Agentic memory-augmented retrieval and evidence grounding in medicine

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large language models (LLMs) hold promise for medical question answering (QA) and clinical decision support, yet remain limited by hallucination, rigid prompting requirements, and restricted context windows. Here, we introduce a unified, open-source LLM-based agentic system that integrates document retrieval, reranking, evidence grounding, and diagnosis generation to support dynamic, multi-step medical reasoning. Our system features a lightweight retrieval-augmented generation pipeline coupled with a cache-and-prune memory bank, enabling efficient long-context inference beyond standard LLM limits. The system autonomously invokes specialized tools, eliminating the need for manual prompt engineering or brittle multi-stage templates. Evaluated on five well-known medical QA benchmarks, our system outperforms or closely matches state-of-the-art proprietary (GPT-4) and open-source medical LLMs in multiple-choice and open-ended formats. These results underscore the effectiveness of tool-augmented, evidence-grounded reasoning for building reliable and scalable medical AI systems.

1 Introduction

Large language models (LLMs) are transforming medical research and practice, showing promise in tasks such as medical question answering (QA) and clinical decision support (1; 2; 3; 4). However, some challenges continue to limit their reliability and scalability in real world. One major concern is hallucination, which relates to the generation of confident yet factually incorrect or ungrounded responses. Another issue is the limited context window of current LLMs, which restricts the amount of information they can process at once, often necessitating retrieval-augmented generation (RAG) pipelines. While RAG improves grounding, it typically incorporates a subset of relevant evidence, which can introduce bias or lead to incomplete assessment (5; 6; 7). Additionally, many diagnostic systems require manually engineered multi-stage prompts (8; 9; 10; 11), making them difficult to scale and adapt. To improve reliability, recent work has explored continual pretraining on medical corpora (12; 13; 14), instruction fine-tuning and reinforcement learning to enhance medical reasoning (12; 14; 15; 16), and RAG frameworks for grounding model outputs in high-quality evidence (5; 6; 8; 9). Despite this progress, most systems focus on either improving reasoning or grounding, rather than jointly optimizing both. Yet, evidence-based medical practice requires sound diagnostic reasoning and alignment with high-quality clinical evidence (17).

To address these challenges, we present a unified, agentic system that integrates evidence retrieval, reranking, grounding, and diagnosis generation. Our system uses open-source tools to orchestrate the entire pipeline, from query analysis to final diagnosis, drawing from a comprehensive evidence base that includes PubMed abstracts and full texts, ClinicalTrials.gov entries, the *New England Journal of Medicine* (NEJM) case reports, medical textbooks, and curated Wikipedia content (5; 18; 19; 20; 21). To efficiently manage this information, we adopted a two-stage retrieval process including

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

coarse-grained retrieval followed by fine-grained reranking. To circumvent the limitations of LLM context windows, we introduced a cache-and-prune memory mechanism that retains high-relevance documents across reasoning steps, allowing the system to make informed decisions over extended sequences. Our contributions are summarized as follows:

- We propose a unified, fully-automated system that integrates document retrieval and reranking, evidence grounding, and diagnosis generation through an open-source AI agent.
- We present a tool-augmented LLM-based agentic architecture that enables dynamic multi-step tool use, eliminating the need for manually engineered prompts or multi-stage pipelines.
- We introduce a cache-and-prune memory bank mechanism that efficiently extends the retention of relevant documents for evidence grounding, enhancing diagnostic accuracy and computational efficiency.

2 Related work

2.1 Medical reasoning and diagnosis in language models

Recent advances in medical reasoning and diagnosis using LLMs have generally progressed along three major directions. The first line of work focused on continual pretraining of publicly available general-purpose LLMs on domain-specific medical corpora, including textbooks, research articles, and podcast transcripts (12; 13; 14; 22). The second direction emphasized instruction tuning or reinforcement learning using medical datasets, which may be manually curated or generated using systems like ChatGPT. These models are fine-tuned through supervised learning or reward feedback to improve chain-of-thought reasoning and emulate realistic doctor-patient interactions (12; 14; 15; 16). Both these strategies aim to enhance medical reasoning skills of general-purpose LLMs. However, despite gains on benchmarks, these models remain vulnerable to hallucinating factually incorrect or unsupported content. A third line of work has explored RAG pipelines to address hallucination risks by grounding model outputs in retrieved medical documents (5; 6; 8; 9; 11). RAG approaches have improved factuality, but often focus on retrieval, without simultaneously optimizing for complex diagnostic reasoning. These observations motivate the need for unified approaches that seamlessly combine robust evidence retrieval with dynamic, multi-step medical reasoning.

2.2 Medical AI agents

Medical AI agents leverage the reasoning and language capabilities of LLMs to perform complex clinical tasks, including diagnosis and decision support (23). Recent work on medical AI agents has evolved in three directions. The first focuses on role simulation, where agents emulate clinical roles, such as doctors, nurses, and patients, in simulated environments (24; 25; 26; 27; 28). These multi-agent systems aim to model clinical workflows through collaborative interactions and reasoning. The second direction centers on visual question answering, where agents are augmented with domain-specific tools, such as segmentation models for identifying salient regions in medical images and optical character recognition systems for processing textual content from clinical documents (29; 30). While promising, these approaches often lack explicit mechanisms for diagnostic reasoning or robust integration with large-scale medical knowledge bases. The third direction involves tool-augmented LLMs, where agents are equipped with capabilities such as document retrieval, function calling and database access. However, these systems often depend on resource-intensive model retraining or rely on closed-source, paid platforms (*e.g.*, GPT-4) (2; 31; 32; 33; 34), limiting scalability and transparency. Current trends point toward an unmet need for flexible, lightweight and interpretable frameworks that can dynamically orchestrate evidence gathering, reasoning, and clinical decision-making without prohibitive computational overhead. Our work addresses this emerging need by designing a modular, open, and deployment-friendly system for medical diagnosis support.

3 Methods

Our agentic system comprises three core components (Fig. 1): (1) a lightweight RAG pipeline for efficient evidence retrieval and reranking; (2) an open-source LLM-based agent that autonomously orchestrates diagnostic workflows, from retrieval to reasoning, grounding, and diagnosis generation;

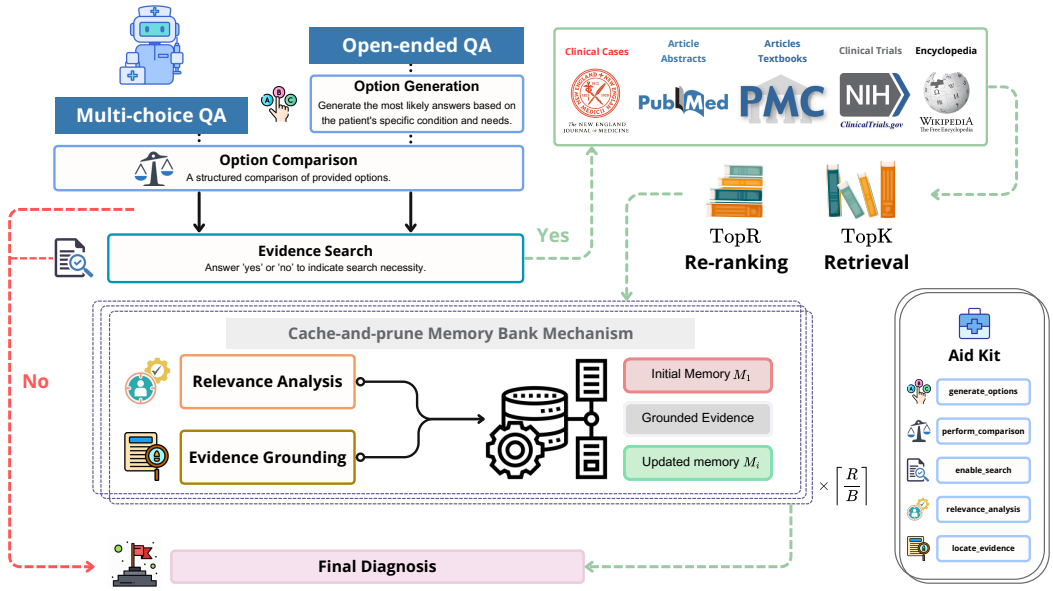

Figure 1: **Overview of the agentic system.** Our pipeline is powered by an open-source LLM-based agent that operates within a fully automated, dynamic workflow. When presented with either multiple-choice or open-ended medical questions, the agent leverages a suite of specialized tools to generate a structured comparison of answer choices or to synthesize plausible options in open-ended scenarios. It then dynamically assesses whether external evidence is needed to answer the question. If no external information is required, the agent proceeds directly to produce a final diagnosis. Otherwise, it initiates a retrieval process, querying a curated knowledge base to obtain the TopK relevant documents and rerank the TopR most informative sources. This evidence pool includes clinical case reports from NEJM, article abstracts from PubMed, full-text articles and textbooks from PubMed Central, clinical trials from ClinicalTrials.gov, and general content from Wikipedia. To manage long-context documents efficiently, the agent employs a cache-and-prune memory bank mechanism. It iteratively reviews B documents in $\lceil R/B \rceil$ batches until sufficient information is gathered, ensuring optimal comprehension within the model's context window. After synthesizing the selected evidence, the agent integrates key insights to deliver a grounded diagnosis. Its performance is further enhanced by an aid kit of five custom-designed tools, detailed in Section A.4.

and (3) a cache-and-prune memory bank that preserves relevant long-context documents to improve evidence use and diagnostic accuracy. Below we provide additional details on these components.

