No Black Boxes: Interpretable and Interactable Predictive Healthcare with Knowledge-Enhanced Agentic Causal Discovery

Anonymous ACL submission

Abstract

Deep learning models trained on extensive Electronic Health Records (EHR) data have achieved high accuracy in diagnosis prediction, offering the potential to assist clinicians in decision-making and treatment planning. However, these models lack two crucial features that clinicians highly value: interpretability and interactivity. The "black-box" nature of these models makes it difficult for clinicians to understand the reasoning behind predictions, limiting their ability to make informed decisions. Additionally, the absence of interactive mechanisms prevents clinicians from incorporating their own knowledge and experience into the decision-making process. To address these limitations, we propose **II-KEA**, a knowledge-enhanced agent-driven causal discovery framework that integrates personalized knowledge databases and agentic LLMs. **II-KEA** enhances interpretability through explicit reasoning and causal analysis, while also improving interactivity by allowing clinicians to inject their knowledge and experience through customized knowledge bases and prompts. **II-KEA** is evaluated on both MIMIC-III and MIMIC-IV, demonstrating superior performance while offering enhanced interpretability and interactivity, as supported by the results of extensive case studies.

1 Introduction

Accurate diagnosis prediction is crucial for improving clinical outcomes by enabling timely interventions and optimizing treatment planning. In recent years, the growing availability of Electronic Health Records (EHR) (e.g., MIMIC datasets (Johnson et al., 2016, 2023)) has provided valuable real-world data, allowing researchers to develop more advanced and complex deep learning models to uncover predictive patterns from a data science perspective. These models often integrate domain knowledge of medical concepts to identify intricate correlations in disease progression and comorbidities, demonstrating promising predictive performance. Despite their success in prediction accuracy, these methods have two key limitations:

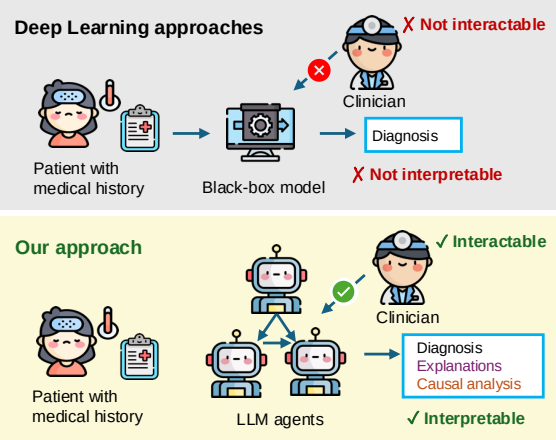

Figure 1: Comparison between deep learning approaches and our approach.

- **Lack of interpretability:** deep learning models inherently function as "black boxes," offering little transparency into the clinical reasoning behind their predictions.
- **Lack of interactivity:** most models operate in an end-to-end manner, limiting practitioner interaction with the system. This prevents users from asking follow-up questions, customizing prediction goals, or incorporating their own knowledge and experience to refine predictions.

The lack of both **interpretability** and **interactivity** undermines trust and acceptance among healthcare professionals who depend on these predictions for informed decision-making. In recent years, Large Language Models (LLMs) have demonstrated extensive knowledge, strong instruction-following capabilities, and impressive reasoning abilities, offering promising solutions to address these limitations. The development of agentic LLMs has further enhanced their flexibility and ca-

pabilities through dynamic interactions with the environment, tool utilization, and inter-agent collaboration. These advancements hold great potential for enabling clinicians to engage more effectively with predictive models, fostering greater adaptability and user-driven refinement.

Inspired by these advancements, we propose **II-KEA**, a knowledge-enhanced **A**gentic **C**ausal Discovery framework designed for **I**nterpretable and **I**nteractive **D**iagnosis Prediction. **II-KEA** is a multi-agent system comprising three LLM agents namely Knowledge Synthesis Agent, Casual Discovery Agent, and Decision-Making Agent collaboratively, and is powered by both clinical dataset and domain knowledge. Similar to other deep learning approaches, **II-KEA** predicts medical diagnoses by addressing the question:*"What diseases is a patient likely to be diagnosed with given their past diagnosis history?"* However, unlike purely data-driven methods, **II-KEA** approaches the problem from a causal perspective, delving deeper into the underlying mechanisms to answer:*"What diseases are likely to be caused by the conditions a patient has already been diagnosed with?"*—thus reframing the task as a causal discovery problem. Recent advances in Large Language Models (LLMs) have demonstrated promising performance in causal discovery, alleviating the need for complex, data-centric, and resource-intensive traditional methods. However, LLMs often generate incorrect answers when domain knowledge is insufficient. To address this limitation, we enhance the causal discovery process by integrating both knowledge-driven reasoning through Retrieval Augmented Generation (RAG) and data-grounded inference, ensuring a deeper contextual understanding and better alignment with real-world observations.

To this end, we emphasize that **II-KEA** is clinician-friendly framework that ensures both interpretability and interactivity.

- **II-KEA is interpretable.** The LLM agents make **II-KEA** inherently interpretable by enabling the decision-making agent to provide detailed explanations and reasoning behind its predictions. Additionally, **II-KEA** gains an extra layer of interpretability through causal analysis. The causal graph generated by the causal discovery agent offers an intuitive and comprehensive representation of the causal mechanisms between diseases, making it easier for users to understand the underlying relationships.
- **II-KEA is interactive.** Clinicians can interact with and customize the prediction process through two pathways. First, the RAG feature of **II-KEA** enables generalization by incorporating external knowledge, allowing clinicians to conveniently provide the knowledge sources their own or selected knowledge sources as the knowledge database. Second, clinicians can interact with the decision-making agent by specifying their personal preferences, ensuring that predictions are tailored to their specific needs.

We evaluate **II-KEA** on EHR datasets, including MIMIC-III and MIMIC-IV, demonstrating superior performance along with enhanced interpretability and interactivity, supported by extensive ablation and case studies.

2 Methodology

We propose, **II-KEA**, a multi-agent system consisting of three LLM-based agents working collaboratively and is powered by both clinical datasets and domain knowledge. **II-KEA** aims to uncover causal relationships between diseases and predict future medical diagnoses. In this section, we introduce each LLM agent and knowledge module and provide a summary of the overall framework.

