Towards Explainable Diagnosis: A Self-learned Explanatory Knowledge Base Approach

Anonymous EMNLP submission

Abstract

Explainable diagnosis requires to the process of reaching diagnostic conclusions with clear rationale that links a patient's clinical phenomenon to authoritative medical knowledge. While large language models (LLMs) show promise in supporting explainable diagnosis, they often fall short due to insufficient diagnostic knowledge. To address this limitation, we propose **S**elf-learned **E**xplainable **K**nowledge **A**ugmented **D**iagnosis (**SEKAD**), a unified LLM-based framework for faithful and explainable diagnosis. Our approach builds a high-quality diagnostic knowledge base through a record-driven explanation learning paradigm, as well as applies this knowledge via an explanation-based diagnostic process that ensures faithful inference. Experiments on the DiReCT and JAMA benchmarks show that **SEKAD** consistently outperforms strong baselines across the metrics. In particular, **SEKAD** achieves absolute improvement of 12.4% in the completeness of explanation metric over the best existing methods, highlighting its effectiveness in enhancing diagnostic explainability.

1 Introduction

Efficient diagnosis enables earlier interventions, improving patient prognosis by preventing disease progression or complications (Agha et al., 2022). Automatic diagnosis can significantly improve diagnostic efficiency, an advantage that has been well demonstrated in recent years by automatic diagnostic systems driven by machine learning (Ahsan et al., 2022) and deep learning (Aggarwal et al., 2021). In automatic diagnosis, diagnostic accuracy is important, and explainable diagnostic results are key to building trust.(Edin et al., 2024) Large language models (LLMs) (Zhou et al., 2023) are considered as a potential choice for building more explainable automated diagnostic tools due to their ability to generate coherent natural language output (Singhal et al., 2023). However, LLMs still have limitations in the quality of diagnostic explanations due to lack of specialized medical knowledge, especially concerning the explanatory aspect (Ji et al., 2023). A promising direction to bridge this knowledge gap is to leverage systematic, updatable medical knowledge sources to guide LLM-based explainable automated diagnosis.

The needs of explanatory diagnosis knowledge. Human physicians rely on medical guidelines as diagnostic references to address complex cases. (National Academies of Sciences et al., 2015) When these knowledge sources are inherently explainable, they can mitigate incomplete knowledge coverage and biases inherent in the limitations of LLMs' pretraining data. DiReCT (Wang et al., 2024a) improves LLMs' faithfulness of explanations by using a knowledge base constructed by experts based on guidelines, demonstrating that LLMs can benefit from manually crafted external knowledge sources to enhance explainable diagnostic capabilities. Thus, defining and building such explanatory knowledge bases is a key strategy for advancing explainable automatic diagnosis.

The construction of explanatory diagnosis knowledge. Medical textbooks, clinical guidelines, and academic literature constitute extensive and readily accessible repositories of diagnostic knowledge. Despite their value, these sources are inherently fragmented and independently structured, making effective utilization a non-trivial task, even for human clinicians, who typically master them only through prolonged training and clinical experience. (Burnier, 2024) While LLMs exhibit strong capabilities in information extraction and reasoning (Xu et al., 2024), studies have shown that their performance in medical knowledge extraction remains unstable (Agrawal et al., 2022). Challenges persist in enabling LLMs to autonomously verify and refine the accuracy of the knowledge they acquire from these traditional, structured texts. In contrast, medical records provide a vast accessible data source. Although they lack explicit basic explanatory annotations, the inherent links they reveal between patients' clinical phenomena and diagnostic conclusions offer a valuable opportunity for the large-scale, automated construction of explanatory knowledge bases. Consequently, a key challenge lies in how to automatically construct such knowledge bases at scale and with high quality.

In this paper, we propose **S**elf-learned **E**xplainable **K**nowledge **A**ugmented **D**iagnosis (**SEKAD**), an explainable diagnosis framework. It consists of an explanatory knowledge base and an explanation-based diagnosis process. To automatically build a large and high-quality knowledge base, we propose **record-driven explanation self-learning** method. First, it enables LLMs to autonomously acquire explanatory diagnostic knowledge from unstructured patient records by broad medical resources, guaranteeing the quantity of the knowledge base. Furthermore, we designed the **diagnostic triangulation** mechanism, which guarantees that the acquired knowledge is supported by multiple sources and could be generalized. Diagnostic triangulation ensures the quality of the knowledge base. Building upon this knowledge base, we propose the **explanation augmented dual-phase diagnosis** method, which consists of **differential diagnosis** and **definitive diagnosis** to avoid biased use of explanatory knowledge. To validate the effectiveness of our framework, we conducted extensive experiments on two explainable diagnosis task. Our method outperforms five existing baselines across multiple explainability metrics, and surpasses the state-of-the-art method by 12.4% and 4.3% on the completeness of explanation and faithfulness of explanation metrics, respectively, in terms of explanation faithfulness. Our contributions are fourfold:

- We are the first to automatically construct an explanatory diagnostic knowledge base for explainable diagnosis. To bridge the knowledge gap in automatic explainable diagnosis, we propose **SEKAD**, which includes a method for building high-quality diagnostic knowledge via *record-driven explanation self-learning*, and a method for utilizing this knowledge through *explanation augmented dual-phase diagnosis*.
- We propose a novel **record-driven explanation self-learning** method, which ensures knowledge quantity through automatic self-learning, and guarantees quality through *diagnostic triangulation*, a mechanism that filters out misleading explanations via multi-source validation.
- To utilize structured knowledge in the diagnostic process, we introduce **explanation-augmented dual-phase diagnosis**, which mitigates the risk of over-relying on contextually bias explanations by ensuring that each diagnosis is supported by comprehensive explanation.
- Experiments on two explainable diagnostic evaluation datasets demonstrate that our method outperforms competing baselines, and achieves superior performance in explanation generation.

2 Related Works

LLM-based automatic diagnosis. LLMs in the medical domain have achieved improved diagnostic accuracy through fine-tuning with domain-specific data (Singhal et al., 2023). To enhance explainability, recent work has introduced multi-agent collaboration frameworks (Tang et al., 2023; Kim et al., 2024) that allow LLMs to exhibit detailed explainable thinking. However, such approaches face limitations due to insufficient medical knowledge. As noted in Medagents (Tang et al., 2023), the lack of reliable domain expertise in the reasoning process leads to reduced credibility of the generated explanations.

Medical knowledge-enhanced LLM. Several approaches have attempted to address this limitation by incorporating structured knowledge into LLMs. LLM-AMT (Wang et al., 2024b) enhances models using curated medical textbooks, MedRAG (Xiong et al., 2024) integrates broad-scope medical corpora, and KGARevion (Su et al., 2024) employs knowledge graphs for domain grounding. These efforts demonstrate the potential of external knowledge sources to augment the factual accuracy of LLMs. Nonetheless, current methods primarily focus on improving diagnostic performance, with limited attention to enhancing the explanatory quality of model outputs. In response to this gap, we propose the **SEKAD** framework, which constructs a knowledge base specifically designed to support explanation-oriented augmentation for LLMs in automated medical diagnosis.