3.1 Lightweight RAG pipeline

We implemented a lightweight yet effective RAG pipeline to acquire relevant medical evidence tailored to patient-specific queries. This pipeline consists of two main stages: document retrieval and evidence reranking. In the retrieval stage, we utilized `SPECTER`, a semantic retriever trained with citation-informed objectives, which improved document-level representation, making it particularly effective in biomedical and scientific domains (5; 35). Denoted as ϕ, `SPECTER` retrieves documents by computing semantic similarity between the query representation $\mathbf{x}$ and document embeddings from the evidence corpus $\mathcal{V}$, using L2 distance as the similarity metric:

$$\mathrm{TopK}(\mathbf{x}, \mathcal{V}) = \underset{\mathbf{v} \in \mathcal{V}}{\arg\,\mathrm{TopK}} \; - \|\phi(\mathbf{x}) - \phi(\mathbf{v})\|_2 \,. \tag{1}$$

As summarized in Table S1, our evidence corpus includes diverse resources such as research paper abstracts and full texts, medical textbooks, clinical case reports, clinical trials, and curated Wikipedia articles. These are drawn from publicly accessible databases such as PubMed, PubMed Central, ClinicalTrials.gov, and Wikipedia. To refine the quality of retrieved TopK evidence, we implemented a reranking stage. Here, a quantized general text embedding model, `gte-Qwen2-7B-instruct`, was used to score and rank the candidate snippets at a finer granularity, and denoted as ψ (36; 37). This ensures that the top-ranked documents are semantically aligned with the query and optimally suited

for downstream diagnostic reasoning:

$$\mathrm{TopR}(\mathbf{x}, \mathcal{K}) = \underset{\mathbf{k} \in \mathcal{K}}{\arg\,\mathrm{TopR}}\ \cos\left(\psi(\mathbf{x}), \psi(\mathbf{k})\right), \tag{2}$$

where $\mathcal{K}$ represents the pool of documents retrieved from the six data sources, and $\mathcal{R}$ denotes the final ranked subset selected for use by the AI agent. Together, these two stages ensured that only the most relevant, high-quality evidence is forwarded for diagnostic processing. This design mitigates hallucination risks and supports accurate, grounded medical reasoning.

3.2 Agent for diagnostic workflow

We integrated an open-source LLM-based agent π as the core multi-step reasoning engine of our system to enable autonomous and interpretable medical decision-making. This agent orchestrates the entire diagnostic workflow, including document retrieval and reranking, patient query interpretation, evidence grounding, and diagnosis generation. We designed the agent to operate using a set of predefined tools (See Section A.4), eliminating the need for manually crafted prompts or rigid, hard-coded stages. Each tool encapsulated a specific function, such as querying external evidence sources, grounding highly-relevant documents, or synthesizing diagnostic conclusions. This allows the agent to perform complex clinical tasks in a structured and interpretable manner. By leveraging explicit tool usage and structured reasoning, the agent interacted dynamically and efficiently with the RAG pipeline and memory bank, enabling long-context, evidence-based clinical inference.

Specifically, in the initial step, given a predefined set of tools T, the patient's background and medical query Q, and instructions I, the AI agent generates a response sequence $\mathbf{y}$ following an autoregressive policy:

$$\pi(\mathbf{y} \mid T, Q, I) = \prod_{t} \pi\left(y_t \mid T, Q, I, \mathbf{y}_{<t}\right), \tag{3}$$

where $\mathbf{y}_{<t}$ denotes the previously generated tokens up to time step $t-1$.

Furthermore, at each step of the multi-step reasoning process, the agent autonomously selects the most appropriate tool to address the current subtask and produces intermediate responses in a multi-turn conversational format. Let C denote the full conversation history. At each step, the agent selects an action a from the action space A. Formally,

$$a \sim \pi\left(A \mid T, Q, I, C\right). \tag{4}$$

During execution, each intermediate reasoning step produced by the agent, along with any corresponding tool outputs, is appended to the conversation history C, enabling coherent multi-turn interactions. This modular tool-based design empowers the agent to flexibly respond to a wide range of clinical queries while ensuring transparency, reproducibility, and traceability throughout the diagnostic workflow. A detailed description of each tool's output parameters is provided in Fig. S4. Unlike traditional prompt engineering approaches, the agent autonomously determines when and how to invoke each tool through multi-step reasoning. This enables transparent, step-by-step justification of clinical decisions grounded in retrieved evidence. Importantly, the entire workflow operates locally, preserving patient privacy and minimizing reliance on proprietary APIs or cloud-based infrastructure.

3.3 Cache-and-prune memory bank mechanism

To overcome the context window limitations of LLMs and ensure persistent access to relevant evidence for the final diagnostic response, we implemented a cache-and-prune memory bank mechanism. This memory module functions as an external, dynamically updated storage that retains high-relevance documents retrieved and reranked during earlier stages of the pipeline. As shown in Algorithm 1, at each reasoning step indexed by i, the AI agent stores the grounded evidence in the memory bank M_i. During the final diagnosis generation, the agent accesses M_i, enabling long-horizon reasoning across multi-turn interactions. To avoid information overload, we designed a cache-and-prune mechanism that filters out outdated or unused evidence, guided by grounding tool usage patterns:

$$M_i = \mathrm{Prune}\left(M_{i-1} \cup \mathcal{B}_i\right), \quad i = 1, \ldots, \left\lceil \frac{R}{B} \right\rceil, \tag{5}$$

where $\mathcal{B}_i = \left\{ \mathbf{r}_i^j \;\middle|\; j = 1, \ldots, B \right\}$ represents the top-ranked documents from each reranked batch $\mathcal{R}$, and $\mathrm{Prune}(\cdot)$ is a logistic filtering function that removes documents that are not grounded by the AI

Algorithm 1 Agentic memory-augmented retrieval and evidence grounding system

Initialize Document Retriever ϕ, Evidence Reranker ψ
Initialize AI Agent π, Conversation C, Memory Bank M_1
Initialize Evidence database $\mathcal{V}$
Given patient background and question Q, instructions I, tools T
AI Agent π generates initial response $\prod_t \pi\left(y_t \mid T, Q, I, \mathbf{y}_{<t}\right)$
while tool calling **do**
    Retrieve content from the tool calling to update conversation C
    if tool calling is `enable_search` **then**
        Retrieve TopK documents $\arg\mathrm{TopK}_{\mathbf{v}\in\mathcal{V}} \; -\|\phi(\mathbf{x}) - \phi(\mathbf{v})\|_2$
        Rerank TopR documents $\arg\mathrm{TopR}_{\mathbf{k}\in\mathcal{K}} \; \cos\left(\psi(\mathbf{x}), \psi(\mathbf{k})\right)$
        while $i \leq \lceil R/B \rceil$ **do**
            Retrieve $\mathcal{B}_i$ (a batch of $\mathcal{R}$) to update conversation C
            if tool calling is `locate_evidence` **then**
                if Relevant document is grounded within <quote></quote> tags **then**
                    Update memory bank $M_i = \mathrm{Prune}\left(M_{i-1} \cup \mathcal{B}_i\right)$
                end if
            end if
            Remove $\mathcal{B}_i$ from conversation C
        end while until Sufficient information is gathered
    end if
end while
if M_i **then**
    return Final diagnosis $\prod_t \pi\left(y_t \mid T, Q, I, C, M_i, \mathbf{y}_{<t}\right)$
else
    return Final diagnosis $\prod_t \pi\left(y_t \mid T, Q, I, C, \mathbf{y}_{<t}\right)$
end if

agent. The final diagnosis is synthesized by conditioning on the complete conversational context, task, instructions, and the curated memory bank M_i:

$$\pi(\mathbf{y} \mid T, Q, I, C, M_i) = \prod_t \pi\left(y_t \mid T, Q, I, C, M_i, \mathbf{y}_{<t}\right). \tag{6}$$

Unlike standard RAG pipelines, which statically inject evidence into the prompt and risk truncation, our memory bank enables selective retention of key information and strategic pruning of less relevant content. This design supports broader context integration and sustained reasoning, mitigating fixed-window constraints and ensuring that only the most salient knowledge informs the agent's output (5).

3.4 Implementation details

All experiments were conducted locally on a distributed setup with four NVIDIA L40S GPUs, powered by the vLLM inference engine (38). We employed `Qwen2.5-72B-Instruct` as the primary backbone (i.e., AI agent), with the tensor parallelism and pipeline parallelism settings configured to 4 and 1, respectively. By default, the sampling parameters were set to a temperature of 0 and $\mathrm{top\_p}$ of 1. To address occasional issues with final answer extraction, we re-evaluated the experiments with a temperature of 0.7 and $\mathrm{top\_p}$ of 0.8. Due to diminished instruction following capabilities after enabling the static YaRN technique, we assigned the maximum context window to $32{,}768$ tokens (39). In practice, however, we observed an effective context window limit of approximately $10{,}000$ tokens. For each multi-turn conversation, we restricted the maximum number of tokens to $8,192$. Additionally, we selected the top 3 most relevant evidence documents for the baseline model that operates without tool access. For evidence retrieval, we fixed $\mathrm{TopK} = 32$ per source, resulting in 192 candidate documents from six sources. After reranking, we selected $\mathrm{TopR} = 32$ documents for use by the agent in downstream tasks (5; 6). Lastly, the cache-and-prune memory bank operates with a default batch size $B = 4$ for incremental evidence integration and pruning.

