MedAssist: LLM-Empowered Medical Assistant for Assisting the Scrutinization and Comprehension of Electronic Health Records

Ran Xu
Emory University
Atlanta, Georgia, USA
ran.xu@emory.edu

Wenqi Shi
UT Southwestern Medical Center
Dallas, Texas, USA
wenqi.shi@utsouthwestern.edu

Jonathan Wang
Emory University
Atlanta, Georgia, USA
jonathan.wang@emory.edu

Jasmine Zhou
Emory University
Atlanta, Georgia, USA
jasmine.zhou@emory.edu

Carl Yang
Emory University
Atlanta, Georgia, USA
j.carlyang@emory.edu

Abstract

Efficiently comprehending diagnosis and treatment plans remains a significant challenge for both medical professionals and patients, particularly when dealing with rare or newly emerging diseases and specific combinations of comorbidities. We present MedAssist, a large language model (LLM)-empowered medical assistant designed to support the scrutinization and comprehension of electronic health records (EHRs). MedAssist leverages two key components: *medical knowledge retrieval*, which retrieves the latest and most comprehensive medical knowledge snippets from the web, and *data retrieval*, which extracts diagnosis and treatment plans for similar patients from existing EHR databases. By integrating these capabilities into user-friendly interfaces, MedAssist bridges critical gaps in medical knowledge accessibility and understanding, and advances patient care in realistic clinical scenarios.

CCS Concepts

• **Computing methodologies** → **Natural language processing**;
• **Applied computing** → **Life and medical sciences**.

Keywords

Electronic Health Records, Retrieval, Large Language Models

ACM Reference Format:
Ran Xu, Wenqi Shi, Jonathan Wang, Jasmine Zhou, and Carl Yang. 2025. MedAssist: LLM-Empowered Medical Assistant for Assisting the Scrutinization and Comprehension of Electronic Health Records. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28–May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/XXXXXX.XXXXXX

1 Introduction

Electronic Health Records (EHRs) are systematically collected across diverse healthcare institutions, covering comprehensive patient information such as diagnoses, medications, and laboratory results.

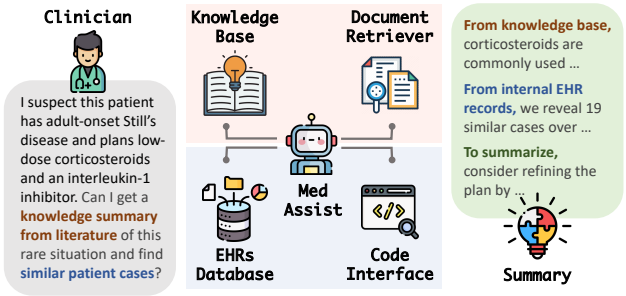

Figure 1: An example of MedAssist. It enables efficient interactions between clinicians and EHR systems and reduce the burden of heavy data engineering with LLMs.

In clinical research and practice, clinicians utilize EHR systems to access relevant cohort data, spanning detailed individual records to population-level insights, to support clinical decision-making and improve the quality and efficiency of healthcare delivery [2].

However, generating comprehensive diagnoses and treatment plans remains a significant challenge due to the reliance on extensive data engineering support to extract information from EHR systems and external knowledge bases. These inefficiencies highlight the need for an automated EHR assistant system capable of streamlining data analysis and improving clinical accuracy, thereby optimizing the efficiency and effectiveness of EHR workflows.

Large language models (LLMs) [5] brings us one step closer to achieving such automated EHR assistants, as with strong text encoding [9, 15], instruction-following [6], and reasoning abilities [7]. These properties enable LLMs to bridge the gap between the complexity of EHR data and actionable insights, facilitating tasks such as extracting relevant information, contextualizing medical codes, and generating personalized patient summaries. Although recent studies have explored adapting LLMs to EHRs, they mainly focus on target disease prediction [12], text-based question answering [3, 13] or data analysis [10]. However, there is a notable lack of a unified framework for systematically integrating LLMs into broader EHR workflows to support clinical decision making.

