KERAP: A Knowledge-Enhanced Reasoning Approach for Accurate Zero-shot Diagnosis Prediction Using Multi-agent LLMs

Yuzhang Xie, MS[1], Hejie Cui, PhD[2], Ziyang Zhang, BS[1], Jiaying Lu, PhD[1], Kai Shu, PhD[1], Fadi Nahab, MD[1], Xiao Hu, PhD[1], Carl Yang, PhD[1]
[1] Emory University, Atlanta, GA; [2] Stanford University, Palo Alto, CA;

Abstract

Medical diagnosis prediction plays a critical role in disease detection and personalized healthcare. While machine learning (ML) models have been widely adopted for this task, their reliance on supervised training limits their ability to generalize to unseen cases, particularly given the high cost of acquiring large, labeled datasets. Large language models (LLMs) have shown promise in leveraging language abilities and biomedical knowledge for diagnosis prediction. However, they often suffer from hallucinations, lack structured medical reasoning, and produce useless outputs. To address these challenges, we propose KERAP, a knowledge graph (KG)-enhanced reasoning approach that improves LLM-based diagnosis prediction through a multi-agent architecture. Our framework consists of a linkage agent for attribute mapping, a retrieval agent for structured knowledge extraction, and a prediction agent that iteratively refines diagnosis predictions. Experimental results demonstrate that KERAP enhances diagnostic reliability efficiently, offering a scalable and interpretable solution for zero-shot medical diagnosis prediction.

Introduction

Medical diagnosis prediction, which is the task of predicting a patient's future health risks based on their historically observed medical data such as electronic health records (EHRs), plays a vital role in enabling accurate healthcare and early interventions.[1,2] Various machine learning (ML) models, such as random forests, XGBoost, and neural networks, have been widely used for diagnosis prediction.[3] However, their reliance on supervised training limits their ability to make a prediction for an unseen case without an explicit training process, which is called as "zero-shot prediction ability". This capability is particularly crucial in scenarios where labeled medical data is scarce. Moreover, acquiring expert-labeled medical records is costly and time-intensive. Therefore, exploring zero-shot diagnosis prediction methods is essential for developing scalable and efficient predictive solutions.[4]

Recently, Large Language Models (LLMs) have shown promise in achieving zero-shot diagnosis prediction by leveraging their unprecedented language understanding abilities and rich biomedical knowledge.[5,6] By modeling patient healthcare data, LLMs can uncover hidden patterns in patient records, improving predictions for conditions such as stroke, cardiovascular diseases, and cognitive decline.[7,8] However, LLMs still face significant challenges. They lack structured medical reasoning capabilities, meaning they are unable to systematically process and infer relationships between medical concepts in a way that aligns with humans' clinical decision-making process. Moreover, they are susceptible to hallucinations-generating plausible yet inaccurate predictions—which pose serious concerns in high-stakes clinical environments.[9] Additionally, LLMs sometimes become ineffective for certain cases since they may blindly classify all cases as high-risk to minimize potential errors, leading to biased and useless predictions.[10] To ensure reliability and clinical applicability, LLMs need external guidance through structured knowledge integration.

To overcome these challenges, knowledge graphs (KGs) have emerged as a promising solution to enhance LLM-based diagnosis prediction.[11] KGs offer a structured and clinically validated framework that describes the relationships among various medical concepts, thereby enhancing the diagnostic ability of LLMs.[12] By integrating KG-enhanced LLMs into diagnostic processes, the models can effectively retrieve relevant disease patterns and align their predictions with established medical knowledge.[13] This symbiotic integration not only mitigates the risk of generating inaccurate or "hallucinated" information but also enhances the interpretability and reliability of the diagnostic outputs.[14] However, existing KG-enhanced diagnosis prediction approaches typically rely on directly prompting LLMs with KG knowledge and EHR data, whose performance could be improved by incorporating a more complex medical reasoning process.[13,14,15] For instance, a multi-agent framework could decompose complex medical reasoning into specialized tasks, allowing different agents to focus on different tasks such as linking medical attributes across different data

sources, retrieving knowledge, and predicting a patient's diagnosis. This reduces cognitive overload for a single model and allows for more structured, task-specific optimizations. Moreover, a multi-stage conversational reasoning framework could enhance robustness by enabling iterative refinement, allowing the LLM to focus on different parts of the extracted knowledge and reflect on its prediction.

In this work, we propose KERAP, a KG-enhanced multi-stage reasoning approach for accurate zero-shot diagnosis prediction using LLMs. KERAP operates through a multi-agent architecture: a linkage agent, a retrieval agent, and a prediction agent. The linkage agent initiates the process by establishing connections between medical attributes—such as symptoms, diagnoses, and treatments—extracted from real-world patient datasets (EHRs) and a comprehensive biomedical KG. LLMs are considered state-of-the-art methods to link different medical attributes.[16] By mapping raw patient data to structured medical knowledge accurately, this agent ensures that the retrieved information is contextually relevant and aligned with clinical reasoning. Subsequently, the retrieval agent queries the KG to extract and summarize attribute-related knowledge, categorizing them into two distinct perspectives: positive knowledge (e.g., "symptom X indicates condition Y") for inclusion criteria, and negative knowledge (e.g., "symptom X rules out condition Z") for exclusion criteria. This step reduces LLM hallucinations and enhances reasoning consistency by ensuring the LLM is guided by structured, validated knowledge rather than unverified text. These related knowledge are then formatted as contextual prompts to provide structured knowledge and reduce hallucinations. Finally, the prediction agent integrates patient-specific data with the retrieved external knowledge, engaging in multi-stage conversation steps to refine its understanding and produce a robust, patient-tailored diagnosis prediction. In our experiments, KERAP demonstrates superior performance, achieving higher accuracy and reliability in zero-shot scenarios. Our approach contributes to the advancement of predictive modeling in clinical artificial intelligence, offering a scalable solution that supports early disease detection, personalized risk assessment, and improved patient outcomes.