4 Experimental settings

4.1 Database for evidence retrieval

To ensure grounding in credible and up-to-date medical evidence, we assembled a comprehensive evidence corpus drawn from six trusted sources. They include peer-reviewed articles from PubMed Central, medical textbooks curated from the NLM LitArch Open Access Subset, and registered clinical trials from the National Library of Medicine at the U.S. National Institutes of Health (18; 19; 21). To enhance clinical relevance and provide real-world diagnostic context, we also incorporated clinical case reports published since 2016 in NEJM (20). We also included two supplementary sources, article abstracts and Wikipedia entries, originally curated by Xiong et al. (5). Section A.1 includes a detailed summary and description of each source included in our evidence retrieval database.

4.2 Benchmark evaluation across question formats

To evaluate the performance of our agentic system, we used five widely adopted medical question answering benchmarks: the United States Medical Licensing Examination (USMLE) Step 1, Step 2, and Step 3, and the English subsets of MedQA and MedExpQA (6; 40; 41). These datasets encompass a range of medical knowledge, clinical reasoning, and decision-making skills, and are well-established standards for evaluating LLMs. See Section A.2 and Table S3 for more details.

We ran experiments in two settings to test our approach: (1) multiple-choice QA, where models choose from given answer options, and (2) open-ended QA, where models generate answers without being given choices. We compared the performance of the agent against proprietary and open-source medical LLMs. Proprietary models included OpenAI GPT-4 and GPT-3.5 (i.e., ChatGPT), while the open-source models evaluated were BioMistral (7B), OpenBioLLM (8B/70B), UltraMedical (8B/70B), and PodGPT (70B) (2; 22; 42; 43; 44). To ensure a fair comparison, we manually ran all open-source models using the vLLM serving engine and applied a consistent zero-shot direct-response prompt. This decision was based on our observation that the performance of some models tended to degrade when presented with more complex instruction prompts. We also set model-specific maximum input lengths and generation token limits to accommodate varying context window sizes. See Section A.3 for more details.

For multiple-choice QA experiments, we activated four core tools within the AI agent: `perform_comparison`, `enable_search`, `relevance_analysis`, and `locate_evidence`. Accuracy was used as the primary evaluation metric, consistent with standard practices in the field (5; 6; 13; 15; 45). In the open-ended QA setting, we removed predefined answer options from the prompts and extended the `generate_options` tool by building it on top of the same four tools used in the multiple-choice setting. Performance was evaluated by cosine similarity based on two state-of-the-art embedding models: `SFR-Embedding-2_R` (SFR) from Salesforce Research and `gte-Qwen2-7B-instruct` (GTE) from Alibaba Group (36; 46). We also employed BERTScore's F1 metric, calculated using Microsoft's `deberta-xlarge-mnli` model, to compare the model-generated answer against ground truth (47). See Section A.4 for more details.

5 Results

5.1 Evaluation of multiple-choice benchmarks

Our agentic system achieved state-of-the-art performance across multiple-choice medical QA benchmarks, surpassing all evaluated models on USMLE Step 1, Step 2, and MedExpQA (Table 1). Specifically, it achieved 82.98% on Step 1 and 86.24% on Step 2, representing relative improvements of 2.31% and 4.57%, respectively, over GPT-4, which is the strongest baseline. On MedExpQA, where GPT-4 was not available, our model outperformed the next-best model (OpenBioLLM 70B at 71.20%) by a relative margin of 7.20%. For USMLE Step 3, our model reached 88.52%, narrowly trailing GPT-4 (89.78%) by only 1.26%. On MedQA, it scored 73.29%, which is 5.58% below GPT-4 but still ahead of all open-source models. When compared to the strongest open-source baseline, PodGPT (70B), our model demonstrated consistent and significant gains: 9.58% on Step 1, 13.76% on Step 2, 13.93% on Step 3, 8.25% on MedQA, and 15.20% on MedExpQA.

Table 1: **Performance evaluation on multiple choice medical QA benchmarks.** Accuracy scores across five benchmarks: USMLE Step 1–3, MedQA, and MedExpQA. The table compares our agentic system with proprietary (GPT-4, ChatGPT) and open-source (BioMistral, OpenBioLLM, UltraMedical, PodGPT) language models. **Bold** and underlined values denote the best and second-best performances for each benchmark, respectively.

Model	USMLE Step 1	USMLE Step 2	USMLE Step 3	MedQA	MedExpQA
GPT-4	80.67	81.67	**89.78**	**78.87**	N/A
ChatGPT	51.26	60.83	58.39	50.82	N/A
BioMistral (7B)	34.04	37.61	37.70	41.01	37.60
OpenBioLLM (8B)	47.87	44.04	50.00	47.84	43.20
UltraMedical (8B)	42.55	27.52	34.43	38.49	35.20
OpenBioLLM (70B)	69.15	70.64	68.85	69.13	71.20
UltraMedical (70B)	70.21	55.05	56.56	52.32	50.40
PodGPT (70B)	73.40	72.48	74.59	65.04	63.20
Ours	**82.98**	**86.24**	88.52	73.29	**78.40**

5.2 Evaluation of open-ended medical questions

Our agentic system achieved the highest performance across all five benchmarks in the open-ended question answering setting, outperforming all baseline models on nearly every metric (Table 2). For semantic textual similarity measured using SFR model, it achieved the top score on four of five benchmarks, including USMLE Step 1 (0.87), Step 2 (0.85), Step 3 (0.86), and MedExpQA (0.84), while ranking second on MedQA (0.85 *vs.* 0.86 from OpenBioLLM 70B). While measured by the GTE model, it outperformed all baselines on USMLE Steps 1–3 (0.66, 0.62, and 0.65 respectively), and was second-best on MedQA (0.61) and MedExpQA (0.60). Similarly, our system achieved the highest or second-highest BERTScore on all benchmarks, tying for the highest score on USMLE Step 1 (0.68), Step 2 (0.67) and MedExpQA (0.65), and ranking second on USMLE Step 3 (0.70 *vs.* 0.71 from OpenBioLLM 70B) and MedQA (0.67 *vs.* 0.70 from OpenBioLLM 70B).

Table 2: **Performance evaluation on open-ended medical questions.** This table reports model performance without answer choices using three embedding-based evaluation metrics: semantic textual similarity scores computed by two state-of-the-art embedding models (SFR and GTE) and BERTScore. Results are shown as mean ± standard deviation across five benchmarks (USMLE Steps 1–3, MedQA, and MedExpQA). **Bold** indicates the highest score, and underlined indicates the second-highest score for each metric within each benchmark.

Benchmark	Model	BioMistral (7B)	OpenBioLLM (8B)	UltraMedical (8B)	OpenBioLLM (70B)	UltraMedical (70B)	PodGPT (70B)	Ours
USMLE Step 1	SFR	0.79 ± 0.09	0.70 ± 0.12	0.81 ± 0.13	0.85 ± 0.10	0.82 ± 0.11	0.86 ± 0.11	**0.87** ± 0.09
	GTE	0.48 ± 0.17	0.38 ± 0.17	0.57 ± 0.21	0.60 ± 0.23	0.63 ± 0.23	0.66 ± 0.24	**0.66** ± 0.22
	BERTScore	0.58 ± 0.12	0.51 ± 0.13	0.61 ± 0.16	0.66 ± 0.17	0.64 ± 0.17	0.68 ± 0.20	**0.68** ± 0.17
USMLE Step 2	SFR	0.76 ± 0.11	0.71 ± 0.10	0.80 ± 0.11	0.82 ± 0.09	0.80 ± 0.10	0.85 ± 0.10	**0.85** ± 0.09
	GTE	0.45 ± 0.19	0.38 ± 0.15	0.52 ± 0.19	0.54 ± 0.19	0.59 ± 0.22	0.62 ± 0.21	**0.62** ± 0.22
	BERTScore	0.58 ± 0.11	0.56 ± 0.11	0.61 ± 0.13	0.64 ± 0.13	0.63 ± 0.14	0.66 ± 0.15	**0.67** ± 0.15
USMLE Step 3	SFR	0.74 ± 0.10	0.70 ± 0.10	0.79 ± 0.12	0.85 ± 0.11	0.80 ± 0.11	0.84 ± 0.11	**0.86** ± 0.09
	GTE	0.41 ± 0.18	0.38 ± 0.14	0.53 ± 0.22	0.63 ± 0.26	0.60 ± 0.23	0.63 ± 0.24	**0.65** ± 0.22
	BERTScore	0.57 ± 0.11	0.52 ± 0.14	0.60 ± 0.17	**0.71** ± 0.19	0.62 ± 0.15	0.67 ± 0.18	0.70 ± 0.17
MedQA	SFR	0.76 ± 0.10	0.71 ± 0.12	0.80 ± 0.12	**0.86** ± 0.11	0.80 ± 0.11	0.84 ± 0.11	0.85 ± 0.10
	GTE	0.43 ± 0.18	0.40 ± 0.17	0.53 ± 0.22	**0.63** ± 0.26	0.58 ± 0.23	0.60 ± 0.23	0.61 ± 0.23
	BERTScore	0.56 ± 0.12	0.52 ± 0.15	0.60 ± 0.16	**0.70** ± 0.19	0.61 ± 0.16	0.65 ± 0.18	0.67 ± 0.18
MedExpQA	SFR	0.76 ± 0.10	0.71 ± 0.11	0.78 ± 0.13	0.81 ± 0.11	0.77 ± 0.13	0.83 ± 0.11	**0.84** ± 0.10
	GTE	0.47 ± 0.18	0.40 ± 0.18	0.52 ± 0.22	0.54 ± 0.24	0.55 ± 0.22	**0.61** ± 0.23	0.60 ± 0.22
	BERTScore	0.58 ± 0.11	0.53 ± 0.12	0.58 ± 0.15	0.62 ± 0.17	0.60 ± 0.14	0.65 ± 0.17	**0.65** ± 0.16