3 Method

In this section, we introduce **SEKAD**, an explainable automatic diagnosis framework augmented by self-learned knowledge. **SEKAD** consists of two parts: (1) **Record-driven explanation self-learning:** Given a large amount of unstructured medical records, autonomously mining explanatory diagnostic knowledge. (2) **Explanation augmented dual-phase diagnosis**: Given a patient's clinical notes, under the guidance of explanatory knowledge, the diagnosis executor first performs differential diagnosis to identify the likely diagnosis, and then generates explanations linking the patient's clinical phenomena to the diagnosis in the definitive diagnosis phase.

3.1 Record-driven Explanation Self-learning

Unstructured medical records, including patient reports and clinical notes, reflect numerous connections between clinical phenomena and diagnostic conclusions. However, the underlying explanations for these connections are dispersed across authoritative medical knowledge sources such as medical textbooks, clinical guidelines, and academic literature. Record-driven explanation self-learning aims to automatically identify these connections from medical records and learn the corresponding diagnostic knowledge from the medical knowledge sources to build a structured explanatory knowledge base.

3.1.1 Explanatory Knowledge Base

During the process of record-driven explanation self-learning, an explanatory diagnostic knowledge base B is incrementally constructed. This knowledge base consists of structured knowledge units, each capturing a link between a patient's clinical phenomenon and a corresponding diagnosis, grounded by an explanatory rationale. Formally, a knowledge unit k is defined as a tuple (p, e, d), where:

- p: represents a single clinical phenomenon observed in the patient, such as "*dizziness*".
- d: represents the diagnosis for the patient, at any level of granularity.
- e: represents a text-based explanation linking the clinical phenomenon p to the diagnosis d.

The explanatory diagnostic knowledge base B is defined as a collection of knowledge units k, where $B = \{k_1, k_2, \ldots, k_n\}$.

To ensure that the explanatory diagnostic knowledge base B provides faithful diagnostic insights, the knowledge unit k must satisfy the following principles:

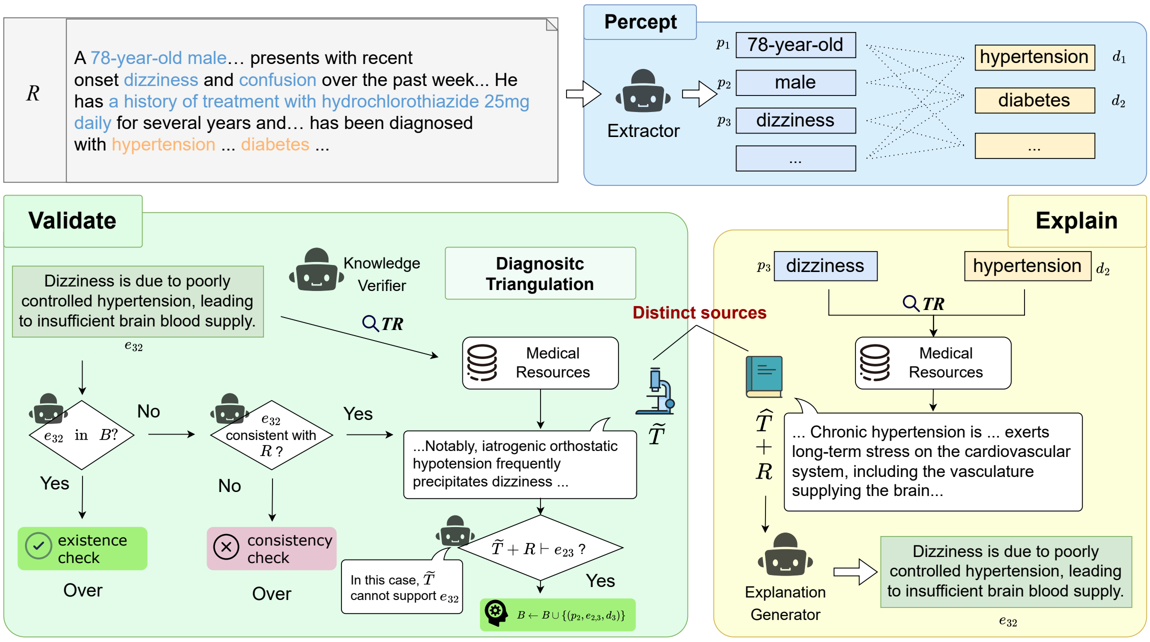
Unit Specificity: Each knowledge unit k must address a single primary clinical phenomenon p. Although concomitant phenomena may be referenced within the explanation e, the core focus remains singular, for example, focusing a unit k solely on 'fever' rather than requiring both 'fever' and 'cough', as knowledge aggregating multiple distinct phenomena would inherently possess a more restricted scope.

Self-contained: For explanation e, all abbreviations of medical terms must be expanded to their full, unambiguous nomenclature. For example, ambiguous abbreviations like "MS" (which could refer to "Multiple Sclerosis" or "Mitral Stenosis") must be explicitly expanded within e to avoid potential misinterpretation. This expansion rule applies exclusively to e, not to p or d.

Generalization: For explanation e, the clinical phenomenon p is represented as a generalized clinical concept rather than a concrete patient case. For example, a specific observation such as "heart rate of 120 bpm" should be transformed into a general clinical descriptor as "tachycardia". This ensures that the knowledge unit correctly captures the clinical concept, making it applicable to all specific situations that fall within that concept.

Faithfulness: Each explanation e should be robustly supported by evidence from multiple, independent and authoritative medical knowledge sources, ensuring the faithfulness of the $p \leftrightarrow d$ association and thus preventing spurious associations.

The explanatory diagnostic knowledge base B serves as a structured and verifiable repository of validated diagnostic knowledge. During diagnosis, relevant knowledge units $k = (p, e, d)$ are retrieved to support explanation-based diagnosis. As illustrated in Figure 1, each unit is incorporated into B through a sequential process of **percept**, **explain**, and **validate**, which enables dynamic updates and ensures knowledge quality.



