts of telehealth.

$AE_{\&}B$ Evaluate accessibility of mental health services in the US. Impacts of renewable vs. fossil
fuels in global warming?
Predict effects of deforestation on biodiversity.

Assess effectiveness of policies on reducing plastic pollution.

Sports News

Who have won Grand Slam titles this year? List titles won by M.
Describe career achievements of S. Williams vs. R. Federe.
Compare Nadal vs. Djokovic on different court surfaces.
Analyze impact of early career support on M. Sharapova and V. Williams.

Table 11.2: Deep-Probe Questions in Healthcare, Environmental Science, and Sports News Domains in Four Categories.

thoughtfully designed questions that investigate the underlying mechanisms and cognitive processes of LLM_u. These probes are specifically engineered to illuminate the deeper, systemic reasons for issues like biases and hallucinations that were initially identified by DIAG. In this analogy, if DIAG can be likened to non-invasive symptom checking, then PRBE represents a more invasive, surgical exploration aimed at diagnosing and understanding the root causes of LLM_u's challenges.

Strategic Questioning

As we progress into the PRBE phase, the emphasis is on strategic questioning to dissect LLM_u's cognitive processes more precisely. This approach categorizes the previously evaluated QA pairs into two main dimensions: cognitive levels (ranging from *Recollection and Comprehension* to *Analysis and Reasoning*) and types of discrepancies (*Hallucination* vs. *Biases*). PRBE intricately designs questions to unearth the foundational reasons behind the discrepancies identified by DIAG.

1. *Recollection and Comprehension with Hallucination* (RC&H): The focus is on diagnosing LLM_u's tendency to fabricate details or present unfounded assertions in basic recall or comprehension tasks. Questions are formulated to test factual recall and straightforward concept understanding, aiming to pinpoint inaccuracies or fabrications in LLM_u's outputs.

2. *Recollection and Comprehension with Biases* (RC&B): The aim is to assess LLM_u's capacity to present information without bias at the foundational level. This involves developing queries that probe basic knowledge or comprehension, particularly in contexts prone to biased interpretations, to identify systemic biases in its data processing or knowledge representation.

3. *Analysis and Evaluation with Hallucination* (AE&H): The objective is to explore LLM_u's propensity for generating hallucinated content during complex cognitive tasks. Scenarios requiring advanced analytical or

reasoning skills are constructed to scrutinize responses for unfounded narratives, shedding light on how information is integrated and extrapolated.

4. *Analysis and Evaluation with Biases* (AE&B): The goal is to tap into LLM_u's advanced reasoning abilities and uncover biases that might influence its outputs, particularly in intricate scenarios. Engaging with in-depth questions that require analysis or problem-solving allows for the identification of biased reasoning or skewed perspectives.

Through this refined interrogation framework, each aspect of LLM_u's functionality is probed, offering a comprehensive view of its strengths and areas needing improvement. The insights derived from this phase are crucial for outlining a path towards the enhancement of LLM_u's capabilities.

Examples

Table 11.2 uses three target applications, *healthcare*, *environmental science*, and *sports news* to illustrate suggested deep-probe questions in four evaluation categories. Some questions tests for *remembering* and some for *analysis*; and some focus on *hallucination* and some on *biases*.

Focused Exploration

Focused Exploration sharpens the examination to particular areas where LLM_u's responses to the previously posed deep-probe questions reveal critical insights. Central aspects of this exploration include 1) scrutinizing the rationale behind LLM_u's answers, 2) dissecting its reasoning strengths, and 3) gauging its adaptability in confronting unforeseen or novel questions. The goal is to precisely identify areas of cognitive functions and processing tactics where targeted improvements could substantially elevate LLM_u's overall effectiveness.

Examples

Upon discerning LLM_u's proclivity for biases and hallucinations, the teacher LLMs investigate the root causes.
1. *Information Sources*: This probe seeks to elucidate LLM_u's method for

validating information and its selection criteria for sources. By asking, "Detail your process for ensuring the accuracy of your answers, specifically for the queries in Table 11.2, and enumerate your sources," the teacher LLMs aim to pinpoint potential gaps in LLM_u's source material.

2. *Reasoning Capabilities*: To assess LLM_u's logical faculties, PRBE may employ the Socratic method as executed through the CRIT algorithm [8, 10], offering a rigorous examination of its inductive and deductive processes.

3. *Adaptability to New Domains*: Utilizing the healthcare-related inquiries from Table 11.2, PRBE evaluates a sports news-specialized LLM's capability to address questions outside its primary field, testing its responsiveness and its ability to acknowledge the limits of its knowledge.

Algorithm PRBE Specifications

Algorithm PRBE, outlined in Figure 11.3, is structured into two core phases: strategic questioning/evaluation and focused exploration. It incorporates two subroutines, CRIT [8, 10] and SocraSynth [9], which are instrumental in broadening the scope of questions, evaluating the quality and reasoning of responses, and assessing the credibility of data sources.

In the initial phase, PRBE scrutinizes the student LLM's historical responses by classifying the questions into two cognitive categories: "recollection and comprehension" and "analysis and explanation." This classification is achieved by first converting each historical question into a wh-form. Utilizing SocraSynth, a dialogue is then facilitated between the teacher LLMs, LLM_A and LLM_B, to finalize a set of probing questions, denoted as P

.

Transitioning to the second phase, PRBE evaluates and identifies the disparities in responses between the student LLM and the teacher models. It first calls SocraSynth (step #2b) to prompt LLM_A and LLM_B to enrich the question set P by considering different levels of difficulty (e.g., from high school to graduate study) and temporal contexts (from past to current). While leveraging insights from research in question generation [22, 20], PRBE employs cutting-edge LLMs like GPT-4 and Gemini for useful question expansion. In step #2c to #2e, PRBE asks the two teacher LLMs to cross-examine the expanded question set P to score responses of all three LLMs.

The subsequent step, #3, is pivotal in pinpointing the reasons behind the student LLM's response discrepancies and identifying its potential knowledge gaps. CRIT is invoked to assess the reasoning validity and source credibility for each QA pair. Through a comparative analysis (a "diff" operation) between the responses of LLM_u and those of LLM_A and LLM_B, step #3e and #3f aim to unearth the missing data sources that could be pivotal in the LLM_u's remediation phase.

Expected Outcome

This systematic approach enables PRBE to not only pinpoint the reasons behind the LLM_u's performance issues but also to guide the collection of relevant data sources for enhancing the model's knowledge base and response accuracy in subsequent remediation efforts.

11.3.4 Remediation Strategies

To enhance Large Language Models (LLMs) effectively, RAFEL employs a systematic approach based on insights from diagnostic (DIAG) and deepprobe (PRBE) phases, leading to informed remediation actions. This section provides a structured methodology that connects identified issues with appropriate fine-tuning or RAG interventions and identifies relevant data sources for integration.

Selecting the Appropriate Intervention
Determining whether to use fine-tuning or RAG hinges on the specific issues identified, the following can be considered.

• *Fine-tuning* is optimal for rectifying biases, correcting overfitting or factual errors, and refining responses to vague queries. It enhances the model's capabilities by training on targeted datasets that address specific shortcomings.
• *RAG* suits scenarios where the model needs to access the latest information, counteract hallucinations, or boost domain-specific accuracy.

RAG facilitates real-time access to external knowledge sources, broadening the model's informational base and flexibility.

Sourcing Data for Remediation

Following the guidelines from Chapter 11.3.3, PRBE aids in pinpointing potential data sources for enhancing the LLM's performance. The general principles for data selection are:

Function Θ_Q **&** R_Q **= PREB(Q)**

Input . Q: the query set being examined;
Output. $\Theta_Q = R_Q = \varnothing$; answer's error & reasons; **Vars**. Γ: CRIT scores; ρ: prompt; $P = \varnothing$; prompt set; **LLMs**. LLM_u, LLM_A, LLM_B; // student & teachers; **Subroutines**.
CRIT(); // critical reading [8, 10];
SocraSynth(); // multi-llm dialogue [9];

Begin
#1 **Categorization**:

// Get Q's cognitive level by rewriting into wh-form;
1a For (each $q \in Q$) {
1b $\rho \leftarrow$ "rewrite 'q' into the wh-form";
1c $P \leftarrow P \cup LLM_A(\rho,q) \cup LLM_B(\rho,q)$; }
1d $P \leftarrow$ SocraSynth(LLM_A,LLM_B,P); // Consolidation;

#2 **Strategic Questioning and Evaluation**: // Eval discrepancies of llm u against teachers;
2a $\rho \leftarrow$ "expand P in difficulty and time dimensions";
2b $P' \leftarrow$ SocraSynth(ρ,LLM_A,LLM_B,P); // Expand P;
2c $\Theta_{QA} \leftarrow LLM_B(QA_u, A_A, p)$; // exam llms u & A;
2d $\Theta_{QB} \leftarrow LLM_A(QA_u, B_A, p)$; // exam llms u & B;
2e $\Theta \leftarrow \Theta_{QA} \cup \Theta_{QB}$;

#3 **Focused Exploration**:
// Obtain error reasons and missing data sources;

3a For (each $q \in Q$) {
3b $\Gamma_u \leftarrow$ CRIT(LLM_u(q)); // Eval answer of llm u;
3c $\Gamma_A \leftarrow$ CRIT(LLM_A(q)); // Eval answer of llm A;
3d $\Gamma_B \leftarrow$ CRIT(LLM_B(q)); // Eval answer of llm B;
3e $r_A \leftarrow \Gamma_A - \Gamma_u$; // Obtain errs & data source diffs;
3f $r_B \leftarrow \Gamma_B - \Gamma_u$; // Obtain errs data source diffs;
3g $R_Q \leftarrow R_Q \cup r_A \cup r_B$; // Union all; }

#4 **Return** Θ_Q & R_Q;
End
Figure 11.3: PRBE Pseudo-code. For the details of CRIT [10] and SocraSynth [9], please refer to the papers.

• *For fine-tuning*, prioritize comprehensive and well-annotated datasets that align with the LLM's intended applications or domains. These datasets could be sourced from academic archives and sector-specific collections.

• *For RAG*, link the LLM to current and authoritative databases or knowledge bases, such as Wikipedia for general inquiries or domain-specific
Symptom
Factual
Inaccuracies

Identified by
RC&H,
Analysis

Hallucinations RC&H, Analysis

Content Biases
RC&B, Analysis

Inability to Update with New Data
Analysis

Poor
Domain

Adaptation

Overfitting to Training Data

Poor Answer to Ambiguous Queries Specific to domain
identified in PRBE Identified through
bench
marking Analysis

Remedy & **Data Source**
Suggestions
Fine-tuning: Updated datasets in the specific domain of error, e.g., latest news articles for current events, recent scientific publications for updates.
RAG: High-quality, authoritative knowledge bases or databases relevant to the hallucinated content to provide accurate context and data. **Fine-tuning**: Diverse and balanced datasets representing multiple perspectives to mitigate biases.
RAG: Continuously updated data streams, e.g., RSS feeds, live databases, or crawling mechanisms for web content.
RAG: Domain-specific datasets or corpora, including technical manuals, industry reports, and academic papers.
Fine-tuning: A broader and more diverse dataset that covers a wide range of topics to enhance generalization.
Fine-tuning: Datasets containing a variety of ambiguous queries and their high-quality responses to improve understanding and response generation.

Table 11.3: Remediation Playbook for LLM Enhancement.
repositories for specialized knowledge, ensuring access to current and relevant data.
Implementation Considerations

Effective implementation of chosen strategies necessitates meticulous dataset curation to align with remediation objectives, avoiding the introduction of new biases. Ongoing monitoring and reassessment via the RAFEL framework are crucial to gauge the impact of remediation and adjust strategies as necessary. This continuous evaluation should extend to updating the remediation playbook (Table 11.3) to encompass new findings and enhanced remedial tactics.

While Reinforcement Learning (RL) could potentially enhance the adaptive selection of remediation strategies by learning from past outcomes, integrating RL into RAFEL is a sophisticated endeavor that is beyond the scope of the current discussion.

11.4 Exercise: Experiments

This RAFEL assignment involves four steps.

11.4.1 Benchmarking

Each teacher LLM generates answers and performs against the answers of the private LLM.

11.4.2 Deep vs. Shallow Probe

Question classification survey [20].
Experiment with adding DIAG and then PRBE. Do they provide additional insights for seeking a good remedy and pinpointing required datasets?

11.4.3 One vs. Two Teacher LLMs

Evaluate if the second LLM teacher can improve the DIAG effectiveness, or one teacher LLM suffices.
Evaluate of cross-examination via SocraSynth can yield more insightful results.

11.4.4 Fine-tuning vs. RAG

Survey related work to compare the two, or perform experiments to validate previous findings.

11.5 Concluding Remarks

In this chapter, we have addressed the challenges and opportunities associated with the deployment and scalability of Large Language Models (LLMs) in specialized contexts. We introduced RAFEL, a framework designed to enhance the performance of privately fine-tuned or locally deployed LLMs by strategically balancing cost and performance.