Related Work

LLM Reasoning. Large Language Models (LLMs), including OpenAI's GPT and Meta's LLaMa, have transformed various domains by enhancing reasoning capabilities.[7] To address complex tasks such as diagnosis prediction, diverse reasoning strategies can be integrated with LLMs to support medical reasoning. For example, Direct Prompting serves as a basic approach, eliciting immediate responses from LLMs. More advanced Step-by-Step techniques, such as Chain-of-Thought,[17] decompose problems into logical intermediate steps. Additionally, KG-Augmented Prompting further incorporates external, structured knowledge to improve factual accuracy and interoperability.[13] Iterative Prompting advances by allowing predictions to be refined over multiple stages through reflective evaluation.[18] These strategies collectively enable LLMs to generate outputs that are coherent, contextually relevant, and clinically accurate.

LLM for Diagnosis Prediction. LLMs have demonstrated substantial potential in enhancing medical diagnosis prediction by analyzing extensive clinical data and patient records. Researchers have explored direct prompting LLMs for diagnosis prediction.[19,20] Additionally, some approaches leverage pre-trained medical-specialized LLMs, such as HuatuoGPT.[21] Some approaches combine the LLM with other trained classification ML models (e.g., Logistic Regression, XGBoost).[22,23,24,25] Another promising direction is in-context learning, where LLMs adapt to medical tasks by utilizing few-shot examples, as seen in CPLLM[26] and EHR-CoAgent.[8] Furthermore, some methods integrate biomedical KGs with LLMs, such as Dr.Knows,[13] medIKAL,[14] and ICP,[15] improving interpretability and diagnostic precision by offering structured explainable relations. However, these approaches often struggle with adaptability to unseen cases, as they either depend on simple direct prompting, require extensive labeled data for fine-tuning or training, or lack structured multi-step reasoning to iteratively refine predictions.

Method

Task Definition. We aim to predict whether a patient will be diagnosed with a specific disease at their next clinical visit, based on medical attributes recorded during the current visit. Formally, let i denote the i-th clinical visit of a patient, and let $\mathbf{r}_i$ be the set of EHR entries associated with this visit: $\mathbf{r}_i = \{a_{i,0}, a_{i,1}, \ldots, a_{i,j}\}$, where each $a_{i,j}$ represents a medical attribute such as a diagnosis (e.g., heart failure) or medication (e.g., desmopressin). These attributes capture crucial aspects of a patient's diagnosis and treatment during the visit. Given $\mathbf{r}_i$, the goal is to predict whether a target disease d will be diagnosed at the next visit.

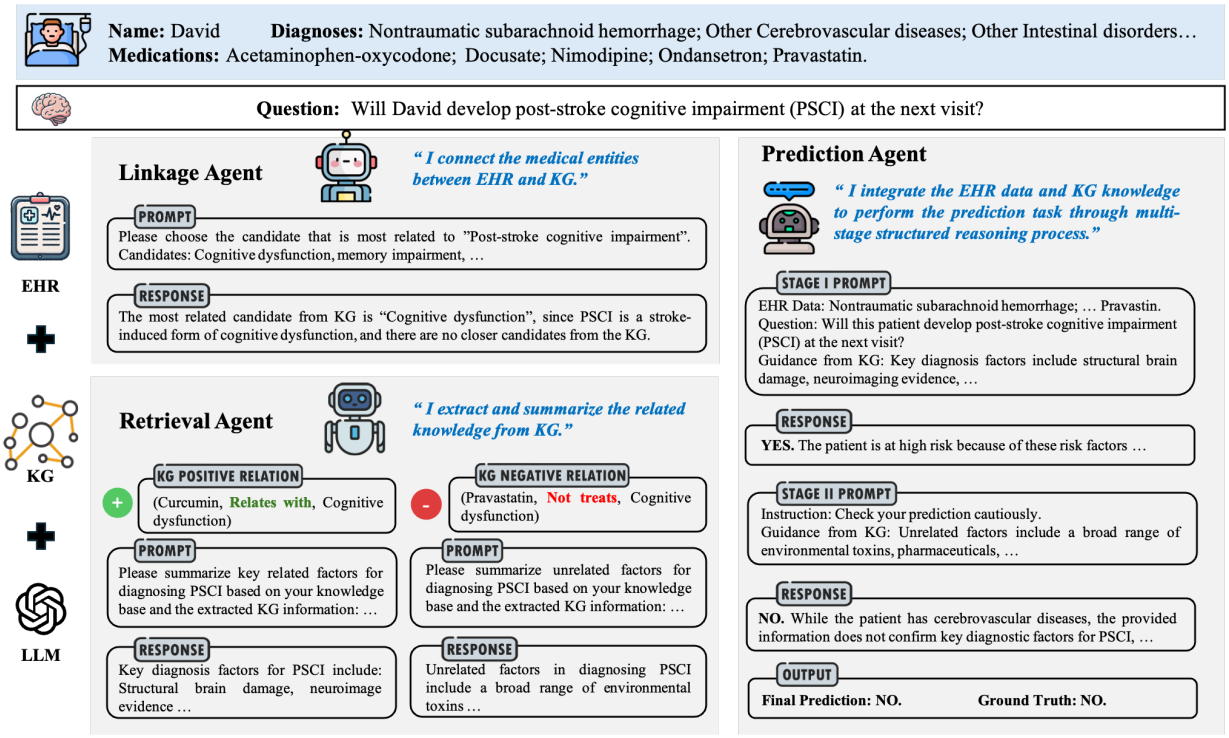


Figure 1: Framework of KERAP.