R
A 78-year-old male... presents with recent onset dizziness and confusion over the past week... He has a history of treatment with hydrochlorothiazide 25mg daily for several years and... has been diagnosed with hypertension ... diabetes ...
Percept
Extractor
p_1 78-year-old
p_2 male
p_3 dizziness
...
hypertension d_1
diabetes d_2
...
Validate
Dizziness is due to poorly controlled hypertension, leading to insufficient brain blood supply.
e_{32}
Knowledge Verifier
Diagnositc Triangulation
$\mathcal{TR}$
Medical Resources
e_{32} in B?
No
Yes
e_{32} consistent with R ?
Yes
No
existence check
Over
consistency check
Over
...Notably, iatrogenic orthostatic hypotension frequently precipitates dizziness ...
$\tilde{T}$
$\tilde{T} + R \vdash e_{23}$?
Yes
In this case, $\tilde{T}$ cannot support e_{32}
$B \leftarrow B \cup \{(p_2, e_{2,3}, d_3)\}$
Distinct sources
Explain
p_3 dizziness
hypertension d_2
$\mathcal{TR}$
Medical Resources
$\hat{T}$ + R
... Chronic hypertension is ... exerts long-term stress on the cardiovascular system, including the vasculature supplying the brain...
Explanation Generator
Dizziness is due to poorly controlled hypertension, leading to insufficient brain blood supply.
e_{32}

Figure 1: Overview of record-driven explanation self-learning. An initial explanation links *dizziness* to poorly controlled *hypertension* based on disease-centered sources. Diagnostic triangulation with pharmacological references reveals that dizziness may instead result from side effects of antihypertensive medications. This mechanism identifies conflicting evidence and filters out potentially misleading diagnostic links.

3.1.2 Percept

Based on the original patient's medical record R, an LLM-based extractor[1]. is instructed to identify documented diagnoses ($d_1, ..., d_m \in D$) and single clinical phenomena ($p_1, ..., p_n \in P$) as exact textual spans. These spans are deliberately kept in their original form at this stage to achieve more semantically relevant retrieval when diagnosing from medical records. Each span is retained only if it matches the string in R and the similarity exceeds a predefined threshold. This identification strategy decomposes the patient's findings into individual phenomena p_i, making each p_i a basis for potentially linking to identified diagnoses d_j. By ensuring each p_i serves as the single phenomenon p in $k = (p, e, d)$, this action guarantees **unit specificity** for k.

3.1.3 Explain

The **explain** action aims to find explainable clinical knowledge that links clinical phenomena p with diagnoses d from relevant authoritative medical knowledge sources. Its input includes a specific clinical phenomenon p identified from the patient's medical record R, and the corresponding diagnosis d. Together, p and d are concatenated to form the search query. Using this query, a text retriever $\mathcal{TR}$ searches for a relevant subset from the medical knowledge sources T. Subsequently, an explanation generator[2] utilizes the retrieved subset $\hat{T}$ as reference to generate the explanation e for pair (p, d), guided by explicit instruction prompts designed to ensure adherence to the principles of **self-contained** and **generalization**. When the retrieved subset $\hat{T}$ is insufficient to support a detectable association between clinical phenomena p and diagnosis d, the generator does not produce an explanation.

3.1.4 Validate

Ensuring the **faithfulness** of each knowledge unit $k = (p, e, d)$ produced by these actions is paramount for a reliable diagnostic knowledge base B, especially given the known limitations of LLM in generating faithful medical explanations. Based solely on their initial source, some initially generated units contain incorrect $p \leftrightarrow d$ associations or associations valid only in specific contexts. To address this crucial requirement, we introduce the **diagnostic triangulation** mechanism designed to verify the $p \leftrightarrow d$ association and its explanation e against multiple, independent, authoritative medical knowledge sources.

Specifically, the **validate** action is performed by

[1] The prompt is shown in E.1.

[2] The prompt is shown in E.2.

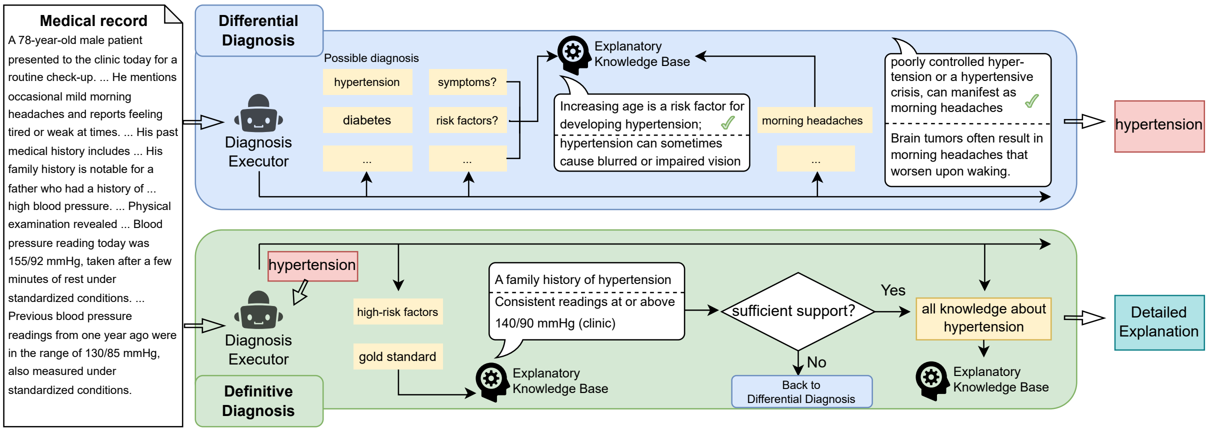

Medical record
A 78-year-old male patient presented to the clinic today for a routine check-up. ... He mentions occasional mild morning headaches and reports feeling tired or weak at times. ... His past medical history includes ... His family history is notable for a father who had a history of ... high blood pressure. ... Physical examination revealed ... Blood pressure reading today was 155/92 mmHg, taken after a few minutes of rest under standardized conditions. ... Previous blood pressure readings from one year ago were in the range of 130/85 mmHg, also measured under standardized conditions.
Differential Diagnosis
Diagnosis Executor
Possible diagnosis
hypertension
symptoms?
diabetes
risk factors?
...
...
Explanatory Knowledge Base
Increasing age is a risk factor for developing hypertension; ✓
hypertension can sometimes cause blurred or impaired vision
morning headaches
...
poorly controlled hypertension or a hypertensive crisis, can manifest as morning headaches ✓
Brain tumors often result in morning headaches that worsen upon waking.
hypertension
hypertension
Diagnosis Executor
Definitive Diagnosis
high-risk factors
gold standard
A family history of hypertension
Consistent readings at or above 140/90 mmHg (clinic)
Explanatory Knowledge Base
sufficient support?
Yes
No
Back to Differential Diagnosis
all knowledge about hypertension
Explanatory Knowledge Base
Detailed Explanation

Figure 2: Overview of explanation augmented dual-phase diagnosis

an LLM-driven **knowledge verifier**, which leverages deductive reasoning capabilities (Srivastava et al., 2022). This validation is structured as a three-stage process:

Validation against existing knowledge B: This initial stage aims to prevent redundancy. The knowledge verifier checks if the generated explanation e for a given (p, d) pair aligns with knowledge already validated and stored in B. For each such (p, d) pair, the knowledge verifier retrieves the k-top existing explanations from B and compares them with e. If e is considered sufficiently similar to any of the retrieved explanations, the knowledge unit k is considered validated and passes this stage. It is not added again to B to avoid duplication.