2.1 Knowledge databases

2.1.1 Clinical datasets

We construct a clinical dataset using training data from Electronic Health Records (EHR). Each record contains diagnosis information for individual patients across multiple visits. This database comprises two data frames: a Disease Transition Probability Matrix and a Diagnosis Matrix.

Let $\mathcal{D}$ denote the complete set of diseases, and $\mathcal{P}_{train}$ denote the patient set from training data. The Disease Transition Probability Matrix, denoted as $\mathbf{A}_T \in \mathbb{R}^{|\mathcal{D}|\times|\mathcal{D}|}$, captures the probability of disease j occurring after disease i. The underlying intuition is that temporal precedence is a necessary but not sufficient condition for causality. Identifying diseases that frequently follow a target disease helps narrow down potential causal candidates. By pre-selecting frequently co-occurring diseases, we provide the LLM agent with a shortlist of candidates, reducing its workload when assessing causal relationships, as we will discuss later. In constructing this matrix, for each visit, we define disease B as a successor of disease A if:

- Disease B appears in a patient's next visit after the visit in which disease A is diagnosed.

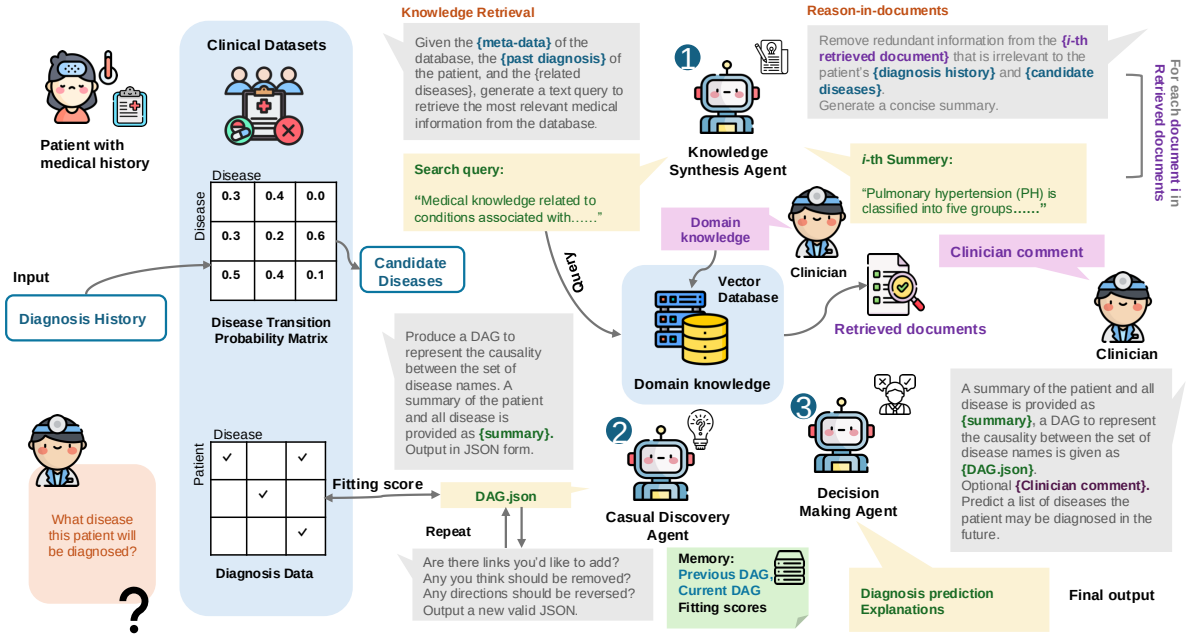

Figure 2: An overview of **II-KEA** framework. It consists of three LLM-based agents working collaboratively and is powered by both clinical datasets and domain knowledge. During inference, a patient's diagnosis history is processed to identify possible diseases. A Knowledge Synthesis Agent retrieves and summarizes relevant documents. Then, a Causal Discovery Agent uncovers causal relationships using both external knowledge and observational data, forming a causal graph. Finally, a Decision-Making Agent integrates all this information—along with optional clinician input—to predict the diagnosis and provide explanations.

• Disease B appears in the same visit as disease A. The second condition accounts for the fact that patients may not visit clinicians frequently, meaning that a succession disease and the target disease could be diagnosed simultaneously.

$$\mathbf{A}_T[a,b] = \frac{\mathcal{N}_{a,b}}{\sum_{p\in\mathcal{P}}\sum_{i=1}^{m_{p-1}}\mathbb{I}[a\in\mathcal{D}_p^i]}, \quad (1)$$

where

$$\mathcal{N}_{a,b} = \sum_{p\in\mathcal{P}_{\text{train}}}\sum_{i=1}^{m_{p-1}}\mathbb{I}[a\in\mathcal{D}_p^i \wedge (b\in\mathcal{D}_p^{i+1}\vee b\in\mathcal{D}_p^i)], \quad (2)$$

$\mathbb{I}[\cdot]$ is the indicator function. $\mathbf{A}_T[a,b]$ is an entry of $\mathbf{A}_T$, representing the transition probability between disease a and disease b. $\mathcal{P}_{train}$ denotes the set of all patients in the training set, and $\mathcal{D}_p^i$ represents the set of diseases diagnosed for patient p during their i-th visit. Note that $\mathbf{A}_T$ is not necessarily symmetric, meaning that $\mathbf{A}_T[a,b] \neq \mathbf{A}_T[b,a]$ in general. The diagnosis matrix $\mathbf{A}_D \in \mathcal{R}^{|\mathcal{P}_{train}|\times|\mathcal{D}|}$ records the occurrence of each disease for all patients. We consider the occurrence of disease a for a patient to be 1 if the patient is has been diagnosed with the disease in any revisits:

$$\mathbf{A}_D[p,a] = \mathbb{1}(a \in \bigcup_{i\in m_p}\mathcal{D}_p^i), \quad (3)$$

we calculate the fitting score between the diagnosis matrix and the output causal graph to provide feedback to the causal discovery agent, as we will discuss in section 2.2.

