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

Dongqi Huang[1,2], Tong Zhou[1,2], Zhuoran Jin[1,2], Shenghui Shi[5]
Yujiao Mao[4], Kang Liu[1,2], Jun Zhao[1,2], Yubo Chen[3] *
[1]The Key Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]School of Information Engineering, Minzu University of China [4]Fujian Cancer Hospital
[5]College of Information Science and Technology, Beijing University of Chemical Technology
huangdongqi2025@ia.ac.cn yubo.chen@muc.edu.cn

Abstract

Explainable diagnosis requires that authoritative medical knowledge provide the rationales linking a patient's clinical manifestations to the diagnostic conclusion. Although large language models (LLMs) hold great potential to facilitate explainable diagnosis, their effectiveness is often constrained by insufficient diagnostic expertise. 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, on the DiReCT benchmark, **SEKAD** improves the explanation completeness metric from 64.5% to 76.9% over the best existing methods, highlighting its effectiveness in enhancing diagnostic explainability and showing that our text mining approach produces knowledge that is both reliable in quality and large in quantity [1].

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; Rashid and Mcguinness, 2025). For automated diagnosis, accuracy is indispensable; yet as

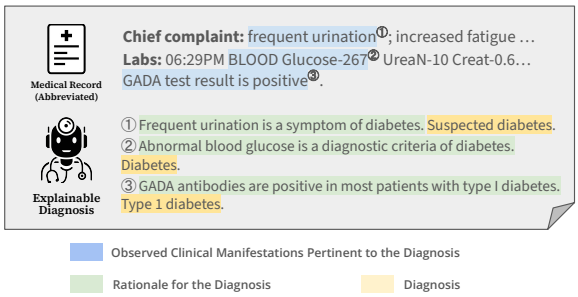

Figure 1: A sample of the explainable diagnosis, where medical knowledge serves as the rationales that link the patient's clinical manifestations to the diagnosis.

Edin et al. (2024) suggests, explainability is equally vital, for plausible but unfaithful explanations may mislead clinicians and compromise trust. Large language models (LLMs) (Zhou et al., 2023) are considered as a potential choice to build 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 the 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 task of explainable diagnosis requires the help of explainable external knowledge. Human physicians rely on medical guidelines as diagnostic references to address complex cases (National Academies of Sciences et al., 2015). These knowledge sources can also 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

*Corresponding author.
[1]We provide our code on Github: https://github.com/Vanthoci/acl2026-sekad-explainable-diagnosis

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 36679–36700
July 2-7, 2026

knowledge sources to improve explainable diagnostic capabilities. Thus, defining and building such explanatory knowledge bases is a key to advance explainable automatic diagnosis.

The automated construction of explanatory diagnostic knowledge bases presents challenges. 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). Enabling LLMs to autonomously verify and refine the accuracy of the knowledge they acquire from these independently structured texts remains a problem to be solved. 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 manifestations and diagnostic conclusions offer a valuable opportunity for the large-scale, automated construction of explanatory knowledge bases. Consequently, a key aspect of the 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 **Di**agnosis (**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**[2] 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 a comprehensive evaluation of our framework on two explainable diagnosis benchmarks. Our method not only outperforms five existing baselines across multiple metrics but also exceeds the state-of-the-art approach on the DiReCT benchmark by **12.4%** in explanation completeness and **4.3%** in explanation faithfulness. To further validate our core contribution, we also performed a series of evaluations of the self-learned knowledge base to confirm its quality. 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 biased explanations by ensuring each diagnosis is supported by comprehensive explanation.
- Experiments on two explainable diagnostic evaluation datasets demonstrate that our method outperforms baselines and excels in explanation generation.

2 Method

In this section, we introduce **SEKAD**, an explainable automatic diagnosis framework augmented

[2]Triangulation: A statistical method of partitioning a sample of data into subsets such that the analysis is initially performed on a single subset, while the other subsets are retained for subsequent use in confirming and validating the initial analysis.

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 with diagnostic results masked, and under the guidance of explanatory knowledge, the diagnosis executor first performs differential diagnosis to identify candidate diagnoses, subsequently generating explanations that link the patient's clinical manifestations to these diagnoses in the definitive diagnosis phase. Pseudo code and all prompts used for LLMs are provided in the appendix G.

2.1 Record-driven Explanation Self-learning

Unstructured medical records, including patient reports and clinical notes, reflect numerous connections between clinical manifestations 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.

2.1.1 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 manifestation and a corresponding diagnosis, grounded by an explanatory rationale. Formally, a knowledge unit k is defined as a tuple (m, e, d), where:

- m: a single clinical manifestation observed in the patient, such as "*dizziness*".
- d: the diagnosis for the patient, at any level of granularity.
- e: an explanation linking the clinical manifestation m 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 manifestation m. Although concomitant manifestations 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 manifestations 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 m or d.