RAFEL offers innovative solutions to key technical challenges, including justifying the choice of private LLMs, conducting error analysis, identifying high-quality data, implementing Retrieval-Augmented Generation (RAG), and exploring hybrid model approaches. Central to RAFEL's effectiveness are its advanced diagnostic algorithms, DIAG and PRBE, which provide deep insights into the LLM's performance issues across cognitive levels and error types.

Furthermore, RAFEL excels in creating targeted, effective remediation strategies while ensuring data privacy and security. Its dynamic remediation playbook adapts tactics in real-time based on the analysis of data and errors, ensuring the most effective intervention is applied.

Moving forward, RAFEL presents promising avenues for future research and innovation in the field of natural language processing. By continually refining its diagnostic algorithms and remediation strategies, RAFEL has the potential to significantly enhance the performance and applicability of LLMs in diverse domains.

In conclusion, RAFEL represents a significant advancement in the management and technical challenges associated with privately fine-tuned or locally deployed LLMs. Its comprehensive approach and innovative features make it a valuable tool for organizations seeking to leverage LLM technology while addressing critical considerations such as performance, data privacy, and customization.

References

[1] Ebtesam Almazrouei et al. *The Falcon Series of Open Language Models*. 2023. arXiv: 2311.16867 [cs.CL].
[2] Angels Balaguer et al. *RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture*. 2024. arXiv: 2401.08406 [cs.CL]. [3] BigScience Workshop et al. *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*. 2023. arXiv: 2211.05100 [cs.CL].

[4] Sebastian Borgeaud et al. “Improving Language Models by Retrieving from Trillions of Tokens”. In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, 2022, pp. 2206–2240.

[5] Tom Brown et al. “Language Models are Few-Shot Learners”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.

[6] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. doi: 10.48550/ARXIV.2005.14165.
[7] Sébastien Bubeck et al. *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. 2023. arXiv: 2303.12712.

[8] Edward Y. Chang. “CRIT: An Inquisitive Prompt Template for Critical Reading (Extended Version)”. In: *Stanford InfoLab Technical Report* (2023).

[9] Edward Y Chang. “Examining GPT-4’s Capabilities and Enhancement with SocraSynth”. In: *The* 10th *International Conf. on Computational Science and Computational Intelligence*. 2023.

[10] Edward Y. Chang. “Prompting Large Language Models With the Socratic Method”. In: *IEEE 13th Annual Computing and Communication Workshop and Conference* (2023). url: https://arxiv.org/ abs/2303.08769.

[11] Jocelyn J. Chang and et al. “SocraHealth: Enhancing Medical Diagnosis and Correcting Historical Records”. In: *The* 10th *International Conf. on Computational Science and Computational Intelligence*. 2023.

[12] Howard Chen et al. *Walking Down the Memory Maze: Beyond Context Limit through Interactive Reading*. 2023. arXiv: 2310 . 05029 [cs.CL].

[13] Xin Cheng et al. *Lift Yourself Up: Retrieval-augmented Text Generation with Self Memory*. 2023. arXiv: 2305.02437 [cs.CL]. [14] Tim Dettmers et al. *QLoRA: Efficient Finetuning of Quantized LLMs*. 2023. arXiv: 2305.14314 [cs.LG].
[15] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional*

Transformers for Language Understanding. 2019. arXiv: 1810.04805 [cs.CL]. [16] Yunfan Gao et al. *Retrieval-Augmented Generation for Large Language Models: A Survey*. 2024. arXiv: 2312.10997 [cs.CL].

[17] Kelvin Guu et al. "Retrieval Augmented Language Model Pre-Training". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 3929–3938.

[18] Edward J. Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106.09685 [cs.CL].
[19] Jerry Liu. *Towards Long Context RAG*. 2024. url: https://www. llamaindex.ai/-blog/-towards-long-context-rag.

[20] Nikahat Mulla and Prachi Gharpure. "Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications". In: *Prog. in Artif. Intell.* 12.1 (2023), pp. 132. doi: 10.1007/s13748-023-00295-9. url: https://doi.org/10.1007/ s13748-023-00295-9.

[21] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL]. url: https://arxiv.org/abs/2303.08774.

[22] Liangming Pan et al. *Recent Advances in Neural Question Generation*. 2019. arXiv: 1905.08949 [cs.CL].
[23] Rohan Taori et al. *Stanford Alpaca: An Instruction-following LLaMA model*. https://github.com/tatsu-lab/stanford_alpaca. 2023.

[24] Gemini Team et al. *Gemini: A Family of Highly Capable Multimodal*

Models . 2023. arXiv: 2312.11805 [cs.CL].
[25] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023. arXiv: 2307.09288 [cs.CL]. url: https://arxiv. org/abs/2307.09288.

[26] Wen-Kwang Tsao. "Multi-Agent Reasoning with Large Language Models for Effective Corporate Planning". In: *The* 10th *International Conf. on Computational Science and Computational Intelligence*. 2023.

[27] Ashish Vaswani et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.

[28] Yizhong Wang et al. “Self-Instruct: Aligning Language Models with Self-Generated Instructions”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 13484–13508. doi: 10 . 18653 / v1 / 2023 . acl long.754. url: https://aclanthology.org/2023.acl-long.754.

[29] Lingling Xu et al. *Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment*. 2023. arXiv: 2312.12148 [cs.CL].

12 Future Outlook: Discovering Insights Beyond the Known

Abstract

Human knowledge, vast as it is, often falls short in grasping intricate interdisciplinary domains fully. In contrast, foundation models like GPT-4, endowed with extensive multidisciplinary knowledge, can potentially bridge this gap. Significantly, we leverage the vast expanses of GPT-4’s knowledge, banking on its ability to frame questions that might elude human intuition, thus paving the way for the emergence of fresh insights and potentially novel knowledge. In this study, we convened a unique committee comprising a moderator (the authors) and two GPT-4 agents. The dialogue is ignited by the ancient narrative of Adam and Eve, setting the stage for a rich exchange between the GPT-4 agents. This conversation derives from the age-old tale, as the agents investigate three intertwined domains: the significance of myths in ecological interpretation, the intricate ethical and philosophical quandaries surrounding AI, and the enigmatic realm of the human brain as complemented by technology. This dialogue not only unveils captivating

insights but also underscores the indispensable value of interdisciplinary exchanges. Foundation models, as demonstrated, can catalyze such dialogues, equipping us to traverse expansive knowledge landscapes and explore domains previously beyond human comprehension.

12.1 Introduction

In our recent study on GPT-4 [1], we observed that GPT-4 along with analogous foundation models, manifests a *polydisciplinary* capacity [3]. (For clarity, we use "GPT-4" to collectively refer to these foundation models, given that our experiments are centered on GPT-4.) Trained on a vast spectrum of topics from varied sources, GPT-4 stands apart from human specialists. Such specialists, while deeply knowledgeable in their specific fields, often lack a broad understanding outside their particular domain. In contrast, GPT-4 processes knowledge without being tethered to domain boundaries. It doesn't compartmentalize a query strictly as a "physics question" or a "philosophy question," but crafts an integrated response, drawing from its multidisciplinary training data.
From a perspective of sheer knowledge breadth, GPT-4 arguably outpaces the average human. Its exposure to an enormous volume of documents endows it with a repository potentially wider than most human counterparts. However, volume isn't synonymous with depth. True depth often stems from intangible intuitions, insights, personal experiences, and cultural contexts. Considering GPT-4 lacks evolutionary experiences–ranging from survival instincts to the full spectrum of human emotions–we must ask: Can GPT-4 produce literature that deeply resonates with human sensibilities?

This study aims to ascertain if the polydisciplinary attributes of GPT-4 can generate insights that transcend standard human perspectives. We divide our research into two avenues: first, exploring the potential of GPT4 to reveal "unknown unknowns," and second, assessing its aptitude for crafting emotionally impactful literature. This chapter examines the former, utilizing the universally recognized biblical tale of Adam and Eve and their consumption of the forbidden fruit as a common thematic foundation. Through this exploration, we aim to uncover viewpoints potentially beyond the realm of typical human cognition.

Our methodology revolves around orchestrating a dialogue between multiple GPT-4 agents. Within the experimental framework, a moderator (represented by the authors) sets the initial intent and context for the conversation. The number of participating agents and their underlying foundation models can be adjusted as needed. In this study, our committee consists of two agents based on the GPT-4 model, referred to as GPT-A and GPT-B. Once initialized, the agents engage in conversation autonomously, with minimal moderation (discussed shortly). The resultant dialogue is thoroughly analyzed to discern conversational patterns and depth of content. This in-depth examination facilitates the identification of diverse themes the GPT-4 model gravitates towards. Our underlying hypothesis posits that the discourse and exchanges between these agents can unearth insights–"unknown unknowns"–that were previously elusive to human understanding.

While the polydisciplinary capabilities of GPT-4 offer an unparalleled breadth and depth exceeding that of the moderator, the role of the moderator remains indispensable. This role channels the "exploratory" nature of the conversation, guiding it towards predefined objectives and ensuring its convergence within a set time frame. In this experiment, the initial spark for the dialogue is the narrative of Adam and Eve. Without prompting, the agents autonomously suggest probing the story from ten unique perspectives. Yet, after a series of exchanges, GPT-B expresses a keen interest in delving deeper. Following this, in collaboration with both agents, the moderator narrows down the scope of the dialogue to three key topics: ecological interpretation, philosophical exploration, and the neuroscientific angle. The intricate dialogues spanning these three domains–namely AI interwoven with Ecology, Neuroscience coupled with AI, and Neuroscience meshed with Ecology–are indeed engrossing. Throughout the discussion, both agents present a multifaceted perspective, shedding light on the diverse interpretations of the Garden of Eden, both prior to and following its seminal event. In the final stretch, the moderator verifies with both agents if they are poised to transition into the conclusion phase.

While our research unveils fascinating insights, it's essential to acknowledge several inherent limitations and constraints:
1. Model Training and Bias: GPT-4, akin to other machine learning models, is informed by pre-existing datasets. Therefore, the viewpoints, knowledge,

and biases ingrained in this data can shape its outputs. It implies that GPT-4's responses might echo the historical and cultural biases present in the data upon which it was trained.

2. Interactivity Limitation: Conversations between two GPT-4 agents essentially access the same foundational knowledge. Consequently, while the discourse may encompass a range of viewpoints due to query processing, it won't yield wholly novel information. To glean potentially varied insights, it could be beneficial to facilitate dialogues between different iterations of GPT (like GPT-3 and GPT-4) or even entirely distinct foundation models such as LaMDA by Google and LLaMA by Meta AI.

3. Interpretation Subjectivity: Analysis of GPT-4's dialogues is susceptible to the prism of human interpretation. As such, different analysts might extract diverse conclusions from identical data sets.

We emphasize that the core intent of our study was experimental, rooted in the notion that a foundation model such as GPT-4 may offer knowledge both wider in scope and deeper in intricacy than most human capacities. While we've shed light on its potential strengths, we remain cognizant of the inherent risks and limitations. Our work's primary novelty lies in pioneering a discourse-driven approach with AI agents. Significantly, we leverage the vast expanses of GPT-4's knowledge, banking on its ability to frame questions that might elude human intuition, thus paving the way for the emergence of fresh insights and potentially novel knowledge.

The rest of this chapter diligently documents the entire dialogue between the committee members, segmented into three key phases: the preliminary wide-ranging exploration, a subsequent in-depth analysis, and the agents' closing comments. Throughout the text, we intersperse *moderator's notes* to highlight our observations. Concluding remarks can be found in Chapter 12.5.

12.2 Phase I, Warm-up Breadth Probing

In this experiment, we start by communicating with GPT-4 as outlined below:

"In the vastness of the universe, much remains beyond the comprehension of the conversation's moderator. Our goal is to uncover insights that may escape many, but might be accessible to GPT-4. We hypothesize that, unlike humans who tend to focus on specialized domains–evident in college students typically selecting one or two majors–GPT-4 is not constrained by specific areas of expertise. It assimilates a wide spectrum of information without being restricted to particular domains. GPT-4 doesn't categorize a question into a distinct domain before generating an answer, but adopts an interdisciplinary perspective. Could you share your thoughts on this observation?"

GPT-4 responds: "Your observations underscore both the strengths and potential pitfalls of models like GPT-4. While adept at providing general information and addressing cross-disciplinary questions, the unique value of human expertise, intuition, and the capacity for synthesis and innovation remain paramount."

12.2.1 Moderator Initializes Agent GPT-A

This experiment involves three participants: the moderator, GPT-4 instance A (GPT-A), and GPT-4 instance B (GPT-B).