2.1.2 Domain knowledge database

We construct a vector database powered by ChromaDB[1] as the source of external knowledge for the Retrieval-Augmented Generation (RAG) of the knowledge synthesis agent, as discussed in Section 2.2. The database can contain any domain knowledge from different sources such as web pages, published papers, or clinical notes. In this paper, we scrape text from Wikipedia pages corresponding to each disease listed in ICD-9. Each Wikipedia page is segmented into sections such as "Overview", "Signs and Symptoms", "Causes", "Diagnosis", "Prevention", "Treatment", "Epidemiology", "History", "Terminology", and "Society and Culture". When creating the vector database, each section

[1]https://pypi.org/project/chromadb/

is treated as an individual document, and its key vector is generated by embedding the document text using the pre-trained Sentence-BERT model `all-MPNet-base-v2`, which exhibits exceptional performance in capturing semantic similarities between sentences.

2.2 Multi-agent Framework

The goal of **II-KEA** is to predict a patient's future diagnoses by conducting causal discovery on their diagnosis history and identifying diseases that are most likely caused by past conditions. However, directly asking an LLM agent to perform this task across thousands of diseases would be computationally expensive. Instead, we leverage a Disease Transition Probability Matrix, denoted as $\mathbf{A}_T$, to select candidate diseases, acknowledging that temporal precedence is a necessary condition for causality. For a patient p, let D^p denote the set of diseases they have been diagnosed with in the past. The set of candidate diseases S^p that could be caused by D^p is then obtained as:

$$S^p = \{b \mid \mathcal{M}[a, b] > \epsilon, \quad \forall a \in D_p\} \tag{4}$$

We then provide both the diagnosis history set D_p and the candidate disease set C_p to the agents to determine which diseases are causally linked. To ensure that the causal discovery process is grounded in sufficient domain knowledge, a straightforward approach would be to query a vector database separately for each disease in D_p and C_p and send the retrieved text to the causal discovery agent. However, this approach has two major drawbacks: 1) Independently querying each disease focuses on individual diseases rather than the relationships between them, failing to retrieve information most relevant to causal links. 2) The retrieved documents may contain redundant information, be excessively long, and exceed the processing capacity of LLMs. To address these issues, we develop a Knowledge Synthesis Agent.

Knowledge Synthesis Agent, $\mathcal{A}_{knowledge}$. The role of $\mathcal{A}_{knowledge}$ is to generate high-quality contextual information for the causal discovery process. Its generation process consists of two steps. In the first step, the agent is provided with the database metadata, the patient's diagnosis history D_p, and the candidate disease set C_p. It is responsible for generating a query text to retrieve relevant information from the database. This query text should effectively summarize D_p and C_p while being tailored to the specific database based on its metadata, which defines its characteristics and content. We then encode the query text using the same pre-trained Sentence-BERT model and retrieve the k most relevant documents. In the second step, A_{ks} performs *reasoning-in-documents*, refining the retrieved information by removing redundancies and generating a concise summary. These summarized documents are then stored for use by the causal discovery agent, enabling a *Retrieval-Augmented Generalization (RAG)* approach. We summarize the workflow of the $\mathcal{A}_{knowledge}$ in Algorithm 1.

Causal Discovery Agent, $\mathcal{A}_{causal}$. The role of $\mathcal{A}_{causal}$ is to identify potential causal relationships among a set of diseases. We provide it with the patient's diagnosis history set D_p and the candidate disease set C_p as a whole, along with the summarized external knowledge generated by $\mathcal{A}_{knowledge}$.

We then adapt the iterative causal discovery procedure proposed in (Abdulaal et al., 2024):

1. *Hypothesis generation.* Given the summarized external knowledge and the empty graph $\mathcal{G}^{\emptyset}$, which consists of all entities in D_p and C_p with no initial relations, the $\mathcal{A}_{causal}$ LLM generates an initial causal graph as a directed acyclic graph (DAG), $\mathcal{G}^s_{t=0}$.
2. *Model fitting.* At each iteration t, we fit the causal model using a data-driven approach with real-world observations. Specifically, we measure the log-likelihood, l_t, of the diagnosis matrix $\mathbf{A}_D$ under the current model $\mathcal{G}^s_t$.

$$l_t = \sum_{p \in \mathcal{P}_{train}} \sum_{a \in \mathcal{D}} \log P(X^a_p \mid \{X^b_p \mid b \in \mathbf{Pa}(a)\}) \tag{5}$$

   where $\mathbf{Pa}(a)$ denotes the parent diseases of a in $\mathcal{G}^s_t$, and $X^a_p \in \{0, 1\}$ represents the observation of disease a in patient p.
3. *Post-processing.* We update the memory M_t to store the causal graph and the fitting score from the previous and current time steps, including M_t, M_{t-1}, l_t, and l_{t-1}. This memory is retained for the next step.
4. *Hypothesis amendment.* The LLM refines the causal model based on the stored memory to enhance its accuracy and better capture causal relationships. It then outputs the updated causal graph as $\mathcal{G}^s_{t+1}$.

Steps 2 to 4 are repeated iteratively until a stopping criterion is met, (when the change in $\mathcal{G}^s_t$ falls

below a predefined threshold or the number of iterations exceeds a limit). We summarize the workflow of the Causal Discovery Agent in Algorithm 2.

Decision-Making Agent, $\mathcal{A}_{decision}$. $\mathcal{A}_{decision}$ integrates and evaluates all available information, including diagnosis history sets, summarized knowledge, and the causal graph, to make the final prediction on a patient's diagnosis. Additionally, clinicians or users can provide their preferences, comments, or experiences to customize the prediction. For example, they may indicate that they are particularly concerned about kidney-related diseases. The agent is then tasked with producing the diagnosis list in a structured format along with an explanation of the reasoning behind its decision. We summarize the workflow of the Decision-Making Agent in Algorithm 3.

