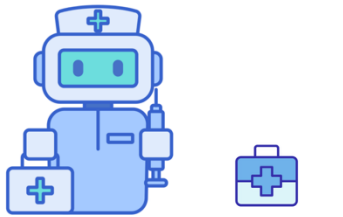
Evidence Retrieval and Grounding in Medicine

Shuyue Jia

Ph.D. Student

Advisor: Dr. Vijaya B. Kolachalama

Aug 9, 2025



Department of Electrical and Computer Engineering

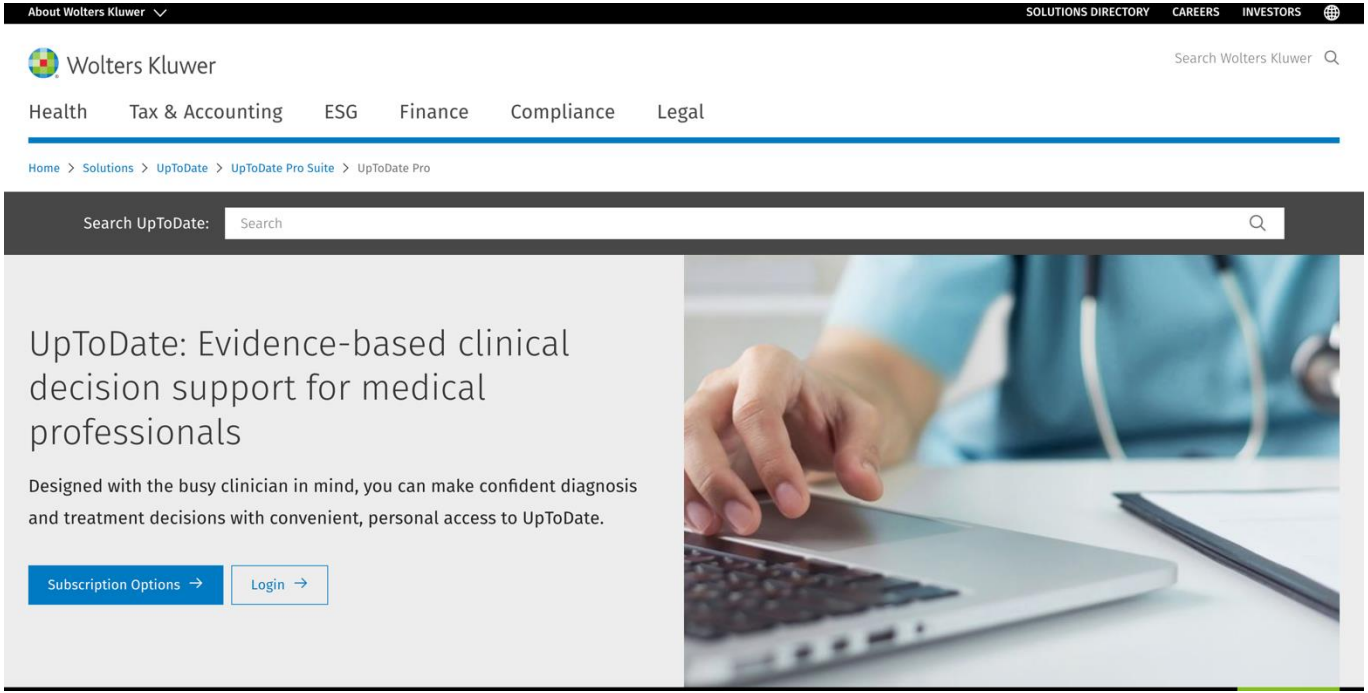
Boston University

Outline

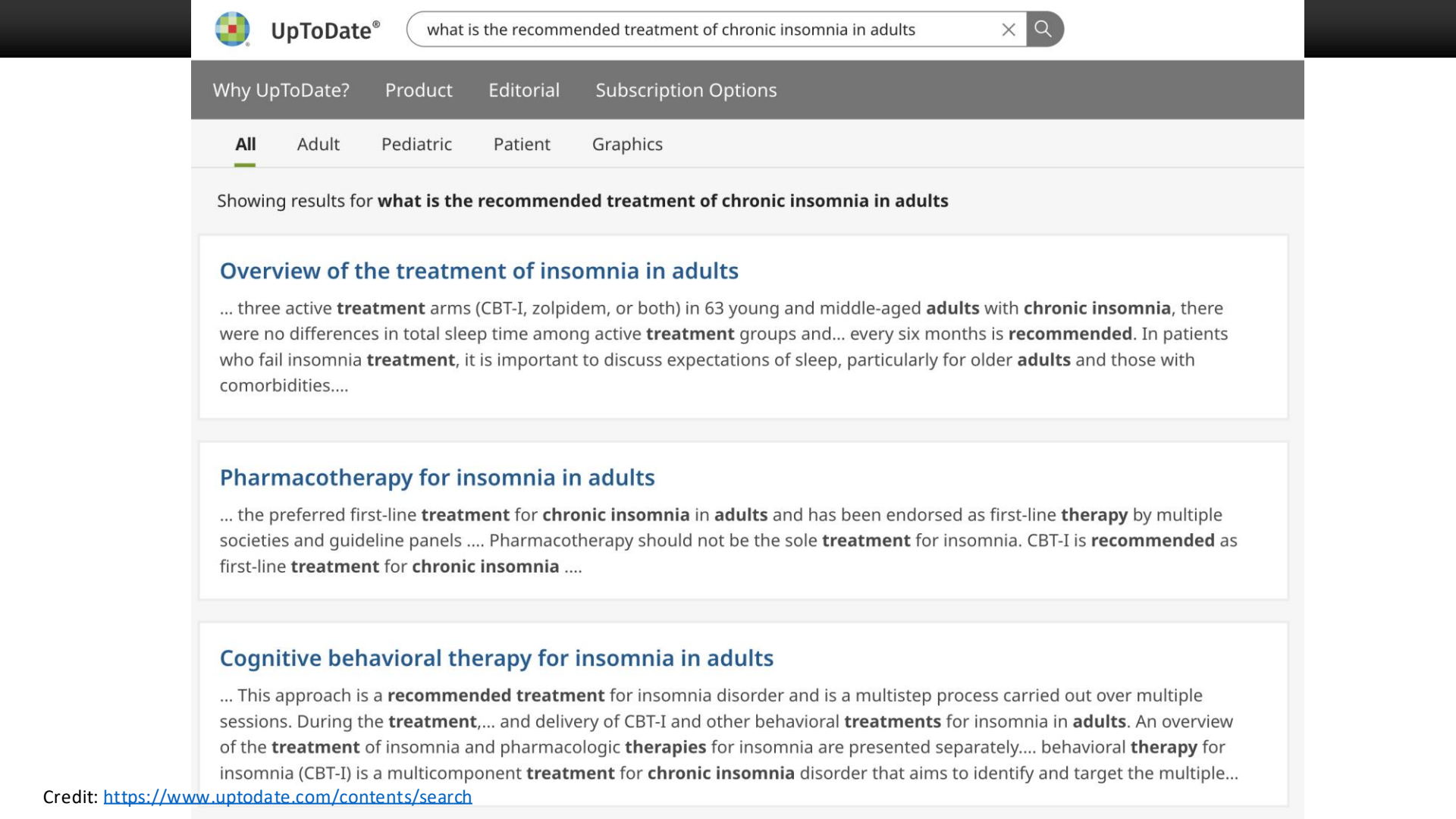
- Background
- Related Works and Motivation
- MedPodGPT
- PodGPT
- Agentic System

Part 1 – Background

- **Traditional search engine: UpToDate** ($6,713 million revenue, 2024)



Credit: https://www.uptodate.com/contents/search



Credit: https://www.uptodate.com/contents/search

< Back

Pharmacotherapy for insomnia in adults

Outline

Select Language

AUTHOR: David N Neubauer, MD
SECTION EDITORS: Ruth Benca, MD, PhD, Joann G Elmore, MD, MPH
DEPUTY EDITOR: April F Eichler, MD, MPH

Literature review current through: **Apr 2025.**
This topic last updated: **May 21, 2025.**

INTRODUCTION

Insomnia is one of the most common symptoms for which adults seek medical advice. While pharmacotherapy should not be the sole treatment for insomnia, medication may be part of an integrated approach that includes behavioral strategies and treatment of relevant comorbidities.

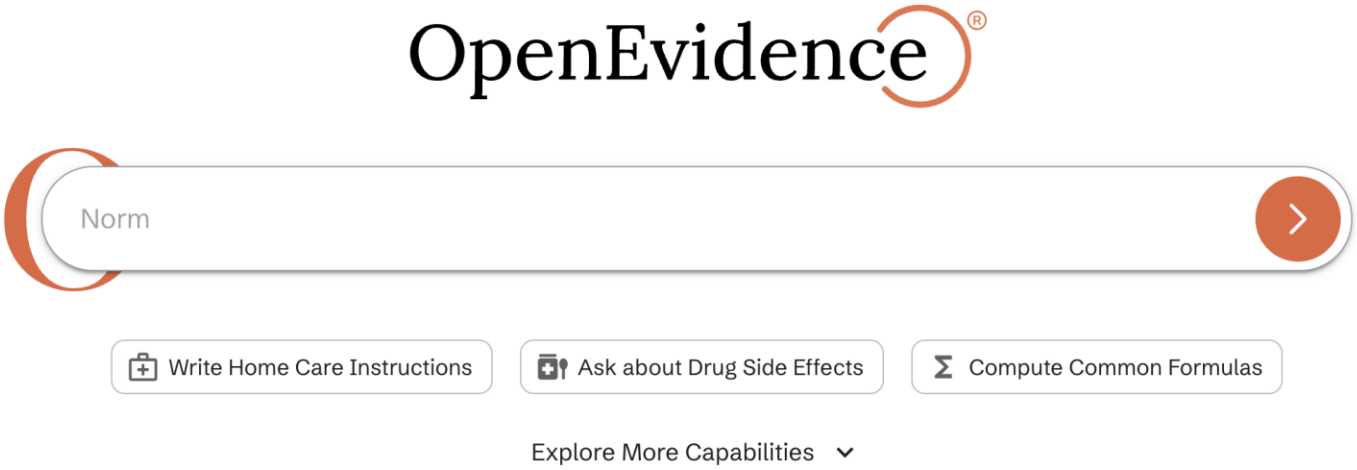
A range of medications and substances are used to treat insomnia in adults, spanning the following categories:

- Medications with regulatory approval for treatment of insomnia
- Off-label prescription medications prescribed for insomnia
- Over-the-counter (OTC) sleep aids (eg, diphenhydramine and doxylamine, alone or in "PM" formulated medications)

Credit: https://www.uptodate.com/contents/search

Part 1 – Background

- **AI-powered search engine: OpenEvidence** ($1 billion valuation, 2025)



Credit: https://www.openevidence.com/

Partnering with OpenEvidence: A Life-Saving Healthcare Revolution

This AI-powered search engine for doctors could be one of the most important companies of the next decade.

Credit: https://www.sequoiacap.com/article/partnering-with-openevidence-a-life-saving-healthcare-revolution/

What is the latest diagnosis of dementia?

Finished thinking

The **latest diagnostic criteria for dementia** are based on a combination of clinical assessment and, increasingly, biomarker evidence, with criteria evolving to reflect advances in the understanding of neurodegenerative diseases.

For all-cause dementia, the National Institute on Aging and the Alzheimer's Association (NIA-AA) recommend that dementia is diagnosed when there is evidence of cognitive or behavioral (neuropsychiatric) symptoms that interfere significantly with the ability to function at work or in usual daily activities. This determination is made through clinical judgment, based on history from the patient and a knowledgeable informant, and supported by cognitive testing. The diagnosis requires impairment in at least two cognitive domains (such as memory, executive function, language, visuospatial skills, or personality/behavior), and the deficits must represent a decline from previous levels of functioning and not be explained by delirium or major psychiatric disorder. The NIA-AA emphasizes that the distinction between mild cognitive impairment and dementia is based on the degree of functional impairment, with dementia requiring significant interference in daily life.[1]

Recent updates from the Alzheimer's Association Workgroup further incorporate biomarkers into the diagnostic process, particularly for Alzheimer's disease. The 2024 criteria define Alzheimer's disease as a biological process, with diagnosis possible in asymptomatic individuals based on abnormal amyloid and tau biomarkers (e.g., amyloid PET, CSF, or plasma phosphorylated tau). An abnormal "Core 1" biomarker is sufficient to establish a diagnosis of Alzheimer's disease, while "Core 2" biomarkers (such as tau PET) provide prognostic information and increase diagnostic confidence. The criteria also recognize the importance of integrating clinical and biological staging, acknowledging that comorbidities and cognitive reserve may influence the clinical presentation.[2-3]

In summary, the most up-to-date consensus is that **dementia is diagnosed clinically by evidence of acquired cognitive decline interfering with daily function, with increasing use of biomarkers for etiological specificity, especially in Alzheimer's disease, as recommended by the National Institute**

In summary, the most up-to-date consensus is that **dementia is diagnosed clinically by evidence of acquired cognitive decline interfering with daily function, with increasing use of biomarkers for etiological specificity, especially in Alzheimer's disease, as recommended by the National Institute on Aging and the Alzheimer's Association.**[2]

Helpful Not Helpful

References

1. The Diagnosis of Dementia Due to Alzheimer's Disease: Recommendations From the National Institute on Aging-Alzheimer's Association Workgroups on Diagnostic Guidelines for Alzheimer's Disease.

McKhann GM, Knopman DS, Chertkow H, et al.

Alzheimer's & Dementia : The Journal of the Alzheimer's Association. 2011;7(3):263-9. doi:10.1016/j.jalz.2011.03.005.

☆ Leading Journal

2. Revised Criteria for Diagnosis and Staging of Alzheimer's Disease: Alzheimer's Association Workgroup.

Jack CR, Andrews JS, Beach TG, et al.

Alzheimer's & Dementia : The Journal of the Alzheimer's Association. 2024;20(8):5143-5169. doi:10.1002/alz.13859.