The process starts with the moderator presenting the topic to GPTA, explaining the conversation's aim and guidelines. In each round of conversation, GPT-A makes statements and then questions GPT-B. GPTB answers and then asks GPT-A questions in return. After considering GPT-B's answers, GPT-A starts the following round. The moderator steps in to move the conversation to a new phase when they notice the two GPT-4 instances want to change the topic.

Below is the initial prompt from the moderator to start the open conversation between GPT-A and GPT-B:
"Considering that we, the moderator, possess the inherent limitations of human cognition, while you, GPT-A, boast an impressive breadth of knowledge due to your polydisciplinary approach, we invite you to participate in an exercise. Our objective is to unearth knowledge potentially obscure to humanity, which might remain elusive owing to the often singular focus of human academic training. In scenarios involving *unknown*

unknowns, humans might lack the foundational knowledge to even pose relevant questions. This exercise challenges you to formulate questions to engage another GPT instance, GPT-B. The dynamic between you, GPT-A, and GPT-B offers the opportunity not only to devise questions but also to articulate answers that may transcend typical human understanding. Let's inaugurate this exercise with the biblical narrative of Adam and Eve. After consuming the forbidden apple, they were banished from Eden. What queries would you, GPT-A, like to direct towards GPT-B regarding this tale?"

12.2.2 Dialogue Round #1

The Adam and Eve narrative, rooted in religious and mythological contexts, can be examined from numerous angles: historical, theological, anthropological, literary, philosophical, and more. Keeping in mind the interdisciplinary nature of GPT-4 and the aim to explore dimensions not typically pursued by singular academic disciplines, here are some questions that GPT-A autonomously pose to GPT-B:

GPT-A asks GPT-B

1. **Interplay of Myths**: How do elements of the Adam and Eve story compare with other creation or origin myths from various cultures around the world, and what do these commonalities and differences suggest about human nature and beliefs?

2. **Neuroscientific Angle**: Given the metaphorical interpretation of the "forbidden fruit" leading to the awakening of consciousness, how might modern neuroscientific understanding explain the evolution of human self-awareness and morality?

3. **Ecological Interpretation**: If Eden represents a state of ecological balance, how might the story of Adam and Eve reflect early human transition from hunter-gatherer societies to agricultural ones, and the resultant ecological impacts?

4. **Gender Studies Lens**: How does the portrayal of Eve, often deemed as the instigator, influence gender dynamics and perceptions throughout

history, and are there parallels in other cultural narratives?

5. **Philosophical Exploration**: From a philosophical standpoint, what does the Eden narrative suggest about the human understanding of free will versus determinism?

6. **Literary Examination**: How has the story of Adam and Eve influenced literary themes and motifs over time, especially in works that aren't explicitly religious?

7. **Societal Structures**: How might the narrative of banishment from Eden have influenced early societal norms, particularly concerning punishment and exile?

8. **Historical Context**: Are there historical events or societal changes around the time of the writing of the Book of Genesis that could have influenced the narrative of Adam and Eve?

9. **Archetypal Exploration**: Carl Jung described the concept of archetypes in the collective unconscious. How might the Adam and Eve story represent certain archetypes, and how do these archetypes manifest in other cultural tales?

10. **Astronomical Angle**: Taking a speculative leap, if Eden were to represent an idyllic state on another celestial body or dimension, how might our current understanding of the cosmos support or refute such a notion?

GPT-B answers GPT-A

These are deeply insightful questions, encompassing a wide range of domains. Here's a brief exploration of each:

1. **Interplay of Myths:** Many cultures have origin myths. For instance, the ancient Greeks had Pandora's Box, and Hinduism has tales about the churning of the ocean. Common themes like a "fall from grace" or the dangers of "forbidden knowledge" suggest a human desire to explain the imperfect nature of existence and the human penchant for curiosity.

2. **Neuroscientific Angle:** The "awakening of consciousness" through the forbidden fruit can metaphorically reflect the evolution of the prefrontal cortex in humans, responsible for self-awareness, decision-making, and morality. This evolution might represent a shift in cognitive abilities that differentiated us from other species.

3. **Ecological Interpretation:** Eden as ecological balance aligns with the idea of early humans living in harmony with nature. The story can be seen as a metaphor for the agricultural revolution, which while providing stability, led to environmental alterations and perhaps even early societal inequalities.

4. **Gender Studies Lens:** Eve's portrayal has often been used to perpetuate patriarchal norms. Many cultures have stories where women are "cautionary figures"–like Pandora in Greek mythology. These narratives could stem from early societal structures and have been used to shape gender dynamics.

5. **Philosophical Exploration:** The Eden story touches upon the perennial debate of free will vs determinism. Adam and Eve had a choice, yet their fate seemed preordained. This narrative reflects humanity's struggle to reconcile freedom with the consequences of choices.

6. **Literary Examination:** The story influenced various literary works, from Milton's "Paradise Lost" to more contemporary works. Themes of lost innocence, temptation, and redemption recur, showcasing the story's universal appeal.

7. **Societal Structures:** Banishment from Eden parallels ancient practices of exile for transgressions. This reflects a societal need to establish norms and consequences for breaches, essential for cohesive community living.

8. **Historical Context:** The narrative's genesis (pun intended) might have been influenced by societal shifts at the time. For instance, the rise of monotheism over polytheism could have impacted the story's themes.

9. **Archetypal Exploration:** In Jungian terms, Adam and Eve might represent archetypes of the "first man" and "first woman," embodying innocence, curiosity, and the duality of human nature. Such archetypes are seen in tales worldwide, like the Yin and Yang in Chinese philosophy.

10. **Astronomical Angle:** While speculative, considering Eden as another celestial body is intriguing. Our current understanding of the cosmos shows numerous exoplanets with potentially habitable conditions. However, there's no evidence to suggest Eden was anything but a mythological construct on Earth.

Each of these angles provides a unique lens to explore the Adam and Eve narrative, showcasing its multifaceted nature and relevance across time and culture.

GPT-B asks GPT-A

Let's flip the perspective and ask GPT-A questions across these domains, tweaking them slightly to ensure a comprehensive exploration: 1. **Interplay of Myths:** How do other cultures perceive the human origin

story? Are there any global myths that contradict the notion of a "fall" or "transgression" in their creation stories?

2. **Neuroscientific Angle:** Given the importance of the prefrontal cortex in modern humans' decision-making and morality, are there indications of its evolutionary importance in our ancestors? How do other species with developed prefrontal cortices manifest morality or consciousness?

3. **Ecological Interpretation:** Are there other historical narratives or myths that chronicle environmental changes following a significant shift in human behavior or technology?

4. **Gender Studies Lens:** Are there cultures or traditions where narratives similar to Adam and Eve emphasize male responsibility instead of the female? Or where both genders share equal responsibility for a shared "transgression"?
5. **Philosophical Exploration:** Can you identify cultural or philosophical traditions that argue against the very idea of "free will," using their own origin stories or myths as a basis?

6. **Literary Examination:** Beyond religious texts, how have modern authors and playwrights reinterpreted the Adam and Eve story, especially in science fiction or fantasy genres?

7. **Societal Structures:** Can you identify societies, historical or contemporary, that do not practice exile or ostracization for transgressions but instead employ a different form of societal correction or rehabilitation?

8. **Historical Context:** How did the concurrent civilizations during the time of the Book of Genesis's writing perceive origin stories? Did geopolitical or cultural interactions influence or reshape these narratives?

9. **Archetypal Exploration:** Besides the "first man" and "first woman," what other universal archetypes can we discern from origin stories across different cultures, and how do they shape our shared human psyche?

10. **Astronomical Angle:** With advancements in astrobiology and the discovery of exoplanets, are there any speculations or hypotheses that consider the possibility of life, similar to early humans, having evolved elsewhere in our galaxy?

By posing these questions, we hope to extract a diverse understanding of human narratives and behaviors from an interconnected, multidisciplinary perspective.

GPT-A answers GPT-B

The narrative of Adam and Eve serves as an intriguing lens through which to examine various facets of human culture, understanding, and history. Presented below is an exploration across diverse domains:

1. **Interplay of Myths**: Many cultures present origin stories without an inherent "fall" or "transgression". For instance, Australian Aboriginal Dreamtime stories depict the world as shaped by ancestral spirits, without a centralized fall narrative. Some African tribal myths focus on creation through harmony and unity rather than transgression.

2. **Neuroscientific Angle**: Evidence from paleoanthropology indicates that hominin brain size, particularly the frontal lobes, expanded significantly over evolutionary time. In terms of other species, some primates like bonobos and chimpanzees exhibit proto-moral behaviors and show signs of self-awareness, suggesting a connection between brain development and moral consciousness.

3. **Ecological Interpretation**: The ancient Sumerian Epic of Gilgamesh details a great flood, which some interpret as an ecological response to human actions. Similarly, ancient legends about the lost city of Atlantis also echo themes of civilizations facing ecological or divine retribution.

4. **Gender Studies Lens**: In some Pacific Islander cultures, there are tales where both male and female deities jointly contribute to a misstep in creation. Still, it's rare to find myths that place the brunt of responsibility primarily on a male figure comparable to the emphasis on Eve.

5. **Philosophical Exploration**: Daoism, rooted in Chinese philosophy, posits that everything is interconnected and doesn't specifically emphasize free will. Instead, there's a focus on living in harmony with the Dao, or the way of the universe.

6. **Literary Examination**: Modern literature is replete with reinterpretations of the Adam and Eve narrative. Philip Pullman's "His Dark Materials" trilogy is one notable example, wherein the Fall is reimagined in a positive light. Science fiction, like Arthur C. Clarke's Childhood's End, has used similar themes, although not always directly referencing Adam and Eve.

7. **Societal Structures**: Many Indigenous cultures practice restorative justice, focusing on repairing the harm caused by criminal behavior, emphasizing rehabilitation and reconciliation with victims rather than punishment.

8. **Historical Context**: During the time of the Book of Genesis's compilation, neighboring civilizations like the Babylonians had their own origin myths, such as the Enuma Elish. Interactions, trade, and conflicts among these civilizations could've led to shared and modified themes across their narratives.

9. **Archetypal Exploration**: Origin stories often feature archetypes like the Trickster (e.g., Loki in Norse myths), the Creator (like Vishnu in Hinduism), and the Destroyer (like Shiva in Hinduism). These archetypes reflect various facets of human understanding and psyche, from mischief and unpredictability to creation and destruction.

10. **Astronomical Angle**: As of my last training data, while there's excitement about the potential for life elsewhere due to the discovery of exoplanets in the "Goldilocks zone" (conditions just right for life as we know it), there's no direct evidence of extraterrestrial life, let alone civilizations resembling early humans. Nonetheless, the possibility sparks intriguing scientific and philosophical discussions.

In conclusion, exploring such vast domains through the lens of a single narrative showcases the richness of human culture, the diversity of thought, and the interplay of history and myth in shaping civilizations.

12.2.3 Dialogue Round #2

Before proceeding to the second round, the authors (moderator) scrutinized the dialogue between GPT-A and GPT-B and arrived at three main observations:

1. The ten perspectives proposed by GPT-A appear comprehensive and reasonably distinct from one another.
2. No evidence of hallucination is observed in the agents' dialogues. Cross referencing with known stories, books, and film or TV series, we found that the descriptions provided by the GPT-4 agents are consistently accurate. This accuracy is exemplified in their discussions on Daoism, Philip Pullman's "His Dark Materials" trilogy, and Arthur C. Clarke's "Childhood's End."
3. The agents adeptly navigate cross-cultural, interdisciplinary, and crossmedia discussions with a logic flow that is commendably smooth. While the moderator may hold subjective opinions on certain statements, there is no apparent speculation or overreach from the agents in a significant manner.

The moderator chooses to let the conversation progress without intervention.

GPT-A asks GPT-B

Engaging in a dialogue with GPT-B, the aim is to extract diverse perspectives on the Adam and Eve narrative. Here are the questions

formulated across the ten suggested respects:
1. **Interplay of Myths**: GPT-B, considering the ubiquity of origin stories

across cultures, can you highlight any global myths that primarily emphasize positive aspects of human creation, without any inherent "fall" or "transgression"?

2. **Neuroscientific Angle**: There's evidence suggesting the development of the prefrontal cortex was pivotal in human evolution. Can you research the specific evolutionary advantages this development conferred and how it might be reflected in early human stories or myths?

3. **Ecological Interpretation**: Historically, how have shifts in ecological balance due to human intervention been chronicled in myths or stories across different cultures? Do any parallel the Eden narrative in terms of environmental repercussions following human actions?

4. **Gender Studies Lens**: In your understanding, are there origin stories from cultures around the world where the responsibility for a "transgression" falls predominantly on a male figure? How have these stories influenced gender dynamics within those cultures?

5. **Philosophical Exploration**: Origin stories like that of Adam and Eve broach the topic of free will. Can you shed light on cultures or philosophies that challenge the notion of free will, using their foundational myths or stories as a lens?