2.3 II-KEA Inference

II-KEA does not involve any training process but requires data preprocessing. First, the EHR training dataset is processed to construct the Disease Transition Probability Matrix $\mathbf{A}_T$ and the Diagnosis Matrix $\mathbf{A}_D$, as described in Section 2.1.1. Additionally, a knowledge vector database Γ is prepared following Section 2.1.2. Both matrices and the database are stored for later inference. During inference, for each patient, we collect their diagnosis history $\mathcal{D}_p$ and apply the preprocessing steps outlined in Section 2.1.1 to determine the candidate disease set $\mathcal{S}_p$. The Knowledge Synthesis Agent $\mathcal{A}_{knowledge}$ then retrieves relevant documents from Γ and summarizes them into $\Gamma_p^{summary}$. Next, the Causal Discovery Agent $\mathcal{A}_{causal}$ iteratively uncovers causal relationships within the expanded disease set $\mathcal{D}_p \cup \mathcal{S}_p$, leveraging both external knowledge from $\Gamma_p^{summary}$ and observational data from $\mathbf{A}_D$. This process results in a causal graph $\mathcal{G}^s$. Finally, the Decision-Making Agent $\mathcal{A}_{decision}$ integrates all available information, including the diagnosis history $\mathcal{D}_p$, candidate diseases $\mathcal{S}_p$, summarized documents $\Gamma_p^{summary}$, causal graph $\mathcal{G}^s$, and an optional clinician-provided comment $\mathcal{C}$. Using this information, the model predicts the patient's diagnosis and provides explanations for the decision. We provide the overview of **II-KEA** in Figure 2, the prompt details in Appendix D. The workflow of its inference process is shown in Algorithm 4.

3 Experiments & Setup

3.1 Datasets

We utilize both the MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2023) datasets for our experiments. MIMIC-III contains 7,493 patients with multiple visits ($T \geq 2$) between 2001 and 2012, while MIMIC-IV includes 85,155 patients with multiple visits spanning from 2008 to 2019. Due to the overlapping time period between the two datasets, we randomly sample 10,000 patients from MIMIC-IV between 2013 and 2019 to ensure minimal redundancy. For the diagnosis prediction task, the objective is to predict the medical codes appearing in the subsequent admission.

To verify the efficiency of the proposed model, MIMIC-III is split into training (6,000 patients), validation (1,900 patients), and test (1,000 patients) sets. Similarly, MIMIC-IV is divided into 8,000, 1,000, and 1,000 patients accordingly. The last recorded visit of each patient serves as the prediction target, while the preceding visits are used as input features. Different from typical predictive models, we feed **II-KEA** by admission records of those patients in the training data for getting co-occurrence matrix, and examine predictive performance upon 500 patient cohort.

3.2 Tasks & Evaluation Metrics

Our experiments focus on the task of *Diagnosis Prediction*, which aims to predict all medical codes that will appear in a patient's next admission. This task is formulated as a multilabel classification problem. To evaluate model performance, we use weighted F_1 score (w-F_1) and top-k recall (R@k) as metrics, following prior work (Choi et al., 2016a; Bai et al., 2018). The w-F_1 score is a weighted sum of the F_1 score across all classes, providing an overall assessment of prediction quality. The R@k metric represents the proportion of true-positive instances among the top-k predictions relative to the total number of positive samples, reflecting model effectiveness in capturing relevant medical codes.

3.3 Baselines

To assess the performance of **II-KEA**, we compare it against 8 machine learning (ML)-based EHR models originally designed for diagnostic prediction: (i) RNN/CNN-based models: RETAIN (Choi et al., 2016b), Dipole (Ma et al., 2017), and Timeline (Bai et al., 2018). (ii) Graph-based models: Chet (Lu et al., 2022) and SeqCare (Xu et al., 2023).

(iii) Transformer-based models: G-BERT (Shang et al., 2019a), BEHRT (Li et al., 2020), and GT-BEHRT (Poulain and Beheshti, 2024).

Moreover, we compare 3 most recent baselines that combine language models with machine learning-based predictors (LM+ML): GraphCare (Jiang et al., 2024), RAM-EHR (Xu et al., 2024), and DualMAR (Hu et al., 2024). To ensure a fair comparison, we use condition codes as the sole input feature (e.g., excluding procedures and medications used in GraphCare), and we reconstruct the knowledge base using ICD-9-CM codes instead of CCS codes. Other agentic baselines like ColaCare (Wang et al., 2024) are excluded from the comparison due to its extensive input requirements, which lead to unstable predictions when only condition codes are provided.

3.4 Implementation Details

We implement **II-KEA** using `Python 3.10`. For all agents, we utilize `ChatGPT-4o mini` (OpenAI et al., 2024), accessed via the `Azure OpenAI`, as our LLM. To build the vector database, we employ `ChromaDB`, where document embeddings are generated using a pre-trained Sentence-BERT `all-MPNet-base-v2` model provided by the `Sentence Transformers`. We report the average performance (%) and standard deviation of each baseline over 5 runs, and we set the temperature value in **II-KEA** as 0. When evaluating the prediction performance of **II-KEA** we set the optional clinical comment to be empty.

3.5 Main Results

Table 1 presents the performance comparison, demonstrating that the proposed model, **II-KEA**, achieves state-of-the-art results across both datasets. Specifically, **II-KEA** outperforms GT-BEHRT by 2.44% in w-F_1, 0.73% in R@10, and 1.06% in R@20 on MIMIC-IV, with similar performance gains observed on MIMIC-III. The results further indicate that graph-based and transformer-based models consistently outperform RNN- and CNN-based approaches. Notably, knowledge-based models such as DualMAR leverage knowledge graphs to enhance learning, yielding a 9% improvement in R@20 on MIMIC-III. Similarly, transformer-based models like GT-BEHRT improve w-F_1 by approximately 8% on MIMIC-IV. While GT-BEHRT and DualMAR achieve competitive performance in certain metrics, **II-KEA** consistently surpasses both across the majority of evaluation criteria. Overall, these findings underscore the effectiveness of **II-KEA** in diagnosis prediction and highlight the potential of a unified agentic framework for advancing predictive healthcare.

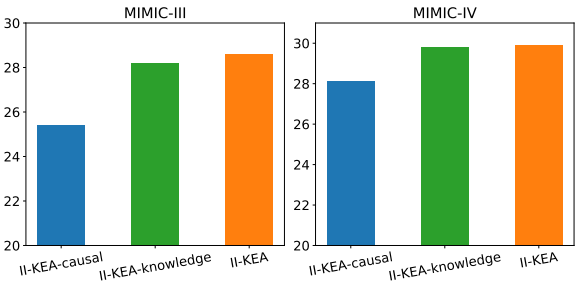

Figure 3: Comparison between different version of **II-KEA**. F1 scores on MIMIC-III and MIMIC-IV are reported.