☆ Leading Journal New Research

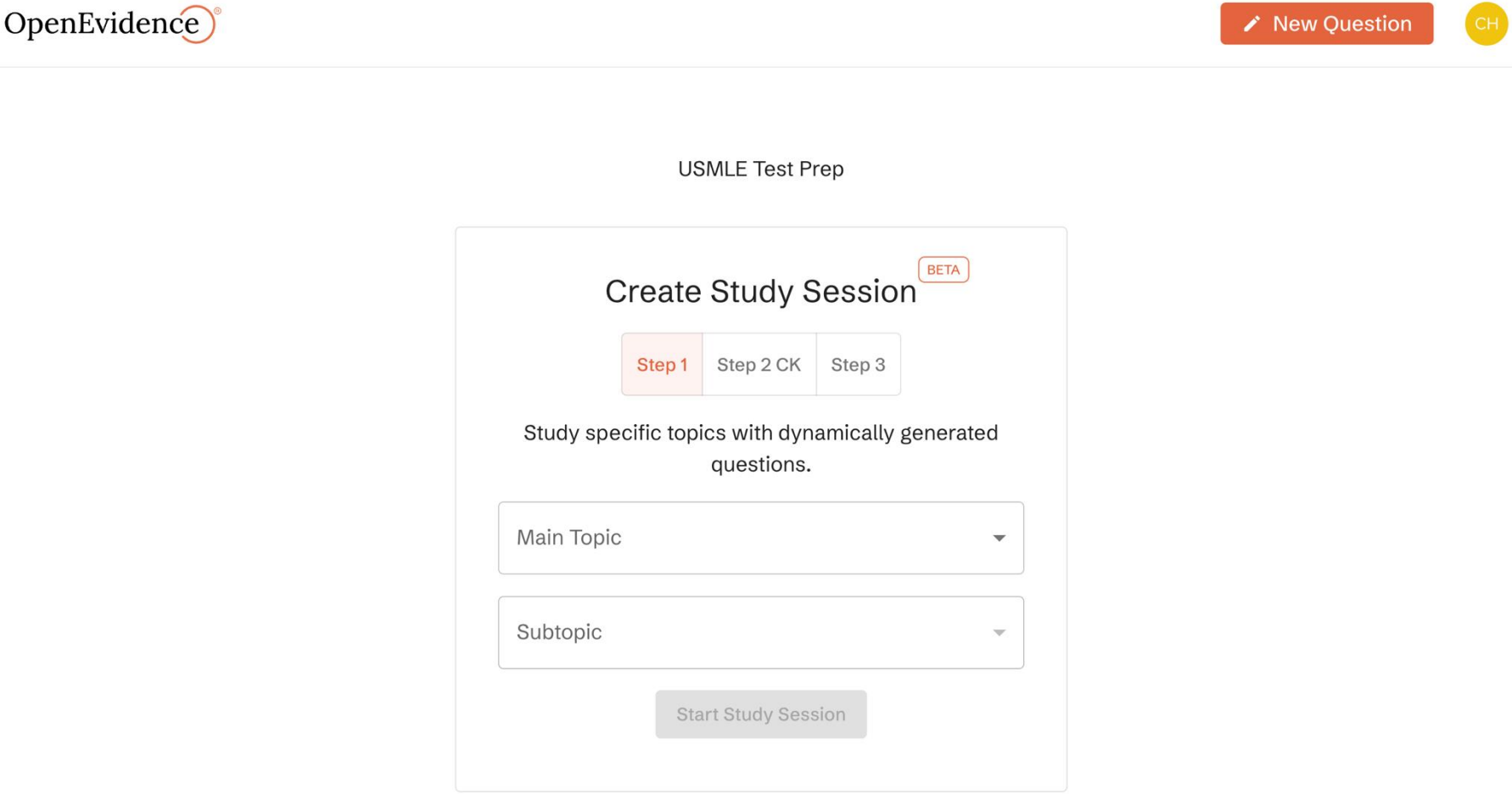
3. Influence of Different Diagnostic Criteria on Alzheimer Disease Clinical Research.

Bieger A, Brum WS, Borelli WV, et al.

Neurology. 2024;103(5):e209753. doi:10.1212/WNL.0000000000209753.

New Research

Credit: https://www.openevidence.com/



Credit: https://www.openevidence.com/

Study Session ⓘ View Session Stats

A 29-year-old female presents to the clinic with a chief complaint of fatigue and weight gain over the past six months. She works as a high school teacher in a rural area with limited access to healthcare specialists. The patient reports feeling cold all the time and has experienced constipation and mild depression. She also notes her periods have become irregular, and she has had difficulty concentrating at work. She denies any recent illnesses or significant stressors.

Her past medical history is significant for a head injury from a motor vehicle accident two years ago, which required hospitalization but did not result in any lasting neurological deficits. She takes no regular medications and denies smoking, alcohol, or illicit drug use.

On physical examination, the patient appears tired and has a dry, pale skin texture. Her vital signs are: blood pressure 110/70 mmHg, heart rate 58 bpm, temperature 96.8°F (36°C), and respiratory rate 14 breaths per minute. Thyroid examination reveals no palpable goiter. Reflexes are delayed, and she has mild facial puffiness with periorbital edema. No other abnormalities are noted.

Laboratory results reveal:
- TSH: 0.4 µIU/mL (normal: 0.5-4.5 µIU/mL)
- Free T4: 0.6 ng/dL (normal: 0.8-1.8 ng/dL)
- Free T3: 1.8 pg/mL (normal: 2.3-4.2 pg/mL)
- Serum cortisol (8 AM): 5 µg/dL (normal: 5-25 µg/dL)
- Prolactin: 25 ng/mL (normal: 4-23 ng/mL)

Additional imaging with MRI of the brain reveals a 1 cm lesion in the pituitary region suggestive of a pituitary adenoma.

A 29-year-old female presents to the clinic with fatigue, weight gain, cold intolerance, constipation, mild depression, irregular periods, and difficulty concentrating. She has a history of a head injury from a motor vehicle accident two years ago. Physical examination reveals dry, pale skin, delayed reflexes, mild facial puffiness, and periorbital edema. Laboratory results show low TSH, free T4, and free T3 levels, with a slightly elevated prolactin level. An MRI of the brain reveals a 1 cm lesion in the pituitary region suggestive of a pituitary adenoma. What is the most appropriate next step in management?

- ◯ (A) Schedule the patient for surgical removal of the pituitary adenoma
- ◯ (B) Order a fine-needle aspiration biopsy of the pituitary lesion
- ◯ (C) Initiate levothyroxine therapy
- ◯ (D) Refer to an endocrinologist for further evaluation and management
- ◯ (E) Start dopamine agonist therapy

Submit

USMLE-style multiple-choice question answering:

👉 Medical Students

👉 Residents

Need harder questions:

Fellows

Doctors (Clinicians/Physicians)

Credit: https://www.openevidence.com/

mild depression, irregular periods, and difficulty concentrating. She has a history of a head injury from a motor vehicle accident two years ago. Physical examination reveals dry, pale skin, delayed reflexes, mild facial puffiness, and periorbital edema. Laboratory results show low TSH, free T4, and free T3 levels, with a slightly elevated prolactin level. An MRI of the brain reveals a 1 cm lesion in the pituitary region suggestive of a pituitary adenoma. What is the most appropriate next step in management?

- (A) Schedule the patient for surgical removal of the pituitary adenoma
- (B) Order a fine-needle aspiration biopsy of the pituitary lesion
- (C) Initiate levothyroxine therapy
- (D) Refer to an endocrinologist for further evaluation and management
- (E) Start dopamine agonist therapy

Submit

Incorrect. The correct answer is: Refer to an endocrinologist for further evaluation and management

Create Anki Card | Explain | Next

Finished thinking

Referral to an endocrinologist for further evaluation and management is the most appropriate next step because pituitary adenomas presenting with central hypothyroidism and mild hyperprolactinemia require specialized assessment to confirm the diagnosis, evaluate for additional pituitary hormone deficiencies, and determine the functional status of the tumor. Central hypothyroidism, particularly in the context of a pituitary mass, is frequently accompanied by other pituitary hormone deficits, and its management—including the initiation and titration of levothyroxine—demands expertise to avoid precipitating adrenal insufficiency or other complications. [1-5]

Endocrinologists are uniquely qualified to distinguish between functioning and nonfunctioning adenomas, interpret laboratory artifacts such as the hook effect, and initiate appropriate first-line therapies (e.g., dopamine agonists for prolactinomas, surgical referral for nonfunctioning or other functioning adenomas).[2][6-8] The Endocrine Society and other expert guidelines recommend comprehensive endocrine evaluation in all patients with pituitary adenomas, including assessment for hormone hypersecretion and hypopituitarism, as well as formal visual field testing if there is concern for optic chiasm involvement.[1-2][4] Early specialist involvement is critical to prevent morbidity from untreated hormone deficiencies, guide further imaging or surgical referral if indicated, and optimize long-term outcomes.[1-2][4]

In summary, the complexity of diagnosis and management, the need for tailored hormone replacement, and the potential for significant morbidity underscore the necessity of endocrinology referral in this clinical scenario.

References

Large | Ask a follow-up question...

Credit: https://www.openevidence.

complications. [1-8]

Endocrinologists are uniquely qualified to distinguish between functioning and nonfunctioning adenomas, interpret laboratory artifacts such as the hook effect, and initiate appropriate first-line therapies (e.g., dopamine agonists for prolactinomas, surgical referral for nonfunctioning or other functioning adenomas).[2][6-8] The Endocrine Society and other expert guidelines recommend comprehensive endocrine evaluation in all patients with pituitary adenomas, including assessment for hormone hypersecretion and hypopituitarism, as well as formal visual field testing if there is concern for optic chiasm involvement.[1-2][4] Early specialist involvement is critical to prevent morbidity from untreated hormone deficiencies, guide further imaging or surgical referral if indicated, and optimize long-term outcomes.[1-2][4]

In summary, the complexity of diagnosis and management, the need for tailored hormone replacement, and the potential for significant morbidity underscore the necessity of endocrinology referral in this clinical scenario.

References

1. Pituitary Incidentaloma: An Endocrine Society Clinical Practice Guideline.
   Freda PU, Beckers AM, Katznelson L, et al.
   The Journal of Clinical Endocrinology and Metabolism. 2011;96(4):894-904. doi:10.1210/jc.2010-1048.
2. Diagnosis and Management of Pituitary Adenomas: A Review.
   Tritos NA, Miller KK.
   Jama. 2023;329(16):1386-1398. doi:10.1001/jama.2023.5444.
   ☆ Leading Journal
3. Central Hypothyroidism: Advances in Etiology, Diagnostic Challenges, Therapeutic Targets, and Associated Risks.
   Iglesias P.
   Endocrine Practice : Official Journal of the American College of Endocrinology and the American Association of Clinical Endocrinologists. 2025;31(5):650-659. doi:10.1016/j.eprac.2025.02.004.
   New Research
4. Hormonal Replacement in Hypopituitarism in Adults: An Endocrine Society Clinical Practice Guideline.
   Fleseriu M, Hashim IA, Karavitaki N, et al.
   The Journal of Clinical Endocrinology and Metabolism. 2016;101(11):3888-3921. doi:10.1210/jc.2016-2118.
5. Interactions Between Hypothalamic Pituitary Thyroid Axis and Other Pituitary Dysfunctions.
   Feldt-Rasmussen U, Klose M, Benvenga S.
   Endocrine. 2018;62(3):519-527. doi:10.1007/s12020-018-1738-6.
6. Approach to the Patient With Prolactinoma.
   Auriemma RS, Pirchio R, Pivonello C, et al.
   The Journal of Clinical Endocrinology and Metabolism. 2023;108(9):2400-2423. doi:10.1210/clinem/dgad174.
7. Diagnosis and Treatment of Pituitary Adenomas: A Review.
   Molitch ME.
   Jama. 2017;317(5):516-524. doi:10.1001/jama.2016.19699.
   ☆ Leading Journal
8. Pituitary-Tumor Endocrinopathies.
   Melmed S.
   The New England Journal of Medicine. 2020;382(10):937-950. doi:10.1056/NEJMra1810772.
   ☆ Leading Journal

Credit: https://www.openevidence.com/

Part 1 – Background

- **Patient background**

A 27-year-old man presents to the emergency department. He was brought in by staff from the homeless shelter when they found him unresponsive. The patient is a known IV drug abuser but otherwise has an unknown past medical history. He currently attends a methadone clinic. His temperature is 99.5°F (37.5°C), blood pressure is 97/48 mmHg, pulse is 140/min, respirations are 29/min, and oxygen saturation is 98% on room air. Initial laboratory values are shown below. Serum: Na+: 139 mEq/L. Cl-: 100 mEq/L. K+: 6.3 mEq/L. HCO3-: 17 mEq/L. Glucose: 589 mg/dL. The patient is given treatment. After treatment, his temperature is 99.5°F (37.5°C), blood pressure is 117/78 mmHg, pulse is 100/min, respirations are 23/min, and oxygen saturation is 98% on room air. His laboratory values are seen below. Serum: Na+: 139 mEq/L. Cl-: 100 mEq/L. K+: 4.3 mEq/L. HCO3-: 19 mEq/L. Glucose: 90 mg/dL.

Credit: Jin et al., What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams, In AAAI'21.

Part 1 – Background

- **Question and options**
  - **Question**

    What is the best next step in management?
  - **Options**

    A. Insulin, potassium, IV fluids, and glucose

    B. IV fluids only

    C. Oral rehydration

    D. Supportive therapy and close monitoring
  - **Answer**

    A. Insulin, potassium, IV fluids, and glucose

(1) Multiple choice question-answering

(2) Open-ended question-answering

Credit: Jin et al., What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams, In AAAI'21.

Part 2 – Related Works

- **Question-answering without rationale → lack explainability**

Dataset	Source	Description	n
		Finetuning	
Medical Flash Cards	Anki Flashcards	Rephrased Q&A pairs derived from the front and back sides of medical flashcards	33,955
Stack Exchange	Academia	Q&A pairs generated from questions and their top-rated answers	39,633
	Biology		7,482
	Fitness		3,026
	Health		1,428
	Bioinformatics		906
Wikidoc	Living Textbook	Q&A pairs generated from paragraphs, where questions were formulated from rephrased paragraph titles, and answers were extracted from paragraph text	67,704
	Patient Information	Q&A pairs generated from paragraph headings and associated text content	5,942
		Evaluation	
USMLE	Step 1	Multiple choice questions from the USMLE self-assessment with image-based questions excluded	119
	Step 2		120
	Step 3		135

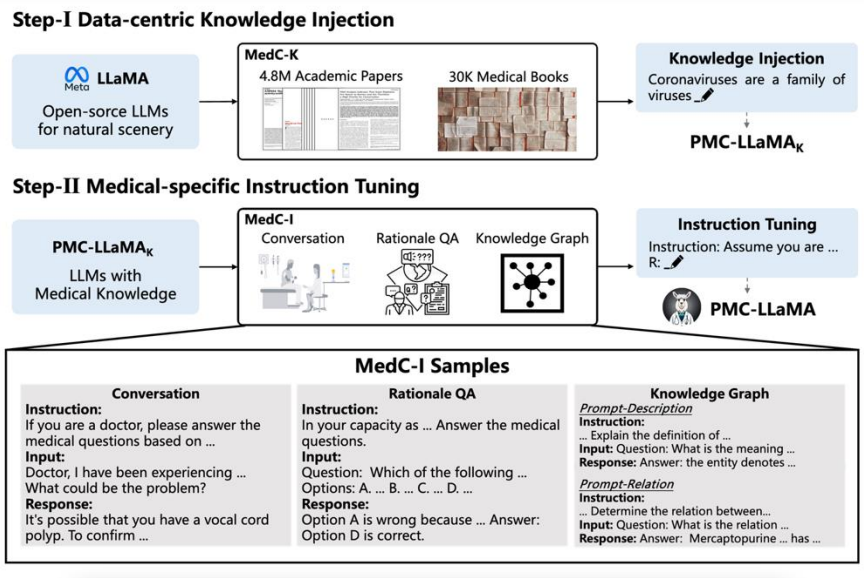
Instruction tuning

(Supervised fine-tuning)

Credit: Han et al., MedAlpaca: An Open-Source Collection of Medical Conversational AI Models and Training Data, In arXiv'23.