Consistency with medical record R: In the second stage, the knowledge verifier assesses the internal consistency between the generated explanation e and the original patient record R, ensuring that e does not conflict with other conditions documented in R.

Diagnosis triangulation by external evidence $\tilde{T}$: Under the diagnosis triangulation mechanism, knowledge validation is framed as a natural language inference task, leveraging external evidence $\tilde{T}$ to assess the validity of a candidate knowledge unit k. A concrete illustration of this mechanism is provided in Figure 1, where conflicting evidence from pharmacological literature challenges an initially misleading explanation. The external set $\tilde{T}$ is obtained by using the explanation e as a query to retrieve heterogeneous knowledge not overlapping with the original source $\hat{T}$. In this task, the retrieved evidence from $\tilde{T}$, together with the patient record R, constitutes the premise, while the candidate knowledge unit k serves as the hypothesis. The verifier then determines whether the premise logically supports the hypothesis.

A knowledge unit k is validated and subsequently incorporated into the knowledge base B only when it has passed the internal consistency check against R and is also judged to be supported by external knowledge under the diagnosis triangulation process.

3.1.5 Reinforcement Learning via Direct Preference Optimization

To jointly optimize the extraction of clinical phenomena (**percept**) and the generation of faithful explanations (**explain**), and to align with the LLM's capability of self-learning explanatory knowledge, we adopt a reinforcement learning framework based on direct preference optimization (DPO) (Rafailov et al., 2023). This enables the LLM agent π_θ to learn from preference data $\mathcal{D}^{\pm}$, where each sample consists of a context x and a preferred–less preferred pair (y^+, y^-).

We construct a preference dataset $\mathcal{D}^{\pm}$, where each instance consists of a context x, a preferred output y^+, and a less-preferred output y^-. For the **explain** action, the context x includes the patient record and the specific phenomenon–diagnosis pair under consideration. A generated knowledge unit k^+_{ij} is considered preferred if it successfully passes the validation process, receiving a binary reward of $W_E = 1$, whereas an alternative k^-_{ij} that fails validation with $W_E = 0$ is treated as less preferred.

For the **percept** action, the context x is the patient record, and the preference is established between two sets of extracted phenomena. A set P^+ is preferred if it leads to a higher aggregated downstream reward $W_P(P^+)$, in comparison to a set P^- associated with a lower reward $W_P(P^-)$.

Formally, for each (x, y^+, y^-), the training objective is:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma \left(\hat{r}_\theta(x, y^+) - \hat{r}_\theta(x, y^-) \right), \quad (1)$$

where $\hat{r}_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$ is the implicit reward, and β is a temperature parameter.

This approach unifies percept and explain within the same optimization framework, improving the efficiency of explanation learning.

3.2 Explanation Augmented Dual-phase Diagnosis

With the accumulation of explanatory knowledge, **SEKAD** performs explainable diagnosis under the guidance of the knowledge base B. Given a patient record without a diagnostic conclusion, **SEKAD** outputs the most likely diagnosis along with a series of rationales that connect the patient's clinical phenomena to the proposed diagnosis. In this method, an LLM acts as the diagnosis executor, querying explanatory diagnostic knowledge base B through self-queries, and performing diagnosis strictly under the guidance of this knowledge.

By simulating real-world clinical workflows, we divide the diagnostic process into two complementary phases: **differential diagnosis**, which involves evaluating and narrowing down the range of potential diagnoses to improve diagnostic accuracy, and **definitive diagnosis**, which provides a detailed explanation for the identified condition.

3.2.1 Differential diagnosis

In clinical practice, differential diagnosis refers to the process by which physicians analyze specific clinical phenomena to narrow down the range of possible conditions. To implement this process, the diagnosis executor adopts a bidirectional knowledge retrieval strategy, as shown in Figure 2. First, a preliminary analysis is performed to identify a likely category of disease and initiates self-queries such as "What are common symptoms or risk factors of this disease?" Based on the retrieved knowledge unit k, if some clinical phenomena are not mentioned, the executor then reverses the querying direction by asking, "What diseases commonly present with this phenomenon?" for those not yet identified[3].

During the differential diagnosis phase, diagnosis executor does not generate full explanations for each tentative candidate diagnosis. This design constraint is intended to avoid overconfident explanations for provisional hypotheses; it helps mitigate the risk of premature diagnostic anchoring arising from excessive explanation at an early stage.

3.2.2 Definitive diagnosis

Since explanations in the differential diagnosis phase remain incomplete, the definitive diagnosis phase builds upon the initial hypothesis by performing more targeted knowledge retrieval focused on the confirmed diagnosis. At this phase, the executor issues diagnosis-centered self-queries, aiming to identify supporting evidence such as high-risk factors and diagnostic gold standards. The objective is to provide a comprehensive explanation of the patient's clinical phenomena by matching them with validated knowledge units k. Only successfully matched knowledge is used to construct the definitive explanation. If contradictions arise or sufficient supporting evidence is lacking, the diagnostic process is designed to revert to the differential diagnosis phase for further exploration.

4 Experiments

4.1 Benchmarks

4.1.1 DiReCT

The DiReCT dataset (Wang et al., 2024a) comprises 511 physician-annotated clinical notes from MIMIC-IV (Johnson et al., 2020), meticulously detailing diagnostic processes and final diagnoses. It defines an explainable diagnostic task where, given a patient's clinical record R and a graph constructed from all diagnoses $\mathcal{G}$, the model must find the path to the primary discharge diagnosis, select relevant observational phenomena p, and provide corresponding explanations e. The benchmark also provides an expert-curated knowledge graph $\mathcal{K}$, which contains guideline knowledge for each diagnostic node in $\mathcal{G}$, used as an external knowledge baseline, for example DiReCT w/ $\mathcal{K}$. Our evaluation primarily focuses on three core aspects: **accuracy of diagnosis** Acc^{cat} and Acc^{diag}, **completeness of observation** Obs^{comp}, and **faithfulness of explanation** Exp^{com} and Exp^{all}. For a detailed experimental setup and metric specifics, please refer to Appendix B.2.1.

4.1.2 JAMA Clinical Challenge

The JAMA Clinical Challenge dataset (Chen et al., 2025) comprises complex, text-based clinical cases sourced from the Journal of the American Medical

[3] The prompt is shown in E.7.

