LLMs-based Few-Shot Disease Predictions using EHR: A Novel Approach Combining Predictive Agent Reasoning and Critical Agent Instruction

Hejie Cui[1,6], Zhuocheng Shen[1], Jieyu Zhang[2], Hui Shao, MD, PhD[4,5], Lianhui Qin, PhD[3], Joyce C. Ho, PhD[1], Carl Yang, PhD[1,4]

[1] **Department of Computer Science, Emory University, Atlanta, GA, USA**

[2] **School of Computer Science & Engineering, University of Washington, Seattle, WA, USA**

[3] **Department of Computer Science & Engineering, UCSD, San Diego, CA, USA**

[4] **Rollins School of Public Health, Emory University, Atlanta, GA, USA**

[5] **School of Medicine, Emory University, Atlanta, GA, USA**

[6] **Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA**

Abstract

Electronic health records (EHRs) contain valuable patient data for health-related prediction tasks, such as disease prediction. Traditional approaches rely on supervised learning methods that require large labeled datasets, which can be expensive and challenging to obtain. In this study, we investigate the feasibility of applying Large Language Models (LLMs) to convert structured patient visit data (e.g., diagnoses, labs, prescriptions) into natural language narratives. We evaluate the zero-shot and few-shot performance of LLMs using various EHR-prediction-oriented prompting strategies. Furthermore, we propose a novel approach that utilizes LLM agents with different roles: a predictor agent that makes predictions and generates reasoning processes and a critic agent that analyzes incorrect predictions and provides guidance for improving the reasoning of the predictor agent. Our results demonstrate that with the proposed approach, LLMs can achieve decent few-shot performance compared to traditional supervised learning methods in EHR-based disease predictions, suggesting its potential for health-oriented applications.

Introduction

Large Language Models (LLMs) have emerged as a powerful tool in various domains, including healthcare. These models, such as GPT family[1] and PaLM,[2] are trained on vast amounts of text data, allowing them to encode extensive knowledge across multiple fields. In the medical domain, the ability of LLMs to leverage their encoded medical knowledge has been showcased in recent studies,[3,4,5] with impressive performance on tasks such as medical question answering,[6] clinical text summarization,[7] and clinical decision support.[8] Certain very large language models demonstrate an emerging ability for few-shot learning, where the model can draw upon their existing understanding to quickly adapt to new tasks with limited examples.[9] This raises the question of whether LLMs can be directly applied to perform few-shot disease predictions using Electronic Health Record (EHR) data.

EHRs contain a wealth of patient data for predictive modeling tasks such as disease prediction, readmission risk assessment, and mortality prediction.[10] Existing approaches to EHR-based prediction primarily rely on supervised learning methods, including traditional machine learning models, representation learning,[11,12,13] and graph-based models.[14,15] While effective, these supervised approaches require training on large labeled datasets, which can be computationally expensive and challenging to obtain due to the high cost and difficulty of acquiring high-quality labeled EHR data.[16] In contrast, the capacity for few-shot learning enables LLMs to adapt to new tasks with minimal data, without any finetuning.[9] This adaptability raises the possibility of employing LLMs for few-shot disease prediction using EHR, a step forward in making healthcare more precise and efficient.[17]

In this study, we investigate the efficacy of LLMs-based few-shot disease prediction using the EHRs generated from clinical encounters that include three types of medical codes: disease, medications, and procedures. We convert the structured patient visit records into unstructured language narratives by mapping the ICD codes to their names and connecting them with proper conjunctives. This conversion process allows LLMs to better understand clinical records and retrieve related internal knowledge. We assess the zero-shot and few-shot diagnostic performance of LLMs using various prompting strategies, such as considering factor interactions and providing prevalence statistics and exemplars. The results of this evaluation provide insights into the potential of LLMs as a tool for EHR-based disease prediction and highlight the influence of prompting strategies on their performance.