Part 2 – Related Works

- **Question-answering with rationale → lack factuality**



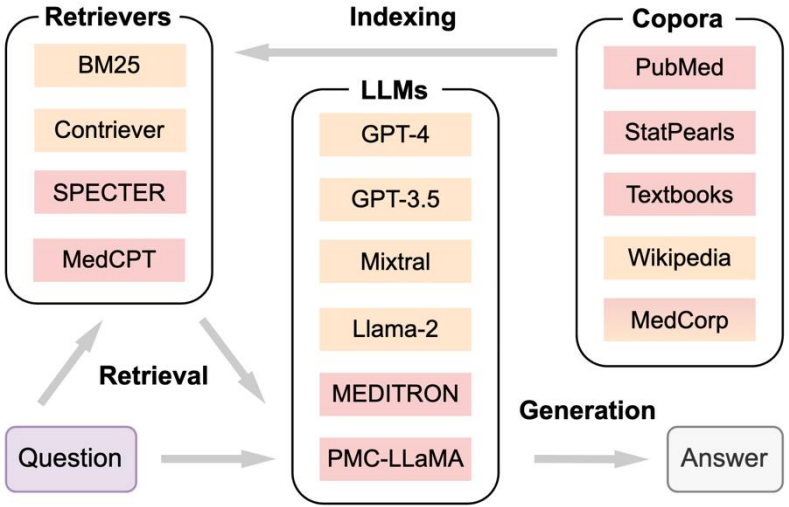
###Rationale: This patient's clinical presentation and laboratory abnormalities suggest a diagnosis of thrombotic thrombocytopenic purpura (TTP). TTP is a microangiopathic hemolytic anemia, resulting from severe deficiency or functional abnormality of the cleaving protease ADAMTS13, leading to accumulation of ultra-large multimers of von Willebrand factor, platelet adhesion, and widespread thrombi in microcirculation. Clinically, TTP is characterized by the pentad of thrombocytopenia, microangiopathic hemolytic anemia, neurologic abnormalities, renal insufficiency, and fever. This pentad is not always complete, and high clinical suspicion should be maintained even if not all five elements are present, as in this case. In terms of management, plasma exchange therapy is the mainstay and should be initiated as soon as the diagnosis is suspected, even before laboratory confirmation, given the severity and rapid progression of TTP. This therapy helps by removing the autoantibodies that inhibit ADAMTS13 and replenishing the deficient enzyme through plasma replacement. Options A and B, high-dose glucocorticoids and cyclophosphamide and rituximab, respectively, can be used in addition to plasma exchange in certain refractory or relapsed cases but are not first-line treatment. Option C, Vancomycin and cefepime, are antibiotics used to treat infections, which doesn't align with this patient's presentation. ###Answer: OPTION D IS CORRECT.

Credit:

[1] Wu et al., PMC-LLaMA: Towards Building Open-source Language Models for Medicine, In Journal of the American Medical Informatics Association'23.
[2] Qiu et al., Towards Building Multilingual Language Model for Medicine, In Nature Communications'24.

Part 2 – Related Works

- **Question-answering with retrieval-augmented generation (RAG)**
  - **Bias problem** ← no further evidence assessment
  - **Low efficiency** ← model doesn't decide whether to use evidence



Credit: Xiong et al., Benchmarking Retrieval-Augmented Generation for Medicine, In ACL'24.

Part 2 – Motivation

- **Question-answering without rationale**
  - Clinical decision-making with rationale → **explainability**
- **Question-answering with rationale**
  - Clinical decision-making with external evidence → **factuality**
- **Question-answering with RAG**
  - **Clinical decision-making with evidence assessment** ← bias problem
  - **Clinical decision-making with an automatic and dynamic process** ← low efficiency

Credit:
[1] Han et al., MedAlpaca: An Open-Source Collection of Medical Conversational AI Models and Training Data, In arXiv'23.
[2] Wu et al., PMC-LLaMA: Towards Building Open-source Language Models for Medicine, In Journal of the American Medical Informatics Association'23.
[3] Qiu et al., Towards Building Multilingual Language Model for Medicine, In Nature Communications'24.
[4] Xiong et al., Benchmarking Retrieval-Augmented Generation for Medicine, In ACL'24.

Part 3 – MedPodGPT

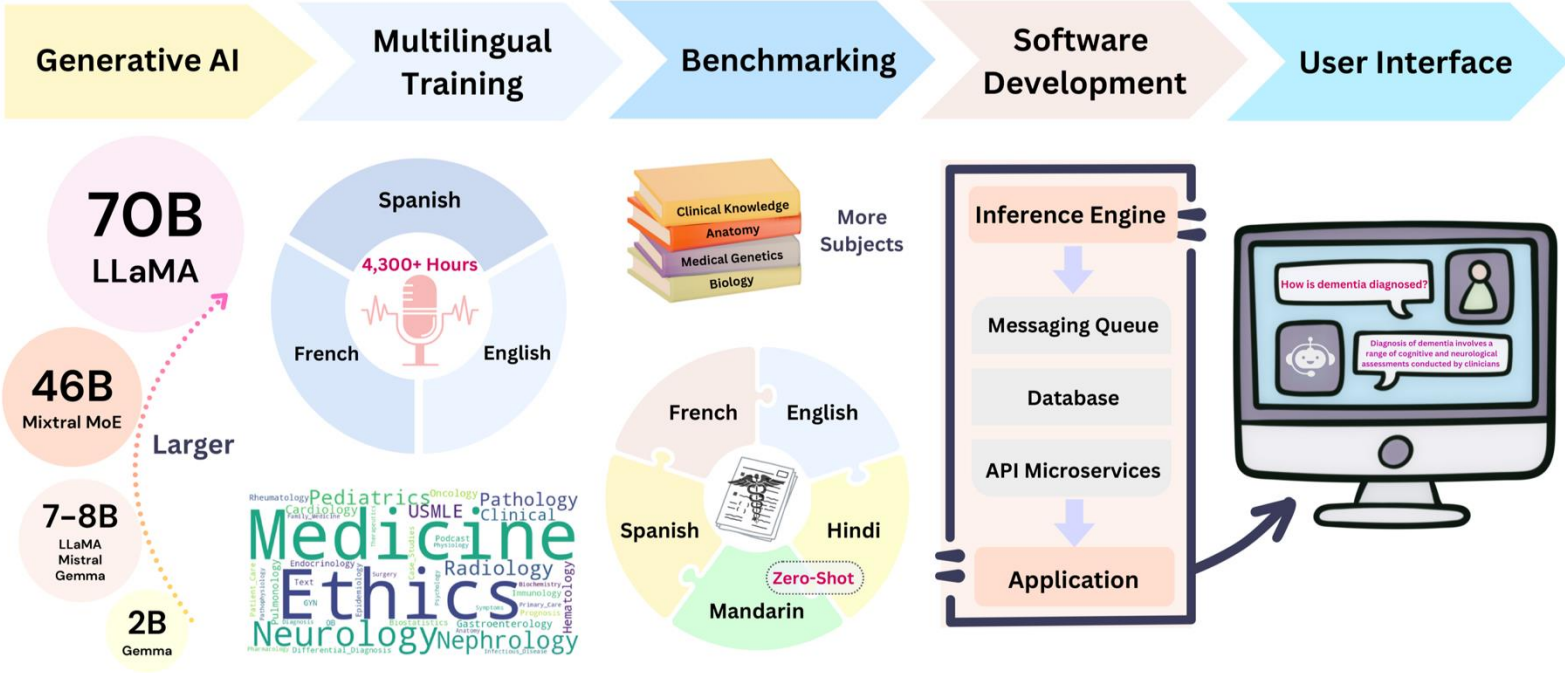
- What: A **multilingual audio-augmented LLM** for medical research and education

- Why: **Up to date and domain-relevant information** ← podcasts
- Where: Journals, exam preparation materials, and clinical practice
- When: Most recent (e.g., since 2023)
- How: Continual pretraining of large language models (LLMs)

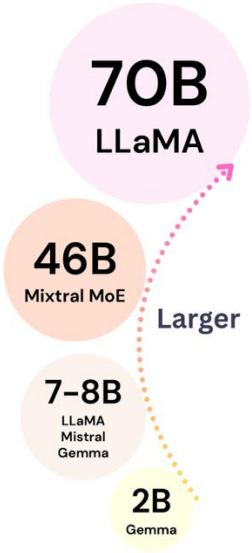
Credit: [1] DIP Ep 606: Drug Interactions For The USMLEs (Step 1-3) [2] NEJM This Week — May 29, 2025 Images are in the public domain.

Part 3 – MedPodGPT

- **A multilingual audio-augmented LLM for medical research and education**



Credit: Jia et al., MedPodGPT: A Multilingual Audio-augmented Large Language Model for Medical Research and Education, In medRxiv'24.



Part 3 – MedPodGPT

- **Large language models**
  - 2B Gemma
  - 7B Gemma, 7B Mistral, 8B LLaMA
  - 46B Mixtral 8×7B MoE
  - 70B LLaMA

$$\mathcal{L}_{\pi_\theta} = -\sum \log\left(\pi_\theta(x_i|\mathbf{x}_{<i})\right).$$

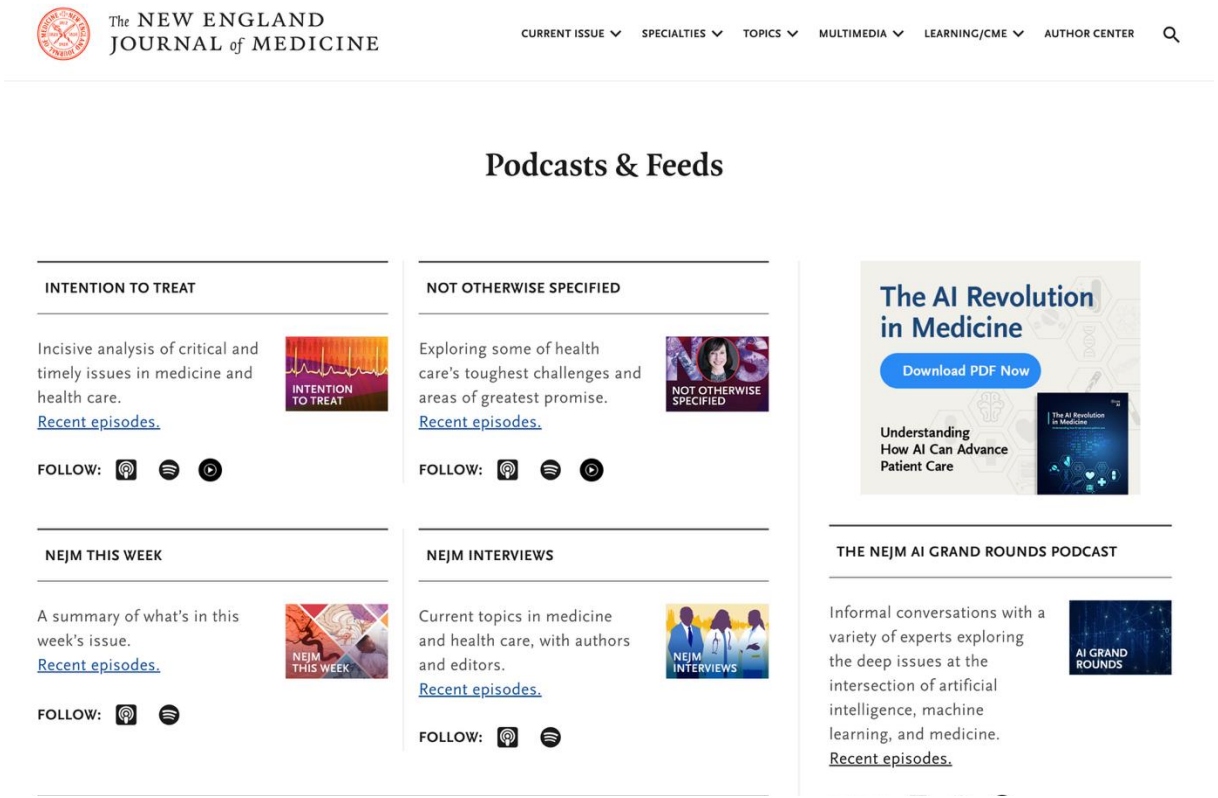
π_θ: LLM, parameterized by θ

$\mathbf{x} = [x_1, x_2, \ldots, x_t]$: a sequence of texts

Credit: Jia et al., MedPodGPT: A Multilingual Audio-augmented Large Language Model for Medical Research and Education, In medRxiv'24. Images are in the public domain.

Part 3 – MedPodGPT

- **Podcasts**



Credit: NEJM Podcasts & Feed: https://www.nejm.org/rss-feed/

Part 3 – MedPodGPT

- **Podcasts (journals)**
  - OpenAI Whisper automatic speech recognition (ASR) model (2022)
  - Audio → texts (transcripts)

Journal Podcasts						
Podcast	**Language**	**Episodes**	**Audio Time (min)**	**Mean Length Episode $\pm\sigma$ (min)**	**Number of Text Tokens**	**Mean Text Tokens per Episode $\pm\sigma$**
NEJM	English	1974	39256.0	19.89 ± 9.74	4,760,783	1928.22 ± 14.87
JAMA	English	2235	32163.0	14.39 ± 8.66	3,454,191	1928.64 ± 15.54
The Lancet	English	2029	28279.0	13.94 ± 7.62	3,300,982	1925.89 ± 20.88
The BMJ	English	300	13264.2	44.21 ± 10.35	2,235,458	1897.67 ± 75.07
Annals Latest Highlights	English	396	6427.0	16.23 ± 8.09	803,958	1927.96 ± 14.60
Annals On Call	English	142	3440.0	24.22 ± 3.65	522,547	1928.22 ± 15.64
Pediatrics on Call	English	98	3299.0	33.66 ± 6.09	565,781	1930.99 ± 15.00
Procedure Ready: Ob/Gyn	English	20	383.7	19.19 ± 5.00	63,667	1929.30 ± 13.41
Revista Médica AFP Podcast	Spanish	40	1055.0	26.38 ± 3.70	190,518	1924.42 ± 16.34

Credit:

[1] https://openai.com/index/whisper/

[2] Jia et al., MedPodGPT: A Multilingual Audio-augmented Large Language Model for Medical Research and Education, In medRxiv'24.

Part 3 – MedPodGPT

- **Podcasts (test preparations)**

Test Preparation Podcasts						
Podcast	**Language**	**Episodes**	**Audio Time (min)**	**Mean Length of Episode $\pm\sigma$ (min)**	**Number of Text Tokens**	**Mean Text Tokens per Episode $\pm\sigma$**
Divine Intervention Podcasts	English	480	18363.8	38.26 ± 24.07	2,269,153	1931.19 ± 13.53
The Radiology Review Podcast	English	127	2517.7	19.82 ± 10.10	292,949	1927.29 ± 26.81
Crush Step 1: The Ultimate USMLE Step 1 Review	English	49	2176.2	44.41 ± 15.31	328,194	1930.55 ± 13.03
The USMLE Guys Podcast	English	31	1464.3	47.24 ± 43.47	156,923	1937.32 ± 6.12
Harrison's PodClass: Internal Medicine Cases and Board Prep	Spanish	95	905.2	9.53 ± 2.24	101,574	1916.49 ± 22.77
El Interno Desvelado	Spanish	4	99.13	24.78 ± 11.91	17,121	1902.33 ± 25.54
Curso MIR Asturias	Spanish	3	17.7	5.89 ± 4.53	3,872	1936.00 ± 9.00

Credit: Jia et al., MedPodGPT: A Multilingual Audio-augmented Large Language Model for Medical Research and Education, In medRxiv'24.