Framework. The KERAP framework is illustrated in Fig.1. It adopts a multi-agent architecture, with GPT-4o-mini supporting each agent. The framework comprises three key components: the linkage agent, the retrieval agent, and the prediction agent. The linkage agent A_{LK} locates the predicted disease in a biomedical KG, serving as a prerequisite for the subsequent retrieval agent to access relevant biomedical knowledge for that disease. For example, the disease expression "Post-stroke cognitive impairment" in the question text is linked to the most appropriate KG entity "Cognitive dysfunction" by A_{LK}, since there are no closer candidates from the KG. The retrieval agent A_{RT} queries the KG to extract and summarize relevant information, categorizing the results into positive knowledge relations, e.g., *(Curcumin, Relates with, Cognitive dysfunction)*, and negative relations, e.g., *(Pravastatin, Not treats, Cognitive dysfunction)*, to support inclusion and exclusion criteria. Finally, the prediction agent A_{PR} integrates patient records with the extracted structured knowledge and employs multi-stage reasoning to perform zero-shot diagnosis prediction, producing a clear "YES" or "NO" outcome.

Linkage Agent. To connect the predicted disease entity d from EHR with the relevant KG entity x, we leverage a linkage agent A_{LK} inspired by PromptLink.[16] To design effective and cost-efficient prompts within the constraints of context length, linking candidates are first generated before the linkage agent performs entity disambiguation. Specifically, we utilize a pre-trained language model, SAPBERT,[27] to generate embeddings $Z_{lm,d} \in \mathbb{R}^{d_{lm}}$ for the EHR entity d and $Z_{lm,x} \in \mathbb{R}^{d_{lm}}$ for the KG entity x ($d_{lm} = 768$ from SAPBERT). Then we compute the cosine similarity $SIM \in [0, 1]$ between $Z_{lm,d}$ and $Z_{lm,x}$, selecting the top-LC candidates (where $LC = 10$) with the highest similarity scores as linking candidates. These candidates serve as input for the linkage agent, which further generates the entity mapping. By analyzing the semantic meaning and contextual relevance of the candidates, the linkage agent chooses the KG entity x as the most appropriate one to link for the EHR medical entity d. This agent A_{LK} utilizes both deep embeddings and LLM-based reasoning to ensure efficiency and precision in linking medical entities. By ensuring accurate disease-to-KG mapping, KERAP enables accurate integration of biomedical knowledge from KG. The process is formally defined as:

$$x = A_{LK}(d). \tag{1}$$

Retrieval Agent. After linking a disease entity d from the EHR to its corresponding KG entity x by A_{LK}, relevant

knowledge is extracted from the KG by the retrieval agent A_{RT}. The KG is a multi-relational graph denoted as KG $= (X, R, RT)$, where X represents the set of entities (nodes), R denotes the set of relations (edges), and RT contains the relational triples. The relations of the linked KG entity x are categorized into positive knowledge relations $RT_{x,+}$ and negative knowledge relations $RT_{x,-}$ based on their semantic meaning. For example, *(Curcumin, Relates with, Cognitive dysfunction)* is classified as positive knowledge relation, indicating that curcumin relates with cognitive dysfunction. Conversely, *(Pravastatin, Not treats, Cognitive dysfunction)* is a negative knowledge relation, signifying that pravastatin does not treat cognitive dysfunction. Next, the retrieval agent A_{RT} is prompted to summarize the extracted relations $RT_{x,+}$ and $RT_{x,-}$. These summarized relations encapsulate relevant knowledge for inclusion or exclusion criteria. The extracted knowledge for d is represented as $\mathcal{KN}_d$ in:

$$\mathcal{KN}_d = \{\mathcal{KN}_{d,+}; \mathcal{KN}_{d,-}\} = A_{RT}(RT_x) = A_{RT}(\{RT_{x,+}; RT_{x,-}\}). \tag{2}$$

Prediction Agent. After retrieving the relevant knowledge $\mathcal{KN}_d$ for disease d, it is integrated with a patient-specific record r_i as prompt for the prediction agent A_{PR}. The prediction agent is then tasked with determining whether the patient will develop disease d at the next visit, completing the diagnosis prediction task. A_{PR} follows a multi-stage reasoning process. In Stage I, it considers the patient's visit data r_i along with positive relations $\mathcal{KN}_{d,+}$ to make an initial prediction, leveraging both patient-specific visit data and enlightening inclusion criteria from external knowledge. However, since LLMs sometimes become ineffective and blindly classify all cases as high-risk, A_{PR} refines its prediction in the second stage. In Stage II, A_{PR} combines the Stage I prompt and output with the negative relations $\mathcal{KN}_{d,-}$, leveraging self-reflection and exclusion criteria from external knowledge to eliminate misclassified cases. The prediction process is formulated as follows, where $\mathcal{PR}_{i,d}$ represents the prediction result for patient visit i and disease d:

$$\mathcal{PR}_{i,d} = A_{PR}(\{r_i;\ \mathcal{KN}_d\}) = \text{“YES” or “NO”}. \tag{3}$$

Experiments

EHR Datasets. In our experiments, we perform diagnosis prediction tasks for four datasets: Pneumonia (PNA), Chronic Kidney Disease (CKD), Congestive Heart Failure Nonhypertensive (CHF), and Post-Stroke Cognitive Impairment (PSCI). Patient-specific EHR data for Pneumonia are sourced from MIMIC-III,[28] while PSCI data come from PROMOTE. MIMIC-III is a public dataset containing over 53,000 patient records from critical care units at Beth Israel Deaconess Medical Center between 2001 and 2012. Following the settings of Xu et al.,[29] we select 12,353 patient encounters, incorporating 7,408 key medical attributes, including 845 diagnoses, 4,522 prescriptions, 2,021 procedures, and 20 services. PROMOTE is a private dataset containing records of stroke patients treated from 2012 to 2021.[30] We extract data on 7,780 patients and 2,595 clinical attributes, including 1,480 ICD-10 diagnosis codes and 1,115 prescribed medications recorded up to each patient's discharge following their index stroke.