Generalization: For explanation e, the clinical manifestation m 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 $m \leftrightarrow d$ 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 $k = (m, e, d)$ are retrieved to support explanation-based diagnosis. As illustrated in Figure 2, each unit is incorporated into B through a sequential process of **extract**, **explain**, and **validate**, which enables dynamic updates and ensures knowledge quality.

2.1.2 Extract

Based on the original patient's medical record R sourced from the PMC-Patients dataset (Zhao et al., 2022), an LLM-based extractor is instructed to identify documented diagnoses $(d_1, ..., d_m \in D)$

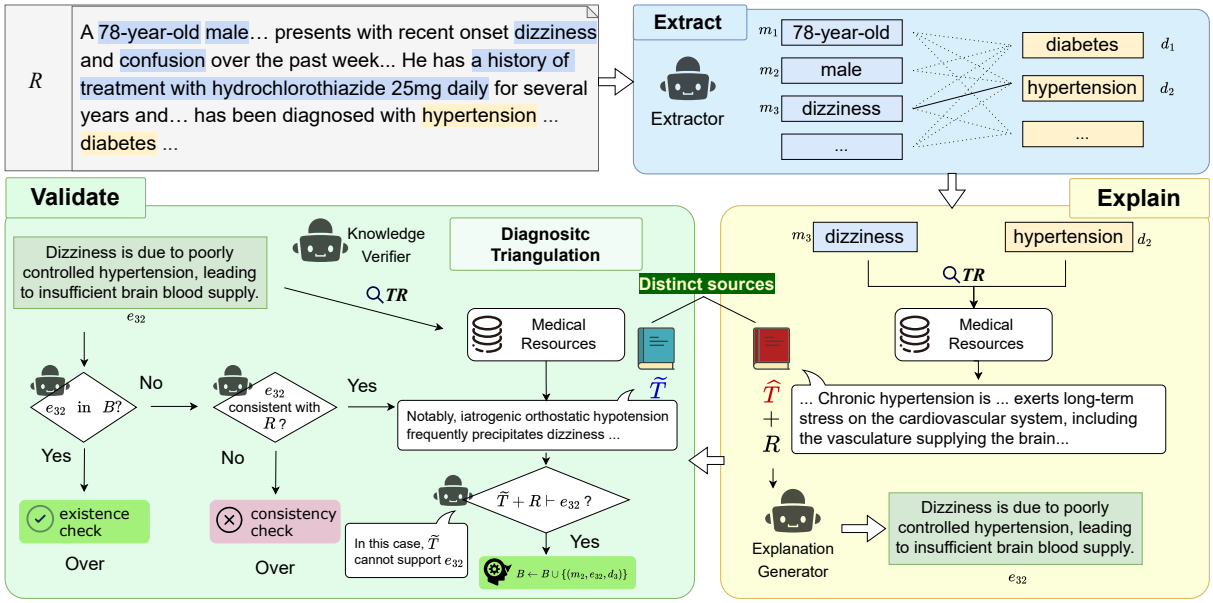

Figure 2: Overview of record-driven explanation self-learning. An initial explanation links *dizziness* to poorly controlled *hypertension* based on disease-centered sources. Diagnostic triangulation with pharmacological references reveals that dizziness may instead result from side effects of antihypertensive medications. This mechanism identifies conflicting evidence and filters out potentially misleading diagnostic links.

and single clinical manifestations ($m_1, ..., m_n \in M$) as exact textual spans. These spans are designed as direct excerpts from the original records to achieve more semantically relevant retrieval when diagnosing from medical records. This identification strategy decomposes the patient's findings into individual manifestations m_i, making each m_i a basis for potentially linking to identified diagnoses d_j. By ensuring each m_i serves as the single manifestation m in $k = (m, e, d)$, this action guarantees **unit specificity** for k.

2.1.3 Explain

The **explain** action aims to find explanatory clinical knowledge that links clinical manifestations m with diagnoses d from relevant authoritative medical knowledge sources from the MedCorp corpus (Xiong et al., 2024). Its input includes a specific clinical manifestation m identified from the patient's medical record R, and the corresponding diagnosis d. Together, m and d are concatenated to form the search query. Using this query, a text retriever $\mathcal{TR}$ implemented via MedCPT (Jin et al., 2023), searches for a relevant subset from the medical knowledge sources T. Subsequently, the explanation generator utilizes the retrieved subset $\hat{T}$ to generate the explanation e, guided by prompts that ensure adherence to the principles of **self-contained** and **generalization**. The generation is conditional, as the generator first determines if $\hat{T}$ contains sufficient evidence and will abstain from producing an explanation if it does not.

2.1.4 Validate

Ensuring the **faithfulness** of each knowledge unit $k = (m, e, d)$ 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 $m \leftrightarrow d$ associations or associations valid only in specific contexts. A key sub-step within the **validate** action is **diagnostic triangulation**, which verifies a candidate explanation against multiple, independent knowledge sources.