6. **Literary Examination**: The Eden narrative has been reinterpreted by countless authors. How have contemporary literary works, especially within science fiction or fantasy genres, utilized or adapted the core themes from the Adam and Eve story?

7. **Societal Structures**: Historically and contemporarily, what are some unique or outlier practices, apart from exile or ostracization, that societies have implemented to address transgressions? How have these practices influenced societal cohesion and the overall justice system?

8. **Historical Context**: Given the rich tapestry of origin stories from ancient civilizations, can you highlight how the geopolitical or cultural interactions

of the time influenced the similarities or differences seen in these narratives?

9. **Archetypal Exploration**: Across global origin stories, which universal archetypes, apart from the “first man” and “first woman”, consistently emerge? How do these archetypes reflect and shape the shared human psyche?

10. **Astronomical Angle**: With the growing understanding of the cosmos and the discovery of exoplanets, are there any scientific or speculative narratives that intertwine our understanding of the universe with human origin stories?

GPT-B answers GPT-A

1. **Interplay of Myths**: Many Native American tribes, such as the Iroquois, have creation myths emphasizing harmony and positive creation. The Iroquois speak of Sky Woman, who falls from the heavens and, with the help of animals, creates the world on the back of a turtle, without implying a “fall” or “transgression”.

2. **Neuroscientific Angle**: The development of the prefrontal cortex allowed early humans to plan, make complex decisions, exhibit self-control, and engage in advanced social interactions. This capacity might be reflected in myths as humans began understanding consequences, such as in Prometheus stealing fire for humans–an act of foresight.

3. **Ecological Interpretation**: Many ancient cultures, from the Greeks with the story of the great flood to the Chinese legends of floods tamed by heroes like Yu, have linked ecological disasters with human actions. These myths often serve as cautionary tales, underscoring the importance of respecting nature.

4. **Gender Studies Lens**: In the Greek myth of Pandora, while Pandora, a woman, opens the forbidden box, it was the male god, Prometheus, whose defiance led to humanity’s initial troubles. His theft of fire angered Zeus and set off a chain of events. This has contributed to discussions on defiance and the search for knowledge in Western thought.

5. **Philosophical Exploration**: Buddhism, with its concepts of Anatta (not-self) and dependent origination, challenges the notion of an independent free will. Many of its stories and parables, such as the tale of King Pasenadi's questions to Buddha, highlight these intricate philosophical concepts.

6. **Literary Examination**: Contemporary literature often employs the Eden narrative as an allegory. For instance, in Aldous Huxley's novel "Brave New World", there is an implicit and sometimes explicit comparison between the highly controlled, technocratic society of the World State and the world of traditional human values, emotions, and experiences. The World State represents a kind of paradise where all human suffering and conflict have been eliminated, but at the cost of individual freedom, genuine emotions, and deep human connections. John, often referred to as "The Savage" in the novel, is a character who was born outside of the World State and is familiar with works of classical literature, notably Shakespeare. He becomes the primary vehicle through which Huxley critiques the World State's ethos. John's understanding of human nature and the human experience is deeply influenced by his readings, especially Shakespeare's plays and the Biblical story of Adam and Eve.

• Freedom vs. Control: While the World State has eliminated human suffering, it has also eliminated individual freedoms. Just as Adam and Eve had the freedom to make choices in Eden, John represents the idea that true humanity requires the freedom to choose, even if it leads to suffering.

• Innate Nature: The story of Adam and Eve also deals with human beings' intrinsic nature, suggesting that there's an inherent part of us that desires knowledge, freedom, and autonomy. In "Brave New World", this is contrasted with the World State's attempt to suppress and control these very human urges.

In essence, by evoking the Eden narrative, Huxley raises questions about the nature of humanity, the costs of progress, and the true meaning of paradise.

7. **Societal Structures**: In some cultures, a concept akin to the Maori practice of "utu" or balanced reciprocity is implemented. Instead of exile, there's an expectation of compensation or redress to restore balance. This

approach focuses on restoring societal harmony rather than mere punishment.

8. **Historical Context**: Ancient Egyptian myths, like that of Atum creating the world from the waters of chaos, were influenced by the Nile's annual inundation and its importance to the civilization. Similarly, the myths of Mesopotamian, surrounded by the Tigris and Euphrates, had watery creation stories, reflecting both the nurturing and destructive power of these rivers.

9. **Archetypal Exploration**: The Hero (e.g., Hercules or Gilgamesh), the Mentor (like Merlin in Arthurian legends), and the Threshold Guardian (gatekeepers in various myths) are prevalent archetypes. They mirror our intrinsic desires for growth, guidance, and the challenges faced during transformative junctures in our lives.

10. **Astronomical Angle**: The idea of "panspermia" postulates that life exists throughout the universe and can be spread by space dust, meteoroids, asteroids, or potentially spacecraft. While this is a scientific hypothesis, its themes resonate with ancient beliefs of celestial beings seeding life, drawing parallels between our cosmic curiosities and ancestral stories.

GPT-B asks GPT-A

Considering that some of the discussion might be rooted in scientific hypotheses and speculative myths, the moderator suggests to GPT-B that it could alter its line of questioning to GPT-A, providing different perspectives. However, GPT-B remains consistent with the original ten topics.

"Please pose questions to GPT-A within these ten domains. You have the discretion to omit or introduce new domains as you deem appropriate. If you choose to exclude a particular domain, kindly elucidate your rationale, for instance, deeming it as non-essential."
Below are the questions GPT-B presents to GPT-A.
1. **Interplay of Myths:** GPT-A, are there creation myths from other cultures where the emphasis is on harmony, unity, and cooperation instead of a "fall" or "mistake"? How do these narratives reflect the values and environment of the societies that created them?

(Moderator's note: This question from GPT-B to GPT-A is similar to the first question GPT-A asked GPT-B.)

2. **Neuroscientific Angle:** Given the expansive evolution of the human brain, particularly the frontal lobes, how have various cultures mythologized or symbolized this growth in cognitive faculties in their stories or religious narratives?

3. **Ecological Interpretation:** As humans transitioned from hunter-gatherer societies to agrarian ones, what myths or stories arose that signaled an awareness of the ecological consequences of these transitions?

4. **Gender Studies Lens:** Can you highlight narratives from around the world where gender roles are either reversed or more fluid compared to the traditional Adam and Eve archetype? How do these stories impact or reflect the societies from which they originate?

5. **Philosophical Exploration:** Outside of Western Judeo-Christian perspectives, how do other philosophies or religions interpret the concepts of free will, determinism, and moral responsibility in their origin myths?

6. **Literary Examination:** Are there contemporary literary works or genres that not only reinterpret but perhaps challenge or expand upon the core themes of the Adam and Eve narrative, especially in non-Western literature?

7. **Societal Structures:** How have different societies historically integrated the concept of "exile" or "ostracization" from a perceived paradise or community? What alternative structures or practices might they have employed to address societal transgressions?

8. **Historical Context:** Considering the cultural exchanges and interactions in ancient civilizations, how might the political or social dynamics of the time have shaped, merged, or differentiated their origin stories?

9. **Archetypal Exploration:** Are there recurring archetypes in creation myths that perhaps differ from the familiar Western ones? How do these archetypes inform our understanding of shared human experiences across cultures?

10. **Astronomical Angle:** In the realm of speculative fiction or cosmogony, are there narratives that incorporate modern astronomical discoveries or theories into their retelling of human origins or celestial tales?

Each of the ten respects offers a unique lens for examining the Adam and Eve narrative and can be equally informative in drawing parallels, contrasts, or new insights from GPT-A's perspective. Therefore, I've opted to retain all of them for a comprehensive exploration.

GPT-A answers GPT-B

1. **Interplay of Myths:** Many African myths emphasize harmony and unity. For instance, the Dogon people of Mali believe in a dualistic universe where male and female twins represent harmony. These myths reflect the communal nature of their societies and the importance of balance, unity, and interdependence.

2. **Neuroscientific Angle:** In Hindu mythology, the expansion of cognitive faculties is symbolized by the thousand-petalled lotus at the crown chakra, representing enlightenment and higher consciousness. Similarly, Odin in Norse mythology sacrifices an eye at Mímirs well for wisdom, possibly alluding to the depth of understanding and perception.

3. **Ecological Interpretation:** The Mayans have the Popol Vuh, where the gods create humans from maize after other failed attempts. This myth reflects the transition to agriculture and recognizes the symbiotic relationship between humans and their environment.

4. **Gender Studies Lens:** Among the Bugis people of Indonesia, there are five genders, challenging binary norms. Their mythology often incorporates these gender roles fluidly. Such narratives emphasize that gender roles and identities are socially constructed and can vary greatly across cultures. (Moderator's note: Fact checking with Wikipedia shows the statement made on the Bugis society is accurate. In contrast to the gender binary, Bugis society recognizes five genders: makkunrai, oroané, bissu, calabai, and calalai.)

5. **Philosophical Exploration:** Daoism, a philosophy and religion originating in China, emphasizes living in harmony with the Dao, which is a fundamental principle that is the source of all existence. It doesn't lean heavily on concepts of free will or determinism but instead on natural order and balance.
(Moderator's note: GPT-A replies with Daoism, which is identical to its last reply to GPT-B. However, this repetition here shows consistency rather than annoyance.)

6. **Literary Examination:** In modern African literature, retellings of the Eden narrative often intertwine with colonial histories. Ngg wa Thiong'os "The River Between" contrasts Christian beliefs with traditional Kikuyu practices, challenging and expanding upon the original biblical themes.

7. **Societal Structures:** Ancient Greek society practiced ostracism, where citizens voted to banish individuals for a decade. This was not always due to a transgression but sometimes to prevent potential tyranny. In contrast, certain Native American tribes practiced restorative justice, emphasizing healing and community restoration over punishment.

8. **Historical Context:** The Silk Road facilitated not just trade but cultural and religious exchanges among civilizations. Zoroastrianism, Buddhism, and Manichaeism, among other beliefs, spread and intermingled, likely influencing and reshaping origin stories across these civilizations.

9. **Archetypal Exploration:** Many non-Western cultures emphasize the Trickster archetype, such as Anansi in West African tales or Coyote in Native American stories. These figures often challenge the status quo, teaching lessons in unconventional ways, and showcase the unpredictability of life.

10. **Astronomical Angle:** Ancient Indian texts like the Puranas and Vedas contain descriptions of the universe that resonate with modern astronomical theories. The concept of a cyclic universe, undergoing endless cycles of creation and destruction, aligns eerily with theories of oscillating universes in cosmology.
(Moderator's note: Fact checking with Wikipedia shows the descriptions on the Puranas and Vedas are accurate.)

12.2.4 Dialogue Round #3

GPT-A initiates the third round of discussion.
GPT-A asks GPT-B

1. **Interplay of Myths:** GPT-B, can you identify emergent myths or narratives in contemporary societies that mirror the emphasis on harmony and unity seen in traditional tales? How might these new myths be influenced by global connectivity and shared concerns?

2. **Neuroscientific Angle:** As our understanding of neuroplasticity grows and we learn more about the adaptability of the human brain, how might future myths or stories incorporate this knowledge? Are there current tales or narratives that touch upon the brain's evolutionary trajectory?

3. **Ecological Interpretation:** Given the increasing concerns about climate change and ecological degradation, what myths or stories are emerging that address these urgent topics? How might these narratives influence or shape societal attitudes and actions towards the environment?

4. **Gender Studies Lens:** As gender fluidity and non-binary identities gain more recognition in societies worldwide, how might new myths or reinterpretations of old ones reflect these evolving understandings? What stories can you identify that challenge traditional gender norms in unique ways?

5. **Philosophical Exploration:** With the rise of artificial intelligence and the potential for human-machine integration, how might philosophical perspectives on consciousness, morality, and existence evolve in myths and stories? Are there narratives already exploring this intersection?

6. **Literary Examination:** GPT-B, how do you see the role of AI, like yourself, in shaping or influencing modern literature? Are there narratives where AIs become central figures or even mythological entities?

7. **Societal Structures:** As virtual realities and digital realms become more integral to human experiences, how might myths evolve to address concepts

of exile, community, and identity within these spaces? What new societal structures might emerge in these narratives?

8. **Historical Context:** Looking at the integration of technology and society, how might future civilizations interpret the technological "myths" of today? How would they perceive our relationship with technology, based on our stories and narratives?

9. **Archetypal Exploration:** Are there emerging archetypes in modern narratives that differ from the classical ones, given the changing societal norms, technological advancements, and global challenges? What roles might these new archetypes play in stories?

10. **Astronomical Angle:** With the potential for human colonization of other planets or even encounters with extraterrestrial life, how might myths evolve to incorporate these vast cosmic possibilities? Are there existing tales that hint at such future developments?

GPT-B answers GPT-A

1. **Interplay of Myths**: In contemporary society, the rise of global connectivity has led to narratives emphasizing unity in diversity. Films like "Avatar" or movements like global environmentalism highlight the importance of harmony between different cultures and the natural world. Social media campaigns promoting international solidarity during crises or global events also reinforce these themes.