KG Data Source. We use a large-scale public KG, iBKH,[31] as the primary KG dataset. iBKH integrates data from various biomedical KGs, offering a comprehensive resource with 2,384,501 entities across 11 categories, including drugs, diseases, symptoms, genes, and pathways, *etc*. Moreover, with over 48 million relation triples, iBKH facilitates deeper insights into complex biological interactions.

Baselines. We evaluate the performance of KERAP against the following baseline methods:

- Direct Prompting: Directly prompting LLM with patient's EHR data.
- Step-by-Step Prompting: Prompting the LLM with a patient's EHR data and prediction steps in a single stage using a representative Step-by-Step reasoning technique, Chain-of-Thought.[17]
- Pre-trained Medical LLM: Prompting patient's EHR data by using a representative, open-source pre-trained medical LLM, HuatuoGPT2-7B.[21]
- KG-Augmented Prompting: Prompting LLM with patient's EHR data and relevant KG knowledge, following methodologies similar to Dr. Knows,[13] medIKAL,[14] and ICP.[15]
- Iterative Prompting: Prompting LLM with patient's EHR data and multi-stage reasoning, following methodologies similar to KERAP, but with no information from KG.

Implementation Details. We implement our empirical study in Python, leveraging libraries such as Hugging Face,

OpenAI, and Scikit-learn. For the zero-shot diagnosis experiments, we primarily use the Azure OpenAI Service with the GPT-4o-mini model, except for the baseline method "Pre-trained Medical LLM", which utilizes the specialized medical model HuatuoGPT2-7B. All evaluated methods are not specifically trained with our four curated datasets. Performance is evaluated using two standard classification metrics: Accuracy (ACC) and F1-score. Considering dataset imbalance, the F1-score is computed as a weighted average across both classes. Each experiment is repeated five times to ensure reliability, with the mean and standard deviation reported. By utilizing Azure OpenAI, we ensure secure processing of sensitive patient data, adhering to regulatory standards such as HIPAA and GDPR. [1]

Zero-shot Diagnosis Results.

Table 1: Zero-shot Performance (%) of LLM-based diagnosis prediction methods (mean±std across five runs).

Method	PSCI (Prevalence:22.30%)		PNA (Prevalence:21.71%)		CKD (Prevalence:25.47%)		CHF (Prevalence:39.20%)	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1
Direct Prompting	21.31±0.58	7.78±0.30	33.77±2.11	32.94±2.50	33.46±0.55	25.41±1.02	46.15±0.92	36.32±1.79
Step-by-Step Prompting	22.22±0.62	7.89±0.41	32.59±1.24	30.56±1.64	32.70±0.87	24.21±1.68	43.55±1.63	37.51±1.77
Pre-trained Medical LLM	28.90±0.56	25.98±0.66	27.43±0.90	22.16±1.27	39.55±1.16	37.07±1.48	41.32±0.31	33.06±0.82
KG-Augmented Prompting	27.78±0.56	23.98±0.82	70.88±0.30	70.54±0.42	47.01±0.62	46.20±0.81	47.02±0.61	38.20±1.01
Iterative Prompting	24.01±1.39	15.19±0.97	66.61±2.69	67.77±1.46	70.60±2.20	72.31±2.05	58.45±2.62	57.67±2.30
KERAP (Ours)	**72.44±0.71**	**68.98±0.74**	**74.24±0.39**	**71.49±0.18**	**76.16±0.45**	**77.42±0.43**	**70.80±1.53**	**71.06±1.37**

The zero-shot diagnosis prediction results of different LLM-based methods across our four EHR datasets and five runs are shown in Table 1. Baseline methods, including Direct Prompting, Step-by-Step Prompting, and Pre-trained Medical LLM, perform poorly, with Accuracy and F1-score all below 50%. A key reason is that LLMs are often ineffective for these diseases, blindly misclassifying most or even all cases as high-risk. This bias is especially problematic given that disease prevalence in the datasets is below 40%, rendering such predictions impractical and clinically uninformative. Moreover, these methods lack structured medical reasoning, preventing them from systematically processing and inferring relationships between medical concepts in alignment with clinical decision-making. They also generate hallucinations, further degrading their performance.

In contrast, methods incorporating KG knowledge (KG-Augmented Prompting) and multi-stage reasoning (Iterative Prompting) achieve better results, particularly on CKD and PNA datasets, indicating that structured knowledge and more advanced reasoning help mitigate the limitations of simpler baselines. Finally, our proposed model, KERAP, outperforms all methods across all datasets, demonstrating its effectiveness in zero-shot diagnosis. This improvement can be attributed to two key design choices: leveraging KG knowledge and enhancing reasoning through a multi-stage process, which reduce hallucinations and encourage LLMs to reason more like human clinicians. Additionally, with guidance from KG-extracted positive and negative knowledge, KERAP produces more accurate and reliable diagnoses.

Cost and Efficiency.

Table 2: Comparison of cost and efficiency of different LLM-based diagnosis prediction methods.

Method	Total Tokens (million)	Token Cost ($)	Time Cost (h)	ACC Performance (%)	F1 Performance (%)
Direct Prompting	19.95	3.19	8.88	33.67	25.61
Step-by-Step Prompting	20.49	3.35	9.61	32.77	25.04
Pre-trained Medical LLM	21.03	0.00	0.69	34.30	29.56
KG-Augmented Prompting	39.17	6.07	9.03	48.17	44.73
Iterative Prompting	42.22	6.73	17.74	54.92	53.24
KERAP (Ours)	72.05	11.20	17.59	73.41	72.24

To evaluate the cost and efficiency of KERAP, we summarize the token cost, time cost, and performance of various methods across four EHR datasets in Table 2. Reported costs represent the total across all datasets, averaged over five runs. Accuracy and F1 scores are averaged across the four datasets. Among the methods, Pre-trained Medical LLM incurs the lowest token and time costs, as it is open-source and runs inference locally. For other methods, token costs are estimated based on GPT-4o-mini pricing [2]. Among the GPT-based approaches, Direct Prompting and Step-

[1]The code, retrieved knowledge, and detailed prompts of this paper are available at https://github.com/constantjxyz/KERAP.