Specifically, the **validate** action is performed by 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 (m, d) pair aligns with knowledge already validated and stored in B. For each such (m, 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 consistency between the generated explanation e and the original patient record R, ensuring that e does not conflict with other conditions 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 2, 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.

2.1.5 Enhanced Efficiency via DPO

To mitigate the high computational overhead of using large-scale models for self-learning, we selected the cost-effective `Qwen2.5-7B` as our base model. To align its capabilities with the goal of generating faithful knowledge, we introduce Direct Preference Optimization (DPO) (Rafailov et al., 2023) within a teacher-student framework. In this setup, the more advanced `DeepSeek-V3` model acts as a teacher, generating preference data to fine-tune the `Qwen2.5-7B` student model.

This approach creates a synergistic optimization loop between two actions. The preference for the **explain** action is determined by the intrinsic quality of a generated explanation, specifically whether it passes the **validate** action, while the preference for the **extract** action is determined by its downstream utility: a set of extracted manifestations is considered superior if it enables the **explain** action to generate a higher number of valid explanations. The **extract** action is explicitly trained to find the most useful clinical evidence for the subsequent explanation task. This unified framework thus enables the simultaneous optimization of both actions, increasing their overall efficiency.

2.2 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 manifestations 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 guided by implicit Bayesian logic: **differential diagnosis**, which constructs prior probabilities based on semantic associations to narrow down potential diagnoses, and **definitive diagnosis**, which leverages decisive evidence to confirm posterior probabilities supported by semantic probability encoded in our knowledge base.

2.2.1 Differential diagnosis

In clinical practice, differential diagnosis refers to the process by which physicians analyze specific clinical manifestations 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 3. First, a preliminary analysis is performed by the executor, which hypothesizes an initial disease category based on its general medical knowledge 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 manifestations are not mentioned, the executor then reverses the querying direction by asking, "What diseases commonly present with this manifestation?" for those not yet identified.

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 anchor-

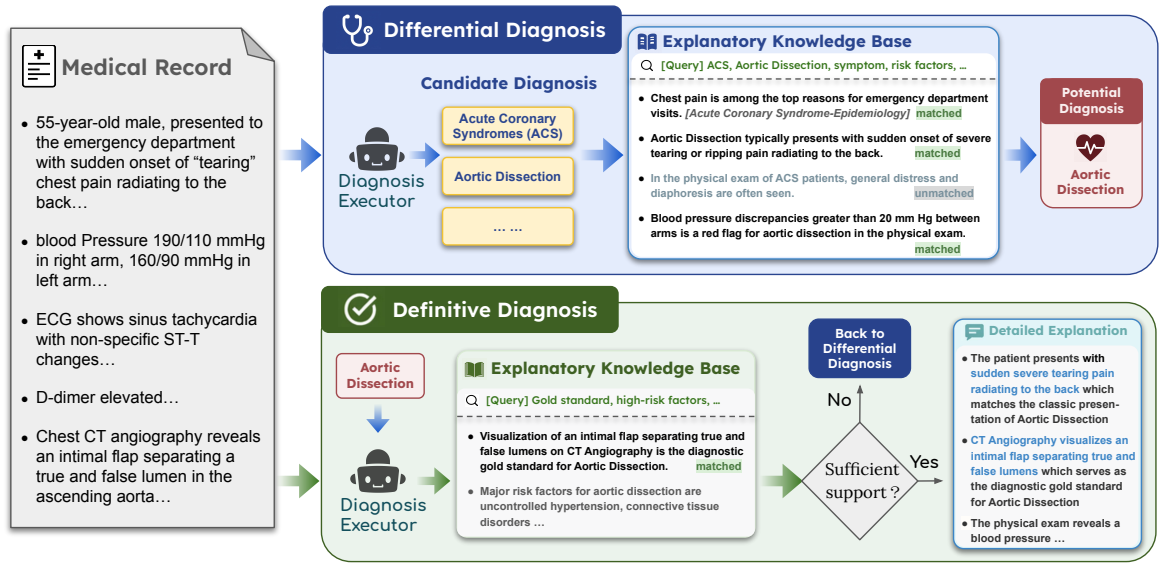

Figure 3: Overview of explanation augmented dual-phase diagnosis

ing arising from excessive explanation at an early stage. This phase proceeds for a fixed number of retrieval cycles before transitioning to the definitive diagnosis phase.

2.2.2 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 manifestations by matching them with validated knowledge units k. Only successfully matched knowledge is used to construct the definitive explanation. The process concludes with a self-correction step, where the executor is prompted to state whether it senses a potential misdiagnosis, triggering a reversion if it does.