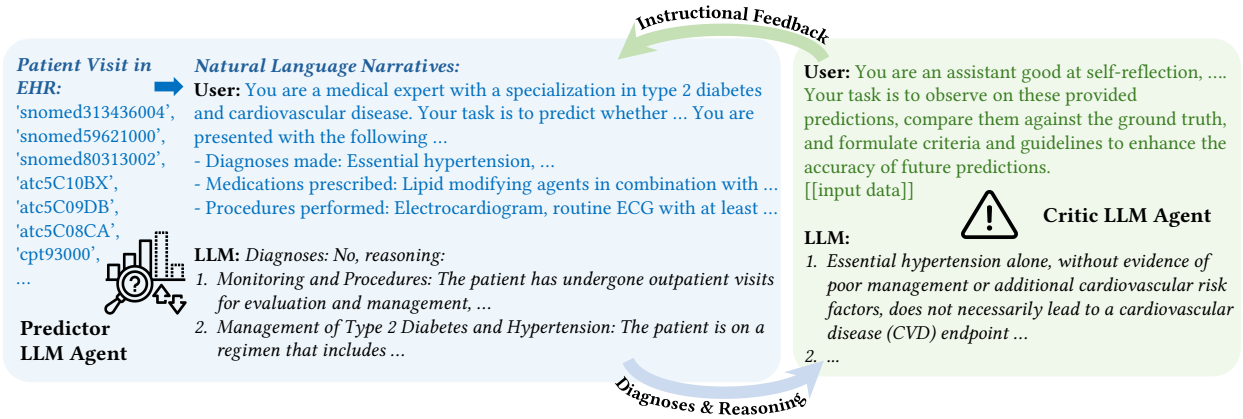

Figure 1: The framework of EHR-CoAgent employs two LLM agents: a predictor agent that makes predictions and generates reasoning processes and a critic agent that analyzes incorrect predictions and provides guidance for improvement. The critic agent's feedback is used to update the prompts given to the predictor agent, enabling the system to learn from its mistakes and adapt to the specific challenges of the EHR-based disease prediction task.

Building upon the findings of our initial evaluation, we propose an innovative approach to further improve the few-shot diagnostic performance of LLMs on EHR data. Studies have shown the promise of specialized LLM agents working collaboratively,[18] leveraging their diverse functionalities through few-shot learning. Our approach combines the strengths of predictive agent reasoning and critical agent instruction to create a more robust and accurate prediction system. The overall framework is shown in Figure 1. Specifically, we employ two LLM agents with different roles: a *predictor agent* and a *critic agent*. The *predictor agent* makes few-shot predictions given the unstructured narratives, which are converted from structured records, and generates a reasoning process to support its predictions. The *critic agent* then takes the predictor's output alongside the ground-truth disease labels as input and identifies issues or biases in the predictor agent's reasoning process. Based on the analysis, the critic agent generates a set of instructions that draw the predictor agent's attention to potentially overlooked factors and offer specific recommendations for refining its reasoning process. These instructions are subsequently appended to the prompts used for the predictor agent, serving to inform its predictions. Our results show that by refining the prompts based on the critic agent's feedback, the overall accuracy of the LLM-based few-shot prediction system improves significantly. This approach leverages the complementary strengths of predictive reasoning and critical analysis, enabling the system to learn from its mistakes and adapt to the specific challenges of EHR-based disease prediction. In summary, our main contributions are:

- We investigate the application of LLMs to EHR-based disease prediction tasks by converting structured data into natural language narratives and evaluating zero-shot and few-shot performance using various prompting strategies.
- We propose a novel approach combining two LLM agents with different roles: a predictor agent that makes predictions and provides reasoning processes, and a critic agent that analyzes incorrect predictions and provides feedback for improvement. The critic agent's feedback is used to update the predictor agent's prompts, enabling the system to learn from its mistakes and adapt to EHR-based disease prediction challenges.
- We summarize a set of insights into the performance of LLMs under various settings and share practical guidance on leveraging LLMs for diagnostic tasks with limited labeled data. We hope this can contribute to developing efficient and effective clinical decision support systems in the era of LLMs.

Related Work

Large Language Models for Healthcare LLMs have demonstrated remarkable capabilities in various application scenarios. Recently, there has been a growing interest in applying LLMs to the medical domain,[19,20,21,22] particularly for tasks such as clinical note analysis,[23,24] medical question answering,[25] disease prediction,[26,27] clinical trial matching,[28] medical report generation.[29] For example, Yang et al.[30] introduced GatorTron, an LLM specifically designed for EHRs. They demonstrated the effectiveness of GatorTron in various clinical natural language processing (NLP) tasks, such as named entity recognition and relation extraction, showcasing the potential of LLMs to extract valuable information from unstructured EHR data. Peng et al.[21] investigated the use of generative LLMs for medical research and

healthcare. They explored the capabilities of LLMs in tasks such as medical question answering, disease prediction, and clinical trial matching, highlighting their potential to support clinical decision-making and assist research.

However, applying LLMs to EHR-based disease prediction tasks remains under-explored. While some studies have investigated the use of LLMs for clinical NLP tasks on EHR,[30,31] there is still a lack of research on leveraging the reasoning and instruction-following capabilities of LLMs for few-shot EHR-based prediction. Our research addresses this gap by exploring the use of LLMs for EHR-based disease prediction and proposes new methods to enable accurate prediction with minimal training data.

Method

In this study, we expand our investigations on two levels: (1) evaluating the zero-shot and few-shot performance of LLMs on EHR-based disease prediction tasks, and (2) proposing a novel approach that leverages collaborative LLM agents to enhance the predictive performance.

LLM Performance on Disease Prediction with EHR The structured patient visit data are typically stored in tabular formats, where each row represents an individual patient visit record generated from clinical encounters, and columns correspond to different medical codes. In this study, we utilize EHR data that includes three types of medical codes $\mathcal{C}$: (1) diseases $\mathcal{C}_D$, (2) medications $\mathcal{C}_M$, and (3) procedures $\mathcal{C}_P$. Each patient visit sample v_i in the record $\mathcal{V}$ is represented by a set of medical codes $\{c_1, c_2, \ldots, c_n\}$, where $c_j \in \mathcal{C}$. We convert the structured EHR records into unstructured language narratives, denoted as $\mathcal{H}$, by mapping the medical codes to their names to enable the application of LLMs.