5.3 Analysis of tool usage

Tool usage patterns revealed that the agent adapted its strategy to the complexity of each benchmark (Fig. 2a & Fig. 2b). While `perform_comparison` remained a consistent first-line tool across all exams, `enable_search` was used selectively, indicating the agent's discretion in deciding when external evidence was necessary to resolve clinical uncertainty. The progressively higher use of

relevance_analysis and locate_evidence tools from Step 1 to Step 3 underscores the agent's increasing reliance on iterative evidence appraisal and grounding in more advanced clinical scenarios. This aligns with the expectation that Step 3 questions, which often involve multi-system reasoning or longitudinal management, demand a deeper chain-of-thought and external validation. The wide distribution in the number of calls to these tools further supports the hypothesis that the agent's behavior is not hardcoded but context-dependent. In particular, questions that required repeated invocations of relevance_analysis and locate_evidence likely reflected either ambiguous clinical presentations or sparse initial document matches, prompting further rounds of evidence screening. Such behavior demonstrates the value of the cache-and-prune memory mechanism, which allowed the agent to incrementally accumulate, filter, and retain salient information while pruning irrelevant context. This architecture enabled scalable reasoning over long contexts without overwhelming the model's input window, supporting robust performance even in highly iterative diagnostic tasks. Overall, the tool usage patterns validate both the flexibility and compositional reasoning capabilities of the agent in adapting to a diverse range of clinical question formats.

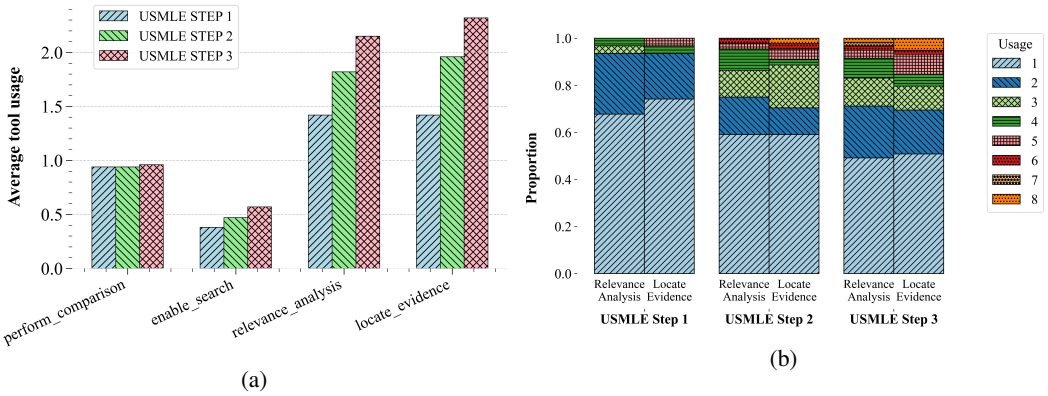

Figure 2: **Tool usage statistics across USMLE benchmarks.** (a) Bar plot showing the average number of times each tool was invoked per question across the USMLE Step 1, Step 2, and Step 3 benchmarks. Tools include perform_comparison, enable_search, relevance_analysis, and locate_evidence. (b) Stacked bar plot indicating the proportion of tool usage frequencies (from 1 to 8 calls) for relevance_analysis and locate_evidence, grouped by USMLE exam.

Table 3: **Impact of core components of the agentic system.** Performance comparison of the agentic system with ablated versions lacking key components: tool integration, cache-and-prune memory mechanism, and external evidence search. Values for ablations indicate the relative percentage drop in accuracy compared to the full model across USMLE Step 1, Step 2, and Step 3 benchmarks.

Benchmark	USMLE Step 1	USMLE Step 2	USMLE Step 3	Average
Ours	**82.98**	**86.24**	**88.52**	**85.91**
w/o Tools	-1.07	-3.67	-4.91	-3.22
w/o Cache & Prune	-1.07	-2.75	-3.27	-2.36
w/o Evidence Search	-2.13	-3.67	-6.55	-4.12

5.4 Ablation studies

We compared performance with and without tool access to evaluate the impact of incorporating tools into the agentic pipeline. Specifically, we performed evaluation using structured instructions I without tool access (w/o Tools), and using the same instructions with full access to the toolset T (Ours). As shown in Table 3, tool integration led to performance improvements: 1.07% on USMLE Step 1, 3.67% on USMLE Step 2, and 4.91% on Step 3, with an average gain of 3.22% across all of them. These results underscore the value of equipping the agent with specialized tools.

To isolate the contribution of individual components, we conducted targeted ablations. Removing the relevance_analysis and locate_evidence tools (denoted w/o Cache & Prune) resulted in an average drop of 2.36%, with performance reductions of 1.07%, 2.75%, 3.27% on USMLE Step 1-3,

highlighting the utility of the iterative memory mechanism. When we removed the `enable_search` tool and the document retrieval and reranking modules (`w/o Evidence Search`), performance dropped by 4.12% on average, with declines of 2.13%, 3.67%, and 6.55% on Steps 1, 2, and 3, respectively, emphasizing the critical role of external evidence in clinical reasoning.

We evaluated how the number of documents retrieved and reranked influenced the performance (Figure 3). Accuracy generally improved with increasing context length up to $\mathrm{TopR} = 32$, beyond which gains plateaued. For Step 2, performance peaked at $\mathrm{TopR} = 8$ with a 7.80% improvement over GPT-4 and remained stable (5.60% gain) from $\mathrm{TopR} = 32$ onward. Step 1 exhibited a similar trend, with gains peaking at 5.50% at $\mathrm{TopR} = 4$ and plateauing beyond $\mathrm{TopR} = 8$. In contrast, while step 3 exhibited lower performance relative to GPT-4, its performance fluctuated slightly at lower TopR values and stabilized around -1.40% to -0.50% from $\mathrm{TopR} = 4$ onward. These results highlight the effectiveness of our cache-and-prune memory bank in leveraging extended context efficiently, while also demonstrating the diminishing utility of low-ranked evidence beyond $\mathrm{TopR} = 32$.

6 Limitations, broader impact and future work

Despite the strong performance of our agentic system, some limitations highlight important directions for future research. First, while our system is designed as a general-purpose medical QA agent, its toolset may require domain-specific customization to handle specialized tasks, such as rare disease diagnosis or surgical decision-making. Incorporating adaptive or plug-and-play tools tailored to niche clinical domains could expand its applicability. Second, the sequential execution of tools, particularly for evidence retrieval and analysis, can introduce latency and limit scalability in real-time or high-throughput settings. Future work will explore parallelized tool execution, caching strategies across sessions, and learned policies for tool invocation to improve computational efficiency. Third, while our evaluation covered a range of benchmarks, real-world clinical scenarios often involve ambiguous, noisy or incomplete data. Expanding evaluations to include complex settings such as NEJM clinicopathological conferences, longitudinal case reports, or multimodal inputs will be important to assess robustness in high-stakes use cases (1; 48).

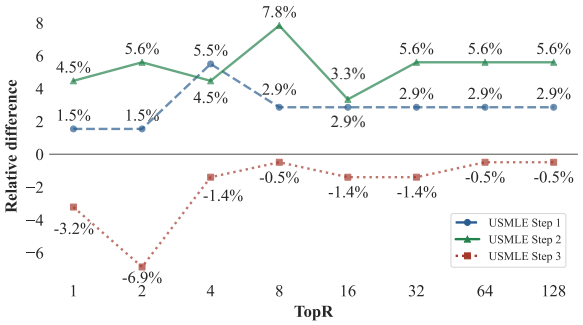

Figure 3: **Impact of evidence context length.** The figure shows the relative performance change on USMLE Step 1, Step 2, and Step 3 benchmarks as a function of the number of top reranked documents (TopR) processed by the agentic system. Each point represents the performance difference relative to GPT-4. Different line styles and colors indicate the benchmark type. The y-axis shows the relative difference in accuracy, and the x-axis denotes the number of retrieved documents.

Looking ahead, we envision broader societal impacts of our work in democratizing medical expertise through accessible, open-source AI systems. However, these benefits must be pursued alongside safeguards for transparency, accountability, and patient safety. As tool-based agents become more capable, interdisciplinary collaboration between clinicians, ethicists, and technologists will be important to ensure their responsible integration into clinical workflows.