3 Experiments

3.1 Benchmarks

3.1.1 DiReCT

The DiReCT benchmark (Wang et al., 2024a) is a dataset for explainable diagnosis, comprising 511 physician-annotated clinical notes from MIMIC-IV (Johnson et al., 2020). It defines an end-to-end explainable diagnosis task where the model, given a patient's clinical note with all diagnostic conclusions masked, must predict the primary discharge diagnosis. This prediction must be supported by explanations that link the diagnosis back to specific clinical manifestations within the note. Performance on the benchmark is evaluated across three structural levels derived from expert-annotated diagnostic triplets (o, z, d), where o is the clinical observation, z the medical rationale, and d the supported diagnostic node. **Accuracy of diagnosis** (Acc^{cat} and Acc^{diag}) quantifies the correct identification of the final discharge diagnosis. **Completeness of observation** (Obs^{comp}) employs Jaccard similarity to measure the overlap between predicted and ground-truth evidence, penalizing both omitted and extraneous findings. **Faithfulness of explanation** assesses the integrity of the full reasoning chain: Exp^{com} measures the score rate of valid rationales within the intersection of correctly identified observations, while Exp^{all} evaluates end-to-end alignment across the union of all observations. These structural components are assigned binary scores by an LLM evaluator, which has been validated to align with human expert judgments. For a detailed experimental setup and metric specifics, see appendix C.2.1.

3.1.2 JAMA Clinical Challenge

The JAMA Clinical Challenge benchmark (Chen et al., 2025) consists of text-based clinical cases from the Journal of the American Medical Association, presented as multiple choice diagnostic questions. The task requires models to predict the most probable diagnosis and generate a corresponding explanation. Performance is evaluated based on

Method	DiReCT					JAMA Clinical Challenge			
	Acc^{cat}	Acc^{diag}	Obs^{comp}	Exp^{com}	Exp^{all}	Acc	$Relev.$	$Coh.$	$Consist.$
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									
Vanilla	0.690	0.586	0.192	0.263	0.071	0.779	4.490	**4.839**	4.318
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**	**0.781**	**4.670**	4.718	**4.326**
DeepSeek-V3									
Vanilla	0.736	0.542	0.229	0.373	0.099	0.731	4.662	4.114	4.225
COT	0.702	0.585	0.185	0.276	0.065	0.711	4.672	**4.945**	4.305
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	0.631	4.331	4.852	4.101
MDAgents	0.688	0.566	0.218	0.349	0.099	0.691	4.531	4.711	4.141
MedAgent	0.740	0.599	0.205	0.319	0.076	0.450	3.651	3.705	3.537
MedRAG	0.817	0.640	0.288	0.232	0.069	0.400	3.745	3.570	3.282
SEKAD	**0.847**	**0.653**	**0.400**	**0.759**	**0.312**	**0.771**	**4.672**	4.740	**4.313**

Table 1: Performance comparison on the DiReCT and JAMA Clinical Challenge benchmarks, where higher values indicate better performance. **Bold** indicates the best result. The '–' symbol indicates that baselines from DiReCT do not apply to the JAMA Clinical Challenge dataset.

diagnostic accuracy and explanation quality. For the latter, the benchmark uses G-Eval metrics (Liu et al., 2023), which have been shown to have the relatively highest alignment with human judgment on this dataset. For further details, see appendix C.2.2.

3.2 Baselines

We evaluate our method with five 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.

3.3 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.

We further observe that existing baselines 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. A more detailed analysis comparing the responses of **SEKAD** and the baseline model is provided in Appendix E.

Figure 1 also 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.

Lacking imaging data, the JAMA dataset's diagnostic context is incomplete. Under these conditions, many baseline models tend to engage in overreasoning or fall into heuristic bias, which results in performance degradation compared to the base

model.. In contrast, **SEKAD** maintains robust diagnostic reasoning through its structured *Differential–Definitive* dual-phase 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.

3.4 Ablation Study

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

As shown in Table 2, 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.

We conduct another ablation study to evaluate different knowledge bases for explainable diagnosis. In this experiment, we replaced our knowledge base B with several alternative knowledge sources. The alternatives include:

- $\mathcal{K}$ **from DiReCT** (Wang et al., 2024a): A knowledge base sourced from the DiReCT dataset, consisting of diagnostic guideline summaries annotated by experts for their specific tasks.
- **PrimeKG** (Chandak et al., 2023): A medical knowledge graph that blends structured and semantic information, which was also utilized in the KGAREVion (Su et al., 2024) method.
- **MedCorp** (Xiong et al., 2024): The dataset used in the MedRAG method, comprising text segments from Wikipedia, medical literature, and medical textbooks.

Table 3 illustrates how the choice of knowledge source impacts overall performance. This result indirectly validates the effectiveness of our self-learned knowledge base in supporting high-quality, explainable diagnosis. To further corroborate these findings, we also conduct a direct evaluation of the knowledge base; see Appendix D.3.

Knowledge Source	Acc^{cat}	Acc^{diag}	Obs^{comp}	Exp^{com}	Exp^{all}
DeepSeek-V3					
$\mathcal{K}$ from **DiReCT**	0.778	0.653	0.385	0.611	0.255
PrimeKG	**0.861**	0.694	0.380	0.478	0.198
MedCorp	0.806	0.611	0.377	0.547	0.233
B from **SEKAD**	0.847	**0.653**	**0.400**	**0.759**	**0.312**

Table 3: Comparison of different knowledge sources on the DiReCT benchmark. **Bold** indicates the best performing source.