Research Project. In this project, we develop MedAssist, a user-guided system designed to help users scrutinize and comprehend diagnoses and treatments, thereby improving patient care. MedAssist features user-friendly interfaces tailored for both clinicians and

WWW Companion '25, April 28–May 2, 2025, Sydney, NSW, Australia.

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/10.1145/XXXXXX.XXXXXX

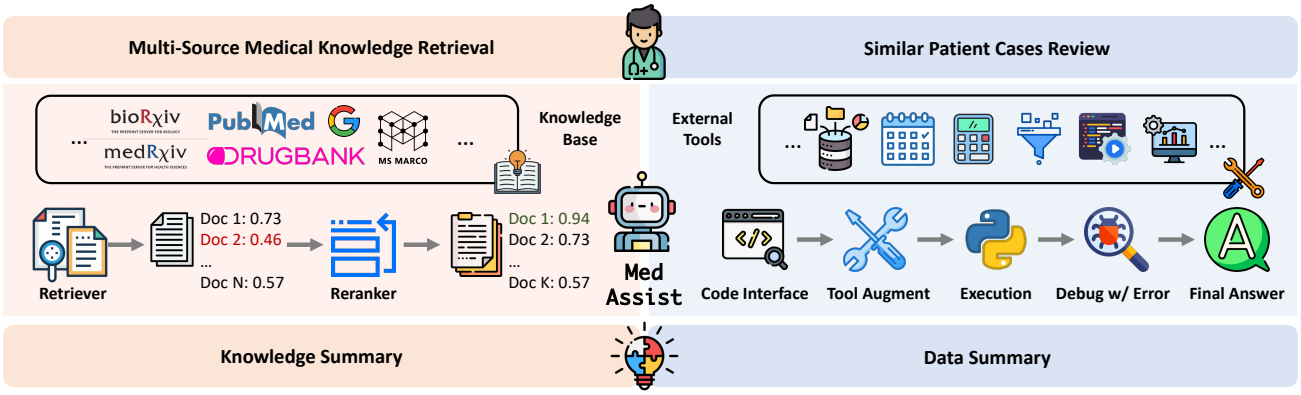

Figure 2: Overview of MedAssist. It incorporates retrieval to extract customized information from *external medical knowledge sources* and *internal health records*. This customized context is then utilized to generate personalized patient summaries.

patients and integrates two key modules: (1) *knowledge retrieval from external sources*, which retrieves relevant literature and knowledge bases to augment patient data. This step provides contextual insights into medical codes. (2) *data retrieval from internal records*, which effectively processes complex user queries to extract and reason over query-specific medical information from existing EHR databases. By integrating external knowledge with local EHR data, MedAssist generates informative and useful summaries to improve patient understanding as well as support clinical decision-making.

Fit with the WWW ecosystem. This demonstration is highly relevant to researchers in Data Mining, Information Retrieval, and Health Informatics, as it showcases the integration of advanced language models with web-based medical data to enhance information retrieval and comprehension. The system leverages web technologies to process and summarize extensive medical literature and EHRs, facilitating improved understanding and decision-making in healthcare. By addressing challenges in accessing and interpreting web-based medical information, MedAssist aligns with the conference's focus on innovative web applications and their societal impact. The core methods of the system will be open-sourced, allowing users to customize them within their local clinical environments.

2 The MedAssist Framework

In this section, we present a detailed description of the MedAssist framework. As depicted in Figure 2, MedAssist encompasses several key steps: (1) *knowledge retrieval from external resources*: it collects diverse knowledge resources, converts them into text format, and uses user input to retrieve tailored knowledge from external sources based on specific queries. (2) *data retrieval from internal records*: it decomposes complex queries into a sequence of manageable actions, utilizing external toolsets to navigate EHRs. It facilitates seamless clinician interaction with EHRs using natural language. (3) *patient summary generation*: it integrates local EHR data with LLMs to produce concise and informative patient summaries, combining insights from both internal and external knowledge sources.