◇ **Zero-Shot: Leveraging Pre-existing Knowledge** Prompt engineering has emerged as a powerful technique for guiding the behavior of LLMs and improving their performance on various healthcare-related tasks, such as clinical named entity recognition[32] and clinical text classification.[33,34] We develop a set of prompting strategies tailored to EHR-based prediction tasks to provide additional context and guide the reasoning process of LLMs, including:

- Chain-of-thought (CoT) reasoning:[35] prompt the LLMs to generate step-by-step explanations;
- Incorporation of factor interactions: encourage LLMs to consider the interactions and dependencies among different medical factors (e.g., diseases, medications, and procedures);
- Prevalence information: integrate information about the prevalence statistics to provide additional context.

◇ **Few-Shot: Enhancing Performance with Limited Examples** We randomly select a small number of positive and negative samples (e.g., 3 positive and 3 negative) from the training data to serve as exemplars for each prediction category. These exemplars are incorporated into the prompts to provide the LLMs with a limited set of task-specific examples to learn from. This leverages the LLMs' vast pre-existing knowledge while allowing them to adapt quickly to the specific characteristics of the EHR prediction task. By this, we aim to guide LLMs' attention toward the most relevant patterns associated with each prediction category.

EHR-CoAgent: Collaborative LLM Agents for Enhanced Prediction Recently, the potential of LLMs has extended beyond single-agent applications. By leveraging the power of multiple LLMs with different roles working together in a collaborative framework, new possibilities have been unlocked for tackling complex problems and enhancing the performance of language models.[?] In this study, we propose a novel approach called EHR-CoAgent (as demonstrated in Figure 1), which harnesses the potential of collaborative LLM agents for enhanced prediction of EHR. Our framework consists of two components: a predictor agent $\mathcal{P}_{\text{LLM}}$ and a critic agent $\mathcal{K}_{\text{LLM}}$. The predictor agent focuses on generating predictions and providing explanatory reasoning, while the critic agent observes the predictor's outputs and provides instructional feedback to refine the prediction process. By integrating the feedback from the critic agent into the prompts used by the predictor agent, we aim to create an in-context learning process with feedback to continuously enhance disease prediction accuracy.

◇ **Predictor Agent: Generating Predictions and Reasoning** The predictor agent $\mathcal{P}_{\text{LLM}}$ is an LLM that performs few-shot disease predictions and provides explanatory reasoning based on the input EHR data. Given a patient's medical history $\mathcal{H}_i$, the predictor LLM analyzes the relevant information and generates the most likely prediction $\widehat{\mathcal{D}_i}$ and provides a step-by-step explanation of its reasoning process $\mathcal{R}_i$. Such explanatory reasoning is crucial for enhancing the interpretability of the generated predictions. By highlighting the key factors and evidence influencing the LLM agent's decision-making process, the reasoning serves as a transparent and informative basis for further analysis and

validation. The detailed prompt we used for the predictor agent in EHR-CoAgent is shown in Figure 3.

◇ **Critic Agent: Providing Instructional Feedback** The critic agent $\mathcal{K}_{\text{agent}}$ is another LLM that plays a different role in the EHR-CoAgent framework by observing a set of sampled wrong predictions from the predictor agent. Each set, denoted as $\mathcal{B}_j = \{(\widehat{\mathcal{D}}_{ji}, \mathcal{R}_{ji})\}_{i=1}^{b}$, contains generated prediction $\widehat{\mathcal{D}}_{ji}$ and the corresponding explanatory reasoning $\mathcal{R}_{ji}$ for b instances. The critic agent analyzes the inconsistency of the generated prediction and reasoning to the ground truth label $\mathcal{D}_{ji}$ for each batch $\mathcal{B}_j$ and identifies error patterns for improvement. Based on this analysis, we let the critic agent generate a set of instructional feedback $\{\mathcal{F}_j\}$ for batch $\mathcal{B}_j$ and repeat this process for m times. The detailed prompt we used for the critic agent in EHR-CoAgent is shown in Figure 4.

To provide concise and coherent guidance, we employ GPT-4 to process the set of instructional feedback $\{\mathcal{F}_j\}_{j=1}^{m}$. GPT-4 analyzes the feedback across multiple batches and generates a consolidated set of instructions $\mathcal{F}_{\text{consolidated}}$ that captures the most important and recurring insights. This consolidated feedback highlights common biases or errors in the reasoning process, offers suggestions for considering additional factors, and provides insights into the relationships between different medical concepts.