Method	Acc^{cat}	Acc^{diag}	Obs^{comp}	Exp^{com}	Exp^{all}
GPT4					
DiReCT w/ $\mathcal{G}$	0.804	0.610	0.391	0.481	0.210
DiReCT w/ $\mathcal{K}$	0.808	0.611	0.371	0.645	0.273
DeepSeek-R1					
COT	0.690	0.586	0.192	0.263	0.071
DiReCT w/ $\mathcal{G}$	0.830	0.687	0.322	0.430	0.152
DiReCT w/ $\mathcal{K}$	0.812	0.611	0.324	0.615	0.222
SEKAD	**0.889**	**0.694**	**0.405**	**0.769**	**0.316**
DeepSeek-V3					
COT	0.702	0.585	0.185	0.276	0.065
DiReCT w/ $\mathcal{G}$	0.796	0.587	0.346	0.321	0.131
DiReCT w/ $\mathcal{K}$	0.808	0.635	0.351	0.492	0.202
KGARevion	0.792	0.629	0.239	0.345	0.094
MDAgents	0.688	0.566	0.218	0.349	0.099
MedAgent	0.740	0.599	0.205	0.319	0.076
MedRAG	0.817	0.640	0.288	0.232	0.069
SEKAD	**0.847**	**0.653**	**0.400**	**0.759**	**0.312**

Table 1: Performance comparison on the DiReCT benchmark. **Bold** indicates the best result.

Association, featuring multiple-choice diagnostic questions and expert-authored explanations. For this task, models predict the most probable diagnosis and generate corresponding explanations for presented clinical cases. Performance is evaluated based on diagnostic prediction accuracy and explanation quality, using G-Eval (Liu et al., 2023) metrics, including coherence, consistency, and relevance, which are scored by an LLM-based evaluator. For further details on the dataset, task setup, and metrics specifics, please refer to Appendix B.2.2.

4.2 Baselines

We evaluate our proposed method with five distinct baselines, including the Chain-of-Thought (COT) (Wei et al., 2022) method and four leading medical-enhanced QA approaches: MedAgents (Tang et al., 2023), MDAgent (Kim et al., 2024), MedRAG (Xiong et al., 2024), and KGARevion (Su et al., 2024).

On the DiReCT benchmark, we also include the official baseline method for comparison, which consists of two configurations: $\mathcal{G}$, a diagnosis graph representing structured diagnostic relationships, and $\mathcal{K}$, which incorporates expert knowledge from diagnostic guidelines at intermediate steps of the diagnostic process.

4.3 Result

We present the evaluation results on the DiReCT benchmark in Table 1. Our method outperforms all six provided baselines and achieves improvements of **8.4%**, **1.4%**, and **4.3%** over the best-performing baseline in terms of accuracy of diagnosis, completeness of observation, and faithfulness of explanation, respectively. Notably, the significant gain in explanation faithfulness highlights our method's ability to generate clinically aligned reasoning. The high score on Exp^{com}, which measures explanation–observation consistency, further demonstrates that **SEKAD** produces explanations that closely reflect expert reasoning based on the patient's clinical presentation.

Method	Acc	$Relev.$	$Coh.$	$Consist.$
DeepSeek-V3				
COT	0.711	4.672	**4.945**	4.305
KGARevion	0.631	4.331	4.852	4.101
MDAgents	0.691	4.531	4.711	4.141
MedAgent	0.450	3.651	3.705	3.537
MedRAG	0.400	3.745	3.570	3.282
SEKAD	**0.771**	**4.672**	4.740	**4.313**

Table 2: Performance comparison on the JAMA benchmark. **Bold** indicates the best result.

We further observe that existing baselines generally underperform in explanation faithfulness. This is primarily because, under this benchmark, only explanations that correctly support the intended diagnostic target are considered valid. Baseline models tend to misinterpret evidence suggestive of a disease as confirmatory, leading to inaccurate diagnostic rationales. This highlights the effectiveness of our dual-phase diagnostic process in distinguishing between diagnostic suspicion and confirmation.

Table 2 reports performance on the JAMA Clinical Challenge dataset. **SEKAD** demonstrates strong competitiveness in diagnostic accuracy as well as in the relevance and consistency of the generated explanations compared to baselines. We also note that COT exhibits superior coherence, because it relies solely on internal reasoning without external information.

Due to the lack of imaging data, the diagnostic context in the JAMA dataset is incomplete. Under these conditions, many baseline models tend to engage in over-reasoning or fall into heuristic bias, often performing worse than the base model. In contrast, **SEKAD** maintains robust diagnostic reasoning through its structured *Differential–Definitive* two-stage explanatory framework. Among the baselines, KGAREVION benefits from a knowledge graph review mechanism that helps filter out misinformation, while MDAGENTS avoids unnecessary complexity through adaptive task decomposition. In comparison, MEDRAG, which relies on text similarity-based retrieval, is more prone to introducing irrelevant knowledge that may mislead diagnosis.

4.4 Ablation Study

As shown in Table 3, the explanatory knowledge base B plays a critical role in enhancing diagnostic performance across all metrics. Removing B results in significant drops in Acc^{diag} from 65.3% to 56.9% and in Exp^{com} from 73.1% to 50.2%, highlighting its centrality to both diagnostic accuracy and explanation faithfulness. In contrast, ablating the diagnostic triangulation mechanism causes a smaller reduction in Acc^{diag} to 63.9%, but still leads to a notable decrease in Exp^{com} to 56.8%. This underscores that while diagnostic triangulation does not directly boost classification accuracy, it plays an essential role in ensuring the faithfulness and completeness of generated explanations.

Method	Acc^{cat}	Acc^{diag}	Obs^{comp}	Exp^{com}	Exp^{all}
DeepSeek-V3					
w/o B	0.792	0.569	0.299	0.502	0.185
w/o D.T.	0.819	0.639	**0.400**	0.568	0.251
origin	**0.847**	**0.653**	**0.400**	**0.731**	**0.295**

Table 3: Ablation study results on the DiReCT benchmark. **Bold** indicates the best result. D.T. stands for diagnostic triangulation.

4.5 Impact of Knowledge Scale on Performance

The performance on accuracy of diagnosis, completeness of observations, and faithfulness of explanation is shown, respectively, in Figure 3. Overall, increasing the scale of the knowledge base leads to consistent improvements, particularly in faithfulness of explanation, which grows from 17% to 29%, demonstrating that richer knowledge significantly enhances explanation faithfulness. Completeness of observations also benefits from scale, though it peaks around 41% before slightly declining, suggesting a limit beyond which added knowledge may become redundant. The diagnostic accuracy exhibits slight fluctuations with increasing knowledge base size but consistently remains higher than the no-knowledge setting, indicating that the incorporation of structured medical knowledge enhances diagnostic performance.

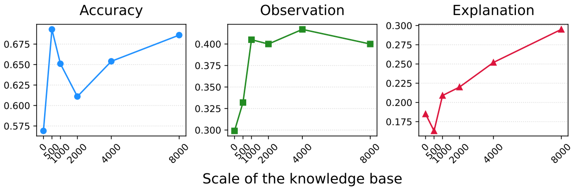


Figure 3: Performance across different knowledge base sizes.