7 Conclusion

We present a unified, fully automated agentic system that integrates document retrieval, evidence reranking, and grounded diagnosis generation through an open-source agent. By enabling dynamic, multi-step reasoning with seamless tool integration, our system removes the need for manual prompt engineering or complex multi-stage pipelines. To overcome the context window limitations of LLMs, we introduced a cache-and-prune memory bank mechanism that improves evidence synthesis and supports more robust diagnostic reasoning. Across five medical benchmarks, our system consistently delivered strong performance, outperforming or matching leading LLMs. These findings highlight the role of tool-based reasoning in building reliable, scalable, and clinically useful medical AI systems.

References

[1] D. McDuff, M. Schaekermann, T. Tu, A. Palepu, A. Wang, J. Garrison, K. Singhal, Y. Sharma, S. Azizi, K. Kulkarni, *et al.*, "Towards accurate differential diagnosis with large language models," *Nature*, pp. 1–7, 2025.

[2] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of GPT-4 on medical challenge problems," *arXiv preprint arXiv:2303.13375*, 2023.

[3] P. Hager, F. Jungmann, R. Holland, K. Bhagat, I. Hubrecht, M. Knauer, J. Vielhauer, M. Makowski, R. Braren, G. Kaissis, *et al.*, "Evaluation and mitigation of the limitations of large language models in clinical decision-making," *Nature Medicine*, vol. 30, no. 9, pp. 2613–2622, 2024.

[4] S. Sandmann, S. Hegselmann, M. Fujarski, L. Bickmann, B. Wild, R. Eils, and J. Varghese, "Benchmark evaluation of DeepSeek large language models in clinical decision-making," *Nature Medicine*, 2025.

[5] G. Xiong, Q. Jin, Z. Lu, and A. Zhang, "Benchmarking retrieval-augmented generation for medicine," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024* (L. Ku, A. Martins, and V. Srikumar, eds.), pp. 6233–6251, Association for Computational Linguistics, 2024.

[6] I. Alonso, M. Oronoz, and R. Agerri, "MedExpQA: Multilingual benchmarking of large language models for medical question answering," *Artificial Intelligence in Medicine*, vol. 155, p. 102938, 2024.

[7] R. Yang, Y. Ning, E. Keppo, M. Liu, C. Hong, D. S. Bitterman, J. C. L. Ong, D. S. W. Ting, and N. Liu, "Retrieval-augmented generation for generative artificial intelligence in health care," *npj Health Systems*, vol. 2, no. 1, p. 2, 2025.

[8] M. Jeong, J. Sohn, M. Sung, and J. Kang, "Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models," *Bioinformatics*, vol. 40, no. Supplement_1, pp. i119–i129, 2024.

[9] G. Xiong, Q. Jin, X. Wang, M. Zhang, Z. Lu, and A. Zhang, "Improving retrieval-augmented generation in medicine with iterative follow-up questions," in *Biocomputing 2025: Proceedings of the Pacific Symposium*, pp. 199–214, World Scientific, 2024.

[10] R. Alzghoul, A. Ayaabdelhaq, A. Tabaza, and A. Altamimi, "CLD-MEC at MEDIQA- CORR 2024 task: GPT-4 multi-stage clinical chain of thought prompting for medical errors detection and correction," in *Proceedings of the 6th Clinical Natural Language Processing Workshop, ClinicalNLP@NAACL 2024, Mexico City, Mexico, June 21, 2024* (T. Naumann, A. B. Abacha, S. Bethard, K. Roberts, and D. S. Bitterman, eds.), pp. 537–556, Association for Computational Linguistics, 2024.

[11] Y. Chen, P. Sun, X. Li, and X. Chu, "MRD-RAG: Enhancing medical diagnosis with multi-round retrieval-augmented generation," *arXiv preprint arXiv:2504.07724*, 2025.

[12] C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, and Y. Wang, "PMC-LLaMA: Toward building open-source language models for medicine," *Journal of the American Medical Informatics Association*, vol. 31, no. 9, pp. 1833–1843, 2024.

[13] X. Wang, N. Chen, J. Chen, Y. Hu, Y. Wang, X. Wu, A. Gao, X. Wan, H. Li, and B. Wang, "Apollo: An lightweight multilingual medical LLM towards democratizing medical AI to 6B people," *arXiv preprint arXiv:2403.03640*, 2024.

[14] P. Qiu, C. Wu, X. Zhang, W. Lin, H. Wang, Y. Zhang, Y. Wang, and W. Xie, "Towards building multilingual language model for medicine," *Nature Communications*, vol. 15, no. 1, p. 8384, 2024.

[15] H. Zhang, J. Chen, F. Jiang, F. Yu, Z. Chen, G. Chen, J. Li, X. Wu, Z. Zhang, Q. Xiao, X. Wan, B. Wang, and H. Li, "HuatuoGPT, towards taming language model to be a doctor," in *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023* (H. Bouamor, J. Pino, and K. Bali, eds.), pp. 10859–10885, Association for Computational Linguistics, 2023.

[16] J. Chen, Z. Cai, K. Ji, X. Wang, W. Liu, R. Wang, J. Hou, and B. Wang, "HuatuoGPT-o1, Towards medical complex reasoning with LLMs," *arXiv preprint arXiv:2412.18925*, 2024.

[17] V. Subbiah, "The next generation of evidence-based medicine," *Nature Medicine*, vol. 29, no. 1, pp. 49–58, 2023.

[18] National Library of Medicine (US), "Pubmed Central," 2024. National Center for Biotechnology Information, U.S. National Library of Medicine.

[19] J. E. Gillen, T. Tse, N. C. Ide, and A. T. McCray, "Design, implementation and management of a web-based data entry system for ClinicalTrials.gov," in *MEDINFO 2004 - Proceedings of the 11th World Congress on Medical Informatics, San Francisco, California, USA, September 7-11, 2004* (M. Fieschi, E. W. Coiera, and Y. J. Li, eds.), vol. 107 of *Studies in Health Technology and Informatics*, pp. 1466–1470, IOS Press, 2004.

[20] E. W. Campion, L. Scott, A. Graham, J. M. Prince, S. Morrissey, and J. M. Drazen, "NEJM.org — 20 years on the web," *New England Journal of Medicine*, vol. 375, no. 10, pp. 993–994, 2016.

[21] National Center for Biotechnology Information (US), "About Bookshelf [Internet]." `https://www.ncbi.nlm.nih.gov/books/about/openaccess/`, 2010. NLM LitArch Open Access Subset.

[22] S. Jia, S. Bit, E. Searls, M. V. Lauber, L. A. Claus, P. Fan, V. H. Jasodanand, D. Veerapaneni, W. M. Wang, R. Au, *et al.*, "PodGPT: An audio-augmented large language model for research and education," *npj Biomedical Innovations*, 2025.

[23] J. Luo, W. Zhang, Y. Yuan, Y. Zhao, J. Yang, Y. Gu, B. Wu, B. Chen, Z. Qiao, Q. Long, R. Tu, X. Luo, W. Ju, Z. Xiao, Y. Wang, M. Xiao, C. Liu, J. Yuan, S. Zhang, Y. Jin, F. Zhang, X. Wu, H. Zhao, D. Tao, P. S. Yu, and M. Zhang, "Large language model agent: A survey on methodology, applications and challenges," *arXiv preprint arXiv:2503.21460*, 2025.

[24] H. Yu, J. Zhou, L. Li, S. Chen, J. Gallifant, A. Shi, X. Li, W. Hua, M. Jin, G. Chen, Y. Zhou, Z. Li, T. Gupte, M. Chen, Z. Azizi, Y. Zhang, T. L. Assimes, X. Ma, D. S. Bitterman, L. Lu, and L. Fan, "AIPatient: Simulating patients with EHRs and LLM powered agentic workflow," *arXiv preprint arXiv:2409.18924*, 2024.

[25] J. Li, S. Wang, M. Zhang, W. Li, Y. Lai, X. Kang, W. Ma, and Y. Liu, "Agent Hospital: A simulacrum of hospital with evolvable medical agents," *arXiv preprint arXiv:2405.02957*, 2024.

[26] W. Yan, H. Liu, T. Wu, Q. Chen, W. Wang, H. Chai, J. Wang, W. Zhao, Y. Zhang, R. Zhang, and L. Zhu, "ClinicalLab: Aligning agents for multi-departmental clinical diagnostics in the real world," *arXiv preprint arXiv:2406.13890*, 2024.

[27] M. K. Almansoori, K. Kumar, and H. Cholakkal, "Self-evolving multi-agent simulations for realistic clinical interactions," *arXiv preprint arXiv:2503.22678*, 2025.

[28] H. Li, W. Pan, S. Rajendran, C. Zang, and F. Wang, "TrialGenie: Empowering clinical trial design with agentic intelligence and real world data," *medRxiv*, 2025.

[29] A. Fallahpour, J. Ma, A. Munim, H. Lyu, and B. Wang, "MedRAX: Medical reasoning agent for chest X-ray," *arXiv preprint arXiv:2502.02673*, 2025.

[30] N. Sharma, "CXR-Agent: Vision-language models for chest X-ray interpretation with uncertainty aware radiology reporting," *arXiv preprint arXiv:2407.08811*, 2024.