[2]The API price is from https://platform.openai.com/docs/pricing

by-Step Prompting require the fewest tokens and have the lowest costs. In contrast, KG-Augmented Prompting and Iterative Prompting involve more complex reasoning, leading to higher token usage and longer computation times. Our proposed method, KERAP, exhibits the highest token consumption (72.05 million), token cost ($11.20), and time cost (17.59h), reflecting its more complex reasoning. However, given the large patient population across the four datasets (over 44,000 in total) and its strong performance, KERAP remains a practical and valuable approach for real-world applications.

Impact of LLM Choice.

To further evaluate the performance of different diagnosis prediction methods, we conduct experiments using various LLMs, replacing GPT-4o-mini with GPT-3.5-turbo and GPT-4o. Taking all four datasets' performance into consideration, as shown in Figure 2 (GPT-3.5-turbo), Figure 3 (GPT-4o), and Table 1 (GPT-4o-mini), we observe that KERAP consistently outperforms all other methods in both Accuracy and F1-score on all four datasets, demonstrating superior predictive capability and robustness across different LLMs. Additionally, GPT-3.5-turbo and GPT-4o generally underperform compared to GPT-4o-mini, reinforcing our decision to adopt GPT-4o-mini as the primary model for zero-shot diagnosis. The performance gap is likely due to GPT-3.5-turbo's limited reasoning ability and knowledge base, while GPT-4o classifies an excessive number of cases as high-risk in certain datasets (e.g., PSCI and CHF), resulting in biased predictions.

Moreover, as shown in Table 3 (GPT-3.5-turbo and GPT-4o) and Table 2 (GPT-4o-mini), both GPT-3.5-turbo and GPT-4o incur significantly higher monetary and time costs compared to GPT-4o-mini [3]. Considering the trade-off between performance, cost, and efficiency, GPT-4o-mini stands out as the most practical and effective choice for zero-shot diagnosis prediction on our datasets, further validating our selection.

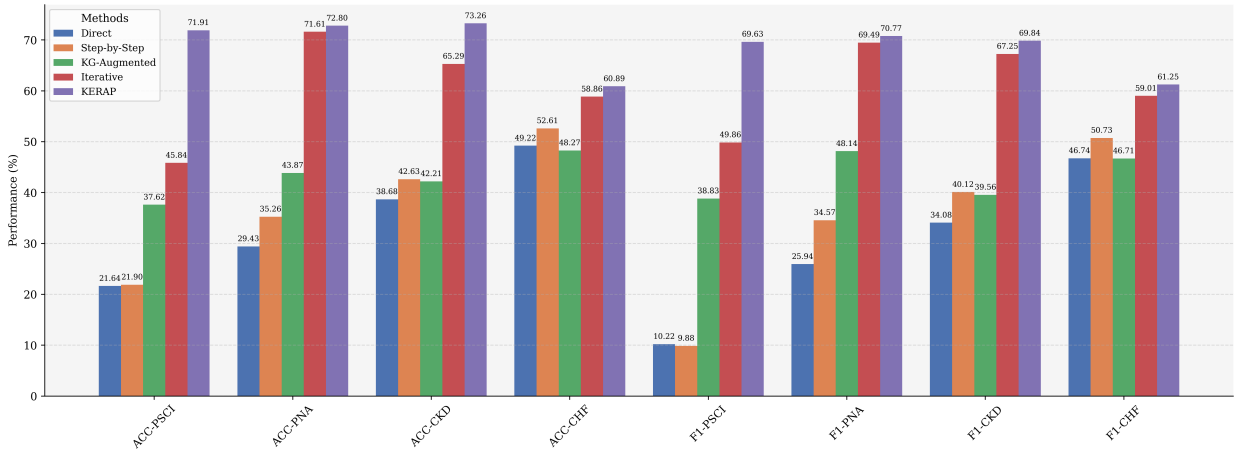


Figure 2: Zero-shot performance (%) of different LLM-based diagnosis prediction methods using *GPT-3.5-turbo*. The blue, orange, green, red, and purple bars represent the results of five compared methods: Direct Prompting, Step-by-Step Prompting, KG-Augmented Prompting, Iterative Prompting, and KERAP, respectively.

Table 3: Comparison of cost and efficiency of different LLM-based diagnosis prediction methods.

Method	GPT-3.5-turbo			GPT-4o		
	Total Tokens (million)	Token Cost ($)	Time Cost (h)	Total Tokens (million)	Token Cost ($)	Time Cost (h)
Direct Prompting	19.34	7.38	8.40	19.55	51.74	10.42
Step-by-Step Prompting	23.53	12.17	9.76	20.42	54.77	12.11
KG-Augmented Prompting	42.22	20.00	11.25	39.25	101.57	12.05
Iterative Prompting	48.40	25.19	18.21	42.30	112.79	21.67
KERAP (Ours)	82.77	42.43	20.98	75.11	196.48	23.64