2.1 Knowledge Retrieval from Medical Corpora

Retrieval serves as an effective tool to inject external knowledge to existing models without expensive parameter update, and have been applied to health domain with success [14, 16]. To leverage the semantic richness of medical codes for enhanced summarization, we retrieve relevant knowledge for each medical code (e.g., diseases, symptoms, medications) in EHRs using its surface name. Intending to ensure comprehensive clinical knowledge coverage, we curate a diverse external corpus comprising PubMed, DrugBank, MeSH, Wikipedia, and MS MARCO. Each knowledge unit is represented as raw text to enable efficient retrieval. For the retrieval process, we employ our developed BMRetriever [15], which archives state-of-the-art performance on a broad suite of biomedical retrieval tasks and can follow human instructions for retrieval well.

Specifically, we first use BMRetriever $R(\cdot)$ to build an index for corpus $\mathcal{M}$ to support retrieval. At runtime, we map each medical code q, paired with a user-defined instruction I (e.g., specifying the type of information required), into an embedding vector aligned with the corpus passage embeddings. The similarity between the query and a passage d is computed as: $f(I, q, d) = R([I, q])^\top R(d)$. For the medical code c_i with the surface name s_i, we retrieve top-N passages $\mathcal{T}_i$ from the corpus $\mathcal{M}$ as

$$\mathcal{T}_i = \underset{d \in \mathcal{M}}{\text{Top-}k}\ f(I_i, s_i, d). \tag{1}$$

To improve the quality of retrieval results, ranking often serves as a intermediate step to filter out low-quality passages [17]. We enhance this process by reranking the top-N retrieved passages using a pre-trained cross-encoder [1][1]. Specifically, for each passage d_i among retrieved k passages, we concatenate medical code c_i and passage p as the input of the pre-trained cross-encoder. For each input, it will output a scalar value between 0 to 1, which is then used as reranking scores for k passages. The top-ranked passages are considered as the external knowledge $\mathcal{K}_i$ for the medical code c_i. In our system, we set $N = 25, k = 5$ to balance between retrieval accuracy and inference latency.

2.2 Data Retrieval from Internal Records

To address complex user queries that require extracting information from internal EHR databases, MedAssist leverages EHRAgent [8] for multi-turn interactive coding with external tools, enabling multi-hop reasoning. We incorporate query-specific medical information

[1] https://huggingface.co/BAAI/bge-reranker-v2-m3

for effective reasoning based on the given query, guiding MedAssist to identify and retrieve the relevant tables and records with few-shot examples. To enable LLMs in complex operations such as calculations and information retrieval, MedAssist integrates various external tools for EHR interaction, detailed as follows:

◇ *Database Loader*: It loads a specific table from the database.
◇ *Data Filter*: It filters the loaded table based on conditions defined by a column name and a relational operator (e.g., "<" or ">").
◇ *Get Value*: It retrieves all values from a specific column or performs basic operations (e.g., mean, max, min, sum) on those values.
◇ *Calculator*: It performs calculations from input strings using the WolframAlpha API[2]. It supports both simple operations (e.g., addition, subtraction, multiplication) and more complex ones (e.g., averages, maximum values).
◇ *Calendar*: It computes the date based on an input and time interval.
◇ *SQL Interpreter*: It executes SQL queries generated from LLMs.

It is worth noting that our toolkits can be easily expanded through natural language tool function definitions in a plug-and-play manner. MedAssist employs LLMs as autonomous agents in a multi-turn conversation with a code executor, iteratively refining code based on execution feedback until reaching an optimal solution.