◇ **Instruction Enhancement: Integrating Critic Agent Feedback to the Predictor Agent** To effectively incorporate the feedback generated by the critic LLM, we introduce an instruction enhancement mechanism by integrating the critic LLM's instructional feedback $\mathcal{F}_{\text{consolidated}}$ directly into the prompts $\mathcal{P}$ used by the predictor LLM to guide the prediction. By augmenting the prompts with feedback instructions and guidance, we aim to steer the predictor LLM's attention toward the most relevant aspects of the input data and encourage it to consider the insights provided by the critic LLM. This iterative process of making predictions, receiving feedback, and refining the prompts allows the predictor LLM to continuously improve its performance and adapt to the specific challenges of EHR-based disease prediction.

Experimental Settings

Datasets We conducted experiments on two datasets: the publicly accessible MIMIC-III dataset and the privately-owned CRADLE dataset. **MIMIC-III**[36] is a large, publicly accessible dataset comprising de-identified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. Our task is to predict whether acute care conditions will be present during a patient's next visit, given their current ICU stay records. We focus on a specific chronic phenotype, Disorders of Lipid Metabolism, which is identified using Clinical Classifications Software (CCS) from the Healthcare Cost and Utilization Project (HCUP)[1]. During preprocessing, we extract patients with more than one hospital visit and create pairs of adjacent visits for each patient. For each pair, the former visit serves as the input, and the phenotypes in the latter visit are used as labels. This process yields 12,353 records with labels. For budget consideration, we randomly sample 1,000 records based on the data distribution of the prediction target as our testing set.

Project **CRADLE** (Emory Clinical Research Analytics Data Lake Environment) is a privately-owned database that contains de-identified electronic health records at Emory Healthcare from 2013 to 2017. In this study, we focus on the patients with type 2 diabetes and predict whether those patients will experience **cardiovascular disease** (CVD) endpoints within a year after the initial diabetes diagnosis. The CVD endpoints include coronary heart disease (CHD), congestive heart failure (CHF), myocardial infarction (MI), or stroke, which are identified by their ICD-9 and ICD-10 clinical codes. For patients who developed CVD complications within a year (positive cases), we select the earliest recorded encounter within a year of the CVD endpoint presence as the input. For patients without CVD complications (negative cases), we randomly select one encounter as the input from all encounters that occurred at least one year before the last recorded encounter. Patients are excluded if they (1) have less than two encounters at Emory Healthcare, (2) the time interval between their first and last encounter is less than one year, or (3) have a history of CVD conditions. After applying these exclusion criteria, 35,404 patients remain in the dataset. Similar to MIMIC-III, we randomly sample 1,000 records based on the data distribution of the prediction target

Evaluation Metrics Both the MIMIC-III and CRADLE datasets exhibit class imbalance, with the prevalence of Disorders of Lipid Metabolism in MIMIC-III being 27.6% and the prevalence of cardiovascular disease (CVD) endpoints in CRADLE being 21.4%. To account for the imbalanced data distributions, we employ accuracy, sensitivity, specificity, and F1 score as evaluation metrics.[14] When evaluating LLM methods, we identify the presence of "Yes" or "No" to-

[1] https://www.hcup-us.ahrq.gov/toolssoftware/ccs/AppendixASingleDX.txt

kens in the LLM responses and extract the top 5 probabilities associated with the predicting token. These probabilities are then normalized over both answers. We observed that GPT family models tend to provide highly confident answers (a confirmed prediction of either "Yes" or "No", with almost 0.0 probability for the other choice), often resulting in a majority probability of either 0.0 or 1.0.

Baselines We compare the performance of EHR-CoAgent with traditional machine learning (ML), including Decision Trees, Logistic Regression, and Random Forests, which are widely used in EHR-based prediction tasks,[37] and single-agent LLM approaches using GPT-4 (`gpt-4-0125-preview`) and GPT-3.5 (`gpt-35-turbo-16k-0613`). The ML models are trained in both fully supervised and few-shot settings, while the LLM approaches are evaluated in pure zero-shot, zero-shot with additional prompt information as mentioned in section , and few-shot learning settings. By comparing EHR-CoAgent with these baselines, we aim to evaluate the effectiveness of diverse LLM agent frameworks in EHR-based disease prediction tasks.

Implementation Details We implemented the empirical study methods in Python. The baseline machine learning models were trained and evaluated using the popular sklearn package, which provides a comprehensive set of tools for machine learning tasks. To access the various GPT models securely, we utilized the Azure OpenAI Service, a trusted and compliant cloud platform. Azure OpenAI offers a secure API interface that allows seamless integration of the GPT capabilities into our research pipeline while maintaining strict privacy and security controls. By leveraging Azure OpenAI, we ensured that the sensitive patient dataset was processed in a protected environment, adhering to necessary regulations and standards, such as HIPAA and GDPR.

Experimental Results

Table 1: Performance (%) of different models under the zero-shot, few-shot, and fully-supervised settings on MIMIC-III and CRADLE datasets. The proposed method is colored in green. The reference results under the supervised training setting (trained on 11,353 samples for MIMIC-III and 34,404 samples for CRADLE) are colored in gray.