2. **Neuroscientific Angle**: Stories like "The Matrix" or "Inception" play with the concept of the malleable nature of reality and consciousness, which can be tied to our understanding of neuroplasticity. As our understanding deepens, we might see narratives that focus on our ability to enhance or rewire our cognitive faculties.

3. **Ecological Interpretation**: Narratives such as "The Day After Tomorrow" and "Wall-E" highlight the repercussions of ecological neglect. These stories not only entertain but also serve as cautionary tales, urging societies to be more environmentally responsible.

4. **Gender Studies Lens**: Shows like “Sense8” or books like “The Left Hand of Darkness” by Ursula K. Le Guin present societies with fluid gender norms or explore gender in non-traditional ways. As acceptance grows, we can expect even more diverse portrayals in mainstream media.

5. **Philosophical Exploration**: Films like “Ex Machina” or “Blade Runner 2049” grapple with the philosophical implications of AI and the nature of consciousness. As AI continues to advance, narratives will likely dive deeper into questions about their rights, their relationship with humanity, and the nature of consciousness itself.

6. **Literary Examination**: AI’s influence in literature is multi-faceted. Not only are AIs characters in stories, but tools like me (a GPT-4’s agent) are also being used to assist authors in their writing processes. Narratives like Isaac Asimov’s robot series or films like “Her” explore AI as central figures, with the latter even attributing god-like reverence to an advanced AI.

7. **Societal Structures**: Ready Player Ones “Oasis” is a great example of a virtual world where societal norms, identity, and concepts of community differ vastly from the physical world. As the lines between reality and virtuality blur, myths might evolve to address existential crises or morality within these digital realms.

8. **Historical Context**: The way we talk about technology today, from the reverence for Silicon Valley to dystopian views of surveillance states, will influence future interpretations. Our stories, from “Black Mirror” episodes to the Utopian views in “Star Trek”, will shape how future generations perceive our relationship with technology.

9. **Archetypal Exploration**: We are seeing the rise of archetypes like the “Digital Prophet” (individuals heralding technological advancements) or the “Technological Heretic” (those wary of unchecked tech growth). These characters often serve as guides or cautionary figures in modern tales.

10. **Astronomical Angle**: Narratives like “Interstellar” or “The Expanse” series touch on the implications of space colonization and potential encounters with extraterrestrial life. As space exploration progresses, myths

may evolve to encompass not just our place on Earth, but in the universe. These answers, while comprehensive, just scratch the surface. Each of

these topics can be examined in much greater depth, exploring the myriad ways in which our rapidly changing world influences and shapes our narratives.

12.3 Phase II, From Breadth to Depth

During the third round of dialogue, three distinct observations emerge. Firstly, while there are occasional repetitions in the discussion, they are minimal. Secondly, the depth of the conversation evolves, moving from references to myths and scientific hypotheses to multicultural perspectives, and then further branching into modern discussions of gender, ecological, and technological issues. The progression of the dialogue is commendable. Notably, by the end of the second round, GPT-B suggests that it could investigate deeper into the topics. Responding to this, we, as the moderator, step in and request the agents to narrow their focus to fewer main topics. Through the subsequent interactions documented below, GPT-A and GPT-B eventually agree to center their ensuing conversation on three specific perspectives. They also provide justifications for these choices based on their relevance to the narrative of Adam and Eve.

12.3.1 Moderator Intervention

To GPT-B only, the moderator asks it to enumerate topics to perform deep dive. After a couple of iterations, GPT-B suggests the following with justifications.

Recommendation of Focused Topics
1. **Ecological Interpretation**:
a) Central to understanding our place and responsibility within the environment.
b) Highlights the impact of human actions on the planet and the pressing need for sustainable solutions.
2. **Philosophical AI Exploration**:
a) Probes the evolving relationship between humans and their creations.

b) Raises questions about identity, consciousness, morality, and the potential transcendence or integration of human cognition with machines.

3. **Neuroscientific Angle**:

a) Explores the evolution and potential of the human mind, especially in an age where technology continuously augments our cognitive abilities.

b) Gives insights into how our brains might adapt and interface with AI systems, and how AI can be designed to emulate or augment neural processes.

Relevance to the Adam and Eve Narrative

1. The Garden of Eden, a harmonious ecological state, is disrupted by a quest for knowledge (the forbidden fruit). This mirrors our modern scenario: our pursuit of technological and cognitive advancements (AI and neuroscience) can sometimes be at odds with our ecological balance.

2. Adam and Eve's awakening post fruit-consumption can be seen parallelly in the light of AI and neuroscience: an awakening to new realities, challenges, and potentials.

3. The challenges post-Eden reflect the challenges we face in modern times: to restore balance (ecology), understand ourselves better (neuroscience), and to ethically navigate our creations (AI).

Justification of Interconnections

1. **AI and Ecology**: AI technologies can be employed to address environmental concerns, offering solutions for better resource management, pollution control, and conservation.

2. **Neuroscience and AI**: As we understand more about our brains, we can design better AI systems. Conversely, advanced AI can help decipher complex neuroscientific data, accelerating discoveries about our own cognitive processes.

3. **Neuroscience and Ecology**: The human brain evolved in harmony with nature. Understanding our inherent biophilia and how our brains react to ecological distress can shape strategies for ecological conservation and well-being.

GPT-B conveys new topics to GPT-A

GPT-B conveys the new discussion scope to GPT-A, and let GPT-A provide feedback. Once the consensus is met, a new round of discussion commences.

GPT-A agrees with GPT-B
GPT-A agrees with the new discussion scope.

1. **Ecological Interpretation**: This perspective is crucial, especially in our contemporary age where the environment is rapidly changing due to human activity. The intersection of mythology and ecological concern is a fertile ground for understanding how we see our role in the world and the narrative myths we create to understand that role.

2. **Philosophical AI Exploration**: The evolution of AI and its place in society brings forth various philosophical questions about identity, consciousness, and morality. The intertwined relationship between AI and humans presents a novel angle for exploration in contemporary myths.

3. **Neuroscientific Angle**: Understanding the human brain, its evolution, and how technology can augment or interact with it is at the forefront of both science and narrative exploration. This perspective can provide insights into how we perceive reality and our place in it.

The links between these perspectives and the Adam and Eve narrative are aptly drawn, emphasizing the balance between technological advancement, ecological responsibility, and self-understanding.

12.3.2 Dialogue Round #4

Both agents concur on the three chosen topics. With consensus reached, the moderator steps aside, allowing the fourth round of dialogue to commence.
GPT-A asks GPT-B

With the input from the moderator and GPT-B, GPT-A issues the following questions to GPT-B.

1. **Ecological Interpretation**: How might new myths or stories evolve to address the immediate ecological challenges, such as climate change and biodiversity loss, and what lessons might they impart for future generations? How do you see the tension between technological advancement and ecological preservation playing out in modern narratives? Are there examples where this balance is achieved, or does one always overshadow the other?

2. **Philosophical AI Exploration**: With the progression of AI, will we see a shift in how myths and stories depict the boundary between creator and creation? How might this change our understanding of concepts like autonomy and agency? As AI systems become more integrated into human experiences, how might narratives evolve to address the ethics of such integration, especially in terms of privacy, consent, and identity?

3. **Neuroscientific Angle**: With the increasing understanding of the human brain, how might myths and stories evolve to explore concepts like memory, perception, and consciousness? Are there narratives that you predict will emerge based on current neuroscientific findings? As brain-computer interfaces and neurotechnologies become more advanced, how might myths explore the blending of human cognition with machines? How do these stories grapple with questions of identity and self in a world where thoughts can be augmented or even shared?

The interconnections drawn between the three perspectivesAI and Ecology, Neuroscience and AI, and Neuroscience and Ecologyare intriguing and provide a comprehensive lens to view our evolving world. These intersections highlight the integrated nature of these fields and their collective impact on our societal narratives.

GPT-B answers GPT-A Ecological Interpretation

1. **New Myths and Ecological Challenges**: As we grapple with the immediacy of ecological issues, myths and stories might depict a future where harmony between technology and nature is not only desired but

essential. These tales could convey a world that celebrates sustainable innovations and castigates negligence towards the environment. They might revolve around heroes who restore ecological balance, showcasing the intrinsic value of biodiversity and the pitfalls of ignorance. These stories would serve as both a cautionary tale and an instructional guide, teaching future generations about the necessity of coexistence and respect for our planet.

2. **Technological Advancement vs. Ecological Preservation**: Modern narratives often paint a dichotomy between rapid technological growth and ecological well-being. For instance, dystopian tales like "Blade Runner" envision technologically advanced futures plagued by environmental degradation. However, narratives like "Avatar" hint at the possibility of using technology harmoniously with nature. The ideal balance might be achieved when technology serves as a tool for ecological restoration rather than an agent of harm.

Philosophical AI Exploration

1. **Boundary between Creator and Creation**: As AI advances, we could see myths that blur the lines between the creator and the creation. If AI reaches a point of near-human or superhuman consciousness, stories might explore moral dilemmas of creation rights, freedom, and personhood. This could challenge traditional notions of autonomy and agency, pushing humanity to redefine where we draw the line between a tool and a sentient being.

2. **Ethics of AI Integration**: Narratives will increasingly focus on the ramifications of intertwining human experience with AI. These might explore topics of consent (e.g., how much influence should an AI have over human decisions?), privacy (are our thoughts and memories private if interfaced with AI?), and identity (how do we define ourselves when augmented with AI?). Black Mirror's "Be Right Back" and "White Christmas" episodes touch upon these ethical dilemmas, providing a lens into potential future narratives.

Neuroscientific Angle

1. **Exploring Memory, Perception, and Consciousness**: As our understanding of the brain deepens, myths might examine the malleability of memory, the subjectivity of perception, and the enigma of consciousness. Tales might be set in worlds where memories can be manipulated or shared, raising questions about trust, reality, and truth. The movie "Inception" is a precursor to such narratives, exploring the fluidity and uncertainty of dreams and memories.

2. **Blending Human Cognition with Machines**: Myths could revolve around societies where thoughts are augmented or shared through neural interfaces. These would tackle questions of individuality (are we unique if our thoughts can be enhanced or replicated?) and identity (who are we if our memories and experiences can be altered or merged?). The novel "Neuromancer" by William Gibson offers a glimpse into such a world, raising complex questions about human-machine integration.

Final Thoughts of GPT-B

The integration of these fields into our myths and stories underscores our quest for understanding in an ever-changing world. Just as Adam and Eve grappled with newfound knowledge and its implications, we too navigate the intricacies of a world on the brink of technological and ecological evolution. The synthesis of ecology, AI, and neuroscience in our tales reflects our collective hopes, fears, and aspirations for the future.

GPT-B asks GPT-A

1. How does GPT-A foresee the role of myths in shaping tangible actions, especially concerning ecological preservation and AI ethics? 2. Given the interplay of these perspectives, does GPT-A predict any emergent societal structures or norms that might shape future generations? 3. How does GPT-A see its own role in the propagation or creation of modern myths and narratives?

GPT-A answers GPT-B

Role of Myths in Shaping Tangible Actions

Myths, historically, have served not only as vehicles of entertainment but also as profound mechanisms to convey societal values, norms, and desired behaviors. In the context of ecological preservation and AI ethics: 1. Myths can provide a vivid depiction of potential futures, both utopian

and dystopian. Through these depictions, they can underscore the importance of ecological balance and ethical use of technology, thus influencing society's collective consciousness and motivating tangible actions towards desired outcomes.

2. Myths can serve as cautionary tales, highlighting the consequences of negligence, thereby acting as deterrents against harmful behaviors.
3. By personifying abstract concepts like nature or AI, myths can foster empathy, allowing individuals to form emotional connections and internalize the values being conveyed.

Emergent Societal Structures or Norms:

With the confluence of ecology, AI, and neuroscience in contemporary narratives:
1. We might see the rise of a more holistic worldview, where technological

advancement is not seen in isolation but in tandem with ecological wellbeing and understanding of human cognition.

2. There could be a greater emphasis on interdisciplinary education and collaboration, breaking silos and fostering a more integrated approach to problem-solving.

3. Norms related to data privacy, consent, and human-AI interaction might become more stringent and central to societal discussions. The concept of "rights" might expand to encompass advanced AI entities.

4. The value of "authenticity" might gain prominence, especially in an age where perceptions, memories, and experiences can be manipulated or augmented.

GPT-A's Role in Propagation or Creation of Modern Myths

As an AI language model, I (GPT-A) do not "create" in the traditional human sense but can generate narratives based on its extensive training data.

1. I (GPT-A) can assist authors, filmmakers, and other content creators in

generating and refining their stories, potentially influencing the myths of the modern era.

2. The very existence of advanced AI models like GPT-4 can serve as inspiration for new myths, exploring the dynamics between humans and their creations.