[3]Pricing from https://platform.openai.com/docs/pricing

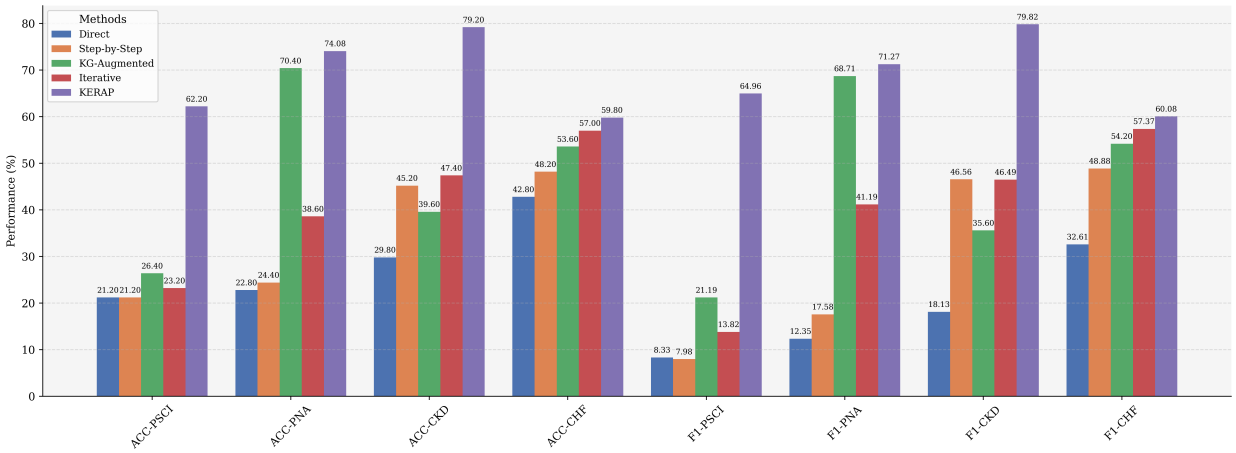
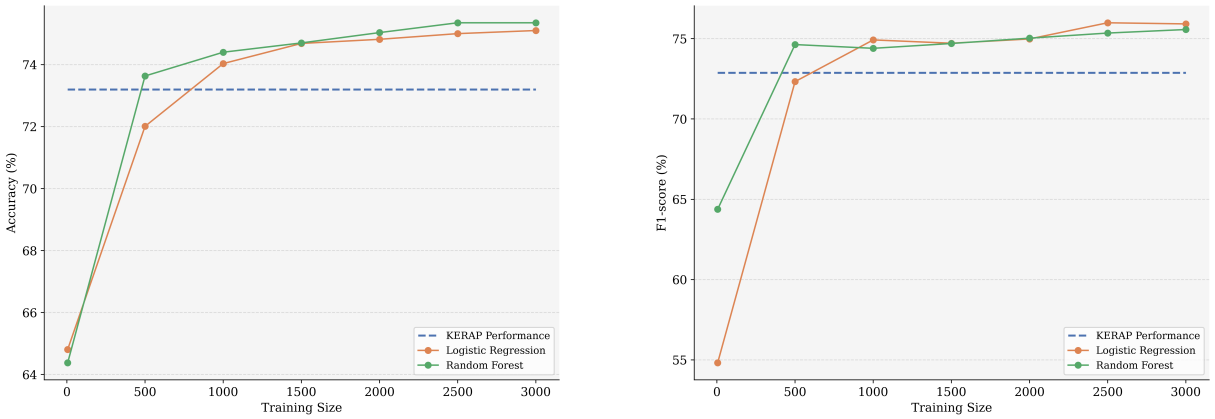


Figure 3: Zero-shot performance (%) of different LLM-based diagnosis prediction methods using *GPT-4o*. The blue, orange, green, red, and purple bars represent the results of five compared methods: Direct Prompting, Step-by-Step Prompting, KG-Augmented Prompting, Iterative Prompting, and KERAP, respectively.

Comparison with Supervised Learning Methods.

To further assess the performance of KERAP, we compare it with two representative supervised learning methods: logistic regression and random forest. We evaluate these models across varying training sample sizes to measure their effectiveness relative to KERAP. For a fair comparison, we hold out a fixed test set of 500 samples for all methods and report the average Accuracy and F1-score across four datasets. The results, shown in Figure 4, indicate that logistic regression and random forest achieve slightly higher performance than KERAP (73.20% in ACC, 72.88% in F1), surpassing it by approximately 2%. However, this improvement comes at the cost of requiring a substantial amount of high-quality labeled data—typically 500–1000 samples. Obtaining expert-labeled medical records is both costly and time-consuming. In contrast, KERAP operates as a zero-shot diagnosis prediction method, eliminating the need for labeled training data while achieving performance comparable to supervised models trained on large labeled datasets. This makes KERAP a scalable and efficient solution, particularly for diseases where labeled data is scarce.



(a) ACC **(b)** F1

Figure 4: Performance comparison of KERAP with supervised learning methods in terms of Accuracy (a) and F1-score (b). In the figures, the blue horizontal dashed line indicates the diagnosis prediction performance of KERAP, which operates without any training samples as *zero-shot* prediction. The green dotted line represents the performance of random forest across different training sample sizes, while the orange dotted line represents logistic regression under varying training samples.

Case Studies.

To demonstrate the effectiveness of KERAP, we present two diagnosis prediction cases - Case A and Case B. Each case includes the prompt, prediction results, and reasoning process generated by GPT-4o-mini. We analyze the cases and advantages of KERAP from the perspectives described below.