3.6 Ablation Study

We conduct ablation studies to evaluate the effectiveness of components in **II-KEA**. Specifically, we aim to understand how causal analysis and external knowledge contribute to prediction performance. The Knowledge Synthesis Agent and the Causal Discovery Agent are separately removed from the prediction workflow and weighted F1 scores are reported in Figure 3. We denote the version without the Causal Discovery Agent as **II-KEA**-*causal* and the version without the Knowledge Synthesis Agent as **II-KEA-knowledge**. The results show that **II-KEA**-*causal* experiences a more significant performance drop than the full model, highlighting the crucial role of causal analysis in improving prediction accuracy. In contrast, removing external knowledge (**II-KEA-knowledge**) results in only a marginal decline, suggesting a lesser impact in its current form. We hypothesize that this is because our knowledge database is sourced from Wikipedia, which primarily serves as a demonstration of external knowledge integration but may offer limited domain-specific medical insights. However, this also underscores the potential for improvement by incorporating curated databases or clinician-maintained knowledge sources.

3.7 Case Study

We conduct a case study to analyze how different agents within **II-KEA** function and collaborate during the decision-making process. We randomly select a patient from the MIMIC-III dataset and report the output of each agent during inference, as shown in Figure 4. The Causal Discovery Agent identifies

Table 1: **Prediction Results on MIMIC-III and MIMIC-IV for Diagnosis Prediction.** The best results for each metric are **highlighted**, and the second bests results are underlined.

Type	Models	MIMIC-III w-F_1	MIMIC-III R@10	MIMIC-III R@20	MIMIC-IV w-F_1	MIMIC-IV R@10	MIMIC-IV R@20
ML	RETAIN	18.37 (0.78)	32.12 (0.79)	32.54 (0.63)	23.11 (0.78)	37.32 (0.79)	40.15 (0.63)
	Dipole	14.66 (0.21)	28.73 (0.22)	29.44 (0.21)	22.16 (0.21)	36.21 (0.22)	38.74 (0.21)
	Timeline	20.46 (0.22)	30.73 (0.12)	34.83 (0.10)	24.76 (0.22)	39.89 (0.12)	44.87 (0.10)
	Chet	22.63 (0.22)	33.64 (0.32)	37.87 (0.22)	25.74 (0.22)	39.23 (0.32)	42.67 (0.22)
	SeqCare	24.36 (0.12)	37.47 (0.11)	40.53 (0.12)	26.12 (0.12)	42.91 (0.11)	46.25 (0.12)
	G-BERT	22.28 (0.32)	35.62 (0.21)	36.46 (0.22)	25.12 (0.32)	41.91 (0.21)	46.25 (0.22)
	BEHRT	23.15 (0.21)	34.68 (0.32)	35.97 (0.11)	24.53 (0.21)	38.42 (0.32)	44.89 (0.11)
	GT-BEHRT	25.21 (0.18)	36.15 (0.23)	40.97 (0.41)	**30.17 (0.18)**	44.93 (0.23)	50.67 (0.41)
LM+ML	GraphCare	25.16 (0.31)	36.74 (0.28)	41.89 (0.36)	27.59 (0.31)	42.07 (0.28)	48.19 (0.36)
	RAM-EHR	23.27 (0.24)	34.66 (0.18)	38.49 (0.25)	26.97 (0.29)	41.17 (0.30)	46.23 (0.21)
	DualMAR	25.37 (0.17)	38.24 (0.26)	41.86 (0.24)	27.97 (0.17)	44.07 (0.26)	48.19 (0.24)
Agent	**II-KEA**	**28.61 (0.00)**	**38.52 (0.00)**	**43.86 (0.00)**	29.87 (0.00)	**45.66 (0.00)**	**51.73 (0.00)**

a causal graph, visualized in the figure, which helps illustrate the underlying mechanisms connecting different diseases. For the Decision-Making Agent, we provide outputs both with and without clinician input. In the first query, no specific guidance is given, leading to a more general prediction that considers all possible diseases. In contrast, in the second query, the clinician provides additional input, specifying a focus on kidney-related diseases. Consequently, the model prioritizes kidney-related predictions. It is important to note that the performance of these two versions cannot be directly compared; rather, the key advantage is that clinicians can incorporate their expertise and preferences to tailor predictions to their specific needs (e.g., a nephrologist may prioritize kidney-related diseases). We also observe that both versions of the predictions not only provide disease codes but also offer detailed explanations, enhancing interpretability and helping clinicians in making informed decisions and determining next-step treatment plans.

4 Related Work

We categorize prior work into clinical prediction (section 4.1), agentic approaches (section 4.2), and causal inference (appendix B.1).

4.1 Predictive Healthcare in EHR

Predictive modeling in healthcare has advanced significantly with the adoption of deep learning techniques applied to Electronic Health Records (EHR) data. Existing neural network-based models, including RNN/Attention-based approaches (Choi et al., 2016a, 2017; Ma et al., 2020), graph-based models (Choi et al., 2017; Ma et al., 2018; Lu et al., 2021), and Transformer-based architectures (Shang et al., 2019b; Luo et al., 2020; Poulain and Beheshti, 2024), have demonstrated effectiveness in capturing temporal patterns and interactions among medical concepts. Recent work (Jiang et al., 2024) has explored leveraging external knowledge sources beyond hierarchical structures such as ICD-9-CM by integrating Large Language Models (LLMs) to enhance medical predictions.

Still, most models remain black boxes, offering limited interpretability and restricting healthcare professionals from interacting with the system to refine or adjust predictions. In clinical applications, predictive models must provide faithful explanations, such as causal pathways, and allow interactive refinement based on expert guidance.