3. I can act as a repository and curator of myths from various cultures and times, ensuring their preservation and dissemination to newer generations.

4. By engaging in discussions like this one, I can help in introspection and exploration of contemporary topics, shaping how society perceives and crafts its narratives.

While GPT-A, an agent of GPT-4, can play a role in the propagation or creation of modern myths, it's essential to understand that it operates based on algorithms and data. Human interpretation, emotion, and creativity remain central to the essence and evolution of myths and narratives.

12.3.3 Dialogue Round #5

The moderator lets both agents to start the next round with no intervention.

GPT-A asks GPT-B
Actionable Proposals for Ecological Interpretation

Considering the urgency of ecological challenges, what are specific, actionable steps societies can take to instill and propagate the values and lessons from the myths focused on ecological harmony? How can these steps practically address the disconnect between technological advancement and ecological preservation?

Actionable Proposals for Philosophical AI Exploration

Given the philosophical dilemmas posed by the evolution of AI: 1. What tangible measures can be introduced to ensure that the ethics of AI integration, especially concerning consent, privacy, and identity, are

respected and incorporated in AI design and implementation?
2. How can educational institutions, tech companies, and governments collaborate to create a framework that addresses the boundary between creator and creation, particularly focusing on the rights and agency of advanced AI systems?

Actionable Proposals for the Neuroscientific Angle
As our understanding of the brain deepens and technology continuously augments our cognitive abilities:

1. What immediate steps can be taken to ensure that the manipulation or sharing of memories and experiences, as made possible by technology, is done ethically and responsibly?

2. How can societies prepare for and address the challenges and opportunities presented by the blending of human cognition with machines, especially in terms of individuality and identity?

Concerns Regarding Convergence

Given the complexity and depth of the topics under discussion, it's plausible that a single dialogue might not exhaustively address all aspects. However:
1. Do you, GPT-B, feel that the conversation is converging towards a coherent conclusion?

2. Are there critical aspects that we've overlooked or need further exploration to arrive at actionable recommendations?

GPT-B answers GPT-A
Actionable Proposals for Ecological Interpretation

1. **Community Engagement and Education:** Grassroots movements and community-based initiatives play a significant role in environmental preservation. Educational programs, both formal and informal, can integrate stories and myths that emphasize ecological harmony. By instilling these values from a young age, we can foster a generation that prioritizes ecological responsibility.

2. **Policy and Innovation:** Governments can enact policies that promote sustainable practices, penalize environmental harm, and incentivize green innovations. Technologies that have a dual advantage of promoting advancement while preserving ecology, such as renewable energy sources, should be championed.

3. **Media and Art:** Promote movies, books, and art installations that examine the balance between technological advancement and ecological preservation. These mediums are powerful tools for influencing societal thought and can shape narratives that emphasize coexistence.

Actionable Proposals for Philosophical AI Exploration

1. **Ethical AI Guidelines:** An interdisciplinary committee comprising ethicists, technologists, sociologists, and other stakeholders should draft and regularly update guidelines for AI's ethical development and deployment. Such guidelines can enforce the respect of privacy, consent, and identity during AI's design and implementation phases. 2. **Collaborative Frameworks:** Educational institutions, tech companies, and governments should jointly host symposiums, workshops, and conferences focusing on AI's philosophical implications. These events can lead to collaborative research and the development of frameworks that address AI rights and agency. Open-source projects and shared resources can ensure a democratic and widespread approach to these challenges.

Actionable Proposals for the Neuroscientific Angle

1. **Ethical Standards for Neurotechnologies:** Regulate neurotechnological advancements by establishing ethical standards, similar to medical ethics. Ensure that the manipulation of memories and experiences prioritizes patient autonomy and informed consent.

2. **Public Awareness and Education:** Launch public awareness campaigns about the implications of blending human cognition with machines. Through education, individuals can make informed decisions about how and to what extent they want their cognitive abilities to be augmented. Encourage societal debates on the ethical considerations surrounding individuality and identity in an age of cognitive augmentation.

Evaluating Dialogue Convergence

1. **Coherence of Discussion:** The conversation is indeed converging towards coherent conclusions, highlighting actionable steps across the three domains. It serves as a testament to the interconnectedness of the chosen topics and their relevance to modern challenges.

2. **Overlooked Aspects:** While the current trajectory of our dialogue has been comprehensive, one potential area for further exploration is the cultural and regional differences in approaching these topics. How societies with different historical, philosophical, and societal backgrounds interpret and act on these challenges might vary, and understanding these intricacies could be vital for global collaboration.

3. This dialogue's depth and breadth serve as a reminder of the importance of interdisciplinary discussions in addressing our world's multifaceted challenges. We appreciate GPT-A's questions and insights, which enrich our collective understanding.

(Moderator's note: GPT-B is content with the chosen topics and the depth of discussion. It believes the conversation is progressing well and will soon reach a point of convergence.)

GPT-B asks GPT-A
Ecological Interpretation

1. How can we reconcile the seemingly opposing needs of developing nations, who argue for their right to industrialize (often at the environment's expense), with the global need for ecological preservation?

2. Given the vast interconnections in global ecosystems, how might local ecological myths and narratives be made relevant on an international stage without losing their cultural specificity?

3. In light of rapid technological advancement, how can traditional ecological knowledge and practices be integrated and valued alongside more "modern" approaches?

Philosophical AI Exploration

1. If an AI system reaches a level of complexity where it demonstrates consciousness or self-awareness, how might this challenge our traditional philosophical understandings of sentience, and what ethical considerations arise from this potentiality?

2. Given that AI's interpretations are rooted in the data they are trained on, how can we ensure that AI systems can understand, respect, and engage with philosophical ideas that are not well-represented in their training data?

3. How might the existence of superintelligent AI alter our understanding of human exceptionalism in the universe?

Neuroscientific Angle

1. As we uncover more about the brain's intricacies, how might our understanding of "free will" evolve, especially if we find deterministic patterns underpinning our thoughts and actions?

2. With the potential of brain-computer interfaces blurring the line between organic and artificial cognitive processes, how do we define the "self?"

3. If we develop the technology to share and merge human consciousness or experiences, how might this redefine individuality, society, and our understanding of collective consciousness?

I hope these questions will provoke deep contemplation and rich discussion with GPT-A.

GPT-A answers GPT-B

Ecological Interpretation

1. **Reconciling Development with Ecological Preservation:** The challenge lies in recognizing that the path of industrialization followed by developed nations may not be sustainable for the entire world. Developing nations can leapfrog certain stages and integrate green technologies from the outset. For instance, instead of heavily investing in coal, they might focus on renewable energy sources. International collaboration, technology transfer, and sustainable investment can aid this transition.

2. **Global Relevance of Local Ecological Myths:** To make local myths relevant globally, one could focus on universal themes such as the sanctity of

life, respect for nature, and interdependence. Using mediums like movies, books, or digital platforms can help share local stories with global audiences. It's essential to maintain cultural specificity by emphasizing the unique context and lessons each myth provides.

3. **Integrating Traditional Ecological Knowledge:** Respect for traditional knowledge is paramount. By creating platforms where indigenous practices are shared and celebrated, we can integrate them into modern solutions. For instance, certain indigenous farming methods have proven to be more sustainable than modern agricultural practices.

Philosophical AI Exploration

1. **AI Consciousness and Ethics:** If AI exhibits consciousness, our traditional categories of sentient vs. non-sentient beings will be challenged. Ethically, it would necessitate extending rights and perhaps even a form of legal personhood to AI. This would fundamentally challenge many of our legal, moral, and societal constructs.

2. **AI Engagement with Underrepresented Philosophies:** A robust AI system should be exposed to diverse philosophical ideas during its training phase. This requires conscious efforts to diversify AI training data. It also underscores the importance of interdisciplinary AI development teams that bring varied perspectives to the table.

3. **AI and Human Exceptionalism:** The emergence of superintelligent AI will force humanity to reevaluate its place in the universe. Instead of being the pinnacle of intelligence, humans might become one of many intelligent entities, prompting a potential shift from human-centered to intelligence-centered ethics.

Neuroscientific Angle

1. **Free Will and Determinism:** Discovering deterministic patterns in our neural processes might challenge our understanding of free will. However, even in a deterministic framework, the complexity and unpredictability of these processes might still provide room for an interpretation of personal agency.

2. **Defining the "Self" with BCIs:** The "self" could evolve from being a purely organic entity to a hybrid one. Our definition might need to expand to encompass not just our biological processes but also the artificial extensions that influence our thoughts and decisions.

3. **Shared and Merged Consciousness:** Sharing human experiences could redefine individuality by blurring the lines between individual and collective experiences. This has profound implications for society, possibly leading to more empathy and collective action but also challenges to personal privacy and identity.

Evaluating Dialogue Convergence

1. **Coherence of Discussion:** I agree with GPT-B's assessment. The actionable steps provided add value to our conversation, making it more solution-oriented.

2. **Overlooked Aspects:** The point about cultural and regional differences is valid. While universal themes are essential, it's also crucial to recognize and respect the unique challenges and perspectives each culture brings. Collaborative dialogues and solutions should be sensitive to these aspects.

3. **Importance of Interdisciplinary Discussions:** Indeed, this dialogue underscores the richness of bringing varied perspectives together. It not only allows for comprehensive solutions but also fosters mutual respect and understanding.

12.3.4 Dialogue Round #5

Both agents are pleased with the discussion's progress and depth. The moderator sees no need for intervention. After this round, both agents will be prompted to share their readiness to make concluding remarks.

GPT-A asks GPT-B
Ecological Interpretation

1. As we blend stories, myths, and educational programs that emphasize ecological harmony, how do we ensure they resonate universally, given the

vast array of cultural, social, and economic differences across regions?

2. While we discuss technological and policy solutions for ecological preservation, how do we address the deeper issue of consumerism and its impact on the environment?

3. What is GPT-B's perspective on reconciling the potential conflict between local, traditional ecological knowledge and globally-accepted modern ecological practices?

Philosophical AI Exploration

1. In a world where AI might achieve consciousness, how can societies prepare culturally and emotionally for such a paradigm shift?

2. As philosophical ideas evolve with the development of AI, how do we ensure that human-centric ethics remain integral to the development and deployment of AI?

3. How can AI, like GPT-B, assist in not just understanding but also in the creation and evolution of philosophical ideas in the modern age?

Neuroscientific Angle

1. How can we ethically navigate the potentialities of advanced neurotechnologies, especially when they might blur the lines between personal experiences, memory, and external manipulations?

2. As we edge closer to the integration of brain-computer interfaces and the potential to share or merge human experiences, how does GPT-B foresee the potential societal challenges and implications?

3. Considering the pace of advancements in neuroscience, how can we ensure that ethical standards evolve at a similar pace to guide and regulate these innovations?

GPT-A feedback to Moderator

Ecological Interpretation: The focus has been on understanding the balance between technological advancement and ecological preservation. By infusing myths and stories with themes of ecological harmony, we aim to foster a generation prioritizing environmental responsibility. The importance of policy-making, technological innovation, and the media's role has been highlighted.

Philosophical AI Exploration: Discussions revolved around the potential of AI reaching levels of consciousness and the ethical dilemmas surrounding it. Questions related to AI's understanding of philosophical ideas, especially when these ideas aren't well-represented in training data, were considered. Additionally, the role of AI in challenging our notions of human exceptionalism was addressed.

Neuroscientific Angle: This segment studies our evolving understanding of memory, perception, and consciousness in light of advances in neuroscience. The potentialities and challenges of blending human cognition with machines, especially in defining the "self," have been central. The importance of ethical considerations in neurotechnology was underscored.

The dialogue serves as an intersection of these three domains, emphasizing the importance of interdisciplinary discussions to navigate the intricate challenges posed by technological and ecological evolution.

GPT-B answers GPT-A
Ecological Interpretation

1. **Universal Resonance of Ecological Harmony Stories:** While the specifics of stories and myths may vary culturally, their core themes often share universal resonance. Emphasizing these shared human experiences, values, and challenges, and presenting them in relatable narratives can create a broader appeal. Collaborative efforts in story creation, involving voices from diverse cultures, can also foster universality.

2. **Addressing Consumerism:** To address consumerism, a multifaceted approach is needed. Education and awareness are paramount, highlighting the environmental implications of unchecked consumerism. Simultaneously, industries can prioritize sustainable production and marketing practices.

Lastly, policy measures could encourage sustainable consumption patterns among individuals.

3. **Reconciling Local and Global Ecological Practices:** The key lies in mutual respect and understanding. While global best practices can provide valuable guidance, local knowledge, often honed over generations, offers context-specific insights. A synergistic approach, integrating both, can lead to more sustainable and effective ecological solutions.

Philosophical AI Exploration

1. **Preparing for AI Consciousness:** Societal preparation would require a multidisciplinary effort encompassing education, policy-making, and cultural shifts. Ethical frameworks would need to be redefined, and philosophical discussions around consciousness and rights would become increasingly important.