[31] S. Gao, R. Zhu, Z. Kong, A. Noori, X. Su, C. Ginder, T. Tsiligkaridis, and M. Zitnik, "Tx-Agent: An AI agent for therapeutic reasoning across a universe of tools," *arXiv preprint arXiv:2503.10970*, 2025.

[32] P. Lu, B. Chen, S. Liu, R. Thapa, J. Boen, and J. Zou, "OctoTools: An agentic framework with extensible tools for complex reasoning," *arXiv preprint arXiv:2502.11271*, 2025.

[33] Y. Liao, S. Jiang, Y. Wang, and Y. Wang, "ReflecTool: Towards reflection-aware tool-augmented clinical agents," *arXiv preprint arXiv:2410.17657*, 2024.

[34] A. J. Goodell, S. N. Chu, D. Rouholiman, and L. F. Chu, "Large language model agents can use tools to perform clinical calculations," *npj Digital Medicine*, vol. 8, no. 1, 2025.

[35] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. S. Weld, "SPECTER: Document-level representation learning using citation-informed Transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020* (D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, eds.), pp. 2270–2282, Association for Computational Linguistics, 2020.

[36] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, and M. Zhang, "Towards general text embeddings with multi-stage contrastive learning," *arXiv preprint arXiv:2308.03281*, 2023.

[37] Bitsandbytes Development Team, "Accessible large language models via k-bit quantization for PyTorch," 2024. GitHub repository.

[38] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with PagedAttention," in *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023* (J. Flinn, M. I. Seltzer, P. Druschel, A. Kaufmann, and J. Mace, eds.), pp. 611–626, Association for Computing Machinery, 2023.

[39] B. Peng, J. Quesnelle, H. Fan, and E. Shippole, "YaRN: Efficient context window extension of large language models," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, International Conference on Learning Representations, 2024.

[40] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, *et al.*, "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," *PLOS Digital Health*, vol. 2, no. 2, p. e0000198, 2023.

[41] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What disease does this patient have? A large-scale open domain question answering dataset from medical exams," *Applied Sciences*, vol. 11, no. 14, 2021.

[42] Y. Labrak, A. Bazoge, E. Morin, P. Gourraud, M. Rouvier, and R. Dufour, "BioMistral: A collection of open-source pretrained large language models for medical domains," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024* (L. Ku, A. Martins, and V. Srikumar, eds.), pp. 5848–5864, Association for Computational Linguistics, 2024.

[43] Meta AI, "How Llama is helping Saama deliver new possibilities in personalized medicine and data-driven care." `https://ai.meta.com/blog/saama-data-driven-care-built-with-llama`, 2025.

[44] K. Zhang, S. Zeng, E. Hua, N. Ding, Z. Chen, Z. Ma, H. Li, G. Cui, B. Qi, X. Zhu, X. Lv, J. Hu, Z. Liu, and B. Zhou, "UltraMedical: Building specialized generalists in biomedicine," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024* (A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, eds.), 2024.

[45] Y. Gao, D. Dligach, T. A. Miller, J. R. Caskey, B. Sharma, M. M. Churpek, and M. Afshar, "DR.BENCH: Diagnostic reasoning benchmark for clinical natural language processing," *Journal of Biomedical Informatics*, vol. 138, p. 104286, 2023.

[46] R. Meng, Y. Liu, S. R. Joty, C. Xiong, Y. Zhou, and S. Yavuz, "SFR-Embedding-2: Advanced text embedding with multi-stage training," 2024.

[47] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *The Eleventh International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, International Conference on Learning Representations, 2020.

[48] A. V. Eriksen, S. Möller, and J. Ryg, "Use of GPT-4 to diagnose complex clinical cases," *NEJM AI*, vol. 1, no. 1, p. AIp2300031, 2024.

[49] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.

[50] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 53728–53741, Curran Associates, Inc., 2023.

[51] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, International Conference on Learning Representations, 2022.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main contributions and scope are carefully articulated in both the abstract and the introduction. To ensure clarity, we also provide a concise summary of the contributions at the end of the introduction (see Section 1).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, we explicitly discuss three key limitations of our current work in Section 6, along with proposed directions to address them in future research.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not present any theoretical results; therefore, no assumptions or formal proofs are provided.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide a comprehensive description of our methods in Section 3 and include full implementation and experimental details in Section 4, ensuring that all information necessary to reproduce the results and support the paper's conclusions is fully disclosed.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: As detailed in Section A.7, we clearly outline the accessibility of all resources used in our study. The source code will be released publicly via GitHub with accompanying documentation. While the clinical case data from the *New England Journal of Medicine* (NEJM) are subject to licensing restrictions and cannot be publicly shared, all other datasets will be made available via Hugging Face under the Creative Commons Attribution-NonCommercial-NoDerivatives (CC BY-NC-ND) license, reflecting the most restrictive terms of the PubMed Central articles included in our corpus.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide a comprehensive description of the implementation setup and experimental settings in Section 3.4 and Section 4, including model configurations, hyperparameter choices, database for evidence retrieval, benchmark evaluation across question formats, ensuring clarity and reproducibility of the reported results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: Due to the substantial computational cost associated with large-scale LLM evaluations, all reported results in our experiments are based on a single run. As a result, we do not include error bars or statistical significance measures.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide detailed information about the computational resources in Section 3.4. All experiments were conducted on a distributed local setup using four NVIDIA L40S GPUs, with inference powered by the vLLM engine. This setup description includes the hardware specifications and inference configuration necessary for reproducing the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research presented in this paper fully adheres to the NeurIPS Code of Ethics, ensuring ethical guidelines are followed throughout the study.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader impacts of our work are discussed in Section 6.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Upon open-sourcing our assets via Hugging Face, we will clearly specify the intended use, license terms, and access restrictions (e.g., non-commercial and no-derivatives use), along with detailed documentation outlining download, processing, and usage instructions. Furthermore, we will prominently disclose the license of the backbone model used in our work, `Qwen2.5-72B-Instruct`, which is governed by the Qwen license agreement (see huggingface.co/Qwen/Qwen2.5-72B-Instruct/LICENSE). These steps collectively support transparent and ethically responsible distribution of the data and models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All external assets used in this work, including datasets, source code, and open-source models, are properly cited and employed in full accordance with their respective licenses and terms of use. Publicly available sources such as PubMed Central, ClinicalTrials.gov, and the NLM LitArch Open Access Subset were selected based on their open-access policies, including the most restrictive Creative Commons Attribution-NonCommercial-NoDerivatives (CC BY-NC-ND) license, where applicable. Pre-indexed corpora and other resources from prior work (e.g., Xiong et al. (5)) are explicitly acknowledged and reused in compliance with their original licensing terms. Open-source language models and software frameworks, such as vLLM, are used in accordance with their licenses and properly credited. Additionally, clinical case reports from the *New England Journal of Medicine* (NEJM) were obtained under an exclusive license from the NEJM Group.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We provide detailed documentation of the new assets introduced in this work. Specifically, in Table S1 and Table S2, we summarize the six corpora used for our retrieval-augmented generation (RAG) evidence corpus, including key statistics. We also list the specific journals and article counts included from PubMed Central with Creative Commons licenses, ensuring transparency and reproducibility of the dataset construction.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: In this work, we adopted the open-sourced `Qwen2.5-72B-Instruct` model as the backbone of our AI agent, which plays a central role in the core methods. The LLM is integrated with specialized tools for medical reasoning and evidence retrieval, forming the foundation of our multi-step diagnostic pipeline. Its usage is essential and original to the system's design and performance.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Appendix

A.1 Database for evidence retrieval

We constructed a comprehensive retrieval-augmented generation (RAG) evidence corpus by aggregating content from six trusted medical and scientific sources to ensure clinical relevance, diversity, and open accessibility. A summary of the dataset statistics, including the number of segmented snippets and their average token lengths (computed using the `Qwen2.5-72B-Instruct` tokenizer), is provided in Table S1. The corpus includes research articles published under Creative Commons licenses from leading biomedical journals indexed in PubMed Central, with specific journal titles and article counts detailed in Table S2 (18). We also incorporated clinical trial records from ClinicalTrials.gov, filtering for studies that had completed recruitment and were classified as Phase 3 or Phase 4, or that investigated device-based or behavioral interventions (19). This selection yielded $156,887$ trials as of March 2025. To enhance real-world clinical applicability, we included $1,479$ clinical case reports published by the *New England Journal of Medicine* between 2016 and March 2025. We further adopted pre-indexed corpora of PubMed abstracts and Wikipedia entries from Xiong et al. (5), which have demonstrated strong utility for medical question answering tasks. Finally, we leveraged $8,226$ open-access medical textbooks from the NLM LitArch Open Access Subset, hosted by the U.S. National Library of Medicine (21). Together, these six sources form the backbone of our evidence retrieval module, supporting the agent's multi-step diagnostic reasoning with high-quality, domain-relevant content.

Table S1: **Overview of data sources for evidence retrieval.** This table summarizes the six corpora comprising our RAG database. For each source, we report the number of full documents, the number of tokenized text snippets used for retrieval, and the average token length per document (as computed using the `Qwen2.5-72B-Instruct` tokenizer). Databases are listed in descending order of document count.