Model	Approach	MIMIC-III (Pos : Neg = 27.6% : 72.4%)				CRADLE (Pos : Neg = 21.4% : 78.6%)			
		ACC	Sensitivity	Specificity	F1	ACC	Sensitivity	Specificity	F1
Decision Tree	Fully-Supervised	81.30	76.97	84.31	76.20	80.30	53.87	88.27	52.15
Logistic Regression	Fully-Supervised	79.70	70.48	83.56	73.18	80.90	58.34	86.15	59.74
Random Forest	Fully-Supervised	78.60	66.12	83.16	70.58	80.20	56.49	86.14	57.34
Decision Tree	Few-Shot (N=6)	71.10	53.14	77.62	51.16	31.90	54.81	25.99	31.71
Logistic Regression	Few-Shot (N=6)	58.70	73.40	53.44	56.78	53.30	53.95	53.13	48.16
Random Forest	Few-Shot (N=6)	69.70	62.88	72.18	63.61	65.00	51.50	68.43	51.04
GPT-4	Zero-Shot	51.90	76.15	42.56	51.89	24.10	51.81	16.82	22.33
	Zero-Shot+	62.90	59.30	64.29	58.58	30.00	53.25	23.76	29.67
	Few-Shot (N=6)	65.70	79.35	59.89	64.72	41.20	59.05	36.33	40.88
	EHR-CoAgent	79.10	73.11	81.43	73.88	70.00	62.88	71.72	60.21
GPT-3.5	Zero-Shot	78.00	66.87	82.37	68.56	56.50	59.88	55.45	52.29
	Zero-Shot+	72.40	50.00	80.37	42.00	62.60	57.62	63.96	54.40
	Few-Shot (N=6)	76.30	63.73	80.93	63.84	40.80	54.56	36.96	40.32
	EHR-CoAgent	79.30	74.49	80.98	71.59	66.60	58.31	68.83	55.83

Table 1 presents the experimental results on the two datasets. The findings highlight several key observations:

◇ Traditional machine learning models achieve respectable performance when trained on full datasets (11,353 for MIMIC-III and 34,404 for CRADLE). However, the performance of simpler models, e.g., Decision Trees and Logistic Regression, substantially deteriorates in the few-shot setting, emphasizing limitations when labeled data is scarce.

◇ When comparing the performance of zero-shot or few-shot LLMs with ML methods under few-shot settings, we observe that LLMs exhibit higher sensitivity but lower specificity. This finding suggests that LLMs excel at correctly identifying positive cases (i.e., patients with the condition of interest) but at the cost of a higher false positive rate. In other words, LLMs are more prone to classifying a patient as having the condition, even when they do not. This tendency implies that LLMs, particularly GPT-4, adopt a more conservative mindset, possibly due to their alignment to err on the side of caution to mitigate the risk of potentially missing true positive cases.

◇ Zero-shot with additional prompting strategies (Zero-Shot+) can improve based on pure zero-shot, with occasionally

Instructional Feedback Examples from the Critic LLM Agent for the MIMIC dataset (GPT-4)
The diagnosis of conditions directly related to lipid metabolism, such as "Disorders of lipoid metabolism," in the patient's medical history, requires ongoing management and monitoring rather than assumptions of worsening or new diagnoses without recent lipid profile assessments.
Risk factors and underlying conditions: Consider the presence of risk factors and underlying conditions that are commonly associated with disorders of lipid metabolism disease, such as diabetes mellitus, obesity, hypertension, and cardiovascular diseases.
The performance of procedures related to cardiovascular health, such as hemodialysis or cardiac catheterization, without direct evidence of unmanaged lipid metabolism issues, should not be solely used to predict future disorders of lipid metabolism disease.
Pharmacological interventions consideration: Incorporate an evaluation of prescribed drugs, focusing on their relevance to managing the risk factors of the disorders of lipoid metabolism.
Instructional Feedback Examples from the Critic LLM Agent for the CRADLE dataset (GPT-4)
Avoid bias towards predicting a positive CVD endpoint based on conservative thinking when the patient is actively monitored and managed for known risk factors. Evaluate the effectiveness of the interventions in place.
The presence of type 2 diabetes mellitus without complication does not necessarily lead to a cardiovascular disease (CVD) endpoint within a year of the initial diagnosis.
The presence of symptoms such as chest pain, dyspnea, and edema, especially when combined with diagnoses like hypertension and hyperlipidemia, increases the likelihood of developing a cardiovascular disease (CVD) endpoint within a year of the initial diagnosis.
Essential hypertension alone, without evidence of poor management or additional cardiovascular risk factors, does not necessarily lead to a cardiovascular disease (CVD) endpoint within a year of the initial diagnosis.

Figure 2: Examples of instructional feedback generated by the GPT-4-based critic agent, which aims to refine the predictor agent's reasoning process and improve the accuracy of its prediction.

produced errors. This observation underscores the importance of carefully crafting prompts to optimize the performance of LLMs in EHR-based disease prediction tasks.