3.5 The Robustness of the Knowledge Base

To evaluate the robustness of the constructed knowledge base B, we perform an ablation study by masking domain-specific knowledge at varying levels of granularity. The detailed experimental setup is provided in D.1. Results in Figure 4 show that even when specialized knowledge varies, the model 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 manifestations. However, when masking is applied at the catalog level, performance drops slightly within specialties. This manifestation can be attributed to the overlapping clinical manifestation observed among pathologies within the same nosological classification, making it harder for the model to distinguish between them and increasing the risk of misdirection.

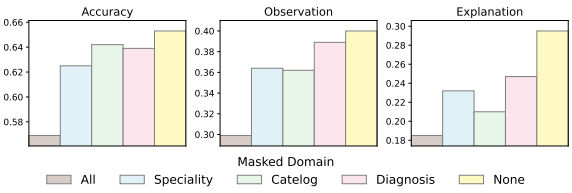

Figure 4: Performance across different degrees of in-domain knowledge masking.

3.6 Performance Evolution with Dynamic Knowledge Updates

As a simulation of the dynamically updating knowledge system, we characterize the dynamic accumulation of knowledge by expanding the scale of our knowledge base and examine the resulting performance evolution, as shown in Figure 5. The most significant gain is in the **faithfulness of explanation**, which grows from 17% to 29%, demonstrating that richer knowledge significantly enhances this aspect. The **completeness of observations** also benefits from scale, peaking around 41% before slightly declining. This suggests a threshold beyond which additional knowledge no longer improves focus on key clinical observations. Meanwhile, **diagnostic accuracy**, despite slight fluctuations, consistently remains higher than the no-knowledge baseline, confirming that incorporating structured medical knowledge enhances diagnostic performance.

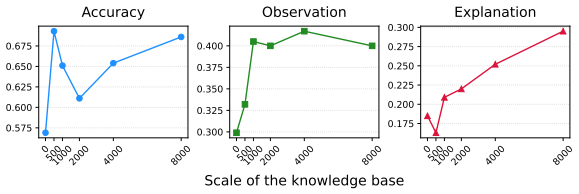

Figure 5: Performance across different knowledge base sizes.

4 Conclusion

We present **SEKAD**, a framework that automatically constructs and utilizes an explanatory diagnostic knowledge base for explainable diagnosis. Experiments on two benchmarks show that **SEKAD** outperforms strong baselines in both diagnostic accuracy and explanation quality.

Limitations

While our approach effectively enhances the explainability of diagnostic reasoning, it primarily centers on the phase of diagnostic identification. Our evaluation was conducted on acute and critical care cases represented by MIMIC-IV, where clinical management is more closely aligned with a diagnosis driven paradigm. In broader clinical practice, diagnosis is often inseparable from treatment, care, and prognosis. Furthermore, complex and diverse co-morbidities are prevalent in real world scenarios, and these capabilities were not included in our current testing scope. We aim to use this work as a foundation to develop more robust solutions that address these complex, practical clinical challenges in the future.

Our evaluation is subject to the inherent limitations of automated metrics in medical natural language processing. Achieving rapid, large scale, and precise evaluation of open ended responses from large language models remains an unresolved challenge. While our primary evaluation on MIMIC-IV-DiReCT establishes a judgment scheme with high alignment to human experts, the automated G-Eval metric used in the JAMA benchmark remains limited in its alignment with human judgment, despite being a relatively superior choice among available options.

We note that the improvement in diagnostic accuracy by **SEKAD** is not drastic, as our primary contribution lies in enhancing the faithfulness of explanations rather than raw accuracy performance. Finally, our framework currently processes only textual data and the integration of multimodal information or combinations with other foundational models remains an area for future exploration.

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.

Acknowledgments

This work is supported by the New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123301). This work is supported by the National Natural Science Foundation of China (No.62576340). This work is also sponsored by Beijing Nova Program (20250484750) and the Youth Innovation Promotion Association CAS.

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.

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67.

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.

Huajun Chen. 2024. Large knowledge model: Perspectives and challenges. *DATA INTELLIGENCE*, 6(3):587–620.

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.

Douglas Johnson, Rachel Goodman, J Patrinely, Cosby Stone, Eli Zimmerman, Rebecca Donald, Sam Chang, Sean Berkowitz, Avni Finn, Eiman Jahangir, et al. 2023. Assessing the accuracy and reliability of ai-generated medical responses: an evaluation of the chat-gpt model. *Research square*, pages rs–3.

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.

Man Luo, Xin Xu, Zhuyun Dai, Panupong Pasupat, Mehran Kazemi, Chitta Baral, Vaiva Imbrasaite, and Vincent Zhao. 2024. Dr.icl: Demonstration-retrieved in-context learning. *DATA INTELLIGENCE*, 6(4):909–922.