Clinicians often pose complex queries that require advanced reasoning across multiple tables and access to a large number of records within a single query. To accurately identify the necessary tables, we first incorporate query-specific medical knowledge into MedAssist to form a detailed understanding of the query under a limited context length. Given a clinical question q and reference EHRs $\mathcal{R} = \{R_0, R_1, \cdots\}$, MedAssist prompts the LLM to produce domain knowledge $B(q)$ most relevant to q, guiding the identification and location of useful references within $\mathcal{R}$.

Following the background assimilation, MedAssist then integrates LLMs and a code executor in a multi-turn conversation for iterative debugging. Initially, MedAssist generates code $C(q)$ that interacts with the EHR database to extract and process relevant data. The generation of $C(q)$ draws upon the database introduction $\mathcal{I}$, tool function definitions $\mathcal{T}$, a set of K-shot examples $\mathcal{E}(q)$, the original query q, and the contextual background knowledge $B(q)$:

$$C(q) = \text{LLM}([\mathcal{I}; \mathcal{T}; \mathcal{E}(q); q; B(q)]). \tag{2}$$

The code executor subsequently extracts and executes $C(q)$ as $O(q) = \text{EXECUTE}(C(q))$. When execution errors or suboptimal outputs occur, the executor provides error feedback, enabling MedAssist to refine the code through continued conversation until achieving accurate query resolution.

2.3 Knowledge Summarization with LLMs

By combining knowledge from external sources and internal health records, MedAssist leverages LLMs' strengths in instruction following [6] and summarization [4] to generate personalized, context-aware patient diagnosis and treatment summaries.

Specifically, the summarization module in MedAssist integrates information from both local EHR data and external medical corpora to generate personalized summaries for each patient. The module structures its input into three key components: (1) external knowledge relevant to the patient's medical codes or queries, (2) the patient's medical history and retrieved EHR data, and (3) user-defined instructions specifying the focus of the summary. Using this structured input, LLM is able to generate informative, actionable summaries to assist in understanding diagnoses and treatment plans. For clinicians, the summaries from MedAssist highlight diagnostic insights, potential treatment strategies, and key clinical considerations, ensuring relevance and brevity for effective decision-making.

3 Demonstration

Implementation Details. The MedAssist demo uses a modular architecture with Python for the backend (FastAPI for APIs and query processing) and React for a user-friendly frontend. The demo utilizes CSV files as internal records to simulate EHRs, which, combined with external knowledge sources, serve as input for the model to generate structured insights and summaries. For deployment, the backend is hosted on AWS Elastic Beanstalk, and the frontend is deployed on AWS S3 with CloudFront for efficient static file hosting and content delivery. We use the HuggingFace to host the retrieval model and use AutoGen 0.2.0 [11] as the interface for communication between the LLM agent and the code executor.

Case Studies. MedAssist provides an interactive system for users to results by leveraging LLMs through the techniques described in Sec. 2. Fig. 3 highlights the system's three main steps, demonstrated through the case of a 14-year-old patient with adolescent idiopathic scoliosis and mild asthma, illustrated as follows:

◇ *Knowledge Retrieval*: Given a query for the treatment of with scoliosis and mild respiratory conditions, MedAssist first searches external knowledge sources like PubMed and bioRxiv. It identifies key evidence supporting bracing and physiotherapy as effective treatments without adverse respiratory effects.
◇ *Data Retrieval*: MedAssist queries the local EHR database to find similar patient profiles, debugging any errors in the process, and retrieves treatments and outcomes from 22 matching cases.
◇ *Retrieval-Augmented Summary Generation*: Combining this information, MedAssist generates a concise summary emphasizing that the proposed treatment plan aligns with established practices while recommending specific protocol requirements (e.g., at least 16 hours per day) and structured follow-up evaluations.

4 Conclusion

In this demonstration, we develop MedAssist, a system for clinicians to easily interact with EHRs. MedAssist reduce the labor of clinicians for creating patient diagnosis and treatment summaries and streamlines the process of integrating patient data with external medical knowledge. This system has the great potential for assisting clinical decision-making by presenting clear and personalized insights. Future work will focus on expanding MedAssist's capabilities for multi-modal data, such as imaging and genomic data, and ensuring its robustness across diverse patient populations.