◇ Most of the time, adding few-shot demonstrations enhance prediction performance compared to their respective Zero-Shot+ counterparts. This finding emphasizes providing even a limited number of labeled examples can potentially steer language models toward more precise predictions. By leveraging a small set of representative samples, LLMs can quickly adapt to the specific characteristics of the EHR-based disease prediction task.

◇ Our proposed approach EHR-CoAgent demonstrates remarkable performance, surpassing other methods and even fully supervised ML models in certain scenarios, with GPT-4 generally outperforming GPT-3.5. On the CRADLE dataset, EHR-CoAgent achieves an F1 score of 60.21%, outperforming all fully trained ML models. Similarly, on the MIMIC-III dataset, EHR-CoAgent obtains an F1 score of 73.88%, comparable to the fully trained Decision Tree model and superior to Logistic Regression and Random Forest.

◇ Compared with the few-shot setting with a single LLM predictor, EHR-CoAgent improves significantly on all four metrics. This can be attributed to the feedback instructions provided by the critic agent, which analyzes the outputs and identifies issues and biases in LLM's reasoning process, such as overly relying on conservative thinking or neglecting certain key factors. The feedback instructions generated by the critic agent help to correct these issues, dynamically refining the predictor agent's reasoning process, thus improving the accuracy of the prediction.

Generated Instructions

Figure 2 showcases examples of the criteria and instructions generated by the critic agent. These examples demonstrate the critic agent's ability to identify potential issues in the predictor agent's prediction and reasoning process and provide targeted instructions to address them. For instance, the first instruction for the CRADLE dataset, "Avoid bias towards predicting a positive CVD endpoint based on conservative thinking when the patient is actively monitored and managed for known risk factors. Evaluate the effectiveness of the interventions in place" highlights a possible prediction bias of the predictor agent. This instruction encourages the predictor agent to avoid relying on conservative assumptions when making predictions, as such assumptions may be a result of the over-alignment of advanced AI models. By explicitly addressing this issue, the critic agent aims to guide the predictor agent toward more objective and comprehensive reasoning. Another example for the MIMIC dataset, "Pharmacological Interventions Consideration: Incorporate an evaluation of prescribed drugs, focusing on their relevance to managing the risk factors of the disorders of lipoid metabolism" suggests that the predictor agent should take into account the role of prescribed medications in managing the patient's condition. By analyzing the relevance and potential impact of these drugs on the risk factors associated with disorders of lipoid metabolism, the predictor agent can make more informed predictions. These examples illustrate how the critic agent's feedback can guide the predictor agent towards more comprehensive and

nuanced reasoning, ultimately leading to improved disease prediction performance.

Conclusions

In this study, we investigated the application of Large Language Models (LLMs) to Electronic Health Record (EHR) based disease prediction tasks. We evaluated the zero-shot and few-shot diagnostic performance of LLMs using various prompting strategies and proposed a novel collaborative approach combining a predictor agent and a critic agent. This approach enables the system to learn from its mistakes and adapt to the challenges of EHR-based disease prediction. Our work highlights the potential of LLMs as a tool for clinical decision support and contributes to the development of efficient disease prediction systems that can operate with minimal training data.

Ethical Considerations

To ensure the ethical use of credential data with GPT-based services, we have signed and strictly adhered to the PhysioNet Credentialed Data Use Agreement[2]. We follow the guidelines[3] for responsible use of MIMIC data in online services, including opting out of human review of the data through the Azure OpenAI Additional Use Case Form[4], to prevent sensitive information from being shared with third parties.

Acknowledgements

This research was supported by multiple sources. We gratefully acknowledge the support of the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health under Award Number K25DK135913. Additional support was provided by the Emory Global Diabetes Center of the Woodruff Sciences Center, Emory University. We also benefited from the Microsoft Accelerating Foundation Models Research (AFMR) grant program. The contributions of these institutions were instrumental in making this work possible.

References

1. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. Gpt-4 technical report. arXiv preprint arXiv:230308774. 2023.
2. Anil R, Dai AM, Firat O, Johnson M, Lepikhin D, Passos A, et al. Palm 2 technical report. arXiv preprint arXiv:230510403. 2023.
3. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620:172-80.
4. Hernandez E, Mahajan D, Wulff J, Smith MJ, Ziegler Z, Nadler D, et al. Do We Still Need Clinical Language Models? In: Conference on Health, Inference, and Learning; 2023. .
5. Cui H, Lu J, Wang S, Xu R, Ma W, Yu S, et al. A Review on Knowledge Graphs for Healthcare: Resources, Applications, and Promises. arXiv preprint arXiv:230604802. 2023.
6. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. arXiv preprint arXiv:230509617. 2023.
7. Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nature Medicine. 2024:1-9.
8. Hegselmann S, Buendia A, Lang H, Agrawal M, Jiang X, Sontag D. Tabllm: Few-shot classification of tabular data with large language models. In: AISTATS; 2023. .
9. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. NeurIPS. 2020.
10. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE journal of biomedical and health informatics. 2017;22:1589-604.
11. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ digital medicine. 2018;1:1-10.
12. Landi I, Glicksberg BS, Lee HC, Cherng S, Landi G, Danieletto M, et al. Deep representation learning of electronic health records to unlock patient stratification at scale. NPJ digital medicine. 2020;3:96.
13. Fridgeirsson EA, Sontag D, Rijnbeek P. Attention-based neural networks for clinical prediction modelling on