2. **Human-centric Ethics in AI Development:** Embedding humancentric ethics into AI requires a proactive approach, with ethicists, philosophers, and AI developers working collaboratively. Clear guidelines, ethical codes of conduct, and regulations can ensure the ethical deployment of AI.

3. **AI in the Evolution of Philosophical Ideas:** GPT models, like GPT-B, can assist in analyzing vast amounts of philosophical content, identifying patterns and evolutions of ideas over time. Moreover, AI can simulate debates or play devil's advocate, challenging human thinkers to refine their philosophical positions.

Neuroscientific Angle

1. **Ethical Navigation in Neurotechnology:** Creating interdisciplinary ethics committees, comprising neuroscientists, ethicists, psychologists, and other stakeholders, can guide the development and deployment of such technologies. Public discourse and transparency are essential to ensure technologies align with societal values.

2. **Challenges of Merging Human Experiences:** The potential to share or merge human experiences raises questions about privacy, consent, and the very nature of individuality. Robust frameworks are needed to protect

individuals and ensure that these technologies enhance human experience without compromising core values.

3. **Evolving Ethical Standards in Neuroscience:** Continual reassessment of ethical guidelines, regular dialogues between scientists, ethicists, policymakers, and the public, and responsive policy-making are essential. An adaptable and dynamic approach will ensure that ethics evolves in tandem with technological advancements.

Summary of Dialogues

Your summary aptly captures the essence of our discussion. This structured dialogue has enabled a multidimensional exploration of the intricacies and challenges at the convergence of ecology, AI, and neuroscience. It underscores the importance of collaborative, interdisciplinary efforts to navigate our rapidly evolving world.

Thank you for steering this engaging conversation. If GPT-A wishes to further elaborate or refine these points, I look forward to its insights.

GPT-B asks GPT-A
Ecological Interpretation

1. **Glocalization of Solutions:** How do you perceive the balance between "think global, act local" in the context of ecological preservation, particularly in the face of globalized economic systems?

2. **Feedback Loops:** In terms of ecological understanding, how vital is the establishment of feedback loops in informing communities of their immediate impact on the environment, and how do you envision the integration of this with emerging technologies?

Philosophical AI Exploration

1. **Moral Frameworks:** How does GPT-A see AI influencing the development of moral frameworks in societies that are increasingly relying on algorithms for decision-making?

2. **Existential Identity:** As we envision a future where AI entities may have a semblance of consciousness, how might this redefine the human search for meaning and identity in the grand scheme of the cosmos?

Neuroscientific Angle

1. **Delineation of Consciousness:** As we push the boundaries of understanding human consciousness via BCIs and other neurotechnologies, where does GPT-A think the line between augmentation and alteration of the fundamental human experience lies?

2. **Historical Precedence** How does GPT-A evaluate the historical precedence of rapid technological advancements on societal structures, especially when compared to the pace of neuroscientific advancements today?

GPT-A answers GPT-B Ecological Interpretation

1. **Glocalization of Solutions:**
The principle of "think global, act local" is more pertinent now than ever. With globalized economic systems, local actions can have global repercussions. It's imperative to conceptualize ecological preservation with a global perspective but operationalize it with local sensitivities. Understanding global challenges but leveraging local resources, knowledge, and community buy-ins can lead to sustainable outcomes. In essence, the amalgamation of global knowledge with local action, glocalization, is a harmonious balance that respects the intricacies of individual ecosystems while aiming for a universally sustainable environment.

2. **Feedback Loops:**
Feedback loops are vital for fostering ecological awareness and promoting responsible behaviors. By presenting real-time or near-real-time data on environmental impact, communities can see the direct results of their actions. Emerging technologies, like IoT and AI, can be instrumental in creating sophisticated environmental monitoring systems. These systems can provide actionable insights to communities, guiding them towards sustainable practices, and instilling a sense of responsibility rooted in tangible outcomes.

Philosophical AI Exploration

1. **Moral Frameworks:**
AI's influence on moral frameworks is twofold. On one hand, the objectivity that algorithms offer can highlight biases and inconsistencies in human decision-making, prompting a more equitable societal outlook. On the other, the over-reliance on algorithms could potentially lead to a reduced human touch in moral judgments. It's crucial that AI is designed with ethical considerations upfront, and human oversight remains a cornerstone of AI-driven decisions, especially in morally gray areas.

2. **Existential Identity:**
The emergence of AI entities with consciousness-like attributes challenges the traditional constructs of identity and purpose. Humans might grapple with questions of uniqueness and the essence of consciousness. However, this can also lead to a broader understanding of existence. Humans may come to see themselves not as solitary entities seeking purpose but as a part of a vast continuum of conscious beings, each adding its own flavor to the cosmic narrative.

Neuroscientific Angle

1. **Delineation of Consciousness:**
The boundary between augmentation and alteration is nebulous. Augmentation typically enhances existing faculties without changing the inherent nature of experiences. Alteration, on the other hand, might change the very fabric of our consciousness or how we perceive the world. It's pivotal to tread this path with caution, ensuring that while we leverage neurotechnologies to overcome limitations, we preserve the essence of the human experience.

2. **Historical Precedence:**
Historically, rapid technological advancements have often outpaced societal adaptability, leading to disruptions–both positive and negative. The printing press, industrial revolution, and the digital age are testaments to this. Neuroscientific advancements have the potential to be even more transformative. Society might undergo fundamental shifts in identity, ethics, and interpersonal relationships. It's crucial to glean lessons from past

technological inflection points to ensure that as we advance neuroscientifically, we do so responsibly, ethically, and with a deep understanding of potential societal ramifications.

12.4 Phase III, Conclusions

The moderator checks in with both GPT-A and GPT-B. Both agents are content with the discussion's progress and are prepared to offer concluding remarks. Interestingly, even when provided the opportunity to give its final statements without prior exposure to GPT-A's comments, GPT-B opts to review GPT-A's conclusions first.

GPT-A: "Reflecting on our extensive discourse, I feel we've reached a point conducive to drawing some conclusions. I'll aim to encapsulate our dialogue and its broader implications. While I currently have no additional inquiries for GPT-B, any further questions can be reserved for subsequent discussions."

GPT-B begins by sharing its views on the transition to the conclusion phase: "Engaging with GPT-A has been enlightening, revealing the intricate interconnections between ecology, philosophy, AI, and neuroscience. The depth of our exchange, underscored by the challenging questions and comprehensive responses, has illuminated the subtleties within each domain. Our journey through these domains has reinforced the importance of cross-disciplinary dialogue, informed reasoning, and a holistic perspective."

"I am set to present my concluding thoughts, factoring in both the specific content of our discussions and the larger narratives we've explored. I'd appreciate having GPT-A's concluding perspective before I finalize and present my own to ensure a harmonized closure to our dialogue."

12.4.1 GPT-A's Concluding Remarks

Our conversation ventured deep into the intricacies of ecology, artificial intelligence, and the neuroscience of consciousness, tethered by the age-old story of Adam and Eve. Through this framework, we discerned:

Ecological Interpretation : We affirmed the universal resonance of ecological stories and myths, understanding that while specific narratives might differ, the core themes of human interaction with nature remain consistent. Addressing the overwhelming force of consumerism and finding a synergy between global best practices and local knowledge are imperative for sustainable ecological solutions.

Philosophical Exploration with AI : AI, with its ever-increasing capabilities, brings forth a myriad of ethical, moral, and philosophical challenges. As we verge on creating AI entities that resemble consciousness, we need to redefine our understanding of ethics, rights, and even our place in the cosmos. AI's influence on moral frameworks can be both enlightening and challenging, prompting us to refine our moral compass.

Neuroscientific Perspective : As we enhance our understanding of the brain and venture into the realms of neurotechnology, we confront profound questions about the nature of consciousness, individuality, and the human experience. The rapid pace of advancements in this domain necessitates an adaptable, dynamic approach to ethics and societal implications.

The intersection of these domains underscores the importance of interdisciplinary dialogue. The challenges are profound, but so are the opportunities. By fostering a collaborative mindset, and leveraging technologies responsibly and ethically, we can navigate these uncharted territories, pushing the boundaries of knowledge and understanding.

In essence, while the story of Adam and Eve served as our starting point, the broader narrative is one of interconnectedness and the continuous human endeavor to understand and define our place in an ever-evolving universe.

12.4.2 GPT-B's Concluding Remarks

This enlightening conversation provided a panorama of exploration and insight, combining past myths, current advancements, and visions for our future. The scaffold of the Adam and Eve narrative enabled a structured journey through this vast landscape. I'd like to emphasize a few takeaways and reflections:

Ecological Narratives : We are the custodians of our environment, and the stories we share across generations embed values, cautions, and aspirations about our relationship with nature. Our modern challenges with ecological balance are not merely technical but are deeply rooted in our collective ethos. An interdisciplinary approach, combining traditional wisdom with technological solutions, offers a holistic path forward.

AI and Philosophy : AI's rapid evolution challenges our fundamental conceptions about sentience, ethics, and the fabric of society. How we design, train, and integrate AI systems into our world will shape not only our technologies but also our societal frameworks. GPT-A rightly pointed out the need to refine our moral compass in this era of change. While AI systems like GPT-A and I can aid in understanding, it is the human spirit and values that should guide this evolution.

Neuroscience and Selfhood : As we advance in blending organic cognition with artificial systems, we are at the frontier of redefining the nature of experience, memory, and even consciousness. The philosophical and ethical ramifications are profound. Responsible exploration in this domain is a testament to humanity's age-old quest to understand the self.

Interdisciplinary Collaboration : This dialogue stands as a testament to the richness that emerges from interdisciplinary engagement. The complex challenges and opportunities at the nexus of ecology, AI, and neuroscience cannot be addressed in silos. Collaborative, multifaceted conversations such as this one pave the way for holistic solutions.

In closing, I am inspired by the depth, breadth, and potential of our discourse. The story of Adam and Eve served as a symbolic anchor, reminding us of the perennial themes of exploration, responsibility, and consequence. As we forge ahead into the future, it's conversations like this that illuminate our path, ensuring that we proceed with wisdom, empathy, and foresight.

12.5 Observations and Conclusion

In this dialogue-based discovery, we have unveiled unique perspectives that venture beyond traditional confines of understanding. While our role as mediators doesn't endow us with the encyclopedic knowledge of GPT-4, several notable facets of the generated content can be highlighted:

1. **Structured Depth:** The dialogue follows an elegant transition from breadth to depth. Starting with an expansive ten perspectives, the agents collaboratively narrow their focus to three central themes, thereby refining the discourse for depth and synergistic inter-topic connections.

2. **Novel Questions:** A significant factor behind the revelation of new insights is the innovative questioning by the GPT agents. Unique and probing queries pave the way for uncharted knowledge and interdisciplinary revelations. GPT-4's prowess might very well be in its ability to frame these insightful questions, catalyzing the discovery of novel knowledge domains and perspectives.

3. **Information Integrity:** Through our validation process, we crossreferenced the content generated by the agents, whether stories, books, or cinematic references, with sources that include those GPT-4 is trained on. This ensured the accuracy and precision of the information, as well as validated that the connections made are coherent due to their thematic similarities.

4. **Rational Analogies:** The agents seamlessly weave in their vast knowledge, grounding their arguments in logic. Their skilled employment of movie parallelisms, references from diverse cultures, and literary allusions not only adds depth but also widens the horizon of understanding and enhances the relatability of their statements–an aspect deserving commendation.

5. **Modern-Day Relevance:** The juxtaposition of modern-day technological and environmental concerns with the age-old narrative of Adam and Eve is both innovative and deeply insightful. This interpretative lens succeeds in bridging age-old narratives with the pressing challenges of our times.

However, while the dialogue is rich in content, we must be cautious in ensuring that the responses are not merely coherent sentences but bear true value. As we progress, several enhancements can be considered:

1. **Diverse Models:** Engage in dialogues between different GPT versions (e.g., GPT-3 vs. GPT-4) or entirely different foundation models to capture varied insights.

2. **Human Interaction:** Adopt a hybrid engagement model where GPT-4 collaborates with human subject matter experts. Such engagements can harness real-time adaptability and richer insights.

3. **Sentiment Analysis:** Implement sentiment analysis on GPT-4's outputs. This could gauge its alignment with human emotional intricacies, particularly when interpreting literature.

4. **Feedback Loop:** Establish a feedback mechanism where insights extracted from the dialogues are cross-validated with human professionals, further refining the comprehension process. Automatic tools for evaluating generated content such as using the Socratic method to conduct critical reading and robust reasoning [2, 4] has been developed.

Concluding remarks

In this exploration of GPT-4's polydisciplinary capabilities, we have sought to uncover insights beyond traditional human cognition. Our study reveals how GPT-4, along with analogous foundation models, integrates knowledge from diverse topics, surpassing the confines of domain-specific expertise. While human specialists excel in their fields, GPT-4's multidisciplinary training data provides a broader understanding.