5 Ethical Use of Data and Informed Consent

In compliance with the PhysioNet Credentialed Health Data Use Agreement, we prohibit sharing confidential patient data with third parties, including via online APIs. To align with Azure OpenAI Service guidelines[3], we opted out of the human review process by

[2] https://products.wolframalpha.com/api

[3] https://physionet.org/news/post/gpt-responsible-use

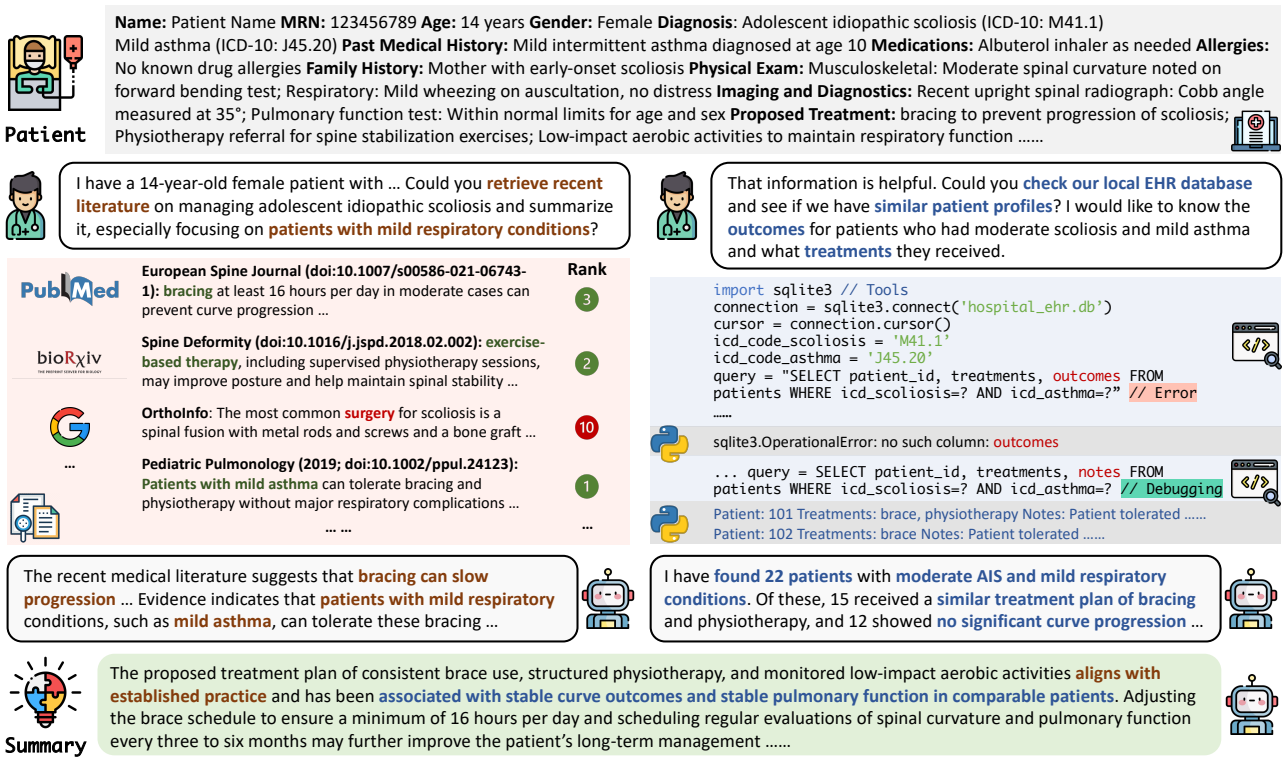

Figure 3: An example of MedAssist on generating informative summaries for a given patient record.

submitting the Azure OpenAI Additional Use Case Form. MedAssist currently uses online biomedical literature (e.g., PubMed) and publicly available EHR data (e.g., MIMIC). Further adoption of data from other resources will be proceeded with proper IRB approvals.