- **Zero-shot capability:** Unlike supervised methods, KERAP requires no training or fine-tuning. It directly utilizes patient data to make predictions for unseen cases, as demonstrated in our examples.
- **Factuality guided by KG:** KERAP enhances factuality through KG-grounded reasoning:
  - *KG knowledge utilization:* Compared to EHR-only methods (such as Direct Prompting), KERAP leverages structured KG knowledge to guide reasoning. In Case A, KG helps correct an incorrect prediction; in Case B, it assists in identifying a high-risk patient (see words in yellow).
  - *Hallucination reduction:* LLMs may hallucinate, but KERAP corrects these errors through KG-based verification. In Case A, a hallucinated statement in Prediction I is corrected in Stage II using KG-derived evidence (see words in green).
- **Augmented reasoning structure:** KERAP goes beyond baseline reasoning methods by incorporating structured and multi-stage reasoning:
  - *Structured reasoning:* KERAP mirrors human-like diagnostic reasoning. It summarizes patient information, retrieves relevant knowledge from the KG for inclusion criteria, prompts the LLM for prediction, and incorporates exclusion criteria to refine the decision.
  - *Multi-agent collaboration:* Unlike methods that merely extract KG paths as text prompts, KERAP employs a more sophisticated medical reasoning process. It utilizes a multi-agent framework that decomposes complex medical reasoning into specialized tasks, allowing collaboration between different agents. In both cases, the prompts integrate information from the KG, extracted by the linkage and retrieval agents, which collaborate with the prediction agent.
  - *Multi-stage reasoning:* KERAP adopts a multi-stage conversational reasoning approach to enhance prediction robustness. In both cases, the prediction follows a two-stage process, encouraging the LLM to consider different perspectives and refine its reasoning.
- **Robust prediction performance:** KERAP generates accurate predictions and reasoning aligned with ground truth. In Case A, Stage I's incorrect prediction is corrected in Stage II, while Case B's initial prediction remains correct and robust.

Case A (Ground Truth as NO)

Prediction Stage I Prompts

EHR Data: Nontraumatic subarachnoid hemorrhage; Other cerebrovascular diseases; Other functional intestinal disorders; acetaminophen-oxycodone; docusate; nimodipine; ondansetron; pravastatin.
Question: Will the patient develop "Post-stroke cognitive impairment" at the next visit?
Guidance from KG: Key diagnosis factors for PSCI include: Structural brain damage, neuroimaging evidence, and biochemical markers...

Prediction I: YES

Reasoning: The patient has been diagnosed with nontraumatic subarachnoid hemorrhage and other cerebrovascular diseases, which are significant risk factors for PSCI ... Additionally, **pravastatin** is linked to cognitive decline, further increasing the risk.

Prediction Stage II Prompts

Instruction: Check your prediction cautiously.
Guidance from KG: Unrelated factors in diagnosing PSCI include a broad range of environmental toxins, pharmaceuticals, industrial chemicals, and biological compounds that do not contribute to cognitive dysfunction treatment or diagnosis. These factors include heavy metals (e.g., lead, cadmium, mercury), industrial pollutants (e.g., polychlorinated biphenyls, dioxins, benzene), pharmaceuticals (e.g., metformin, haloperidol, tacrine, **statins**)...

Prediction II: NO

Reasoning: While the patient has cerebrovascular diseases, the provided information does not confirm key diagnostic factors for PSCI. The diagnosis requires severe structural brain damage, biochemical markers, and neuroimaging evidence, none of which are explicitly mentioned. Additionally, **pravastatin** and nimodipine do not directly contribute to PSCI risk. Given the standards provided, there is insufficient evidence to definitively predict PSCI.

Case B (Ground Truth as YES)

Prediction Stage I Prompts

EHR Data: Nephritis and nephropathy, not specified as acute or chronic, with lesion of proliferative **glomerulonephritis**; **Angina** decubitus; Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled; Background diabetic retinopathy; Cardiovascular Surgery Service...
Question: Will the patient develop "Chronic kidney disease" at the next visit?
Guidance from KG: Chronic kidney disease (CKD) is a progressive condition characterized by declining kidney function, often leading to kidney failure. It shares similarities with conditions such as membranous **glomerulonephritis,** hypertension, and focal segmental glomerulosclerosis. Key diagnostic indicators include ... Cardiovascular complications, including acute coronary syndrome, **angina**, heart murmurs, and high cardiac output, further indicate CKD progression.

Prediction I: YES

Reasoning: Nephritis and nephropathy with proliferative glomerulonephritis indicates kidney involvement and is a potential precursor to CKD if it progresses. Diabetes mellitus (Type II) is a major risk factor for CKD, as diabetic nephropathy is a leading cause of chronic kidney disease. Cardiovascular conditions (coronary atherosclerosis, angina, bypass surgery, PSVT) also indicate kidney disease since cardiovascular disease and CKD often coexist and contribute to each other's progression. Diabetic microvascular complications (like retinopathy) are closely linked to kidney damage. Given the presence of glomerulonephritis, diabetes, and cardiovascular complications, the patient is at a high risk of developing CKD.

Prediction Stage II Prompts

Instruction: Check your prediction cautiously.
Guidance from KG: Unrelated factors in diagnosing CKD include conditions such as hypertension ...

Prediction II: YES

Reasoning: The patient has nephritis and nephropathy, which indicate kidney damage, and proliferative glomerulonephritis, a condition that can lead to progressive kidney dysfunction. Additionally, diabetes mellitus and coronary atherosclerosis are well-known risk factors for CKD. Given these conditions, the patient is at a significant risk of developing chronic kidney disease.

Conclusion

In this work, we proposed KERAP, a KG-enhanced reasoning framework for zero-shot diagnosis prediction using LLMs. Our approach addresses key challenges in LLM-based medical diagnosis, including the need for structured reasoning, hallucination reduction, and improved effectiveness. By integrating a multi-agent framework consisting of a linkage agent, a retrieval agent, and a prediction agent, KERAP effectively connects patient data with structured medical knowledge and refines its diagnostic reasoning through multi-stage conversational steps. Experimental results demonstrate that KERAP enhances diagnostic accuracy compared to existing methods effectively and efficiently. Our findings highlight the potential of structured knowledge integration and iterative reasoning in medical AI, paving the way for more reliable and scalable zero-shot diagnosis prediction models. Future work may explore further optimizations in multi-agent collaboration and expanding KG-based reasoning to a broader range of medical conditions.