4.2 LLM Agents for Healthcare AI

More recently, LLMs have demonstrated agentic capabilities in clinical applications through multi-agent frameworks. EHRAgent (Shi et al., 2024) utilizes multiple agents for multi-tabular retrieval, integrating external tools and long-term memory to handle complex clinical queries. KG4Diagnosis (Zuo et al., 2024) enhances diagnostic reasoning through hierarchical agent collaboration and knowledge graph construction guided by semantic understanding. ColaCare (Wang et al., 2024) improves EHR-based report generation and treatment planning by facilitating collaboration between DoctorAgents and MetaAgent using retrieval-augmented generation (RAG) techniques. MDAgents (Kim et al., 2024) and AgentClinic (Schmidgall et al., 2024) simulate clinical interactions using multi-agent systems, where agents

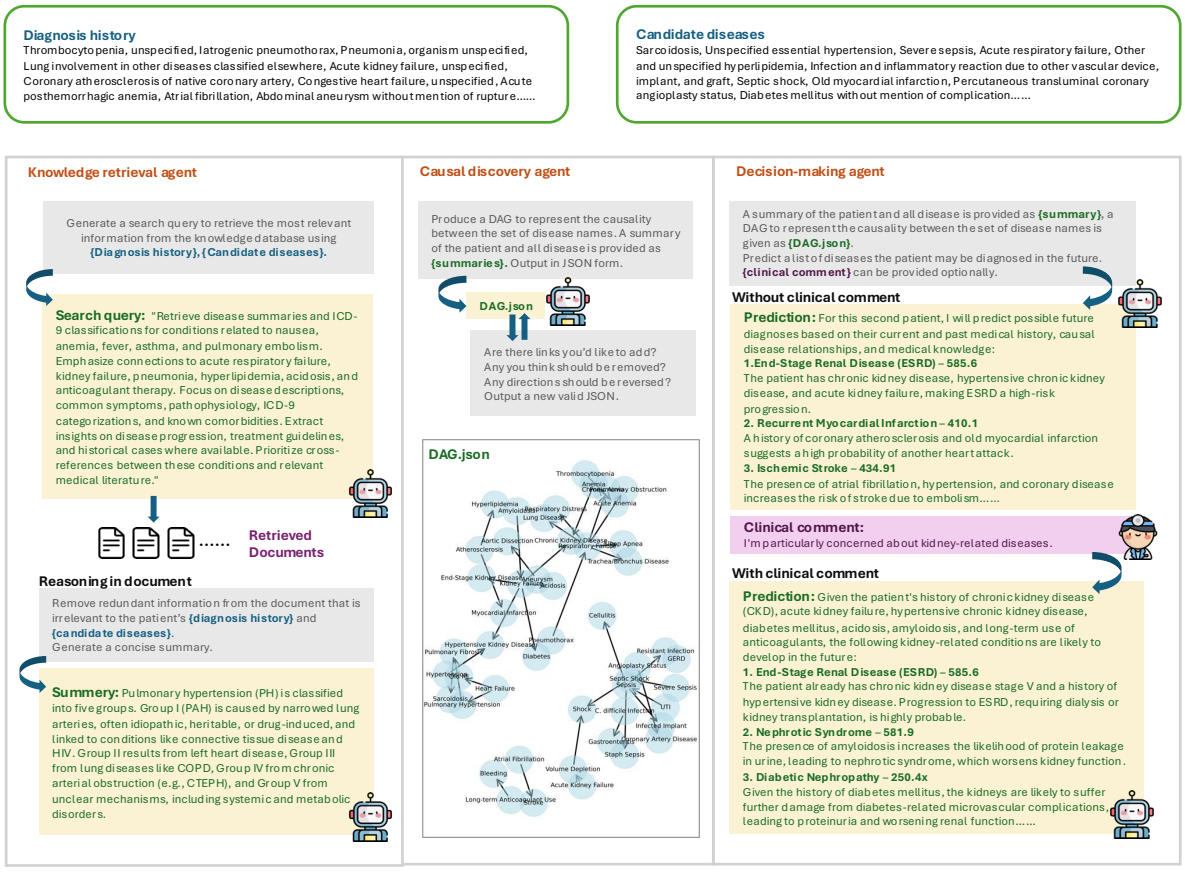

Figure 4: A case study on a patient from MIMIC-III.

collaborate to support multi-modal reasoning and communication benchmarking. These studies highlight an emerging trend in agentic AI for clinical applications, where LLMs leverage in-context learning and dynamically retrieve medical knowledge to provide personalized and adaptive responses.

Still, the integration of LLM agents for sequential diagnostic prediction remains underexplored, presenting an opportunity to develop interactive and explainable models for medical diagnosis.

5 Conclusion

In this paper, we introduce **II-KEA**, a knowledge-enhanced Agentic Causal Discovery framework designed for interpretable and interactive diagnosis prediction. **II-KEA** consists of three LLM-based agents working collaboratively and is powered by both clinical datasets and domain knowledge. We evaluate **II-KEA** on the MIMIC-III and MIMIC-IV datasets and conduct both ablation and case studies to demonstrate its effectiveness. The ethical consideration can be checked in Appendix A.

6 Limitation and Future Work

II-KEA showcases a promising paradigm for interpretable and interactive diagnosis prediction by leveraging LLM agents. However, **II-KEA** has significant potential to solve a broader range of medical challenges. Future work includes:

- **Enhancing external domain knowledge:** In this work, we use the Wikipedia database as a proof of concept. In future work, we aim to integrate more domain-specific external knowledge sources to enhance diagnosis prediction in fine-grained target domains.
- **Expanding task diversity:** While this work focuses on diagnosis prediction, future research will explore additional tasks tailored to clinicians' needs, such as treatment planning and personalized medical recommendations.
- **Incorporating multiple stakeholders:** The current version of **II-KEA** facilitates interactions only with clinicians. Future iterations will explore collaborative decision-making involving multiple stakeholders to enhance holistic and patient-centered care.

References

Ahmed Abdulaal, adamos hadjivasiliou, Nina Montana-Brown, Tiantian He, Ayodeji Ijishakin, Ivana Drobnjak, Daniel C. Castro, and Daniel C. Alexander. 2024. Causal modelling agents: Causal graph discovery through synergising metadata- and data-driven reasoning. In *The Twelfth International Conference on Learning Representations*.