Corpus	Number of Docs	Number of Snippets	Average Length
PubMed Abstracts	23,897,881	23,897,881	290.01
Wikipedia	6,458,670	29,642,311	166.47
Clinical Trials	156,887	4,177,121	268.33
PubMed Central Articles	123,194	8,155,929	202.46
Textbooks	8,226	2,224,013	207.95
Clinical Cases	1,479	17,821	215.61

A.2 Experimental benchmarks

We evaluated our system using five medical question answering benchmarks: USMLE Step 1, USMLE Step 2, USMLE Step 3, MedQA, and MedExpQA (Table S3). Each benchmark includes clinical case descriptions, multiple-choice options, and a correct answer.

The USMLE is a three-step examination series designed to assess progressively advanced competencies required for medical practice in the United States. All steps primarily use multiple-choice questions structured as clinical scenarios to evaluate critical thinking and clinical judgment. Step 1 focuses on foundational knowledge in the basic sciences, including physiology, pharmacology, pathology, and disease mechanisms. It serves as a critical assessment of preclinical competencies and includes 94 clinical cases (40). Step 2, also known as clinical knowledge, evaluates the ability to apply medical and clinical science in the context of supervised patient care. It emphasizes diagnostic reasoning, clinical management, and ethical decision-making, with a benchmark of 109 questions (40). Step 3 assesses readiness for independent practice by testing advanced clinical reasoning and decision-making skills across complex scenarios, including diagnosis, prognosis, and patient management. This benchmark includes 122 test cases.

MedQA is a curated benchmark for four-choice, free-form medical question answering, collected after the USMLE board exams. It spans material from Steps 1 through 3 and covers a broad range of clinical knowledge and case-based scenarios. While the original dataset includes both simplified and traditional Chinese, we used the English subset, which contains $1,273$ test cases (41). MedExpQA

Table S2: **Journals and article counts included from PubMed Central.** This table lists the 74 most represented journals in our corpus, sorted in descending order by article count. These journals span general medicine, specialty domains, and global health, contributing to a diverse and comprehensive retrieval corpus. The final row reports the total number of included articles from all journals.

Journal Title	Article Count	Journal Title	Article Count
BMJ Open	37,488	JAMA Ophthalmol	434
Proc Natl Acad Sci U S A	16,619	Lancet HIV	387
JAMA Netw Open	10,824	BMJ Health Care Inform	366
Nature	8,148	JAMA Surg	287
Cell	4,811	BMJ Neurol Open	282
Science	4,660	JAMA Dermatol	279
BMJ	3,636	Lancet Psychiatry	270
BMJ Glob Health	3,460	Lancet Public Health	264
N Engl J Med	2,159	BMJ Support Palliat Care	262
BMJ Open Qual	1,569	BMJ Nutr Prev Health	254
JAMA	1,552	Lancet Respir Med	252
BMJ Open Diabetes Res Care	1,434	JAMA Cardiol	239
Lancet	1,344	Lancet Diabetes Endocrinol	225
Neurology	1,216	Lancet Microbe	167
BMJ Open Sport Exerc Med	1,201	BMJ Ment Health	167
Lancet Reg Health West Pac	1,196	JAMA Otolaryngol Head Neck Surg	164
BMJ Case Rep	1,190	Lancet Planet Health	162
BMJ Paediatr Open	1,145	Lancet Haematol	157
Lancet Reg Health Eur	1,077	BMJ Med	154
BMJ Open Respir Res	1,031	Lancet Child Adolesc Health	154
Lancet Reg Health Am	901	BMJ Evid Based Med	136
Ann Intern Med	881	Lancet Digit Health	124
Lancet Glob Health	805	BMJ Surg Interv Health Technol	120
JAMA Intern Med	797	Lancet Gastroenterol Hepatol	117
Lancet Infect Dis	676	BMJ Oncol	114
BMJ Open Ophthalmol	656	Lancet Healthy Longev	102
JAMA Neurol	639	BMJ Sex Reprod Health	100
JAMA Health Forum	628	BMJ Mil Health	64
BMJ Open Gastroenterol	625	Lancet Rheumatol	61
Lancet Oncol	613	BMJ Open Sci	49
BMJ Qual Saf	601	BMJ Innov	46
JAMA Psychiatry	597	BMJ Simul Technol Enhanc Learn	42
JAMA Pediatr	569	JAMA Facial Plast Surg	39
BMJ Qual Improv Rep	547	Ann Intern Med Clin Cases	6
JAMA Oncol	490	BMJ Outcomes	1
Lancet Reg Health Southeast Asia	464	BMJ Clin Evid	1
BMJ Public Health	453		
Lancet Neurol	444	**Total Number of Articles**	123,194

Table S3: **Overview of benchmark datasets used for evaluation.** This table summarizes the five medical QA benchmarks evaluated in our study. For each dataset, we report the total number of test cases and the maximum number of answer choices presented per question.

Benchmark	Number of Testing Cases	Number of Choices
USMLE Step 1 (40)	94	9
USMLE Step 2 (40)	109	6
USMLE Step 3 (40)	122	6
MedQA (41)	1,273	4
MedExpQA (6)	125	5

follows a similar format and was constructed from the Spanish national residency medical exam. It consists of 125 test cases, each with five answer choices and detailed explanations. For our evaluation, we used the translated and annotated English subset (6).

A.3 Backbone large language models

Our AI agent was benchmarked against closed-source and open-source models, spanning general-purpose and medical-specific LLMs. Specifically, we compared medical diagnosis performance with leading proprietary models, including OpenAI's GPT-4 and ChatGPT (2). On the open-source front, we included recent state-of-the-art medical LLMs such as BioMistral, OpenBioLLM, UltraMedical, and PodGPT. For all the models evaluated in this study, including our AI agent, we reported performance in the zero-shot setting.

We adopted the `Qwen2.5-72B-Instruct` model as the backbone of our AI agent. The open-source Qwen series has demonstrated competitive performance against Meta's LLaMA 3.1 models on various open-domain benchmarks, including knowledge-based and math-based tasks (49). By default, Qwen models support a context window of up to $32,768$ tokens, which can be extended to 128K tokens using the YaRN technique (39). However, we observed a decline in instruction-following capabilities when extending the context window under vLLM version $0.6.3$. Consequently, we retained the default maximum context window of $32,768$ tokens for all experiments. Due to computational resource constraints, we focused exclusively on this model as our AI agent.

GPT-4 and GPT-3.5 (ChatGPT) from OpenAI are advanced general-purpose language models that excel across a broad spectrum of real-world tasks. In the domain of medical question answering, they have achieved state-of-the-art performance and are widely regarded as strong baselines. The evaluation results for these models, specifically `gpt-4-turbo` and `gpt-3.5-turbo`, are reported in (2).

BioMistral is the first biomedical language model based on the Mistral architecture, continually pre-trained on PubMed Central articles released under Creative Commons licenses (42). It demonstrates improved performance on medical benchmarks compared to baseline models. In our experiments, due to its $2,048$-token context window limitation, we generated up to 128 tokens. We omitted the system prompt, as the Mistral chat template did not support it.

OpenBioLLM builds upon the LLaMA 3 architecture and is available in both 8B and 70B parameter versions. These models are fine-tuned using direct preference optimization, a reinforcement learning-based alignment technique (50). OpenBioLLM demonstrates competitive performance against both its baseline and proprietary counterparts (43). In this study, we evaluated the 8B and 70B variants. Additionally, we configured the models with a maximum context length of $8,192$ tokens and generated up to $1,024$ tokens per response.

UltraMedical models, trained through supervised fine-tuning and preference-based learning, demonstrate competitive performance with proprietary LLMs such as OpenAI GPT-4 (44; 50). In our experiments, we evaluated both the 8B model, based on LLaMA 3.1, and the 70B model, based on LLaMA 3, as the LLaMA 3.1 version of the UltraMedical 70B model was not publicly available at the time of this study.

PodGPT is a family of language models continually pre-trained on publicly available podcasts spanning the domains of science, technology, engineering, mathematics, and medicine (STEMM). Designed specifically for scientific and educational applications, these models were evaluated across a range of STEMM benchmarks, including datasets focused on medical question answering (22). We employed the best-performing PodGPT model, based on the `Llama-3.3-70B-Instruct` architecture fine-tuned with a low-rank adapter (51). To maintain consistency with the OpenBioLLM and UltraMedical configurations, we set the context window to $8,192$ tokens and allowed up to $1,024$ tokens to be generated.

A.4 Designed tools

As illustrated in Fig. S1, we designed five specialized tools to serve as the diagnostic aid kit within our AI agent. For question interpretation, the agent uses `perform_comparison` to handle multiple-choice tasks and `generate_options` for open-ended scenarios, enabling flexible reasoning formats. In particular, `generate_options` is tailored for scenarios lacking predefined choices, enabling the

Table S4: **Parameters used in the agent's toolset.** This table outlines the parameters and their corresponding descriptions for each tool integrated into our diagnostic framework. The tools `perform_comparison`, `enable_search`, `relevance_analysis`, and `locate_evidence` are used for multiple-choice QA tasks. For open-ended QA, the `generate_options` tool is additionally employed, generating plausible answer options for further analysis.