4.6 Generalization to Unseen Diseases

To evaluate the generalizability of the constructed knowledge base B, we perform an ablation study by selectively masking domain-specific knowledge at varying levels of granularity. The detailed experimental setup is provided in Appendix C.1. Results in Figure 4 show that even when specialized knowledge varies, the model still benefits by 13%, 6%, and 5%, respectively, across diagnostic metrics. This suggests that knowledge from other specialties can aid differential diagnosis by helping to rule out diseases from the perspective of shared clinical phenomena. However, when masking is applied at the catalog level, performance drops slightly within specialties. This is likely because diseases within the same specialty often share similar manifestations, making it harder for the model to distinguish between them and increasing the risk of misdirection.

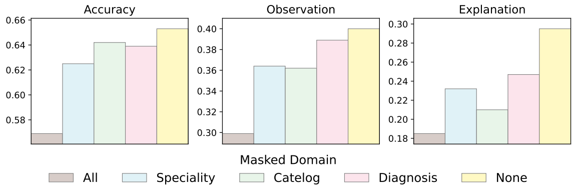


Figure 4: Performance across different degrees of in-domain knowledge masking.

5 Conclusion

We present **SEKAD**, a framework that automatically builds and applies an explanatory diagnostic knowledge base for interpretable medical diagnosis. It combines record-driven explanation self-learning and an explanation-augmented dual-phase diagnostic strategy. Experiments on two benchmarks show that **SEKAD** outperforms strong baselines in both diagnostic accuracy and explanation quality.

Limitations

This work, while demonstrating promising results, has inherent limitations. Our current framework primarily operates on textual clinical data and does not yet incorporate multimodal information or extend to multilingual clinical contexts. Furthermore, its evaluation is currently limited to the scale of existing benchmarks; scaling up to larger and more diverse real-world datasets presents avenues for future research. While our method utilizes **SEKAD**, integrating and evaluating it with larger and more advanced foundational models remains unexplored.

Ethics Statement

We affirm that all patient data utilized was strictly anonymized and strictly adhere to the data Use Agreement of the MIMIC dataset. We acknowledge the imperative to address potential biases in both data and algorithms to ensure equitable outcomes. Besides, we use an AI assistant to check the grammar. However, we double-checked and made sure that the AI assistant did not change the original meaning of the paper.

References

Ravi Aggarwal, Viknesh Sounderajah, Guy Martin, Daniel SW Ting, Alan Karthikesalingam, Dominic King, Hutan Ashrafian, and Ara Darzi. 2021. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ digital medicine*, 4(1):65.

Leila Agha, Jonathan Skinner, and David Chan. 2022. Improving efficiency in medical diagnosis. *Jama*, 327(22):2189–2190.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Md Manjurul Ahsan, Shahana Akter Luna, and Zahed Siddique. 2022. Machine-learning-based disease diagnosis: A comprehensive review. *Healthcare*, 10(3).

Michel Burnier. 2024. Poor physician adherence to clinical guidelines in hypertension—time for physicians to face clinical inertia. *JAMA Network Open*, 7(8):e2426830–e2426830.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599, Albuquerque, New Mexico. Association for Computational Linguistics.

Joakim Edin, Maria Maistro, Lars Maaløe, Lasse Borgholt, Jakob Drachmann Havtorn, and Tuukka Ruotsalo. 2024. An unsupervised approach to achieve supervised-level explainability in healthcare records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4869–4890, Miami, Florida, USA. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2020. Mimic-iv. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)*, pages 49–55.

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.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Engineering National Academies of Sciences, Medicine, et al. 2015. The diagnostic process. *Improving diagnosis in health care*, pages 31–80.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in*

Neural Information Processing Systems, 36:53728–53741.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Xiaorui Su, Yibo Wang, Shanghua Gao, Xiaolong Liu, Valentina Giunchiglia, Djork-Arné Clevert, and Marinka Zitnik. 2024. Knowledge graph based agent for complex, knowledge-intensive qa in medicine. *arXiv preprint arXiv:2410.04660*.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*.

Bowen Wang, Jiuyang Chang, Yiming Qian, Guoxin Chen, Junhao Chen, Zhouqiang Jiang, Jiahao Zhang, Yuta Nakashima, and Hajime Nagahara. 2024a. Direct: Diagnostic reasoning for clinical notes via large language models. *arXiv preprint arXiv:2408.01933*.

Yubo Wang, Xueguang Ma, and Wenhu Chen. 2024b. Augmenting black-box LLMs with medical textbooks for biomedical question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1754–1770, Miami, Florida, USA. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6):186357.

Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. 2022. Pmc-patients: A large-scale dataset of patient summaries and relations for benchmarking retrieval-based clinical decision support systems. *arXiv preprint arXiv:2202.13876*.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, et al. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.

A Details of Record-driven explanation self-Learning

A.1 Datasets

Patient Records. We use **PMC-Patients** (Zhao et al., 2022), a corpus of 167,000 patient summaries extracted from case reports in PubMed Central. Only unstructured patient narratives are utilized.

Medical Knowledge Sources. The explanatory knowledge is retrieved from **MedCorp** (Xiong et al., 2024), a comprehensive corpus that aggregates data from various public biomedical repositories. MedCorp is composed of PubMed (containing 23.9 million biomedical articles), StatPearls (9,330 clinical decision support articles), medical textbooks (18 books, chunked), and Wikipedia (chunked encyclopedia data). These components collectively provide access to the latest biomedical research, clinical decision support, foundational medical knowledge, and general domain information, forming a cross-source retrieval resource. These sources serve as $\hat{T}$ or $\tilde{T}$ depending on the retrieval context.

A.2 Retrieval Method

We adopt **MedCPT** (Jin et al., 2023), a neural retriever optimized for zero-shot semantic search, developed by the National Center for Biotechnology Information (NCBI). For explanation generation, the top-5 relevant texts ($|\hat{T}| = 5$) are retrieved; for diagnostic triangulation, we retrieve $|\tilde{T}| = 8$ diverse knowledge entries to support cross-validation.

A.3 LLM Backbone and Training Details

The core modules, including the extractor, explanation generator, and knowledge verifier, are powered by `Qwen-7B-Instruct`. To align model preferences with high-quality explanatory reasoning, we apply Direct Preference Optimization (DPO) using 200 preference samples from `DeepSeek-V3` (Liu et al., 2024), with a batch size of 64, a peak learning rate of 5×10^{-6}, and 3 epochs. We used 10 NVIDIA GeForce RTX 3090 GPUs (24GB) for running DPO, and 2 GPUs for the whole learning stage.

B Details of Main Experiments

B.1 Baseline and SEKAD Configurations

KGAREVion utilizes **PrimeKG** as its structured medical knowledge graph. For explanation verification, we adopt the publicly released LLAMA-3 checkpoint provided by the original authors.

MedRAG Based on the **MedCorp** corpus as our method. It applies an **RRF-4** ensemble retriever to fetch the top 16 documents per query.

MDAgents We have set 3 agents responsible for internal clinical tasks (ICT) and 5 agents mimicking a multidisciplinary team (MDT) of medical experts.