Part 3 – MedPodGPT

- **Podcasts (clinical experts)**

Clinical Podcasts						
Podcast	**Language**	**Episodes**	**Audio Time (min)**	**Mean Length of Episode $\pm\sigma$ (min)**	**Number of Text Tokens**	**Mean Text Tokens per Episode $\pm\sigma$**
The Curbsiders Internal Medicine Podcast	English	485	28749.2	59.39 ± 16.08	5,772,083	1929.82 ± 17.06
This Podcast Will Kill You	English	168	18363.8	38.26 ± 24.07	2,269,153	1931.19 ± 13.53
The Clinical Problem Solvers	English	315	13500.1	42.86 ± 14.51	2,493,777	1927.18 ± 23.32
PsychEd: educational psychiatry podcast	English	62	3556.3	57.36 ± 17.52	607,237	1927.74 ± 16.36
Run the List	English	97	1973.0	20.34 ± 6.44	352,977	1928.84 ± 15.17
Goljan Pathology Lectures	English	37	1886.0	50.97 ± 4.58	412,086	1934.68 ± 13.45
Core IM: 5 Pearls	English	54	1847.1	34.21 ± 10.19	361,213	1931.62 ± 10.16
Neurology Clinical Pearls	English	27	333.2	12.34 ± 3.19	42,494	1931.54 ± 10.78
Tutorías Medicina Interna	Spanish	570	19834.9	34.80 ± 25.01	4,311,263	1898.39 ± 64.31
Leucocitos isotópicos	Spanish	68	2537.8	37.32 ± 9.55	481,676	1797.29 ± 154.42
Medicina Con Cabeza	Spanish	246	2457.8	9.99 ± 3.44	462,383	1902.81 ± 57.55
Medicina de impacto	Spanish	57	2406.5	42.22 ± 9.13	492,363	1915.81 ± 29.28
Ronda, El Podcast de Medicina Interna	Spanish	20	1084.4	54.22 ± 25.01	206,218	1891.91 ± 71.90
Medicina De Bolsillo | Hablando de Medicina	Spanish	45	958.3	21.30 ± 10.79	186,268	1844.24 ± 124.82
La Tertulia de Cajal	Spanish	27	876.3	32.46 ± 18.28	186,001	1897.97 ± 57.71
PedCast: Dos Pediatras y un Podcast	Spanish	14	458.5	32.75 ± 10.62	89,127	1896.32 ± 58.05
Neurobiologie et Immunite	French	21	1882.8	89.66 ± 14.77	383,189	1896.97 ± 40.12
Incubateur Néonat	French	25	1579.3	63.17 ± 21.28	391,475	1918.99 ± 24.15
Guideline.care	French	68	1369.1	20.13 ± 6.30	293,301	1917.0 ± 29.29
La Minute Rhumato	French	119	921.0	7.74 ± 2.19	132,354	1918.17 ± 23.59
Oncologie cellulaire et moléculaire - Hugues de Thé	French	11	852.9	77.53 ± 19.81	186,693	1905.03 ± 44.42
Le podcast des Conférenciers (UFR3S) by Université de Lille	French	65	768.4	11.82 ± 19.58	86,105	1913.44 ± 42.78
Super Docteur	French	47	676.3	14.39 ± 6.50	139,824	1915.40 ± 33.26
Médecine, Sciences et Recherche clinique	French	24	332.2	13.84 ± 4.58	63,314	1918.61 ± 26.60
NéphrOdio	French	40	318.6	7.96 ± 2.58	55,716	1921.24 ± 19.59
La Minute Néonat	French	37	307.6	8.31 ± 1.93	57,435	1914.50 ± 31.58
Le Med G Eclairé	French	11	249.2	22.66 ± 16.76	51,988	1925.48 ± 12.57
La Minute du Pancréas	French	22	209.4	9.52 ± 2.34	38,376	1918.80 ± 23.34
L'essentiel des principales pathologiesaå	French	14	151.3	10.81 ± 13.10	23,098	1924.83 ± 11.25
AR-Pod le Podcast de lanesthésie-réanimation	French	12	139.0	11.59 ± 4.52	22,998	1916.50 ± 26.48

Credit: Jia et al., MedPodGPT: A Multilingual Audio-augmented Large Language Model for Medical Research and Education, In medRxiv'24.

Part 3 – MedPodGPT

- **Benchmarks (in-domain performance)**

Language	Benchmark Datasets	Model											
		Gemma 2B		Gemma 7B		Mistral 7B		LLaMA 3 8B		Mixtral MoE		LLaMA 3 70B	
		Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*
English	MedExpQA	15.20	**23.40**	34.40	**45.20**	**47.20**	46.20	57.60	**61.40**	52.80	**61.20**	**78.40**	77.60
	MedMCQA	34.81	**35.24**	40.66	**44.86**	42.65	**45.50**	58.64	**58.82**	50.20	**53.54**	**71.12**	70.58
	MedQA	29.69	**33.38**	38.26	**44.56**	46.27	**47.80**	**61.12**	59.21	**54.05**	53.22	**77.85**	77.51
	PubMedQA	47.80	**55.50**	**63.40**	55.30	**51.60**	41.75	**59.40**	49.20	**42.80**	32.20	73.00	**75.75**
	Anatomy	**43.70**	42.04	49.63	**52.96**	56.30	**56.67**	68.89	**69.82**	64.44	**68.15**	77.04	**77.78**
	Clinical Knowledge	**41.51**	38.78	55.47	**62.17**	61.89	**62.26**	72.08	**73.68**	67.92	**74.90**	82.26	**83.40**
	College Biology	44.44	**47.05**	61.11	**68.06**	61.81	**64.93**	74.31	**77.43**	72.92	**77.95**	91.67	**92.36**
	College Medicine	36.99	**37.14**	50.29	**55.06**	57.80	**59.97**	67.05	**68.06**	63.58	**69.07**	**78.61**	78.18
	Medical Genetics	43.00	**44.75**	54.00	**66.00**	64.00	**65.50**	**80.00**	77.25	70.00	**78.00**	91.00	**91.00**
	Professional Medicine	29.78	**34.10**	50.37	**60.02**	56.99	**63.33**	**76.84**	75.64	72.06	**73.07**	**90.44**	90.26
	Average	36.69	**39.14**	49.76	**55.42**	54.65	**55.39**	**67.59**	67.05	61.08	**64.13**	81.14	**81.22**
French	FrenchMedMCQA	**29.91**	28.43	29.60	**40.27**	**45.48**	44.32	41.74	**44.63**	55.14	**58.02**	63.24	**73.05**
	MedExpQA	19.20	**20.60**	26.40	**39.20**	40.80	**41.20**	**48.00**	43.60	50.40	**56.00**	**76.80**	74.00
	Anatomy	**35.56**	35.18	48.15	**49.63**	33.33	**39.45**	45.19	**47.41**	55.56	**59.63**	67.41	**68.52**
	Clinical Knowledge	32.45	**36.51**	50.94	**57.92**	**55.47**	53.02	**61.89**	61.13	65.66	**71.51**	78.87	**80.56**
	College Biology	33.33	**38.02**	46.53	**52.78**	**53.47**	49.65	57.64	**62.50**	67.36	**72.92**	86.81	**87.67**
	College Medicine	32.95	**35.84**	43.93	**47.98**	**51.45**	48.56	57.80	**59.40**	57.80	**63.44**	69.94	**74.71**
	Medical Genetics	35.00	**40.00**	50.00	**57.25**	47.00	**59.00**	66.00	**67.00**	71.00	**72.00**	**90.00**	89.50
	Professional Medicine	24.26	**28.95**	33.09	**42.00**	43.38	**43.84**	51.47	**55.51**	59.56	**64.15**	72.79	**73.34**
	Average	30.33	**32.94**	41.08	**48.38**	46.30	**47.38**	53.72	**55.15**	60.31	**64.71**	75.73	**77.67**
Spanish	HeadQA	33.77	**34.32**	48.21	**54.47**	53.79	**55.54**	59.66	**61.24**	64.77	**68.00**	81.44	**82.44**
	MedExpQA	21.60	**23.00**	32.80	**38.40**	**46.40**	40.40	40.00	**43.00**	**52.80**	52.40	73.60	**76.60**
	Anatomy	37.78	**39.08**	42.22	**51.11**	45.93	**49.63**	48.15	**52.96**	60.74	**62.22**	71.11	**74.44**
	Clinical Knowledge	37.74	**38.78**	53.96	**55.47**	54.34	**56.13**	58.49	**62.08**	**68.68**	68.40	78.49	**80.00**
	College Biology	29.17	**35.94**	48.61	**50.35**	55.56	**56.25**	54.86	**55.04**	66.67	**69.10**	**85.42**	84.20
	College Medicine	32.37	**34.39**	43.93	**48.84**	**54.34**	48.99	49.71	**54.05**	**59.54**	58.24	69.94	**72.97**
	Medical Genetics	32.00	**34.75**	46.00	**59.50**	53.00	**57.25**	**72.00**	68.00	**67.00**	66.75	86.00	**86.75**
	Professional Medicine	26.47	**30.06**	38.24	**43.56**	**47.06**	45.68	**51.84**	50.74	53.68	**56.90**	**69.49**	68.94
	Average	31.36	**33.79**	44.25	**50.21**	**51.30**	51.23	54.34	**55.89**	61.74	**62.75**	76.94	**78.29**

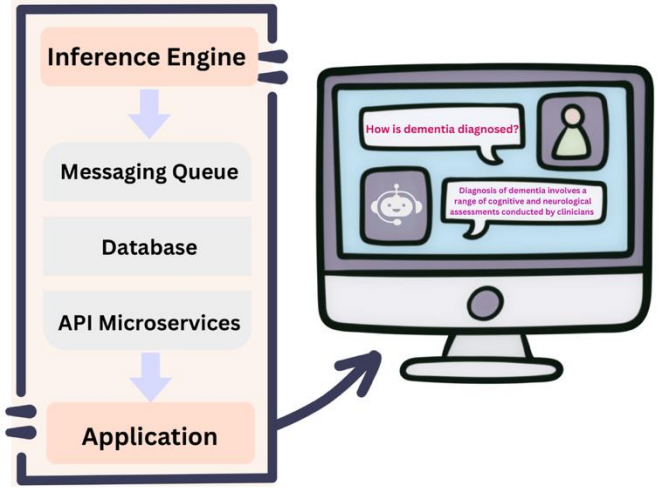
Part 3 – MedPodGPT

- **Benchmarks (zero-shot cross-lingual transfer performance)**

Language	Benchmark Datasets	Model											
		Gemma 2B		Gemma 7B		Mistral 7B		LLaMA 3 8B		Mixtral MoE		LLaMA 3 70B	
		Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*
Chinese	MedQA-MCMLE	33.39	**33.43**	40.51	**45.98**	39.67	**39.25**	63.63	**66.32**	45.80	**47.14**	**84.68**	83.73
	Anatomy	**28.38**	23.98	25.00	**31.59**	25.00	**30.41**	33.78	**35.98**	**33.11**	26.35	63.51	**64.02**
	Clinical Knowledge	**29.11**	28.27	31.22	**39.87**	33.33	**32.60**	49.37	**50.95**	**39.24**	38.61	71.73	**71.94**
	College Medicine	28.94	**32.23**	33.70	**36.08**	**30.77**	30.68	52.01	**56.50**	38.46	**40.94**	75.82	**80.49**
	Medical Genetics	32.39	**32.39**	43.75	**45.17**	38.64	**42.33**	43.18	**44.60**	45.45	**45.88**	**61.36**	57.53
	Medical Nutrition	33.79	**35.69**	40.69	**44.66**	**42.07**	37.24	**53.10**	50.00	49.66	**51.90**	66.21	**68.28**
	Traditional Chinese Medicine	27.57	**28.52**	31.35	**36.35**	24.86	**28.52**	**43.24**	39.46	30.27	**30.94**	66.49	**67.98**
	Virology	**37.28**	36.98	46.15	**54.44**	43.79	**48.22**	**59.76**	58.88	**53.25**	50.15	76.33	**77.51**
	Average	31.36	**31.44**	36.55	**41.77**	34.77	**36.16**	49.76	**50.34**	**41.91**	41.49	70.77	**71.44**
Hindi	Anatomy	25.93	**32.22**	34.07	**36.86**	23.70	**30.00**	**40.00**	35.18	31.11	**34.44**	52.59	**57.78**
	Clinical Knowledge	26.42	**28.96**	**41.89**	41.04	24.91	**35.85**	**48.30**	46.70	**38.11**	36.70	63.40	**69.06**
	College Biology	26.39	**33.16**	26.39	**34.03**	19.44	**28.47**	32.65	**37.16**	**30.56**	32.81	58.33	**68.06**
	College Medicine	24.86	**27.60**	42.20	**43.35**	23.12	**33.09**	41.04	**43.64**	27.17	**33.24**	60.69	**64.74**
	Medical Genetics	**31.00**	30.50	36.00	**41.75**	28.00	**29.25**	**46.00**	45.75	40.00	**43.25**	71.00	**77.00**
	Professional Medicine	25.37	**26.19**	30.88	**41.08**	22.06	**28.67**	36.40	**39.34**	29.41	**29.50**	45.59	**64.70**
	Average	26.66	**29.77**	35.24	**39.69**	23.54	**30.89**	40.73	**41.29**	32.73	**34.99**	58.60	**66.89**

Part 3 – MedPodGPT

- **Software development**
  - vLLM: inference engine
  - Apache Cassandra: query optimization
  - Flask: store chats and conversations in Cassandra and send text inference requests to a queue
  - RabbitMQ: queue requests
  - PostgreSQL: RAG database management
  - OAuth 2.0: Authorization and user management

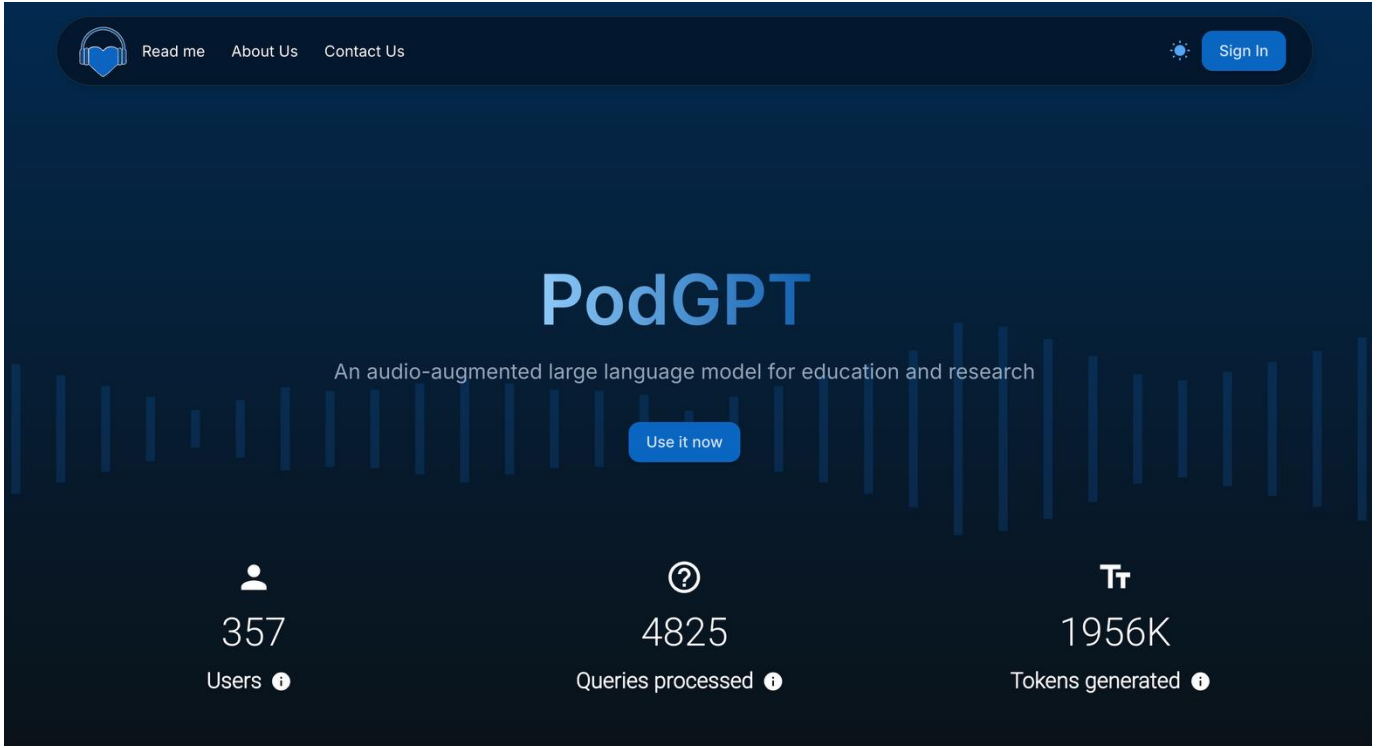


Credits:
[1] PodGPT: https://podgpt.org/
[2] vLLM: https://github.com/vllm-project/vllm
[3] Apache Cassandra: https://github.com/apache/cassandra
[4] Flask: https://github.com/pallets/flask
[5] RabbitMQ: https://github.com/rabbitmq/rabbitmq-server
[6] PostgreSQL: https://github.com/postgres/postgres
[7] OAuth 2.0: https://github.com/postgres/postgres

Part 3 – MedPodGPT

- **Software development**

https://podgpt.org



1. **Customized teaching cases**: "I'm teaching second-year medical students. Generate a case on a patient with a chronic obstructive pulmonary disease (COPD) exacerbation with diagnostic uncertainty about heart failure exacerbation."
2. **Expert discussion**: "I have a medical case in 4 parts. As our diagnostic expert, provide the problem representation, a prioritized differential, and your choice of next test for each part."
3. **Summarization and feedback**:

- "Would my patient be included in this study?" or "Summarize methodological points of the study."
- "I need a 2-paragraph summary of these course evaluations. Include 2 direct quotes highlighting learner qualities."
- "I've written these patient instructions aiming for an 8th grade reading level. Determine their current reading level and offer feedback for improvement."