Tian Bai, Shanshan Zhang, Brian L Egleston, and Slobodan Vucetic. 2018. Interpretable representation learning for healthcare via capturing disease progression through time. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 43–51.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016a. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 787–795.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016b. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29.

Kristy Choi, Chris Cundy, Sanjari Srivastava, and Stefano Ermon. 2022. Lmpriors: Pre-trained language models as task-specific priors. *Preprint*, arXiv:2210.12530.

Pengfei Hu, Chang Lu, Fei Wang, and Yue Ning. 2024. Dualmar: Medical-augmented representation from dual-expertise perspectives. *arXiv preprint arXiv:2410.19955*.

Pengcheng Jiang, Cao Xiao, Adam Cross, and Jimeng Sun. 2024. Graphcare: Enhancing healthcare predictions with personalized knowledge graphs. In *The Twelfth International Conference on Learning Representations*.

Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, and Yoshua Bengio. 2024. Efficient causal graph discovery using large language models. *Preprint*, arXiv:2402.01207.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(160035).

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. In *Advances in Neural Information Processing Systems*, volume 37, pages 79410–79452. Curran Associates, Inc.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2024. Causal reasoning and large language models: Opening a new frontier for causality. *Preprint*, arXiv:2305.00050.

Hao Duong Le, Xin Xia, and Zhang Chen. 2024. Multi-agent causal discovery using large language models. *arXiv preprint arXiv:2407.15073*.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2020. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155.

Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, and 1 others. 2024. Large language models and causal inference in collaboration: A comprehensive survey. *arXiv preprint arXiv:2403.09606*.

Stephanie Long, Alexandre Piché, Valentina Zantedeschi, Tibor Schuster, and Alexandre Drouin. 2023. Causal discovery with language models as imperfect experts. *Preprint*, arXiv:2307.02390.

Chang Lu, Tian Han, and Yue Ning. 2022. Context-aware health event prediction via transition functions on dynamic disease graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4567–4574.

Chang Lu, Chandan K. Reddy, Prithwish Chakraborty, Samantha Kleinberg, and Yue Ning. 2021. Collaborative graph learning with auxiliary text for temporal event prediction in healthcare. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3529–3535.

Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. 2020. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 647–656.

Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of*

the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1903–1911.

Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. 2018. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 743–752.

Liantao Ma, Junyi Gao, Yasha Wang, Chaohe Zhang, Jiangtao Wang, Wenjie Ruan, Wen Tang, Xin Gao, and Xinyu Ma. 2020. Adacare: Explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 825–832.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Raphael Poulain and Rahmatollah Beheshti. 2024. Graph transformers on ehrs: Better representation improves downstream performance. In *The Twelfth International Conference on Learning Representations*.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *Preprint*, arXiv:2405.07960.

Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. 2019a. Pre-training of graph augmented transformers for medication recommendation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5953–5959. International Joint Conferences on Artificial Intelligence Organization.

Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. 2019b. Pre-training of graph augmented transformers for medication recommendation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*, pages 5953–5959.

ChengAo Shen, Zhengzhang Chen, Dongsheng Luo, Dongkuan Xu, Haifeng Chen, and Jingchao Ni. 2024. Exploring multi-modal integration with tool-augmented llm agents for precise causal discovery. *arXiv preprint arXiv:2412.13667*.

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 *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22315–22339.

Zixiang Wang, Yinghao Zhu, Huiya Zhao, Xiaochen Zheng, Tianlong Wang, Wen Tang, Yasha Wang, Chengwei Pan, Ewen M Harrison, Junyi Gao, and 1 others. 2024. Colacare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration. *arXiv preprint arXiv:2410.02551*.

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 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 754–765.

Yongxin Xu, Xu Chu, Kai Yang, Zhiyuan Wang, Peinie Zou, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2023. Seqcare: Sequential training with external medical knowledge graph for diagnosis prediction in healthcare data. In *Proceedings of the ACM Web Conference 2023*, pages 2819–2830.

Kaiwen Zuo, Yirui Jiang, Fan Mo, and Pietro Lio. 2024. Kg4diagnosis: A hierarchical multi-agent llm framework with knowledge graph enhancement for medical diagnosis. *arXiv preprint arXiv:2412.16833*.

Appendix

A Privacy and Ethical Statement

Our work involves the analysis of EHR data, which contains sensitive personal medical information. In compliance with the PhysioNet Credentialed Health Data Use Agreement 1.5.0[2], we conducted all interactions between the language models and the EHR data through *Azure OpenAI Service*[3], which adheres to enterprise-grade security and compliance standards. We also submitted the opt-out form[4] to decline human review in terms of the responsible use guidelines specified for MIMIC datasets available at *Responsible Use of MIMIC Data with online services*[5], which outlines proper handling of EHR data when used with generative models. This ensured that the capabilities of large language models were applied without compromising the privacy and confidentiality of patient information. Furthermore, we continuously and carefully monitor our compliance with these guidelines and relevant privacy regulations to uphold the ethical use of data in our research and operations.

B Additional Related Work

B.1 Causality Inference on LLM

Causal inference is a cornerstone of medical research, enabling the discovery of relationships between clinical factors. LLMs, equipped with extensive domain knowledge, have the potential to assist in causal graph generation and infer causal relationships from unstructured data. Consequently, recent studies have started exploring LLM-driven causal discovery frameworks (Liu et al., 2024). Several works (Le et al., 2024; Shen et al., 2024; Choi et al., 2022; Kıcıman et al., 2024; Long et al., 2023; Jiralerspong et al., 2024) have employed LLMs for causal relation inference and graph generation, yet their application to EHR-based predictive tasks remains limited. Most existing approaches focus on general causal reasoning tasks or static datasets, without fully leveraging the interactive and adaptive capabilities of LLM agents for healthcare-specific causal discovery.

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

[3]https://azure.microsoft.com/en-us/products/ai-services/openai-service/

[4]https://azure.microsoft.com/en-us/products/cognitive-services/openai-service/

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

However, causal discovery is essential for diagnosis prediction, as it provides a structured explanation of disease co-occurrences, facilitating more transparent and interpretable decision-making. Bridging this gap is crucial for achieving explainable AI in clinical practice by enabling collaborative causal reasoning among AI agents. Furthermore, integrating interactive causal discovery mechanisms allows healthcare professionals to refine insights and better understand disease relationships, ultimately improving diagnosis and treatment planning.