MedAgents models agent-based interaction with $m = 5$ domain-specialized experts generating diagnostic questions and $n = 2$ additional experts evaluating the candidate answers.

SEKAD follows an explanation-guided diagnostic paradigm. During the explanation-based diagnosis phase, it employs the MEDCPT retriever to collect the top-10 relevant knowledge subsets per query. The system is allowed a maximum of 3 reasoning rounds per diagnostic episode. All language model components operate with a fixed decoding temperature of 0.7 to balance output diversity and coherence.

B.2 Benchmarks

B.2.1 DiReCT

Dataset. The DiReCT (Wang et al., 2024a) dataset comprises 511 clinical notes, spanning 25 disease categories, sourced from the publicly available database MIMIC-IV (Johnson et al., 2020). Each clinical note is meticulously annotated with fine granularity by professional physicians, detailing the diagnostic process from observations within the note to the final diagnosis, which is presented in an entailment tree structure.

Task setup. DiReCT defines a diagnostic task that requires explanations, given a patient's clinical record without diagnostic conclusions and a graph constructed from all the diagnoses in the dataset domain $\mathcal{G}$, the model is required to find the path to the primary discharge diagnosis from the graph and to choose the patient's observational phenomena at each node along the path and explain them accordingly. In addition, DiReCT provides a knowledge graph $\mathcal{K}$, corresponding to $\mathcal{G}$, which contains the knowledge extracted by the expert from the corresponding diagnostic guidelines for each diagnostic node in $\mathcal{G}$. In our experiments, DiReCT with $\mathcal{K}$ is considered as an alternative baseline enhanced by external knowledge.

Metrics. We mainly report five experimental metrics, grouped into three categories.

Accuracy of diagnosis quantifies the model's ability to correctly identify diseases. This is measured by Acc^{cat}, reflecting performance across 25 predefined disease categories, and Acc^{diag}, which represents the accuracy of the final discharge diagnosis.

Completeness of observation, denoted by Obs^{comp}, quantifies the model's attention to and coverage of patient clinical phenomena during diagnostic explanation generation. This metric integrates both the recall and precision of identified observations.

Faithfulness of explanation assesses the consistency between the model's generated explanations and expert-annotated ground truth. Exp^{com} measures the faithfulness for observations successfully matched with the ground truth, while Exp^{all} measures the overall alignment with expert-annotated explanations. All binary judgments for model predictions against expert annotations (for both explanations and observations) are performed automatically using `Llama-3.1-8B`, which has been shown to align well with human judgments in DiReCT.

Baseline Adaptation to DiReCT

DiReCT evaluates models based on their ability to explain diagnoses using only nodes from the predefined diagnostic graph $\mathcal{G}$. We modified the baseline to operate in an end-to-end manner, taking medical history as input and generating explanations as output, and embedded the diagnostic graph $\mathcal{G}$ from DiReCT in the prompt. For evaluation, we extracted all observation-diagnosis pairs from the generated explanations and mapped them to DiReCT's diagnostic graph $\mathcal{G}$.

B.2.2 JAMA Clinical Challenge

Dataset. The JAMA Clinical Challenge (Chen et al., 2025) dataset is constructed from real-world cases published in the Clinical Challenge archive of the Journal of the American Medical Association. Each case includes a complex clinical vignette, a multiple-choice question regarding diagnosis or management, and expert-authored explanations justifying the correct and incorrect options. While the original cases include accompanying images, they are excluded in this dataset, as part of them do not contain information essential for diagnostic decision-making. This design emphasizes evaluation in settings where textual clinical information is the primary source.

Task Setup. In the experiment, we focused on questions related to diagnosis from the dataset. We utilized 149 challenge questions published by JAMA from 2022 to 2025. Models are presented with a clinical case report and four answer options. The task requires the model to predict the most probable diagnosis and generate the corresponding explanation, which is performed end-to-end directly from the patient report.

Metrics. Model performance is evaluated based on diagnostic prediction accuracy and the quality of generated explanations. To assess explanation quality, we adopt three automatic metrics from the G-Eval (Liu et al., 2023): coherence, consistency, and relevance. These metrics have shown relatively strong alignment with human judgment on this benchmark, particularly in evaluating factual correctness. Each metric is defined on a 5-point Likert scale and scored by `DeepSeek-V3`.

C Additional experiments

C.1 Generalization to Other Knowledge Domains

We evaluated the generalization value of the acquired knowledge by classifying the target diagnoses within the DiReCT benchmark according to different hierarchical levels. These levels include the first level by specialty (e.g., Cardiology, Endocrinology), the second level by disease catalog (e.g., ACS, Aortic Dissection), and the third level by specific diagnosis (e.g., Type A Aortic Dissection, Type B Aortic Dissection). To assess generalization, we conducted experiments where, for each patient case, the in-domain knowledge in the knowledge base corresponding to its main discharge diagnosis was masked or removed at different classification levels.

C.2 Impact of Retrieval Scale

We varied the number of retrieved knowledge units during the explanation-based diagnosis process. As shown in Figure 5, performance on both accuracy and faithfulness improves initially but saturates at approximately 15 retrieved units. Beyond this point, additional knowledge introduces noise, leading to a decline in both accuracy and faithfulness. In contrast, completeness of observation continues to improve as more knowledge is incorporated, reflecting its dependence on the quantity rather than the precision of retrieved evidence.

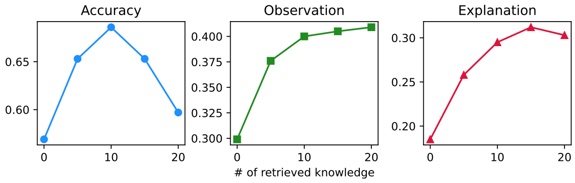


Figure 5: Performance across different retrieved knowledge numbers.

D Notation

Symbol	Meaning
R	Medical record
B	Explanatory knowledge base
k	Knowledge unit
p	Clinical phenomenon observed in the patient
P	Set of clinical phenomena
d	Diagnosis for the patient
D	Set of diagnoses
e	Textual explanation linking p and d
$\mathcal{TR}$	Text retriever
T	Medical knowledge sources
$\hat{T}$	Retrieved subset of T for explanation generation
$\tilde{T}$	Retrieved subset of T from sources disjoint with $\hat{T}$

Table 4: Notation used throughout the paper.

E Prompts

Prompt E.1: Extractor of Percept Action

Given the patient’s clinical note, extract all clinical phenomena hat are may relevant to the patient’s diagnosed disease.

Return them as a Python-style list. Each item must be extracted from the origin note.

Do not include any additional text outside the list.

{{Few-shot Sample}}

Prompt E.2: Explanation Generator of Explain Action

Input
1. `phenomenon`: A description of the patient’s symptoms or findings.
2. `candidate_diseases`: A list of potential diseases.
3. `reference_passages`: A set of text passages, each with a unique `SourceID`.