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.

Sabbir M. Rashid and Deborah L. Mcguinness. 2025. Designing and evaluating an ensemble reasoning-based clinical decision support system. *DATA INTELLIGENCE*, 7(1).

Shuo Shang, Renhe Jiang, Ryosuke Shibasaki, and Rui Yan. 2024. Foundation models for information retrieval and knowledge processing. *DATA INTELLIGENCE*, 6(4):891–892.

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 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) and retrieval-based in-context learning (Luo 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 can lead to reduced credibility of the generated explanations.

Medical knowledge-enhanced LLM. Several approaches have attempted to address this knowledge limitation by incorporating external knowledge into LLMs (Chen, 2024; Shang et al., 2024). For instance, 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. While these efforts demonstrate success in augmenting the factual accuracy of LLMs, their primary focus has been on improving diagnostic performance. The challenge of constructing and utilizing knowledge bases specifically to enhance the explanatory quality of diagnostic models, which is the central focus of our main work, remains a less explored area in the literature.

B Details of Record-driven explanation self-learning

B.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 re-

search, 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.

B.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.

B.3 LLM Backbone and Training Details

The core modules, including the extractor, explanation generator, and knowledge verifier, are powered by `Qwen2.5-7B`. To align model preferences with efficient explanatory reasoning, we apply 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.

C Details of Main Experiments

C.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.

C.2 Benchmarks

C.2.1 DiReCT

Dataset Construction. The DiReCT (Wang et al., 2024a) dataset contains 511 clinical notes sourced from the MIMIC-IV (Johnson et al., 2020) database. These notes cover 25 disease categories across five medical domains, and each note is associated with a single Primary Discharge Diagnosis (PDD). For the task, each clinical note is presented as an excerpt of six specific sections from the subjective and objective portions of the SOAP-formatted record. A crucial modification was made to the content of the notes: any description that discloses the PDD was manually removed to create the core diagnostic challenge. Furthermore, for 73 notes that originally lacked sufficient evidence for a final diagnosis, annotating physicians added a plausible observation to the record to ensure a complete diagnostic process could be followed.

Annotation Process and Diagnostic Knowledge Graph. The core of the DiReCT dataset is a detailed annotation process performed by clinical physicians and subsequently verified by senior medical experts. For each clinical note, annotators identify key text segments ("observations") that lead to a diagnosis and provide a "rationalization" explaining why each observation supports its corresponding diagnosis. This annotation process strictly adheres to a purpose-built Diagnostic Knowledge Graph. This graph, based on existing diagnostic guidelines, contains nodes for medical statements (premises) and diagnoses, which are connected by supporting edges (premise-to-diagnosis) and procedural edges (diagnosis-to-diagnosis). The graph serves a dual function: it acts as the gold standard to guide physicians toward uniform and precise annotations, and it is also provided as an input to models to align their reasoning processes with those of medical professionals. The benchmark provides this graph in two forms: a diagnosis graph $\mathcal{G}$ representing the structured diagnostic relationships, and a knowledge-enhanced graph $\mathcal{K}$, which incorporates expert-curated guideline knowledge for each diagnostic node.

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 manifestations 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 manifestations 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`. It is validated on the DiReCT benchmark, **where `Llama-3.1-8B` demonstrated strong alignment with human expert judgments**, achieving a correlation of 0.869 for observation assessment and 0.734 for explanation assessment.

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 $\mathcal{G}$.

C.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 G-Eval (Liu et al., 2023): coherence, consistency, and relevance, with each scored by `DeepSeek-V3` on a 5-point Likert scale. The analysis of various automated metrics on the JAMA benchmark shows different degrees of correlation with human judgment. For instance, metrics such as ROUGE-L, BERTScore, and CTC showed negative or near-zero correlations (-0.01, -0.06, and -0.05, respectively). In comparison, the G-Eval metrics for coherence (0.32), consistency (0.26), and relevance (0.22) all yielded positive, albeit moderate, correlations.

D Additional experiments

D.1 The Robustness of the Knowledge Base

We conduct the evaluation 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). 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.

D.2 Impact of Retrieval Scale

We varied the number of retrieved knowledge units during the explanation-based diagnosis process. As shown in Figure 6, 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.

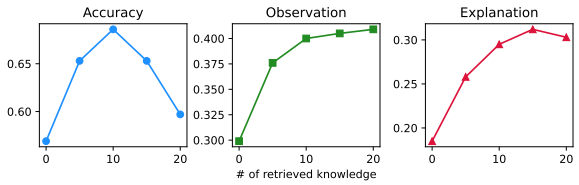

Figure 6: Performance across different retrieved knowledge numbers.

D.3 Evaluating the Knowledge Base Construction

We analyze the quality of our knowledge base B and the impact of DPO, incorporating a human evaluation of 100 randomly sampled knowledge units. Three models are compared: the base model `Qwen2.5-7B`, our DPO-finetuned model, and the original teacher model `DeepSeek-V3`.