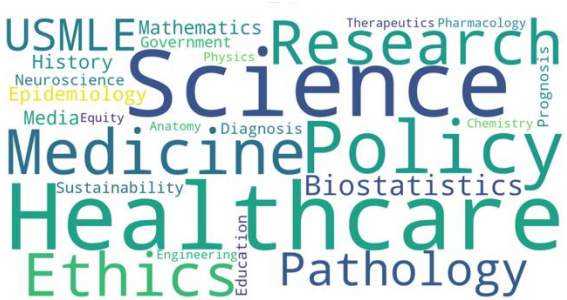
Credit: Rodman et al., Using Generative Artificial Intelligence in Medical Education, In Academic Medicine'25.

Part 3 – MedPodGPT

- **Take away**
  - **Podcasts** are valuable sources of **up to date** and **domain-relevant** information
  - **Continual pretraining of LLMs on podcasts** can enhance overall **model performance**
  - Improve **in-domain (language)** performance
  - Improve **zero-shot multilingual transfer** capabilities

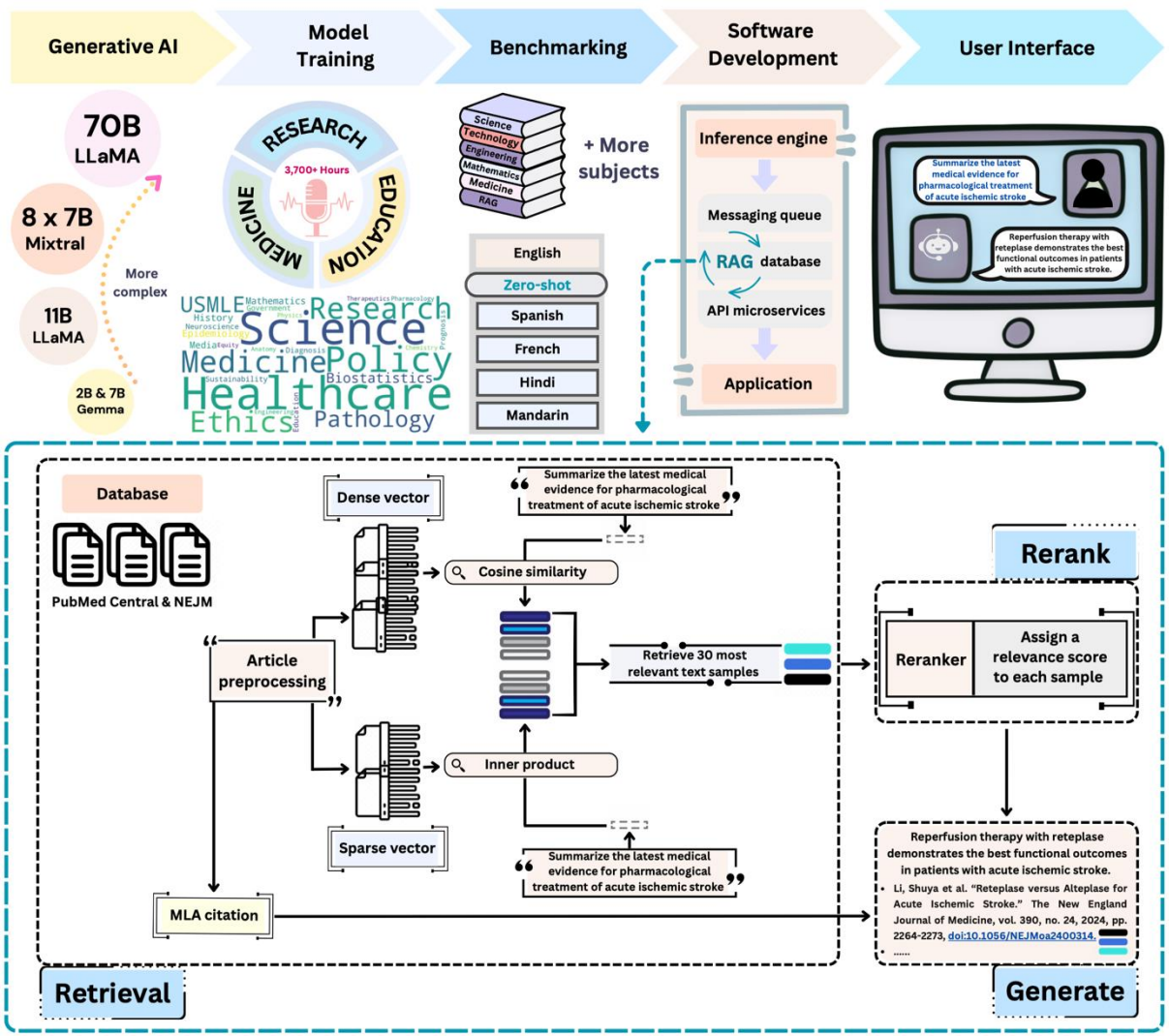
Part 4 – PodGPT

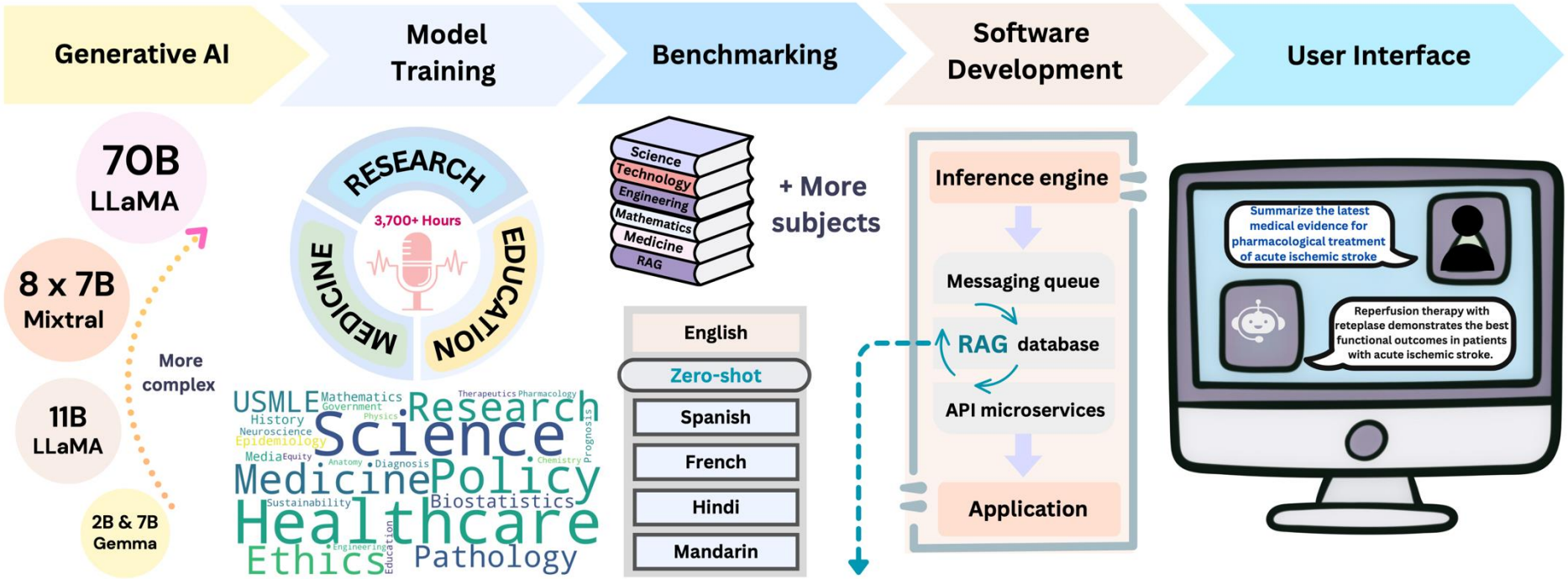
- What: An **audio-augmented LLM** for research and education
  - Fields: Science, technology, engineering, mathematics, and medicine (STEMM)



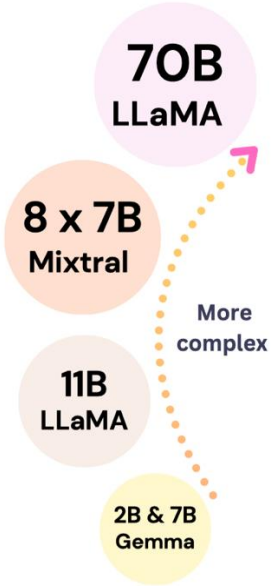
- Why: (1) **Up to date information** ← podcasts (2) **Factual information** ← journal articles
- Where: Creative Commons Attribution (CC BY) podcasts from journals/experts, and NEJM
- When: Most recent (e.g., since 2023)
- How: Continual pretraining of LLMs

Credit: Jia et al., PodGPT: An Audio-augmented Large Language Model for Research and Education, In Nature npj Biomedical Innovations'25.


Generative AI
Model Training
Benchmarking
Software Development
User Interface
70B LLaMA
8 x 7B Mixtral
11B LLaMA
2B & 7B Gemma
More complex
RESEARCH
EDUCATION
MEDICINE
3,700+ Hours
Science
Technology
Engineering
Mathematics
Medicine
RAG
+ More subjects
English
Zero-shot
Spanish
French
Hindi
Mandarin
Inference engine
Messaging queue
RAG database
API microservices
Application
Summarize the latest medical evidence for pharmacological treatment of acute ischemic stroke
Reperfusion therapy with reteplase demonstrates the best functional outcomes in patients with acute ischemic stroke.
Database
PubMed Central & NEJM
Dense vector
Sparse vector
Article preprocessing
Cosine similarity
Inner product
Retrieve 30 most relevant text samples
MLA citation
Rerank
Reranker
Assign a relevance score to each sample
Retrieval
Generate



Credit: Jia et al., PodGPT: An Audio-augmented Large Language Model for Research and Education, In Nature npj Biomedical Innovations'25.



Part 4 – PodGPT

- **Large language models**
  - 2B Gemma
  - 7B Gemma, 7B Mistral
  - Quantized 70B LLaMA (11B)
  - 46B Mixtral 8×7B MoE
  - 70B LLaMA

Llama 3
From Meta

Gemma

MISTRAL AI_

Credit: Jia et al., PodGPT: An Audio-augmented Large Language Model for Research and Education, In Nature npj Biomedical Innovations'25.

Part 4 – PodGPT

- **Podcasts**

Podcast	Episodes	Audio time (min)	Mean length episode $\pm\sigma$ (min)	Number of text tokens	Mean text tokens per episode $\pm\sigma$
NEJM This Week	457	13,300.17	29.10 ± 2.92	2,029,219	4440.30 ± 419.35
NEJM Interviews	654	9,223.44	14.10 ± 8.15	1,732,879	2649.66 ± 1522.35
NEJM Core IM | Internal Medicine Podcast	170	5,285.72	31.09 ± 10.08	1,077,154	6336.20 ± 2093.18
NEJM Curbside Consults	74	1,977.46	26.72 ± 11.04	408,189	5516.07 ± 2522.46
NEJM Clinical Conversations	108	1,829.66	16.94 ± 4.13	320,968	2971.93 ± 774.89
NEJM Leadership Conversations	100	1,765.76	17.66 ± 4.38	306,490	3064.90 ± 759.09
NEJM AI Grand Rounds	24	1,459.11	60.80 ± 14.84	303,499	12645.79 ± 3107.35
NEJM Intention to Treat	40	997.22	24.93 ± 5.27	169,632	4240.80 ± 954.41
NEJM Not Otherwise Specified	20	836.03	41.80 ± 15.71	146,718	7335.90 ± 2880.36
TWiV: This Week In Virology	1,186	104,089.57	87.77 ± 30.30	20,188,268	17022.15 ± 5785.86
TWiP: This Week In Parasitism	245	21,129.03	86.24 ± 15.35	4,381,511	17883.72 ± 3814.75
TWiM: This Week in Microbiology	320	20,641.50	64.50 ± 10.36	3,667,519	11461.00 ± 2121.78
TWiEVO: This Week In Evolution	100	8,845.09	88.45 ± 10.99	1,756,480	17564.80 ± 2517.80
IMMUNE	93	6,918.84	74.40 ± 19.67	1,363,176	14657.81 ± 3888.47
TWiN: This Week In Neuroscience	53	3,581.13	67.57 ± 10.40	644,712	12164.38 ± 2078.49
Matters Microbial	62	3,418.33	55.13 ± 11.33	637,647	10284.63 ± 2357.42
Infectious Disease Puscast	65	2,298.71	35.36 ± 5.96	415,145	6386.85 ± 1126.15
Urban Agriculture	29	2,171.63	74.88 ± 17.58	443,859	15305.48 ± 4214.82
On The Wards: On The Pods Medical Podcast for Doctors	245	5,915.14	24.14 ± 8.00	1,175,307	4797.17 ± 1781.74
Digital Campus Podcast	64	3,169.44	49.52 ± 5.85	605,977	9322.72 ± 1528.99
emDOCs.net Emergency Medicine (EM) Podcast	112	1,576.72	14.08 ± 4.30	303,672	2711.35 ± 864.29
Policy in Plainer English Podcast	73	1,089.20	14.92 ± 7.52	208,580	2857.26 ± 1509.80
Open Minds ... from Creative Commons	21	803.28	38.25 ± 12.79	145,624	6619.27 ± 2722.90
What is Global Health?	18	486.47	27.03 ± 10.02	87,653	4869.61 ± 2054.00
Consilience Sustainability In Progress (SIP) Podcast	9	403.95	44.88 ± 17.89	70,461	7829.00 ± 3535.75
Research Pulse: Future Focussed Health Insights	16	177.79	11.11 ± 2.29	33,672	2104.50 ± 476.85
Our People: Central to Healthcare	9	161.15	17.91 ± 7.42	31,545	3505.00 ± 1474.57

Credit: Jia et al., PodGPT: An Audio-augmented Large Language Model for Research and Education, In Nature npj Biomedical Innovations'25.