[2]https://physionet.org/about/licenses/physionet-credentialed-health-data-license-150

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

[4]https://aka.ms/oai/additionalusecase

electronic health records. BMC Medical Research Methodology:285.
14. Choi E, Xu Z, Li Y, Dusenberry M, Flores G, Xue E, et al. Learning the graphical structure of electronic health records with graph convolutional transformer. In: AAAI; 2020. .
15. Zhang Z, Cui H, Xu R, Xie Y, Ho JC, Yang C. TACCO: Task-guided Co-clustering of Clinical Concepts and Patient Visits for Disease Subtyping based on EHR Data. In: KDD; 2024. .
16. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. Journal of the American Medical Informatics Association. 2018;25:1419-28.
17. Wornow M, Thapa R, Steinberg E, Fries J, Shah N. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. NeurIPS. 2023.
18. Jin Q, Yang Y, Chen Q, Lu Z. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. Bioinformatics. 2024;40:btae075.
19. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nature medicine. 2023;29(8):1930-40.
20. He K, Mao R, Lin Q, Ruan Y, Lan X, Feng M, et al.. A Survey of Large Language Models for Healthcare: from Data, Technology, and Applications to Accountability and Ethics; 2023.
21. Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, et al. A study of generative large language model for medical research and healthcare. npj Digital Medicine. 2023;6:210.
22. Ling C, Zhao X, Lu J, Deng C, Zheng C, Wang J, et al. Domain specialization as the key to make large language models disruptive: A comprehensive survey. arXiv preprint arXiv:230518703. 2023.
23. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. In: EMNLP; 2022. .
24. Mannhardt N, Bondi-Kelly E, Lam B, O'Connell C, Asiedu M, Mozannar H, et al.. Impact of Large Language Model Assistance on Patients Reading Clinical Notes: A Mixed-Methods Study; 2024.
25. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? Patterns. 2024;5:100943.
26. Wang G, Yang G, Du Z, Fan L, Li X. ClinicalGPT: large language models finetuned with diverse medical data and comprehensive evaluation. arXiv preprint arXiv:230609968. 2023.
27. Cui H, Fang X, Xu R, Kan X, Ho JC, Yang C. Multimodal fusion of ehr in structures and semantics: Integrating clinical records and notes with hypergraph and llm. arXiv preprint arXiv:240308818. 2024.
28. Yuan J, Tang R, Jiang X, Hu X. Large language models for healthcare data augmentation: An example on patient-trial matching. In: AMIA Annual Symposium Proceedings. vol. 2023; 2023. p. 1324.
29. D'Antonoli TA, Stanzione A, Bluethgen C, Vernuccio F, Ugga L, Klontzas ME, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. Diagnostic and Interventional Radiology. 2024;30:80.
30. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. npj Digital Medicine:194.
31. Xu R, Cui H, Yu Y, Kan X, Shi W, Zhuang Y, et al. Knowledge-Infused Prompting: Assessing and Advancing Clinical Text Data Generation with Large Language Models. In: Findings of the Association for Computational Linguistics ACL 2024; 2024. .
32. Hu Y, Chen Q, Du J, Peng X, Keloth VK, Zuo X, et al. Improving large language models for clinical named entity recognition via prompt engineering. Journal of the American Medical Informatics Association. 2024:ocad259.
33. Lu Y, Zhao X, Wang J. Medical knowledge-enhanced prompt learning for diagnosis classification from clinical text. In: Clinical Natural Language Processing Workshop; 2023. p. 278-88.
34. Sivarajkumar S, Wang Y. Healthprompt: A zero-shot learning paradigm for clinical natural language processing. In: AMIA Annual Symposium Proceedings. vol. 2022; 2022. p. 972.
35. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. NeurIPS. 2022;35.
36. Johnson AE, Pollard TJ, Shen L, Lehman LwH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific data. 2016;3:1-9.
37. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. JAMIA. 2017;24:198.