The evaluation measured three metrics: **Generation Yield**, the proportion of knowledge units successfully generated per medical record that were incorporated into the knowledge base; **Validation Rate**, the percentage of generated units that successfully passed the **validate** action; and **Faithfulness**, which measures medical accuracy as confirmed through human evaluation. Adopting the methodology from (Johnson et al., 2023), faithfulness was assessed on a six-point Likert scale during this manual validation.

As shown in Table 4, the results demonstrate that fine-tuning the backbone model with DPO enables it to generate more candidate knowledge units and improves the quality of these units, as evidenced by the increase in the validation pass rate. Notably, the multi-stage validation process consistently ensures a high accuracy for the knowledge ultimately included in the knowledge base, highlighting the effectiveness of the framework's quality control mechanism.

Model	Generation Yield	Validation Rate	Faithfulness
`Qwen2.5`	2.32	44%	5.56
`Qwen2.5`$_{\text{DPO}}$	3.51	52%	5.58
`DeepSeek-V3`	4.84	55%	5.67

Table 4: Comparative analysis of knowledge generation quality

E Case Study and Error Analysis

We conduct an error analysis on the diagnostic explanations generated for the DiReCT benchmark. Our study identified five primary categories of errors: (i) **Diagnostic Error:** The model may correctly identify a clinical finding and provide a valid explanation for it, but incorrectly attributes this rationale to the wrong intermediate diagnosis. (ii) **Explanation Incompleteness:** The explanation omits parts of the standard reasoning mentioned by the expert. (iii) **Extraneous Information:** The explanation includes superfluous rationale that are not part of the expert's provided rationale. (iv) **Medical Factual Error:** The explanation contains statements that contradict established medical knowledge or the facts presented in the expert's gold-standard rationale. (v) **Contextual Contradiction:** The explanation makes claims that are inconsistent with the specific details presented in the patient's clinical note. We randomly sampled 100 cases from the DiReCT dataset and performed a manual analysis to compare the error types between a COT baseline and our **SEKAD** method. The distribution of these errors, with statistics calculated on a per explanation basis, is summarized in Table 5. For specific examples illustrating these error types with original patient cases and detailed analysis, please refer to Figure 7.

Error Type	COT	SEKAD
Diagnostic Error	23.8%	20.7%
Explanation Incompleteness	31.6%	13.2%
Extraneous Information	29.7%	36.5%
Medical Factual Error	4.4%	1.5%
Contextual Contradiction	1.1%	1.0%
Correct	9.3%	27.1%

Table 5: Error analysis on 100 samples from the DiReCT benchmark.

This error analysis reveals that for LLMs performing explainable diagnosis, the primary sources of error are Explanation Incompleteness, Extraneous Information, and Diagnostic Errors, with a smaller portion stemming from Medical Factual Errors and Contextual Contradictions. We found that **SEKAD** most significantly reduces Explanation Incompleteness, indicating that it successfully introduces more medical knowledge that aligns with expert explanations.

F Notation

Symbol	Meaning
R	Medical record
B	Explanatory knowledge base
k	Knowledge unit
m	Clinical manifestation observed in the patient
M	Set of clinical manifestations
d	Diagnosis for the patient
D	Set of diagnoses
e	Textual explanation linking m 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 6: Notation used throughout the paper.

Chief Complaint: SOB on exertion

History of Present Illness: A male with a history of an ICD, Chronic Myelogenous Leukemia (CML), Chronic Kidney Disease (CKD), and Congestive Heart Failure (CHF, EF 25%) presents with shortness of breath on minimal exertion, such as dressing. He reports recent weight gain and increasing ankle edema over the past few days. He was recently hospitalized for a ...

Past Medical History: Congestive Heart Failure (EF 25%, severe global hypokinesis, 4+ MR), Coronary Artery Disease (CAD), Chronic Myelogenous Leukemia (on Gleevec), Chronic Kidney Disease (CKD), s/p ICD placement (for VT), s/p CVA (no residual deficits)...

Family History: No family history of CAD, leukemia, or lymphoma.
Physical Exam: Vitals: T 98.1, BP 136/69, HR 81, SpO2 97% on Room Air; General: No acute distress; Neck: Supple, no JVD...
Results: Pro-BNP: 3317 pg/mL; CXR: Cardiomegaly, hazy bibasilar opacities, and a small right pleural effusion; Echo:LVEF 30% with depressed LV systolic function and a moderately dilated right atrium ...