Acknowledgement

This research was partially supported by the Emory Global Diabetes Center of the Woodruff Sciences Center, Emory University. This research was partially supported by the Texas Advanced Computing Center. This research was also supported by the National Institute Of Diabetes And Digestive And Kidney Diseases of the National Institutes of Health under Award Number K25DK135913, and the National Science Foundation under Award Numbers IIS-2312502 and NCS-2319449. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *ACL Findings* (2024).
[2] Michelle R Hribar, Sarah Read-Brown, Isaac H Goldstein, Leah G Reznick, Lorinna Lombardi, Mansi Parikh, et al. 2018. Secondary use of electronic health record data for clinical workflow analysis. *JAMIA* 25, 1 (2018), 40–46.
[3] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, et al. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. In *NeurIPS*.
[4] Philippe Laban, Wojciech Kryściński, Divyansh Agarwal, Alexander Richard Fabbri, Caiming Xiong, Shafiq Joty, and Jason Wu. 2023. Measuring LLM ability at factual reasoning through the lens of summarization. In *EMNLP*. 9662–9676.
[5] OpenAI. 2023. GPT-4 technical report. *arXiv* (2023).
[6] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, et al. 2022. Training language models to follow instructions with human feedback. *NeurIPS* (2022), 27730–27744.
[7] Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Haotian Sun, Hang Wu, Carl Yang, and May Dongmei Wang. 2024. MedAdapter: Efficient Test-Time Adaptation of Large Language Models Towards Medical Reasoning. In *EMNLP*. 22294–22314.
[8] Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce Ho, Carl Yang, and May Dongmei Wang. 2024. Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *EMNLP*. 22315–22339.
[9] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving Text Embeddings with Large Language Models. In *ACL*. 11897–11916.
[10] Hao Wu, Yinghao Zhu, Junyi Gao, Zixiang Wang, Xiaochen Zheng, Ling Wang, Wen Tang, Yasha Wang, Ewen M Harrison, Chengwei Pan, and Liantao Ma. 2024. EHRFlow: A Large Language Model-Driven Iterative Multi-Agent Electronic Health Record Data Analysis Workflow. In *KDD-AIDSH*.
[11] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2024. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversations. In *COLM*.
[12] Zhenbang Wu, Anant Dadu, Michael Nalls, Faraz Faghri, and Jimeng Sun. 2024. Instruction Tuning Large Language Models to Understand Electronic Health Records. In *NeurIPS (Datasets and Benchmarks)*.
[13] Ran Xu, Hui Liu, Sreyashi Nag, Zhenwei Dai, Yaochen Xie, Xianfeng Tang, Chen Luo, Yang Li, Joyce C Ho, Carl Yang, et al. 2024. SimRAG: Self-Improving Retrieval-Augmented Generation for Adapting Large Language Models to Specialized Domains. *arXiv preprint arXiv:2410.17952* (2024).
[14] Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Bowen Jin, May Dongmei Wang, Joyce Ho, and Carl Yang. 2024. RAM-EHR: Retrieval Augmentation Meets Clinical Predictions on Electronic Health Records. In *ACL*. 754–765.
[15] Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Yanqiao Zhu, May Dongmei Wang, Joyce C. Ho, Chao Zhang, and Carl Yang. 2024. BMRetriever: Tuning Large Language Models as Better Biomedical Text Retrievers. In *EMNLP*. 22234–22254.
[16] Muchao Ye, Suhan Cui, Yaqing Wang, Junyu Luo, Cao Xiao, and Fenglong Ma. 2021. Medpath: Augmenting health risk prediction via medical knowledge paths. In *WWW*. 1397–1409.
[17] Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. RankRAG: Unifying Context Ranking with Retrieval-Augmented Generation in LLMs. In *NeurIPS*.