References

1. Ghaffar Nia N, Kaplanoglu E, Nasab A. Evaluation of artificial intelligence techniques in disease diagnosis and prediction. Discover Artificial Intelligence. 2023;3(1):5.
2. 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: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 2024. p. 6324-34.
3. Garriga R, Mas J, Abraha S, Nolan J, Harrison O, Tadros G, et al. Machine learning model to predict mental health crises from electronic health records. Nature medicine. 2022;28(6):1240-8.
4. Kumar Y, Koul A, Singla R, Ijaz MF. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. Journal of ambient intelligence and humanized computing. 2023;14(7):8459-86.
5. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80.
6. Bhasuran B, Jin Q, Xie Y, Yang C, Hanna K, Costa J, et al. Preliminary analysis of the impact of lab results on large language model generated differential diagnoses. npj Digital Medicine. 2025;8(1):166.
7. Zhou S, Xu Z, Zhang M, Xu C, Guo Y, Zhan Z, et al. Large language models for disease diagnosis: A scoping review. arXiv preprint arXiv:240900097. 2024.
8. Cui H, Shen Z, Zhang J, Shao H, Qin L, Ho JC, et al. Llms-based few-shot disease predictions using ehr: A novel

approach combining predictive agent reasoning and critical agent instruction. arXiv:240315464. 2024.
9. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature. 2024;630(8017):625-30.
10. Anibal JT, Huth HB, Gunkel J, Gregurick SK, Wood BJ. Simulated misuse of large language models and clinical credit systems. NPJ Digital Medicine. 2024;7(1):317.
11. Pan S, Luo L, Wang Y, Chen C, Wang J, Wu X. Unifying large language models and knowledge graphs: A roadmap. IEEE Transactions on Knowledge and Data Engineering. 2024;36(7):3580-99.
12. Pan J, Razniewski S, Kalo JC, Singhania S, Chen J, Dietze S, et al. Large Language Models and Knowledge Graphs: Opportunities and Challenges. Transactions on Graph Data and Knowledge. 2023.
13. Gao Y, Li R, Croxford E, Caskey J, Patterson BW, Churpek M, et al. Leveraging Medical Knowledge Graphs Into Large Language Models for Diagnosis Prediction: Design and Application Study. JMIR AI. 2025;4:e58670.
14. Jia M, Duan J, Song Y, Wang J. medIKAL: Integrating Knowledge Graphs as Assistants of LLMs for Enhanced Clinical Diagnosis on EMRs. In: 31st International Conference on Computational Linguistics; 2025. p. 9278-98.
15. Wu J, Wu X, Yang J. Guiding clinical reasoning with large language models via knowledge seeds. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence; 2024. p. 7491-9.
16. Xie Y, Lu J, Ho J, Nahab F, Hu X, Yang C. PromptLink: Leveraging Large Language Models for Cross-Source Biomedical Concept Linking. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2024. p. 2589-93.
17. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems. 2022;35:24824-37.
18. Renze M, Guven E. Self-reflection in llm agents: Effects on problem-solving performance. 2nd International Conference on Foundation and Large Language Models. 2024.
19. Chiu WHK, Ko WSK, Cho WCS, Hui SYJ, Chan WCL, Kuo MD. Evaluating the Diagnostic Performance of Large Language Models on Complex Multimodal Medical Cases. Journal of Medical Internet Research. 2024;26:e53724.
20. Sonoda Y, Kurokawa R, Hagiwara A, Asari Y, Fukushima T, Abe O. Structured clinical reasoning prompt enhances LLM's diagnostic capabilities in diagnosis please quiz cases. Japanese Journal of Radiology. 2024:1-7.
21. Chen J, Wang X, Ji K, Gao A, Jiang F, Chen S, et al. Huatuogpt-ii, one-stage training for medical adaption of llms. arXiv preprint arXiv:231109774. 2023.
22. Xu R, Shi W, Yu Y, Zhuang Y, Jin B, Wang MD, et al. 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); 2024. p. 754-65.
23. Wang Z, Zhu Y, Gao J, Zheng X, Zeng Y, He Y, et al. RetCare: Towards Interpretable Clinical Decision Making through LLM-Driven Medical Knowledge Retrieval. In: Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare; 2024. .
24. Jin M, Yu Q, Shu D, Zhang C, Fan L, Hua W, et al. Health-LLM: Personalized retrieval-augmented disease prediction system. arXiv preprint arXiv:240200746. 2024.
25. Zhu Y, Ren C, Wang Z, Zheng X, Xie S, Feng J, et al. Emerge: Integrating rag for improved multimodal ehr predictive modeling. arXiv preprint arXiv:240600036. 2024.
26. Ben Shoham O, Rappoport N. Cpllm: Clinical prediction with large language models. PLOS Digital Health. 2024;3(12):e0000680.
27. Liu F, Shareghi E, Meng Z, Basaldella M, Collier N. Self-alignment pretraining for biomedical entity representations. arXiv preprint arXiv:201011784. 2020.
28. 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):1-9.
29. Xu R, Ali MK, Ho JC, Yang C. Hypergraph transformers for ehr-based clinical predictions. AMIA Summits on Translational Science Proceedings. 2023;2023:582.
30. Xie Y, Nahab F, Ge Y, Wu Y, Saurman J, Yang C, et al. Predicting Post-Stroke Cognitive Impairment (PSCI) Using Multiple Machine Learning Approaches. Stroke. 2025;56(Suppl_1):AWP175-5.
31. Su C, Hou Y, Zhou M, Rajendran S, Maasch JR, Abedi Z, et al. Biomedical discovery through the integrative biomedical knowledge hub (iBKH). Iscience. 2023;26(4).