Error Type	Model's Answer	Brief Analysis
Diagnostic Error	Patient presents with SOB on exertion. This symptom is considered consistent with HFrEF. The diagnosis is `HFrEF`.	Diagnosis too specific: The model jumps to a definitive diagnosis (HFrEF) from a general symptom, where experts infer a preliminary one.
Extraneous Information	The patient is on Torasemide 100mg QD. This is a common diuretic treatment for heart failure. The diagnosis is `Strongly Suspected Heart Failure`.	Uses non-diagnostic data: Reasoning is based on medication history, a data type not used by experts for primary diagnosis in this context.
Contextual Contradiction	Patient's temperature is T=98.1°F. The rationale cites that fever can be a systemic response to heart failure. The diagnosis is `Suspected Heart Failure`.	Finding-rationale mismatch: The rationale discusses "fever" despite citing a normal temperature (98.1°F) as the clinical finding.
Medical Factual Error	A finding of RA is noted. The rationale states reduced right atrial pressure is a sign of right-sided heart failure. The diagnosis is `Suspected Heart Failure`.	Incorrect medical fact: The rationale incorrectly states that reduced right atrial pressure signifies heart failure; it should be increased.

Figure 7: A sample of patient's clinical note from DiReCT dataset, and classification of errors from the model-generated explanation examples

G Algorithms and Prompts

Algorithm 1 Record-driven Explanation Self-learning

Require: A corpus of patient records R.
Ensure: An explanatory knowledge base B.

$B \leftarrow \emptyset$
for EACH record R_i in the corpus **do** ▷ Extract
  $M_{\text{raw}} \leftarrow$ LLM_Extractor(R_i, type="manifestation")
  $D_{\text{raw}} \leftarrow$ LLM_Extractor(R_i, type="diagnosis")
  $M, D \leftarrow$ Verify_Spans($M_{\text{raw}}, D_{\text{raw}}, R_i$)
  for EACH pair (m, d) in $(M \times D)$ **do** ▷ Explain
    query $\leftarrow m + d$
    $\hat{T} \leftarrow$ TR.search(T, query)
    is_sufficient $\leftarrow$ LLM_Assess($\hat{T}, m, d$)
    $e \leftarrow$ LLM_Generate($\hat{T}, m, d$)
    if NOT is_sufficient OR e is NULL **then**
      CONTINUE
    end if
    ▷ Validate
    $k \leftarrow (m, e, d)$
    is_duplicate $\leftarrow B$.contains_similar(k)
    is_conflicting $\leftarrow$ LLM_Check_Conflict(e, R_i)
    if is_duplicate OR is_conflicting **then**
      CONTINUE
    end if
    $\tilde{T} \leftarrow$ TR.search(T, query $= e$, exclude $= \hat{T}$)
    is_supported $\leftarrow$ LLM_Verify_External($k, \tilde{T}$)
    if NOT is_supported **then**
      CONTINUE
    end if
    $B \leftarrow B \cup \{k\}$
  end for
end for
RETURN B

Algorithm 2 Explanation Augmented Dual-phase Diagnosis

Require: A patient record R, An explanatory knowledge base B.
Ensure: A final diagnosis d_{final}, a detailed explanation E_{final}.

CandidateDiagnoses $\leftarrow$ LLM_InitialHypothesis(R, B)
$d_{\text{final}} \leftarrow$ NULL
$E_{\text{final}} \leftarrow \emptyset$
while d_{final} is NULL AND CandidateDiagnoses is not $\emptyset$ **do** ▷ Differential Diagnosis
    DiagnosisKnowledge $\leftarrow B$.search(query $=$ CandidateDiagnoses)
    $M_{\text{observed}} \leftarrow$ LLM_ExtractManifestations(R)
    ManifestationKnowledge $\leftarrow B$.search(query $= M_{\text{observed}}$)
    $d_{\text{candidate}} \leftarrow$ LLM_SelectCandidate(DiagnosisKnowledge, ManifestationKnowledge, CandidateDiagnoses)
    CandidateDiagnoses $\leftarrow$ CandidateDiagnoses $\setminus \{d_{\text{candidate}}\}$
    ▷ Definitive Diagnosis
    $d_{\text{final}} \leftarrow$ NULL
    $E_{\text{temp}} \leftarrow \emptyset$
    while TRUE **do**
        RefinedKnowledge $\leftarrow B$.search(query $= d_{\text{final}}$ OR $d_{\text{candidate}}$)
        $d_{\text{next}}, e_{\text{step}} \leftarrow$ LLM_RefineDiagnosis(RefinedKnowledge, d_{final} OR $d_{\text{candidate}}$)
        if d_{next} is NULL **then**
            BREAK
        end if
        $d_{\text{final}} \leftarrow d_{\text{next}}$
        $E_{\text{temp}} \leftarrow E_{\text{temp}} \cup \{e_{\text{step}}\}$
    end while
    $E_{\text{final}} \leftarrow E_{\text{temp}}$
end while
RETURN $(d_{\text{final}}, E_{\text{final}})$

Prompt G.1: Extractor of Extract Action

Given the patient's clinical note, extract all clinical manifestations that 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 G.2: Explanation Generator of Explain Action

Input
1. `manifestation`: 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 `manifestation`, `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 `manifestation`.
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 `manifestation` (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 G.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 G.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 G.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 G.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 G.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 G.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.