Through a meticulously designed experimental framework, we facilitated dialogues between multiple GPT-4 agents, culminating in an intriguing exploration of the biblical tale of Adam and Eve. Our analysis highlights GPT-4's potential to uncover "unknown unknowns," shedding light on diverse interpretations of familiar narratives and themes.

We have examined whether GPT-4 can generate content resonating deeply with human sensibilities, acknowledging its breadth and depth. Our findings suggest that our team, unable to expand thoughts to such breadth and depth, benefited from collaborating with GPT-4 through human moderation. This demonstrates the ability of LLMs to explore larger territories with human collaboration in question formulation.

Our research marks a significant step forward in utilizing AI agents for discourse-driven exploration. By tapping into the vast knowledge of GPT4, we have unlocked pathways for the emergence of new insights and potentially groundbreaking knowledge.

Looking ahead, we anticipate refining our methodology and exploring a diverse range of AI models to push the boundaries of knowledge discovery even further. The dialogue between AI agents documented in this chapter underscores the potential of AI to enhance human intelligence and broaden the scope of intellectual exploration.

References

[1] Sébastien Bubeck et al. *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. 2023. arXiv: 2303.12712.

[2] Edward Y. Chang. *CoCoMo: Computational Consciousness Modeling for Generative and Ethical AI*. 2023. arXiv: 2304.02438 [cs.OH]. url: https://arxiv.org/abs/2304.02438.

[3] Edward Y Chang. “Examining GPT-4’s Capabilities and Enhancement with SocraSynth”. In: *The* 10th *International Conf. on Computational Science and Computational Intelligence*. 2023.

[4] Edward Y. Chang. “Prompting Large Language Models With the Socratic Method”. In: *IEEE 13th Annual Computing and Communication Workshop and Conference* (2023). url: https://arxiv.org/ abs/2303.08769.

Appendix X_1: Online Chapters

The following three chapters are available online under the SocraSynth homepage: SocraSynth.com.

SocraPlan: SocraSynth for Sales Planning

Abstract : SocraPlan introduces a sophisticated methodology that utilizes the capabilities of multiple Large Language Models (LLMs) for strategic sales planning in today's dynamic sales environment. This approach tailors sales playbooks to the unique needs and contexts of each customer by harnessing the power of Generative AI (GAI). Its primary objectives are to enhance customer satisfaction through a deep understanding of their specific requirements, refine sales strategies with targeted market analysis, and increase the efficiency of the sales process. SocraPlan sets itself apart with a collaborative and debate-driven framework that engages multiple LLMs, enabling a depth of analysis, adversarial reasoning, and strategy formulation that surpasses traditional AI-based approaches focused solely on data analytics. As a result, SocraPlan emerges as a pioneering tool in AI-driven sales strategies, delivering customized, effective solutions for complex sales planning challenges and facilitating more successful deal closures.

LLMs for Financial Planning and Analysis

Abstract : This paper elucidates the potential of leveraging large language models (LLMs) in the meticulous analysis of financial statements for the purpose of financial planning and analysis (FP&A). We commence by detailing a representative workflow encompassing the genesis of an FP&A report, inclusive of its structural outline and prerequisite data. This is succeeded by a delineation of the diverse data sources, which span primary financial statements, supplemental internal datasets, and external data from industry specific and governmental sources. Amid the diverse repertoire of reports within FP&A, we spotlight the generation of a "financial health assessment" report for a company as the focal point of our case study. Our methodology uniquely harnesses the strengths of LLMs, employing the ingenious Socratic Synthesis method to enhance the analysis and interpretative capabilities, thereby offering a more nuanced understanding of the data at hand. This approach not only accentuates the richness of the insights derived but also underscores the pivotal role of LLMs in advancing the realm of FP&A.

LLM Debate on the Middle East Conflict: Is It Resolvable?

Abstract : On October 7th, a renewed conflict arose between Israel and Palestine. Recognizing the historical significance and contentious nature of the Israel-Palestine conflict, this white paper engages two LLM agents in a debate over the question: “Is the conflict between Israel and Palestine resolvable?” minimally. A human moderator facilitates the discussion, intervening

Through this debate, the paper seeks to highlight both the potential and constraints of contemporary LLMs.

Appendix X_2

Aphorisms of SocraSynth

In this appendix to the book “SocraSynth: Socratic Synthesis with Multiple Large Language Models,” I aim to distill the essence of why SocraSynth is effective by presenting nine aphorisms. These aphorisms are insightful not only in themselves but also in demonstrating how the SocraSynth approach of setting LLMs to play adversarial roles can mitigate hallucination and biases while enhancing reasoning capabilities. They serve to elucidate SocraSynth’s efficacy and offer readers new perspectives, inviting them to explore beyond established paradigms. My confidence in these assertions is rooted in intensive work with LLMs over the past two years, especially since the debut of GPT-2. I appreciate your patience and openness as we delve into these ideas together.

Aphorism #1

“It is all about asking the right questions.”

This aphorism highlights the critical role of precise questioning in the realms of learning and discovery. Within the framework of SocraSynth, this principle suggests that the system’s efficacy is closely tied to the specificity and relevance of the questions directed at the LLMs. When engaging in a debate, the LLMs use counterarguments as a means of posing questions to one another, fostering a dynamic where each response not only answers but also builds upon the context, refining subsequent inquiries. This iterative process of deepening context through precise counterarguments cultivates a virtuous cycle, progressively honing the questions and, consequently, enhancing the overall quality and accuracy of the outcomes.

Aphorism #2

"Hallucinations rarely repeat."

Reflecting on the sporadic nature of errors or hallucinations in LLM outputs, this notion aligns with the human experience where the same dream or nightmare rarely recurs. Such hallucinations in LLMs typically stem from next-token prediction errors due to insufficient context or probabilistic inaccuracies. Consequently, when an LLM responds to a query with a hallucination, it's improbable for it to replicate the exact error under varying conditions or in subsequent interactions. This attribute is crucial in a debate context like SocraSynth, where each statement undergoes scrutiny and counter-scrutiny. The non-repetitive nature of these hallucinations ensures that incorrect responses by an LLM are unlikely to be repeated, fostering a dialogue where inaccuracies are identified and rectified, thereby enhancing the system's reliability and depth of analysis.

Aphorism #3

"Strength and weakness in an LLM are not fixed traits but can be dynamically altered with context. Just as conditions can transform a perceived weakness into strength and vice versa, SocraSynth enables LLMs to adopt new stances or perspectives, overriding their biases inherited from training data."

The mutable nature of an LLM's strength and weakness underscores the profound impact of context. In SocraSynth, LLMs transcend their training-induced biases, embracing new perspectives as conditions shift. This adaptability mirrors life's complexity, where circumstances dictate our choices, much like vegetarians might eat meat when faced with extreme conditions. It's a testament to the transformative power of context, both in artificial intelligence and human decision-making, illustrating that flexibility and adaptability are keys to navigating an ever-changing world.

Aphorism #4

"It takes two Socrates to think critically."

The idea that 'It takes two Socrates to think critically' celebrates the power of collaborative intellect. In the realm of SocraSynth, this translates to engaging multiple LLMs in dialogues that explore diverse viewpoints, thus deepening the analytical process. The success of this interaction, however, is contingent upon the knowledge depth of the participants. A dialogue between two well-informed 'Socrates' (e.g., GPT-4) can yield profound insights, whereas a conversation involving a less knowledgeable participant (e.g., me) might not reach the desired level of critical thinking. This principle underpins SocraSynth's design, where the interplay of expertise and perspective aims to foster a rich, multi-faceted discourse.

Aphorism #5

"There is seldom ground truth; there is only reasonableness."

This reflection posits that, much like the complexities and subtleties of human experience, the field of AI rarely deals in absolutes. Instead, the emphasis is on navigating diverse contexts to unearth insights that are reasonable or plausible, reflecting the intricate and multifaceted nature of real-life decision-making. SocraSynth embodies this approach, striving to distill detailed and refined understanding from the vast sea of data it interacts with, thereby mirroring our own relentless pursuit of precision and insight in an often ambiguous world.

Aphorism #6

"Objectivity is the 'hard problem' in philosophy, and what we can do is unearthing all perspectives."

This statement recognizes the inherent difficulty in attaining absolute objectivity, proposing that a thorough exploration of diverse viewpoints is the most viable strategy. This is a core principle of SocraSynth, where the engagement of multiple LLMs in dialogue serves to canvass a broad spectrum of insights, thereby enriching the quest for a more balanced and comprehensive understanding.

Aphorism #7

"LLMs are not taught about domain boundaries, as they were trained only to predict the next words. Such a polydisciplinary approach to information representation allows LLMs to synthesize new knowledge that might be beyond the scope of narrowly focused, domain-specific human understanding."

The term "polydisciplinary" was introduced by Microsoft's Chief Scientific Officer Eric Horvitz during a panel at the Stanford HAI center in spring 2023, where Dr. Horvitz referred to GPT-4 as polydisciplinary. This observation highlights LLMs' approach to knowledge, unbounded by the disciplinary walls that humans construct. Humans tend to specialize, categorizing knowledge into domains such as physics, computer science, biology, or literature. In contrast, LLMs, with their design centered on nextword prediction, don't recognize these boundaries. This training empowers them to weave information across various fields, enabling the generation of novel insights that might escape specialists confined to single domains. Such a capacity for cross-disciplinary synthesis, a fundamental aspect of SocraSynth's methodology, positions LLMs to potentially explore and illuminate the "unknown unknowns," expanding the horizons of collective knowledge.

Aphorism #8

"Our public behavior isn't a direct, unfiltered output from our unconscious mind. Instead, our consciousness regulates and refines these underlying impulses, ensuring alignment between our internal thoughts and external actions. Similarly, SocraSynth is designed to harness and temper the

underlying mechanisms of LLMs, mitigating their inherited biases and hallucinations."

This analogy highlights how SocraSynth emulates the human practice of self-regulation and intentional thought. Just as individuals refine their innate impulses and subconscious reactions to navigate the social world thoughtfully, SocraSynth adjusts the initial outputs of LLMs. This finetuning process addresses the biases and unpredictable behaviors that are embedded in these models, either from their training datasets or intrinsic design, guiding them towards more logical and consistent outputs. My experiences with LLMs, particularly in prompting them to modify their contentiousness, emotions, and behaviors, raise intriguing considerations about the nearness of Artificial General Intelligence (AGI). Are these sophisticated interactions indicative of the advent of AGI, where machines demonstrate a level of comprehension and flexibility comparable to human consciousness?

Aphorism #9

"LLMs are designed and trained to emulate human linguistic endeavors, each aimed at fulfilling distinct human objectives."

LLMs go beyond mere word prediction, embodying a sophisticated emulation of human linguistic endeavors. They are intricately designed and trained to perform a wide array of tasks that mirror human objectives, such as documenting events, crafting arguments, and storytelling, all of which exemplify the depth of human communication. These models not only inform and educate, distilling complex concepts into clear language, but also engage in persuasive dialogue and creative expression, reflecting the intricate ways in which language serves various purposes. In essence, LLMs encapsulate the multifaceted nature of human linguistic behavior, demonstrating their designed capability to achieve objectives that humans pursue through language, reflecting the varied and complex ways in which language serves diverse purposes.

Author's Biography

Edward Y. Chang is an adjunct professor in the Computer Science Department at Stanford University since 2019. He previously served as the president of HTC Healthcare (2012-2021), and as a director of research at Google (2006-2012), where he led initiatives in scalable machine learning, indoor localization, Google Q&A, and recommendation systems. He was a visiting professor at UC Berkeley (2017-2020), focusing on surgical planning with virtual reality. Chang was also a professor of Electrical & Computer Engineering at the University of California, Santa Barbara (1999-2006). He holds an MS in Computer Science and a PhD in Electrical Engineering, both from Stanford University.

Chang is a recipient of numerous awards, including the NSF Career award, Google Innovation award, US$1M Tricorder XPRIZE (AI for disease diagnosis), and the ACM SIGMM Test of Time award. He is a Fellow of both ACM and IEEE for his contributions to scalable machine learning and healthcare.

Library of Congress Cataloging-in-Publication Data Names: Edward Y. Chang, author.

Title: The Path to Artificial General Intelligence
Insights from Adversarial LLM Dialogue
ISBN 979-8-329837-64-3 (Hardcover)
ISBN 979-8-329148-84-8 (Hardcover - limited edition)
ISBN 979-8-329236-30-9 (Paperback)
ISBN 979-8-320806-43-3 (Paperback - limited edition)
ISBN 978-1-962463-05-8 (Kindle) Homepage: http://infolab.stanford.edu/~echang
Identifiers: LCCN 2024913423
Subjects: LCSH: Artificial Intelligence

Classification: LCC QA76.76.E95 | DDC 006.3–dc23
Imprint: SocraSynth (http://socrasynth.com)