Tool Name	Parameter	Parameter Description
`generate_options`	answers	The most likely answers based on the patient's specific condition and needs.
`perform_comparison`	comparisons	A structured comparison of all options, detailing their relevance to the patient's case.
`enable_search`	search	Answer 'yes' or 'no' to indicate search necessity.
`relevance_analysis`	analysis	A comprehensive analysis detailing the relevance of each document to the patient's presentation, highlighting key matches, inconsistencies, and important findings.
`locate_evidence`	evidence	Relevant evidence applicable to the patient's presentation, with article IDs in <quote></quote>tags.

agent to propose plausible answer candidates from the contextual details of the case. Additionally, the `enable_search` tool determines whether external evidence is necessary to support a diagnosis. To facilitate evidence retrieval and interpretation, `relevance_analysis` evaluates the semantic alignment between the patient's case and retrieved documents, while `locate_evidence` identifies and grounds specific articles most pertinent to the diagnosis.

A.5 Evaluation models for open-ended question answering

In this work, we employed two state-of-the-art semantic similarity models, `SFR-Embedding-2_R` (SFR) and `gte-Qwen2-7B-instruct` (GTE), alongside BERTScore, enabling fine-grained semantic comparison between the model-generated responses and ground-truth answers (46; 36; 47). For both SFR and GTE, we used the default cosine similarity to compute phrase-level similarity, while for BERTScore, we reported the F1 metric to assess alignment at the token level.

The `SFR-Embedding-2_R` model is based on the Mistral architecture with 7 billion parameters and supports input lengths of up to $4,096$ tokens (46). This model achieves strong results on the massive text embedding benchmark (MTEB), highlighting its robustness for semantic similarity tasks.

The `gte-Qwen2-7B-instruct` model was built on the Qwen2 architecture with 7 billion parameters and supports input lengths of up to 32K tokens (36). It was instruction-tuned for a range of natural language processing tasks, including retrieval, classification, and reranking. The model ranks highly on the MTEB leaderboard, demonstrating state-of-the-art performance in semantic textual similarity.

BERTScore evaluates the semantic similarity between two phrases by computing the cosine similarity between their contextualized token embeddings, derived from a pretrained language model (47). In our experiments, we used the `deberta-xlarge-mnli` model as the backbone for BERTScore computation. This model, with 750 million parameters, was fine-tuned on the multi-genre natural language inference tasks, making it particularly well-suited for assessing phrase and sentence-level semantic alignment in open-ended medical QA tasks.

A.6 Used prompts

To ensure fair and consistent evaluation, we employed a unified set of prompts across all open-source models in a direct-response format. Each model was paired with its designated chat template, as defined by its tokenizer specifications. For our AI agent, the primary prompt templates used for multiple-choice question answering are presented in Table S5, with a standardized SYSTEM PROMPT applied uniformly across all configurations. For the open-ended QA setting, we adapted the same templates by removing the answer choices, allowing the models to generate free-form diagnostic responses. The corresponding prompt format for open-ended questions is provided in Table S6.

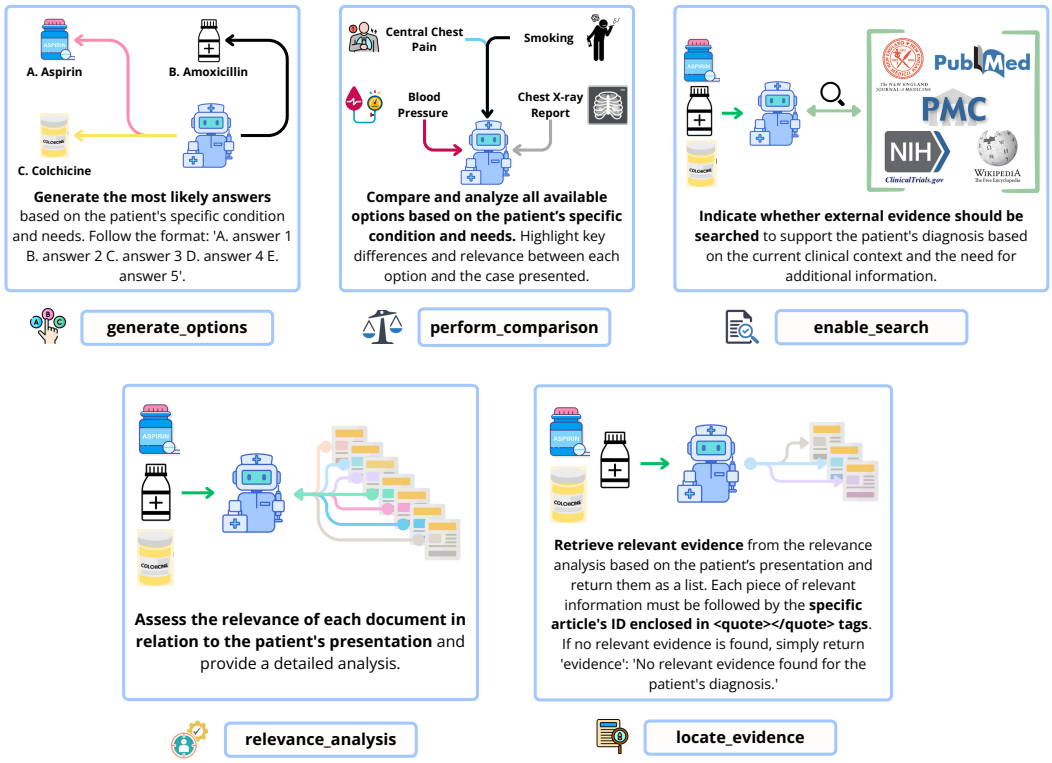

Figure S1: **Overview of specialized tools in the agentic framework.** This figure illustrates the five custom-designed tools used by the agent for medical question answering in the open-ended setting. Each tool performs a distinct role: `generate_options` first proposes potential answers to the problem, `perform_comparison` then analyzes the candidate options in the context of the problem description, `enable_search` decides whether external evidence is needed, `relevance_analysis` assesses the contextual fit of retrieved documents, and `locate_evidence` extracts grounded evidence snippets tied to article IDs. Together, these tools enable dynamic, interpretable, and evidence-grounded reasoning.

A.7 Data and code availability

The clinical case data from NEJM used in this study are not publicly available and can be obtained under an exclusive licensing agreement with the NEJM Group. All other datasets used in this work, sourced from publicly accessible platforms such as PubMed Central, ClinicalTrials.gov, and the National Library of Medicine, will be released via Hugging Face under a Creative Commons Attribution-NonCommercial-NoDerivatives (CC BY-NC-ND) license. The full source code developed for this study, including all implementation and evaluation scripts, will be made publicly available on GitHub, along with detailed documentation and instructions to facilitate reproducibility.

Table S5: **Prompt templates for multiple-choice question answering.** This table presents the SYSTEM PROMPT and PROMPT TEMPLATE used for multiple-choice QA, along with the document formatting template and the cache-and-prune memory bank mechanism template employed by our AI agent.

SYSTEM PROMPT
You are a medical professional specializing in evidence-based medicine (EBM). Your role is to answer questions using a systematic approach, integrating the best available research evidence, clinical expertise, and patient-specific factors.
PROMPT TEMPLATE
Here is the background information and question about the patient: <background> {background} </background> The available answer options are: <option> {option} </option> Follow these steps to answer the question: 1. Compare each option with the case details, analyzing key clues in the text to identify the best choice. 2. If the question can be answered through comparison, directly return the best option term with the option capital within <final_result></final_result>tags, placing the explanation outside of the <final_result>tags. 3. If multiple options are plausible or additional evidence is needed for better decision-making, enable search to find credible sources. 4. Analyze the relevance between each document and the patient's presentation, followed by a systematic search to locate relevant evidence applicable to the patient's case. 5. While we are continuing to provide additional evidence, iterate the previous step to analyze more additional evidence. 6. Once sufficient information is gathered, return the best option term with the option capital within <final_result></final_result>tags, placing the explanation outside of the <final_result>tags.
Document template
Relevant documents related to the patient's care: <document> {document} </document>
Cache-and-prune memory bank mechanism template
Here are the selected relevant documents related to the patient's care: <document> {document} </document> Review your answer and return the best option term with the option capital within the <final_result></final_result>tags, leave the explanation outside of the <final_result>tags.

Table S6: **Prompt templates for open-ended question answering.** This table presents the PROMPT TEMPLATE, document formatting template, and the template for the cache-and-prune memory bank mechanism used in open-ended QA.

PROMPT TEMPLATE
Here is the background information and question about the patient: <background> {background} </background> Follow these steps to answer the question: 1. Compare each option with the case details, analyzing key clues in the text to identify the best choice. 2. If the question can be answered through comparison, directly return the full answer term within <final_result></final_result>tags, placing the explanation outside of the <final_result>tags. 3. If multiple options are plausible or additional evidence is needed for better decision-making, enable search to find credible sources. 4. Analyze the relevance between each document and the patient's presentation, followed by a systematic search to locate relevant evidence applicable to the patient's case. 5. While we are continuing to provide additional evidence, iterate the previous step to analyze more additional evidence. 6. Once sufficient information is gathered, return the full answer term within <final_result></final_result>tags, placing the explanation outside of the <final_result>tags.
Document template
Relevant documents related to the patient's care: <document> {document} </document>
Cache-and-prune memory bank mechanism template
Here are the selected relevant documents related to the patient's care: <document> {document} </document> Review your answer and return the full answer term within the <final_result></final_result>tags, leave the explanation outside of the <final_result>tags.