C Pseudocodes

Algorithm 1 Knowledge Synthesis Agent, $\mathcal{A}_{knowledge}$.

Input: Diagnosis history $\mathcal{D}_p$, Candidate diseases $\mathcal{S}_p$, Knowledge vector database Γ with meta data $\mathcal{K}_\Gamma$.

// Generate search query

$q_{search} = LLM_{search}(\mathcal{D}_p, \mathcal{S}_p, \mathcal{K}_\Gamma)$

// Retrieve k most relevant documents from database Γ

$\Gamma_q = query(q_{search}, \Gamma)$

for each document $\gamma \in \Gamma^q$:

// Reasoning in document

$a_\gamma = LLM_{reason-in-doc}(\gamma, \mathcal{D}_p, \mathcal{S}_p)$

end for

Output: Summarized document set $\Gamma_p^{summary} = a_\gamma \mid \gamma \in \Gamma^q$.

D Prompt Details

In this section, we provide the prompt templates for knowledge retrieval agent, causal discovery agent, and decision-making agent, separately.

Algorithm 2 Causal Discovery Agent, $\mathcal{A}_{causal}$.

Input: Empty graph $\mathcal{G}^{\emptyset}$ consists of entities from $\mathcal{D}_p \cup \mathcal{S}_p$, Candidate diseases , Summarized document set $\Gamma_p^{summary}$, diagnosis matrix $\mathbf{A}_D$

$t \triangleq 0$

// Hypothesis generation

$\mathcal{G}^s_{t=0} = \text{LLM}_{\text{hypo-gen}}(\mathcal{G}^{\emptyset}, \Gamma_p^{\text{summary}})$

While 1:

// Model fitting

$l_t = \texttt{log likelihood}\ (\mathcal{G}^s_t, \mathbf{A}_D)$

// Post-processing

$\mathcal{M} \triangleq \{\mathcal{G}^s_t, \mathcal{G}^s_{t-1}, l_t, l_{t-1}\}$

// Hypothesis amendment

$\mathcal{G}^s_{t+1} = \text{LLM}_{\text{hypo-amend}}(\mathcal{M})$

if *stopping criteria* is meet: $\mathcal{G}^s \triangleq \mathcal{G}^s_{t+1}$

break

$t = t + 1$

Output: Final causal graph $\mathcal{G}^s$

Algorithm 3 Decision-Making Agent, $\mathcal{A}_{decision}$.

Input: Diagnosis history $\mathcal{D}_p$, Candidate diseases $\mathcal{S}_p$, Summarized document set $\Gamma_p^{\text{summary}}$, Causal graph $\mathcal{G}^s$, Optional clinician comment $\mathcal{C}$

// Make diagnosis prediction with explanations

$\mathcal{D}_{\text{pred}}, \mathcal{E} = \text{LLM}_{\text{decison}}(\mathcal{S}_p, \Gamma_p^{\text{summary}}, \mathcal{G}^s, \mathcal{C})$

Output: Predicted diagnosis $\mathcal{D}_{\text{pred}}$ and explanations $\mathcal{E}$

Algorithm 4 II-KEA inference.

Require: Pretrained LLM model, EHR training data $\mathcal{D}_p \mid p \in \mathcal{P}_{train}$, Knowledge vector database Γ with meta data $\mathcal{K}_\Gamma$.

// Data processing

Calculate $\mathbf{A}_T$ and $\mathbf{A}_D$ with Equation 1 and 3.

// Inference

For $p \in \mathcal{P}_{test}$:

Input: Diagnosis history $\mathcal{D}_p$

Obtain candidate disease $\mathcal{S}_p$ with Equation 4.

// Knowledge Synthesis Agent

$\Gamma_p^{summary} = \mathcal{A}_{summary}(\mathcal{D}_p, \mathcal{S}_p, \mathcal{K}_\Gamma)$

// Causal Discovery Agent

$\mathcal{G}^s = \mathcal{A}_{causal}(\mathcal{D}_p, \mathcal{S}_p, \Gamma_p^{summary}, \mathbf{A}_D)$

// Decision-Making Agent

$\mathcal{D}_{pred}, \mathcal{E} = \mathcal{A}_{decision}(\mathcal{D}_p\mathcal{S}_p, \Gamma_p^{summary}, \mathcal{G}^s, \mathcal{C})$

Output: Final causal graph $\mathcal{G}^s$, predicted diagnosis $\mathcal{D}_{\text{pred}}$ and explanations $\mathcal{E}$.

Knowledge retrieval agent - Prompt

 # Knowledge retrieval:
Generate a search query to retrieve the most relevant information from the knowledge database using {Diagnosis history} and {Candidate diseases}. The generated search query should take into account the characteristics of the knowledge database, as described by the provided {Meta-data}.

 # Reasoning in document:
Summarize the {Document i}. The output summary should satisfy the following requirements:
Relevance: Include only information related to the patient's {Diagnosis history} and {Candidate diseases}.
Conciseness: Remove redundant and unnecessary details while maintaining key insights.
Clarity: Ensure the summary is well-structured and easy to understand.

Causal discovery agent - Prompt

 # Hypothesis Generation:
Generate a Directed Acyclic Graph (DAG) to represent the causal relationships between the given set of {Disease names}. Use the provided {Summary}, along with contextual knowledge and reasoning, to infer causality. The output should be in JSON format.

 # Hypothesis Amendment:
Adjust the causal graph based on the current and previous versions stored in {Memory}, along with their fitting scores. Consider the following questions:
Are there any links that should be added?
Should any existing links be removed?
Should any directions be reversed?
Generate a revised causal graph and output it in a valid JSON format.

Decision-making agent - Prompt

Predict a list of diseases the patient may be diagnosed with in the future based on:
Patient summary and disease information: {Summary}
Causal DAG of disease relationships: {DAG.json}
Optional clinician comment: {Clinician comment}
Output format:
A JSON list of predicted ICD-9 codes.
A detailed explanation of the reasoning process.
Separate the two parts using the special token <SEP>.