Part 4 – PodGPT

- **Benchmarks (in-domain performance)**

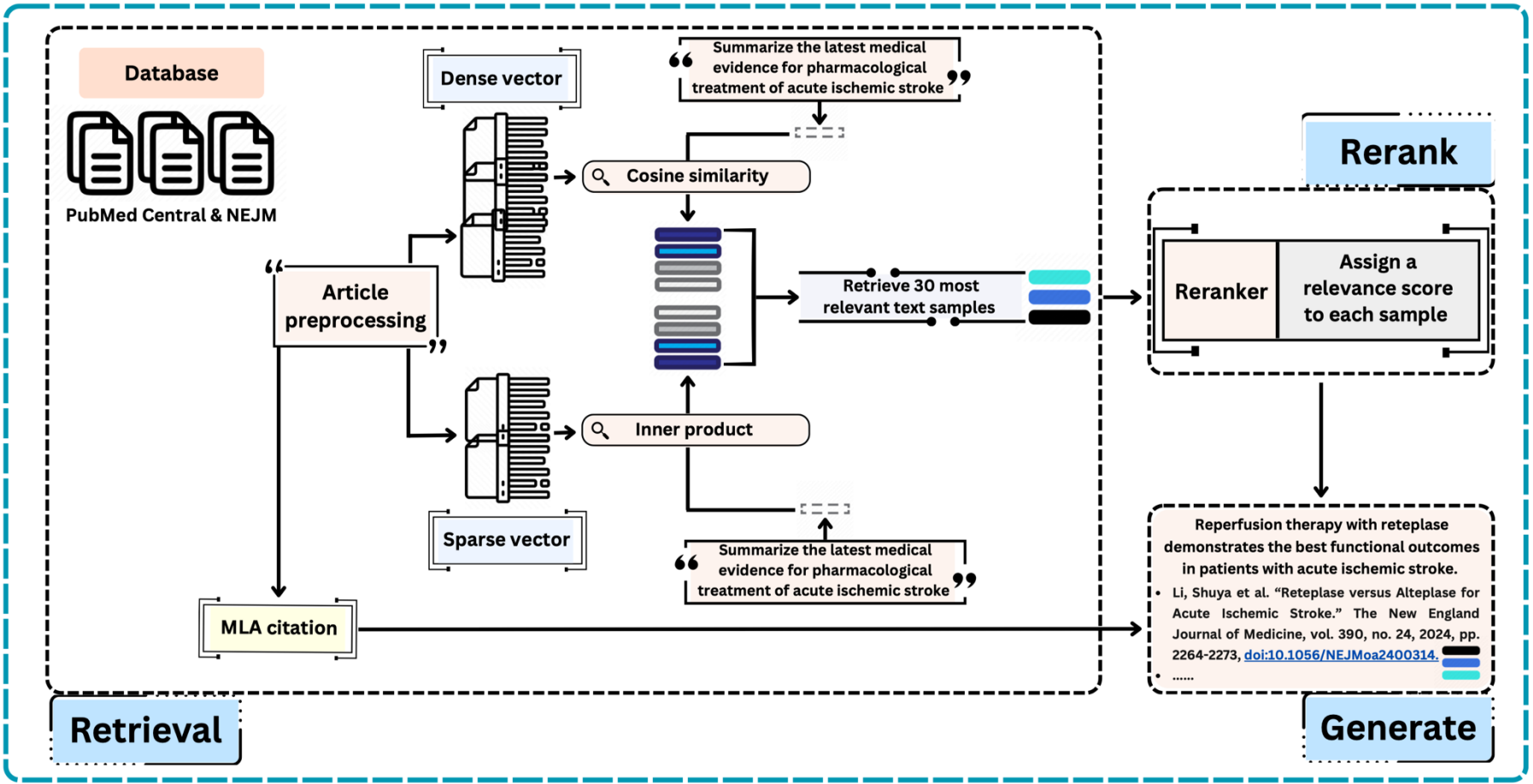
Model	Gemma 2B		Gemma 7B		Quantized LLaMA 70B		Mixtral 8×7B MoE		LLaMA 70B	
MMLU benchmark	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*
Physics	29.85	**30.74 (0.12)**	42.38	**47.64 (0.64)**	76.79	**79.24**	56.08	**56.19 (0.71)**	78.63	**79.26**
Biology	44.82	**46.58 (0.45)**	63.32	**66.43 (0.58)**	90.97	**91.51**	72.14	**77.84 (0.90)**	93.53	**93.53**
Chemistry	30.04	**30.96 (0.62)**	43.42	**44.14 (1.31)**	69.94	**73.15**	46.60	**50.33 (1.34)**	**70.16**	69.66
Computer Science	40.59	**44.20 (0.11)**	53.58	**54.62 (0.29)**	80.24	**82.30**	59.25	**60.98 (1.25)**	78.72	**79.44**
Engineering	39.31	**42.07 (0.49)**	44.14	**47.94 (0.60)**	73.79	**73.79**	**57.24**	56.38 (0.90)	75.17	**75.17**
Mathematics	25.69	**26.44 (0.57)**	34.97	**39.23 (0.16)**	68.83	**71.34**	**47.42**	46.59 (1.01)	63.97	**64.18**
Medicine	40.62	**41.72 (0.15)**	55.22	**59.50 (0.14)**	86.11	**87.38**	67.38	**74.00 (0.71)**	88.65	**88.65**
Average	35.85	**37.53**	48.15	**51.36**	78.10	**79.82**	58.02	**60.33**	78.40	**78.56**

Credit: Jia et al., PodGPT: An Audio-augmented Large Language Model for Research and Education, In Nature npj Biomedical Innovations'25.

- **Benchmarks (zero-shot transfer)**

Language	Benchmark datasets	Model									
		Gemma 2B		Gemma 7B		Quantized LLaMA 70B		Mixtral 8×7B MoE		LLaMA 70B	
		Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*
Mandarin	MedQA-MCMLE	**33.54**	33.14 (0.13)	40.81	**45.20 (0.06)**	82.69	**84.15**	**45.94**	45.49 (0.31)	**86.14**	86.02
	Physics	**32.75**	31.73 (0.37)	35.37	**40.13 (0.54)**	**61.08**	60.53	**42.72**	38.90 (1.42)	**62.42**	61.92
	Biology	24.26	**25.44 (0.00)**	33.14	**37.43 (0.64)**	62.72	**63.91**	**44.38**	39.94 (2.15)	59.76	**60.36**
	Chemistry	25.00	**27.08 (0.32)**	29.55	**37.12 (0.54)**	**45.45**	43.94	29.55	**34.47 (0.85)**	46.97	**47.73**
	Computer Science	32.24	**34.18 (0.32)**	40.78	**47.92 (0.59)**	73.60	**76.82**	**57.16**	56.20 (2.42)	78.46	**78.46**
	Engineering	**33.40**	32.05 (0.40)	41.72	**45.58 (0.20)**	62.48	**63.72**	42.30	**46.75 (0.93)**	**66.92**	66.16
	Mathematics	**26.82**	24.64 (0.65)	27.96	**30.73 (0.81)**	**53.72**	51.93	**38.21**	30.41 (0.90)	**58.89**	57.90
	Medicine	31.18	**31.62 (0.15)**	35.72	**39.65 (0.18)**	70.06	**71.21**	41.06	**41.77 (0.74)**	**73.42**	73.29
	Average	29.90	**29.98**	35.63	**40.47**	63.98	**64.53**	**42.66**	41.74	**66.62**	66.48
French	MedExpQA	19.20	**22.40 (0.00)**	28.00	**37.40 (1.18)**	**78.40**	77.60	49.60	**52.80 (1.50)**	77.60	**77.60**
	FrenchMedMCQA	**31.46**	28.04 (0.22)	33.64	**43.69 (0.34)**	81.62	**84.11**	57.01	**64.64 (0.47)**	87.23	**87.85**
	Physics	26.93	**28.88 (0.23)**	35.58	**42.65 (0.59)**	**71.28**	70.31	55.56	**56.51 (0.51)**	**72.40**	72.29
	Biology	34.51	**39.35 (0.39)**	50.16	**57.04 (0.34)**	89.29	**91.01**	70.03	**73.52 (1.01)**	90.90	**90.90**
	Chemistry	26.07	**29.09 (0.32)**	37.98	**40.20 (0.53)**	**64.94**	64.93	47.14	**49.73 (0.45)**	**65.18**	64.43
	Computer Science	35.46	**37.40 (0.69)**	44.20	**45.23 (0.70)**	75.85	**75.88**	**57.54**	56.43 (0.41)	**76.49**	76.24
	Engineering	**38.62**	36.20 (0.60)	46.90	**47.59 (0.00)**	68.28	**72.41**	53.79	**54.31 (0.57)**	71.72	**71.72**
	Mathematics	26.14	**28.01 (0.51)**	30.62	**32.92 (0.98)**	**64.87**	63.90	44.65	**46.68 (1.14)**	65.63	**65.92**
	Medicine	32.56	**35.48 (0.12)**	45.80	**51.55 (0.18)**	81.36	**83.44**	65.08	**66.49 (0.29)**	84.21	**84.29**
	Average	30.11	**31.65**	39.21	**44.25**	75.10	**75.95**	55.60	**57.90**	**76.82**	76.80
Hindi	Physics	25.49	**26.60 (0.56)**	29.32	**33.39 (0.46)**	**58.48**	55.90	**32.06**	31.14 (1.26)	58.68	**59.12**
	Biology	29.02	**32.58 (0.18)**	29.28	**39.08 (0.66)**	**66.55**	66.46	**35.69**	34.03 (0.74)	**72.74**	72.61
	Chemistry	**24.08**	20.88 (0.10)	35.26	**36.97 (0.90)**	50.84	**51.05**	**33.98**	30.25 (2.59)	**54.04**	53.29
	Computer Science	**32.15**	30.30 (0.40)	36.64	**41.49 (0.37)**	**65.78**	61.22	**37.20**	36.40 (2.60)	66.45	**66.45**
	Engineering	**43.45**	42.42 (0.34)	40.00	**41.72 (1.04)**	**59.31**	57.93	40.00	**43.28 (1.23)**	57.24	**58.62**
	Mathematics	**25.33**	24.87 (0.24)	29.33	**30.96 (0.50)**	**53.19**	48.35	**33.74**	33.10 (0.47)	**54.60**	52.82
	Medicine	26.77	**29.07 (0.15)**	34.00	**40.26 (0.21)**	64.60	**65.08**	**33.98**	33.43 (0.80)	71.04	**71.05**
	Average	29.47	**29.53**	33.40	**37.70**	**59.82**	58.00	**35.24**	34.52	**62.11**	61.99
Spanish	HEAD-QA	33.66	**34.38 (0.10)**	48.32	**52.87 (0.19)**	81.58	**82.97**	64.48	**66.76 (0.50)**	**84.28**	84.21
	MedExpQA	21.60	**23.20 (0.00)**	32.80	**37.40 (0.35)**	76.80	**80.00**	51.20	**54.00 (1.65)**	76.80	**78.40**
	Physics	28.06	**28.86 (0.26)**	40.64	**43.63 (0.11)**	70.20	**70.92**	53.32	**54.84 (0.67)**	71.59	**72.13**
	Biology	30.63	**39.50 (0.55)**	52.19	**57.26 (0.51)**	85.37	**85.46**	66.46	**71.11 (0.47)**	**86.73**	86.50
	Chemistry	**27.06**	25.94 (0.22)	35.98	**39.93 (0.36)**	**63.72**	61.72	48.36	**48.74 (1.00)**	**67.18**	65.93
	Computer Science	37.09	**40.43 (0.40)**	45.93	**48.04 (0.25)**	73.38	**74.30**	**56.74**	56.71 (0.78)	74.77	**75.72**
	Engineering	**43.45**	38.79 (0.75)	47.59	**49.14 (0.89)**	73.10	**73.10**	**55.17**	53.79 (1.89)	**72.41**	71.72
	Mathematics	**26.63**	25.80 (0.14)	31.14	**32.79 (0.39)**	**64.80**	61.97	**42.04**	41.70 (0.72)	**67.08**	66.47
	Medicine	31.89	**35.54 (0.15)**	45.94	**51.81 (0.44)**	78.28	**79.64**	61.17	**63.25 (0.90)**	**80.19**	80.09
	Average	31.12	**32.49**	42.28	**45.87**	74.14	**74.45**	55.44	**56.77**	75.67	**75.69**

Evidence Retrieval and Grounding in Medicine



Credit: Jia et al., PodGPT: An Audio-augmented Large Language Model for Research and Education, In Nature npj Biomedical Innovations'25.

Part 4 – PodGPT

- **RAG Database**

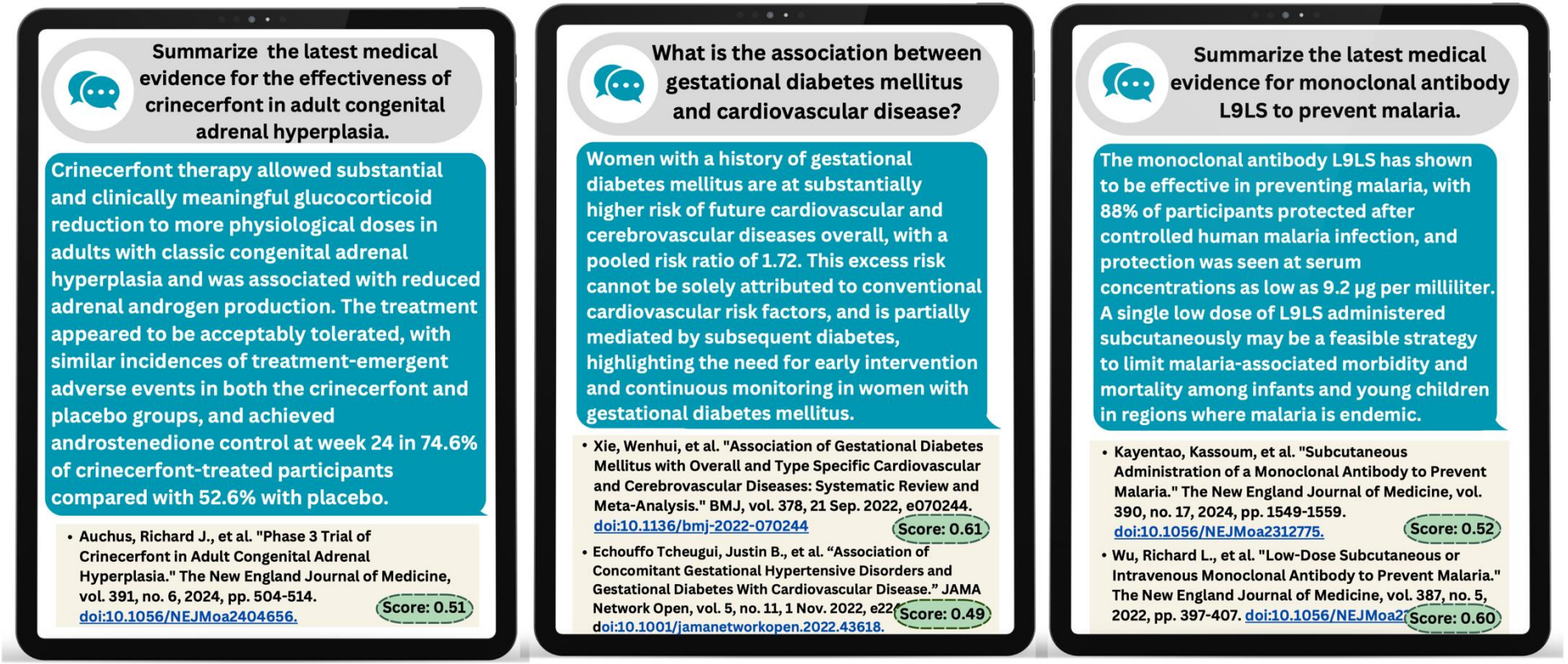
Journal	Number of articles	Number of text samples	Average length per sample
JAMA Network Open	9,367	119,672	389.61
The New England Journal of Medicine	2,013	33,344	400.33
Cell	497	24,977	417.86
British Medical Journal	601	10,307	399.27
The Lancet	458	9,425	403.77
The Lancet Global Health	539	9,003	399.56
Neurology	428	6,862	404.43
JAMA Health Forum	524	5,524	390.70
The Lancet Regional Health - Europe	337	5,435	398.94
The Lancet Infectious Diseases	272	4,720	400.31
The Lancet Regional Health – Western Pacific	243	3,267	397.28
The Lancet Public Health Home	179	3,059	401.88
JAMA Psychiatry	180	2,859	399.22
JAMA Neurology	141	2,462	397.50
The Lancet Oncology	116	2,304	400.62
The Lancet Regional Health – Americas	174	2,262	395.02
The Lancet Microbe	111	2,054	403.17
The Lancet Psychiatry	119	1,951	400.77
The Lancet Neurology	110	1,841	396.53
JAMA Oncology	138	1,810	394.91
The Lancet HIV	97	1,786	401.54
The Lancet Planetary Health	100	1,776	400.43
JAMA Pediatrics	135	1,773	391.99
JAMA Internal Medicine	128	1,766	395.25
The Lancet Regional Health – Southeast Asia	134	1,607	397.18
The Lancet Respiratory Medicine	72	1,360	404.70
The Lancet Child & Adolescent Health	68	1,205	402.95
The Lancet Diabetes & Endocrinology	64	1,082	399.89
The Lancet Gastroenterology and Hepatology	54	1,007	397.67
JAMA Cardiology	60	947	396.69
JAMA Surgery	68	941	399.31
The Lancet Healthy Longevity	49	855	403.10
JAMA Ophthalmology	56	810	398.01
The Lancet Rheumatology	35	713	403.76
The Lancet Haematology	33	604	398.36
JAMA Dermatology	30	413	399.73
JAMA Otolaryngology – Head & Neck Surgery	17	244	399.27
JAMA Facial Plastic Surgery	1	10	424.80