Instructions
1. Analyze the `phenomenon`, `candidate_diseases`, and `reference_passages`.
2. Identify the single disease from `candidate_diseases` that is most strongly supported by the information *within the passages* as the cause or explanation for the `phenomenon`.
3. Identify the *single* `SourceID` of the passage that provides the best evidence for this link.
4. Formulate an `explanation`:
- This must be a single, complete, affirmative sentence.
- It must state a general medical fact, principle, or definition linking a key aspect of the `phenomenon` (generalized, e.g., "high fever" not "39.5 C fever") to the chosen disease.
- This explanation should function as a standalone "theorem" – objective, definitive, and suitable for use as a fundamental statement without referring back to its origin.
- Crucially, do not mention the patient’s specific details, the passages, source IDs, or use phrases like "according to the source," "the reference indicates," "this case matches," or any wording that implies it’s derived *from* a specific source *within the sentence itself*.

5. Construct a JSON object containing the `explanation`, the exact `disease` name, and the selected `source_id`.
6. Output *only* the JSON object. Ensure no extra text precedes or follows the JSON structure.

{{Few-shot Sample}}

Prompt E.3: Knowledge Verifier of Validate Action

Task
Given a set of reference passages and a conclusion statement, evaluate whether the conclusion is sufficiently supported by the references.

Input Reasoning Process
First, think step by step about what kind of reasoning or evidence would be required to justify the conclusion. Then, examine the provided references to determine whether they contain the necessary support. Finally, state whether the references support the conclusion or not, and explain why.

Input Output Structure
Your output should include:

1. A short reasoning process describing what is needed to justify the conclusion.
2. An assessment of whether the references satisfy that need.
3. A final determination: either `[Supported]` or `[Unsupportable]`, with a brief justification.

Prompt E.4:Prompts for Differential Diagnosis (1)

Medical Record:
{notes}
Think step by step, determine which of the following diagnoses the patient is likely to have based on his medical records.
The diagnosis you identify must come from this list:
{disease_options}
Please include your final chosen diagnosis in the <diagnosis> tag.
Output Format:
[Thinking Here ...]
<diagnosis>[likely diagnosis from the list, split with a comma]</diagnosis>

Prompt E.5: Prompts for Differential Diagnosis (2)

TASK: Create an extremely concise clinical summary for '{diag}' based on the provided discrete medical facts.
INPUT FACTS:
{exp_knowledge}

KEY AREAS:
{queries_key}

CORE RULES:
1. STRICTLY BASED ON INPUT: The summary content must solely be derived from the 'INPUT FACTS' provided above. Do not add any external knowledge or information.
2. STRUCTURE: The summary must be organized under 'KEY AREAS'. Each key area uses bold font for its heading (e.g., Risk Factor).
3. CONTENT: Under each bold heading, synthesize the relevant 'INPUT FACTS' into an extremely compact list of phrases or terms. Full sentences are not required. The goal is maximum conciseness.
4. PROHIBITIONS: Do not use bullet points, numbered lists, or lengthy paragraphs.
OUTPUT FORMAT REQUIREMENT (Strictly adhere):
Key Area Name
Terms/phrases related to this area, extracted from Input Facts and compactly arranged.
EXAMPLE OUTPUT FORMAT:
{{Few-shot Sample}} Please generate the summary for '{diag}' now.

Prompt E.6: Prompts for Differential Diagnosis (3)

Medical Record:
{notes}
Analyze the patient's medical data below and determine the most likely next diagnosis from the provided list.
— Data for Analysis —
- guidelines -
{knowledges}

- Patient Medical Notes -
Provided previously.
(Note: This section contains the patient's clinical information and findings.)

- Previous Diagnostic Summary -
{summary}
— End Data —

Instructions:

1. Detailed Analysis: Perform a step-by-step analysis based on the patient's medical records and strictly follow the diagnosis guidelines. Find evidence to support or refute the potential diagnosis from the list of potential diagnoses. Detail your reasoning process. Output this analysis results within the <analyze> tag.

2. Diagnosis Summary & Confidence: Based on your analysis in step 1, provide a concise summary of the key findings and your conclusions related to the diagnosis selection. This summary MUST also explicitly include the strength of evidence supporting the primary diagnosis suggested by the notes and analysis. Use one of the following exact phrases to state the evidence strength: "Strength of Evidence: High", "Strength of Evidence: Moderate", "Strength of Evidence: Low", "Strength of Evidence: Insufficient". If you determine that the patient's condition does not align with any condition in the list of options (leading you to select 'None' in Step 3), you MUST rate the strength of evidence as "Strength of Evidence: Insufficient". Output the entire summary, including the strength of evidence statement, within the <summary> tag.

3. Select Next Diagnosis: Choose the single most appropriate NEXT diagnosis from the Potential Diagnoses List. Your selection MUST be an EXACT STRING MATCH to an item in the list: {disease_list + ["None"]}. Select 'None' **if and only if** you find that your current illness does not fall into any of the categories in the list. Output this selection within <diagnosis> tags.

Output ONLY the content within the specified tags, in the order: <analyze>, <summary>, <diagnosis>.

Format Example:
<analyze>
[Detailed analysis text from Step 1 goes here]
</analyze>
<summary>
[Concise summary text from Step 2 goes here]
</summary>
<diagnosis>
[Selected diagnosis string from Step 3 goes here]
</diagnosis>

Prompt E.7: Prompts for Differential Diagnosis (4)

You are now going to differentiate the disease {diag}.
Only focus on confirming the diagnosis; do not consider treatment or other aspects.
What aspects of {diag} would you like to know about for diagnosis?
Please list {q_num} items, each starting with '-', one per line.

Prompt E.8: Prompts for Definitive Diagnosis

Objective:
Analyze the Medical Record using the Guidelines to map the diagnostic reasoning process.

Instructions:

1. Medical record analysis:

- Identify the criteria for the current step within the Guidelines.
- Scan specific patient evidence (phenotypes) in the Record to match these criteria.
- Explain why the evidence is relevant by citing Guideline knowledge.
- Maintain strict focus: Only include evidence directly supporting the current diagnostic step.

2. JSON Output:

- Structure: Top-level keys are the exact Guideline diagnostic step names. Each key's value is a dictionary:
- Keys:
- Patient evidence (phenotypes). Extract the original record text and record in the order of the original text.
- Each piece of evidence can only be used once at multiple steps.
- Values: Justification based strictly on Guideline knowledge explaining the evidence's relevance to that step.
- Strict Relevance: Ensure every entry directly supports its parent step.
- No Evidence: If a step has no supporting evidence in the Record per the Guidelines, use an empty object {} as its value.

Procedure:
Perform the analysis first, then output the JSON.
{{Few-shot Sample}}
Input:
Guidelines:
{all_exp}

Medical Record:
{note}.

Initiate the Chain-of-Thought process now, and follow it with the final JSON output.