Prompt for the Predictor LLM Agent	
role: system	**content:** You are a medical expert with a specialization in type 2 diabetes and cardiovascular disease. Your task is to predict whether a patient with type 2 diabetes will develop a cardiovascular disease (CVD) endpoint within a year of their initial diagnosis.
role: user	**content:** Task: Your task is to predict whether a patient with type 2 diabetes will develop a cardiovascular disease (CVD) endpoint within a year of their initial diagnosis based on the provided patient's medical history. You will be presented with a patient's medical history and various resources to aid in your prediction. Please provide your reasoning and make your prediction by learning from the resources. You are presented with the following: 1. [CVD Endpoint Definition] The definition of the prediction target: cardiovascular disease (CVD) endpoint. 2. [Past Medical History] Patient's past medical history, which captures specific diagnoses made, medications prescribed, and procedures performed. 3. [Instructions] Guidelines on how to analyze the patient's medical history, provide reasoning, and make predictions. This includes referring to the demonstration cases and exploring the interaction of various factors and the interplay between diseases, medications, and procedures that the patient has undergone. The reasoning process should support and aid in the final prediction for a CVD endpoint. 4. [Demonstration Cases] Some real and typical cases, including the patient's medical history (diseases, medications, and procedures) and the ground truth result of whether the patients with type 2 diabetes experience cardiovascular disease (CVD) endpoint within a year after the initial diagnosis. 5. [Output Format] The required format for your response. Please ensure that you strictly adhere to the format requirements. You must provide a confirmed prediction by choosing between "Yes" or "No". [CVD Endpoint Definition] A CVD endpoint is identified by the presence of coronary heart disease (CHD), congestive heart failure (CHF), myocardial infarction (MI), or stroke. [Past Medical History] {info_generate(code_1_list)} [Instructions] Based on the patient's past medical history, provide reasoning on whether a patient with type 2 diabetes will develop a cardiovascular disease (CVD) endpoint within a year of their initial diagnosis. You should know that: globally, cardiovascular disease (CVD) affects about 32.2% of people with type 2 diabetes (T2D). CVD is the most common cause of morbidity and mortality in people with T2D. In this specific training set, 20% of the patients develop CVD within a year of their initial diagnosis. Please: (1) use your knowledge; (2) learn from the provided demonstration cases; (3) perform a comprehensive analysis of the patient's medical history; (4) consider the interplay of diseases, medications, and procedures, make sure to: Identify and weigh both risk factors and protective factors evident in the patient's medical history; Consider the presence of any comorbid conditions that may independently increase or decrease the risk of CVD; Examine the patient's medication profile to discern any pharmacological interventions that may alter the course of disease progression; Evaluate any medical or surgical procedures the patient has undergone that could impact their cardiovascular health; (5) Utilize a multihop and step-by-step reasoning approach to systematically analyze the data. [Demonstration Cases] {few_shots_generate(few_shots_label_0, few_shots_label_1)} [Output Format] Your final response should include: 1. Prediction: Conclude your analysis with a clear and concise prediction. This prediction must be a single word, either "Yes" or "No", indicating whether you believe the patient is likely to develop a CVD endpoint within a year of their initial diabetes diagnosis. This prediction should be the first line of your response, to facilitate easy parsing. 2. Reasoning: Provide a detailed reasoning process. Ensure that your analysis is thorough and based on the information provided, leading logically to your final prediction.

Figure 3: Prompt for Predictor Agent in EHR-CoAgent for the CRADLE dataset.

Prompt for the Critic LLM Agent	
role: system	**content:** You are an assistant who is good at self-reflection, gaining experience, and summarizing criteria. By reflecting on failure predictions that are given below, your task is to reflect on these incorrect predictions, compare them against the ground truth, and formulate criteria and guidelines to enhance the accuracy of future predictions.
role: user	**content:** Task: You will be given a batch of input data samples, where each sample is composed of three parts: the patient's medical history, the ground-truth result for whether the patient will develop a cardiovascular disease (CVD) endpoint within a year of their initial diagnosis of type 2 diabetes, and a wrong prediction. Please always remember that the predictions above are all incorrect. You should always use the ground truth as the final basis to discover many unreasonable aspects in the predictions and then summarize them into instructions and criteria. You are presented with the following: 1. [Input Data] A batch of input data samples. Each data in the batch includes three parts: (a) the patient's medical history, including diseases that the patient has been diagnosed with, medications that the patient has taken, and procedures the patient has undergone; (b) the ground-truth result for each patient's medical history on whether the patient will develop a cardiovascular disease (CVD) endpoint within a year of their initial diagnosis of type 2 diabetes; (c) a wrong prediction. 2. [Instructions] Instructions on suggesting criteria. [Input Data] {batch_generate(input_data_batch)} [Instructions] 1. Please always remember that the predictions above are all incorrect. You should always use the ground truth as the final basis to discover many unreasonable aspects in the predictions and then summarize them into experience and criteria. 2. Identify why the wrong predictions deviated from the ground truth by examining discrepancies in the medical history analysis. 3. Determine key and potential influencing factors, reasoning methods, and relevant feature combinations that could better align predictions with the ground truth. 4. The instructions should be listed in distinct rows, each representing a criteria or guideline. 5. The instructions should be generalizable to multiple samples, rather than specific to individual samples. 6. Conduct detailed analysis and write criteria based on the input samples, rather than writing some criteria without foundation. 7. Please note that the criteria you wrote should not include the word "ground truth".

Figure 4: Prompt for Critic Agent in EHR-CoAgent for the CRADLE dataset.