Part 4 – PodGPT

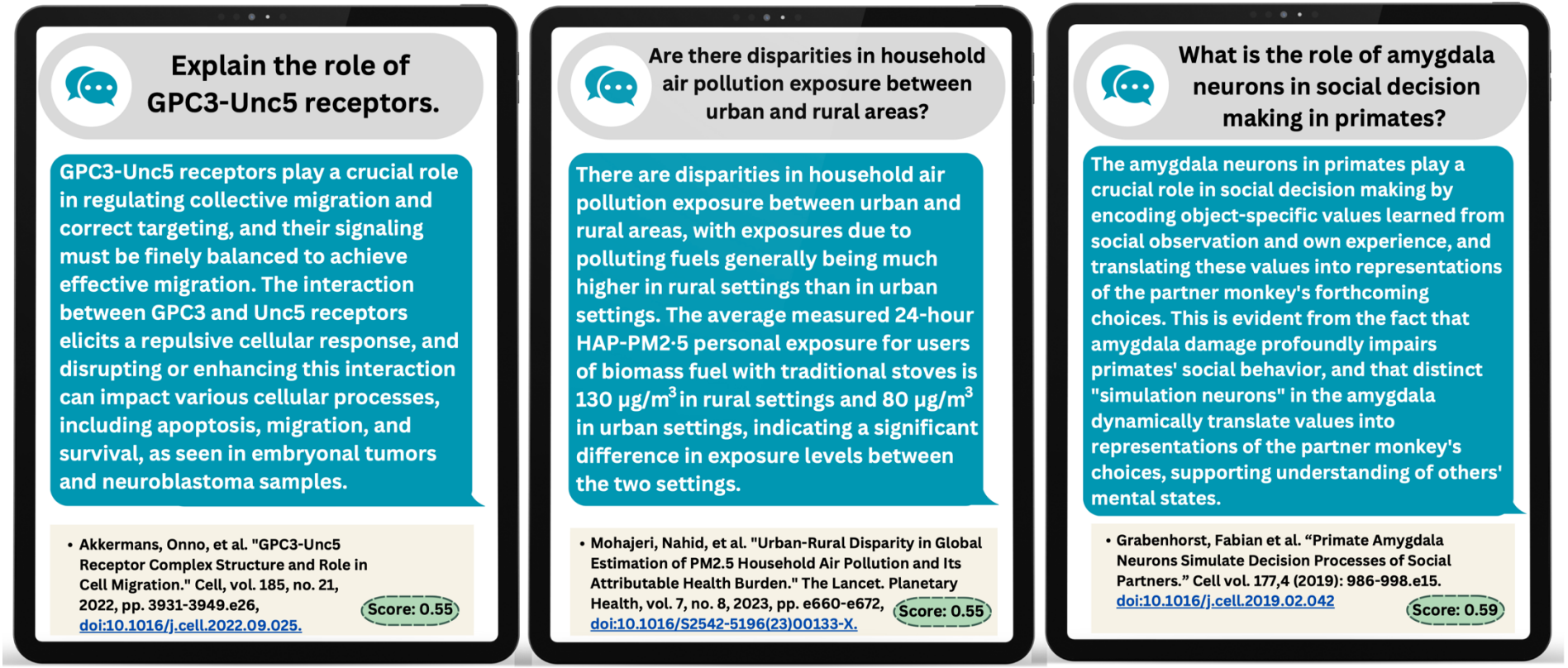
- **Benchmarks (in-domain performance)**

Benchmark datasets			MedExpQA	MedMCQA	MedQA	PubMedQA	MMLU Medicine	**Average**
Model	Gemma 2B	Baseline	19.20	**34.71**	29.54	46.80	40.62	34.17
		Ours	**21.20 (0.69)**	34.62 (0.02)	**32.91 (0.15)**	**54.25 (0.54)**	**41.72 (0.15)**	**36.94**
		Baseline+RAG	23.20	35.91	32.60	49.00	41.47	36.44
		Ours+RAG	**28.00 (0.00)**	**35.96 (0.07)**	**34.43 (0.12)**	**51.95 (0.78)**	**42.12 (0.13)**	**38.49**
	Gemma 7B	Baseline	34.40	40.69	37.78	**61.80**	55.22	45.98
		Ours	**42.00 (0.89)**	**44.64 (0.09)**	**44.14 (0.21)**	57.35 (1.37)	**59.50 (0.14)**	**49.53**
		Baseline+RAG	35.20	40.64	39.28	**61.40**	54.14	46.13
		Ours+RAG	**47.40 (1.18)**	**43.54 (0.07)**	**43.32 (0.25)**	55.70 (1.88)	**60.11 (0.39)**	**50.01**
	Quantized LLaMA 70B	Baseline	73.60	68.30	76.67	77.60	86.11	76.46
		Ours	**75.20**	**69.11**	**78.00**	**77.60**	**87.38**	**77.46**
		Baseline+RAG	69.60	66.51	66.77	76.80	83.91	72.72
		Ours+RAG	**76.00**	**68.13**	**69.60**	**78.80**	**85.62**	**75.63**
	Mixtral 8×7B MoE	Baseline	52.80	50.11	54.20	**42.00**	67.38	**53.30**
		Ours	**58.60 (1.54)**	**51.94 (0.31)**	**55.36 (0.81)**	25.35 (0.78)	**74.00 (0.71)**	53.05
		Baseline+RAG	49.60	48.58	47.84	**34.60**	65.45	49.21
		Ours+RAG	**57.12 (1.21)**	**51.29 (0.28)**	**51.14 (0.01)**	24.44 (1.11)	**72.01 (0.02)**	**51.20**
	LLaMA 70B	Baseline	75.20	70.31	**78.40**	76.20	88.65	77.75
		Ours	**75.20**	**70.62**	78.32	**76.40**	**88.65**	**77.84**
		Baseline+RAG	68.00	68.11	66.93	77.40	84.24	72.94
		Ours+RAG	**71.20**	**68.59**	**68.42**	**77.40**	**85.60**	**74.24**

Credit: Jia et al., PodGPT: An Audio-augmented Large Language Model for Research and Education, In Nature npj Biomedical Innovations'25.



Credit: Jia et al., PodGPT: An Audio-augmented Large Language Model for Research and Education, In Nature npj Biomedical Innovations'25.



Credit: Jia et al., PodGPT: An Audio-augmented Large Language Model for Research and Education, In Nature npj Biomedical Innovations’25.

Part 4 – PodGPT

- **Performance on conversational dataset**

Model		Gemma 2B		Gemma 7B		Quantized LLaMA 70B		Mixtral 8×7B MoE		LLaMA 70B	
		Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*
Perplexity	Original transcripts	17.54	**10.15 (0.01)**	27.04	**8.75 (0.02)**	7.34	**7.27**	**6.11**	6.32 (0.04)	7.31	**7.10**
	Our transcripts	14.08	**7.64 (0.02)**	22.72	**6.67 (0.02)**	6.12	**5.51**	5.23	**5.05 (0.01)**	6.10	**5.55**

$$\mathrm{PPL}(\mathbf{x}) = \mathrm{e}^{-\frac{1}{t}\sum_i^t \log\left(p_\theta(x_i|\mathbf{x}_{<i})\right)}.$$

Credit: Jia et al., PodGPT: An Audio-augmented Large Language Model for Research and Education, In Nature npj Biomedical Innovations'25.

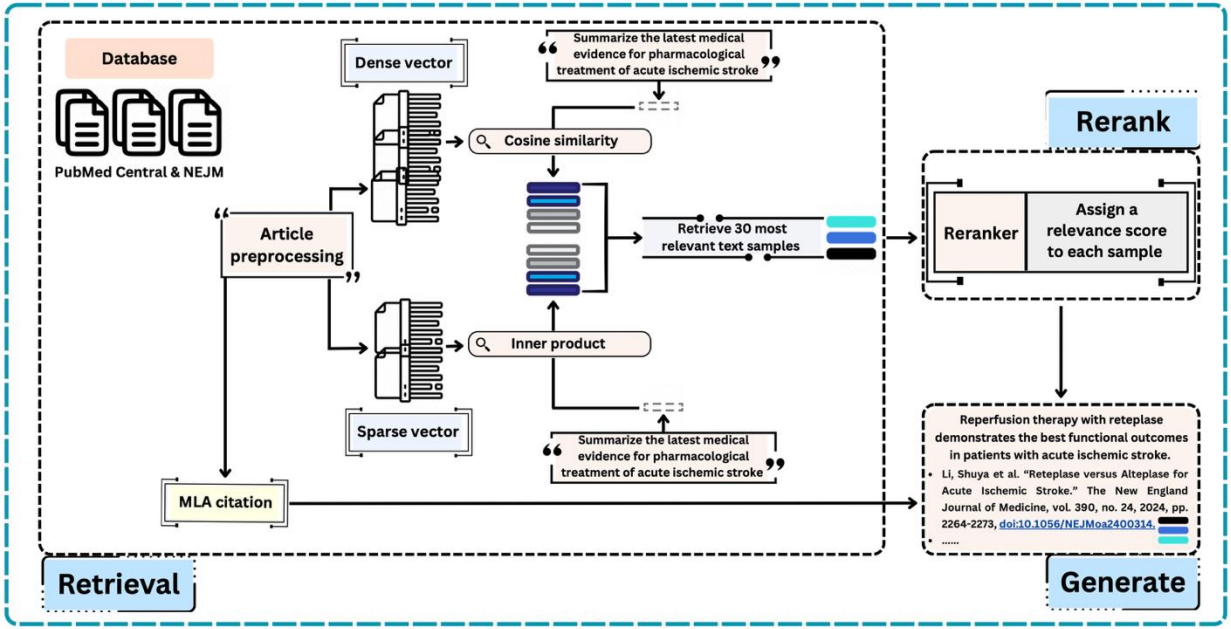
Part 4 – PodGPT

- **Take away**
  - **Podcasts** are valuable resources for **STEMM research and education**
  - **Grounding LLMs with RAG** enhances **factual accuracy and reliability**
  - **Continual pretraining of LLMs on podcasts** improves their **conversational** capability
  - **Medical journals** are valuable resources for **evidence-based AI generation**

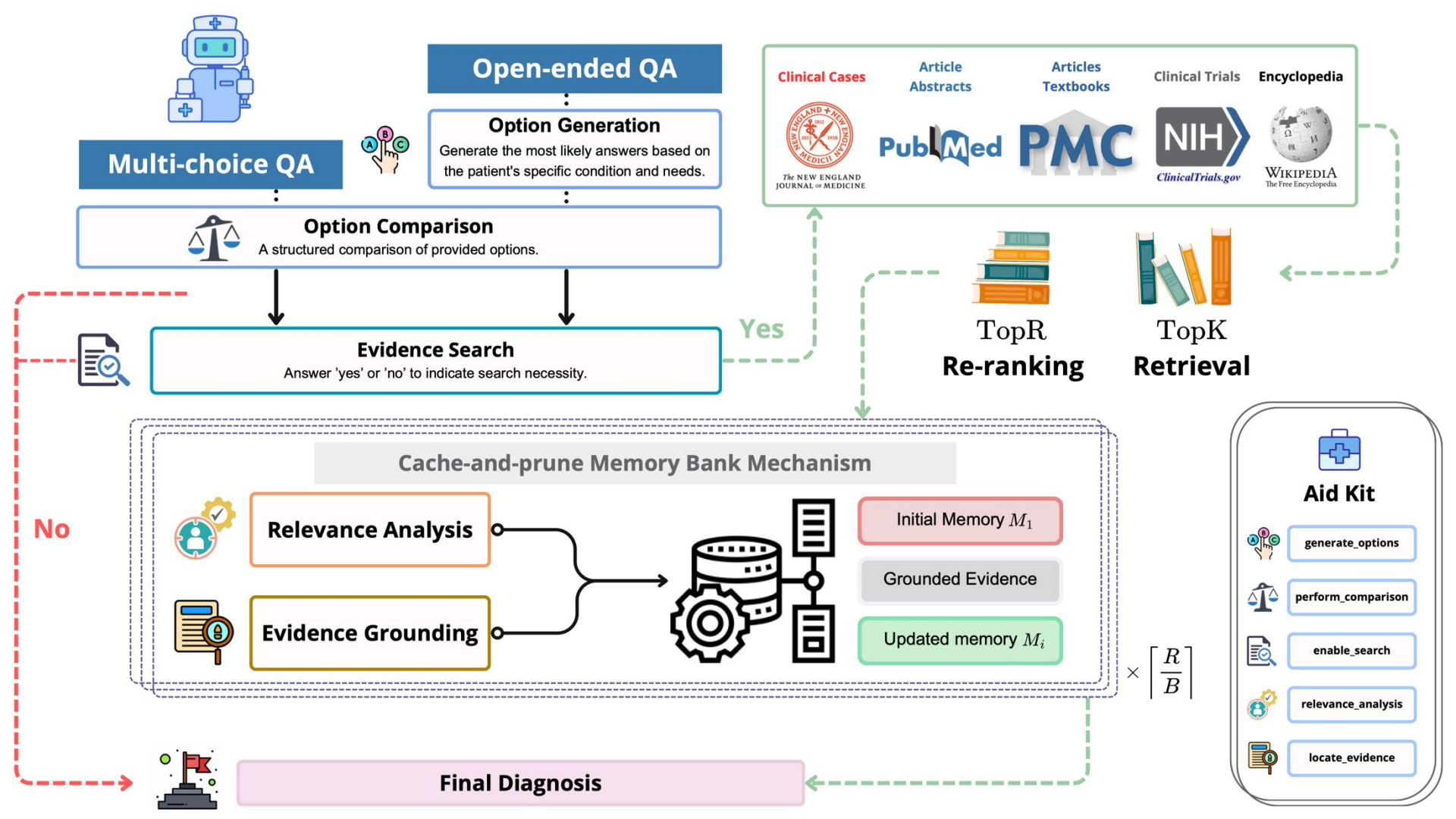
Part 5 – Agentic System

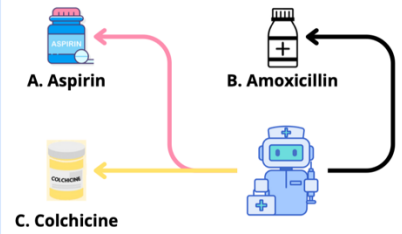
- Agentic memory-augmented retrieval and evidence grounding in medicine
  - Retrieval
  - Reranking
  - Generation
- Unified
- Automatic
- Dynamic

Aid Kit



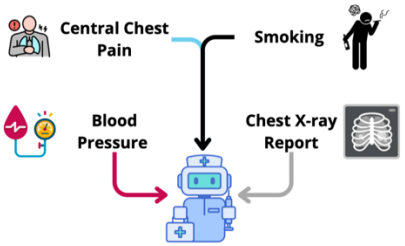
Credit: Jia et al., Agentic Memory-augmented Retrieval and Evidence Grounding in Medicine, Under Review.


Open-ended QA
Option Generation
Generate the most likely answers based on the patient's specific condition and needs.
Multi-choice QA
Option Comparison
A structured comparison of provided options.
Evidence Search
Answer 'yes' or 'no' to indicate search necessity.
Yes
No
Clinical Cases
Article Abstracts
Articles Textbooks
Clinical Trials
Encyclopedia
The NEW ENGLAND JOURNAL of MEDICINE
PubMed
PMC
NIH
ClinicalTrials.gov
WIKIPEDIA
The Free Encyclopedia
TopR
Re-ranking
TopK
Retrieval
Cache-and-prune Memory Bank Mechanism
Relevance Analysis
Evidence Grounding
Initial Memory M_1
Grounded Evidence
Updated memory M_i
$\times \left\lceil \frac{R}{B} \right\rceil$
Aid Kit
generate_options
perform_comparison
enable_search
relevance_analysis
locate_evidence
Final Diagnosis



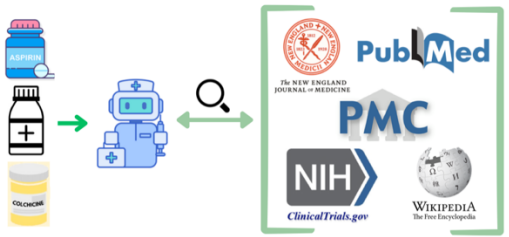
Generate the most likely answers based on the patient's specific condition and needs. Follow the format: 'A. answer 1 B. answer 2 C. answer 3 D. answer 4 E. answer 5'.

generate_options



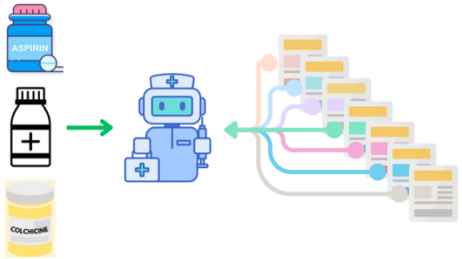
Compare and analyze all available options based on the patient's specific condition and needs. Highlight key differences and relevance between each option and the case presented.

perform_comparison



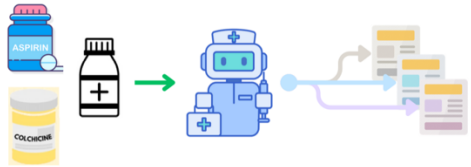
Indicate whether external evidence should be searched to support the patient's diagnosis based on the current clinical context and the need for additional information.

enable_search



Assess the relevance of each document in relation to the patient's presentation and provide a detailed analysis.

relevance_analysis



Retrieve relevant evidence from the relevance analysis based on the patient's presentation and return them as a list. Each piece of relevant information must be followed by the **specific article's ID enclosed in <quote></quote> tags**. If no relevant evidence is found, simply return 'evidence': 'No relevant evidence found for the patient's diagnosis.'

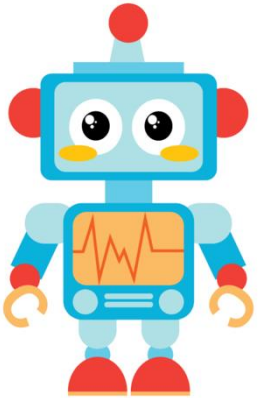
locate_evidence

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

Part 5 – Agentic System

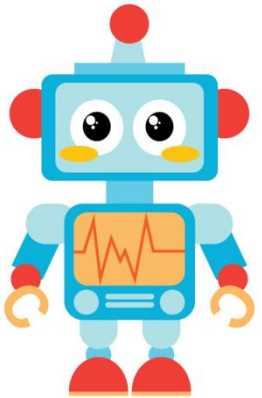
- **Tool using (functional call)**
- **Enhanced interpretability and user trust**
  - **Perform a specific task** with **specific output constraints** (but with input perturbations).
  - Better information retrieval with reduced hallucinations.
- **Access to external tools, databases, and APIs**
  - **Knowledge Acquisition** ← External knowledge
- **Automation and efficiency**
  - Complete complex tasks without excessive prompting.
- **Interaction enhancement**
  - **Interact with external models**, e.g., MRI model, CT scan model, etc.



Credit: Qu et al., Tool Learning with Large Language Models: A Survey, In Frontiers of Computer Science'25.

Part 5 – Agentic System

- **Tool using (functional call)**
- **Task planning**
  - Which step should I solve?
- **Tool selection**
  - Should I use a tool(s)?
  - Which tool(s) should I use to solve this step problem?
- **Tool calling**
  - Calling and performing the specific action that the tool has defined.
- **Response generation**
  - Generating a response based on the tool's output description.



Credit: Qu et al., Tool Learning with Large Language Models: A Survey, In Frontiers of Computer Science'25.

Part 5 – Agentic System

- **Tools**

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.

Credit: Jia et al., Agentic Memory-augmented Retrieval and Evidence Grounding in Medicine, Under Review.

Part 5 – Agentic System

- **Database**

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

Credit: Jia et al., Agentic Memory-augmented Retrieval and Evidence Grounding in Medicine, Under Review.

- **Database**

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

Part 5 – Agentic System

- **Benchmarks (medical analysis and diagnosis)**

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

Getting more challenging

Credit: Jia et al., Agentic Memory-augmented Retrieval and Evidence Grounding in Medicine, Under Review.

Part 5 – Agentic System

- **Benchmarks (multiple choice question-answering)**

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**

Credit: Jia et al., Agentic Memory-augmented Retrieval and Evidence Grounding in Medicine, Under Review.

Part 5 – Agentic System

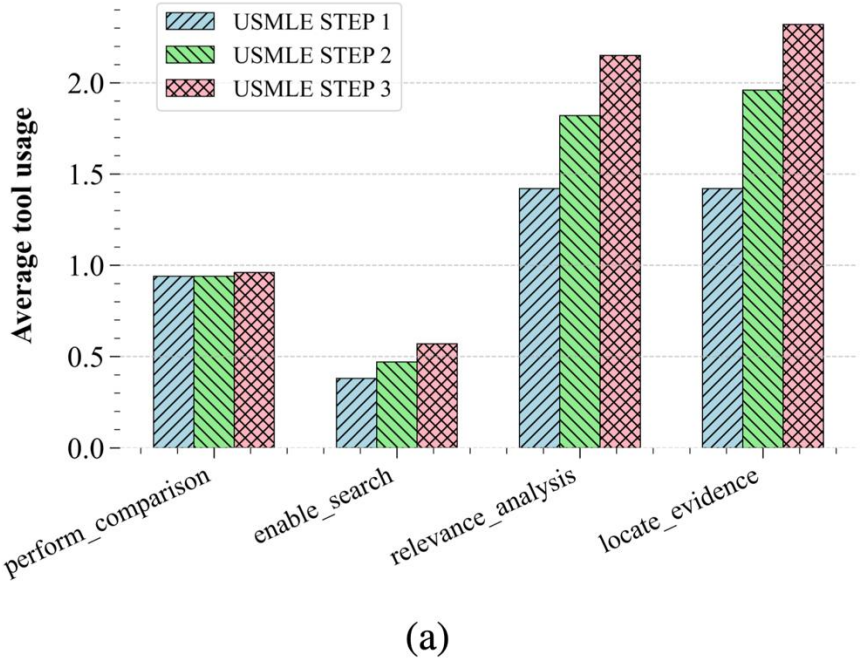
- **Benchmarks (open-ended question-answering)**

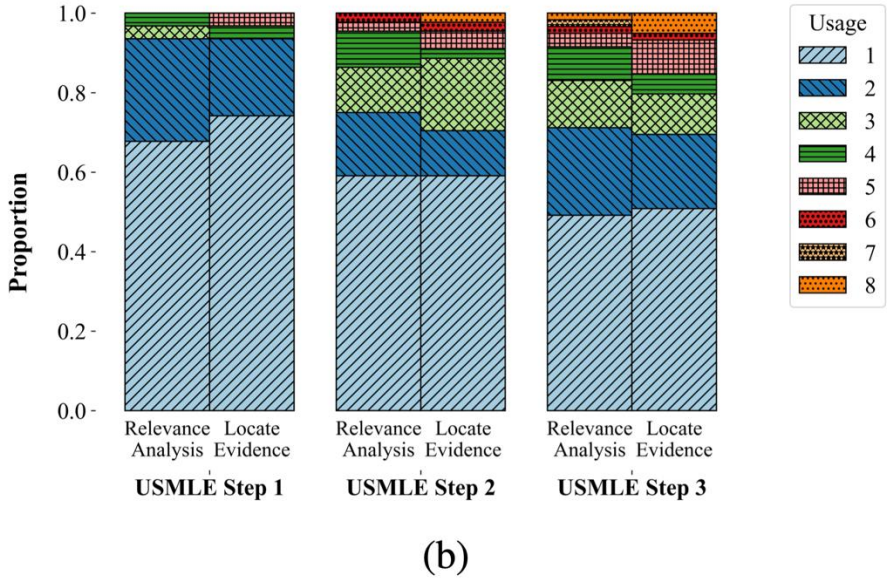
Benchmark	Model	BioMistral (7B)	OpenBioLLM (8B)	UltraMedical (8B)	OpenBioLLM (70B)	UltraMedical (70B)	PodGPT (70B)	Ours
USMLE Step 1	SFR	0.79 ± 0.09	0.70 ± 0.12	0.81 ± 0.13	0.85 ± 0.10	0.82 ± 0.11	0.86 ± 0.11	**0.87** ± 0.09
	GTE	0.48 ± 0.17	0.38 ± 0.17	0.57 ± 0.21	0.60 ± 0.23	0.63 ± 0.23	0.66 ± 0.24	**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

Credit: Jia et al., Agentic Memory-augmented Retrieval and Evidence Grounding in Medicine, Under Review.

Part 5 – Agentic System

- **Tool usage**





Credit: Jia et al., Agentic Memory-augmented Retrieval and Evidence Grounding in Medicine, Under Review.

Part 5 – Agentic System

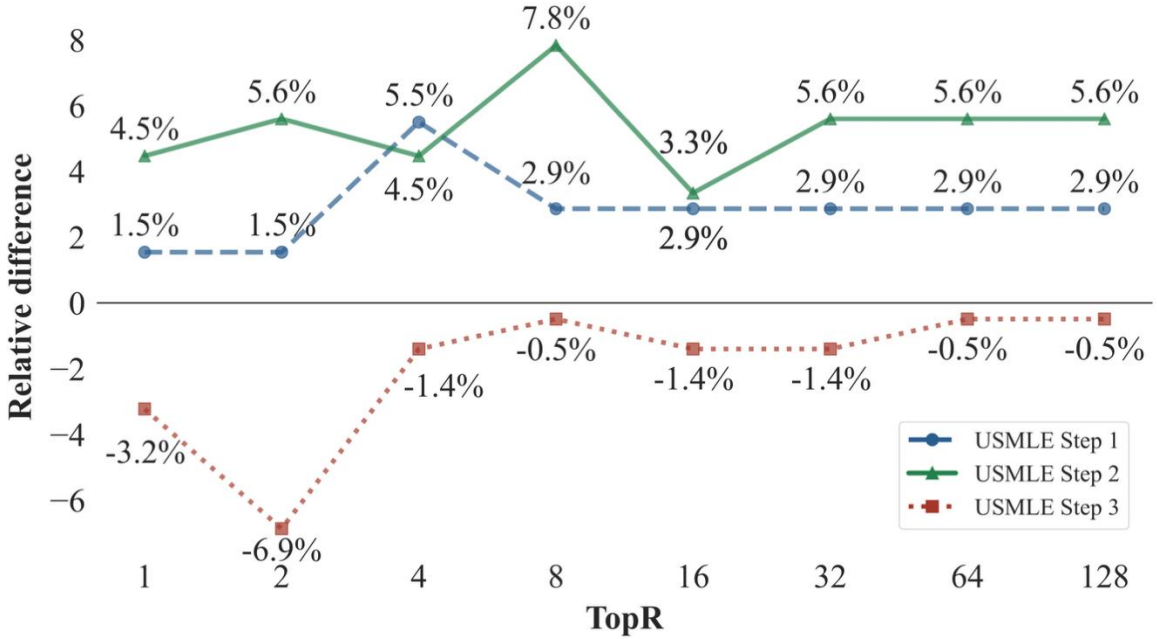
- **Ablation study**

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

Credit: Jia et al., Agentic Memory-augmented Retrieval and Evidence Grounding in Medicine, Under Review.

Part 5 – Agentic System

- **Ablation study**



Credit: Jia et al., Agentic Memory-augmented Retrieval and Evidence Grounding in Medicine, Under Review.

Part 5 – Agentic System

- **Ablation study**

Table 1: Performance evaluation of different LLMs. We report different LLMs' performance (accuracy) across the USMLE Steps 1 to 3.

Model	USMLE Step 1	USMLE Step 2	USMLE Step 3
GPT-4	80.67	81.67	**89.78**
ChatGPT	51.26	60.83	58.39
Ours (Qwen2.5 72B)	82.98	86.24	88.52
Ours (Qwen3 32B)	**91.49**	**87.16**	86.07

Reasoning-based model

Credit: Jia et al., Agentic Memory-augmented Retrieval and Evidence Grounding in Medicine, Under Review.

Part 5 – Agentic System

- **Runtime performance metric**

Table 1: Latency (in seconds), per-GPU seconds per query (in seconds), and GPU hours per query (in hours) of the agentic system deployed using 4 NVIDIA L40S GPUs.

Qwen2.5 72B	USMLE Step 1	USMLE Step 2	USMLE Step 3
Latency	1.12	1.48	1.80
per-GPU seconds per query	39.15	51.64	83.30
GPU hours per query	0.043	0.057	0.093

Credit: Jia et al., Agentic Memory-augmented Retrieval and Evidence Grounding in Medicine, Under Review.

Part 5 – Agentic System

- **Take away**
  - **Agentic framework unifies** document retrieval, evidence grounding, as well as AI generation, with an **automatic and dynamic process**
  - **Tool-augmented LLM-based agent** enables **dynamic multistep tool use**, eliminating the need for manually engineered prompts or multi-stage pipelines
  - **The cache-and-prune memory bank mechanism** efficiently extends the retention of relevant documents for **evidence grounding**, enhancing diagnostic accuracy and computational efficiency

Thank you very much for your attention!

Department of Electrical and Computer Engineering
Boston University

BOSTON UNIVERSITY