

MultiDx: A Multi-Source Knowledge Integration Framework towards Diagnostic Reasoning

Yimin Deng^{1,2*}, Zhenxi Lin³, Yejing Wang², Guoshuai Zhao^{1†}, Pengyue Jia², Zichuan Fu², Derong Xu², Yefeng Zheng⁴, Xiangyu Zhao^{2†}, Li Zhu^{1†}, Xian Wu^{3†}, Xueming Qian¹

¹Xi'an Jiaotong University, ²City University of Hong Kong,

³Tencent Jarvis Lab, ⁴Westlake University

dymanne@stu.xjtu.edu.cn, guoshuai.zhao@xjtu.edu.cn, xianzhao@cityu.edu.hk

zhuli@xjtu.edu.cn, kevinxwu@tencent.com

Abstract

Diagnostic prediction and clinical reasoning are critical tasks in healthcare applications. While Large Language Models (LLMs) have shown strong capabilities in commonsense reasoning, they still struggle with diagnostic reasoning due to limited domain knowledge. Existing approaches often rely on internal model knowledge or static knowledge bases, resulting in knowledge insufficiency and limited adaptability, which hinder their capacity to perform diagnostic reasoning. Moreover, these methods focus solely on the accuracy of final predictions, overlooking alignment with standard clinical reasoning trajectories. To this end, we propose MultiDx, a two-stage diagnostic reasoning framework that performs differential diagnosis by analyzing evidence collected from multiple knowledge sources. Specifically, it first generates suspected diagnoses and reasoning paths by leveraging knowledge from web search, SOAP-formatted case, and clinical case database. Then it integrates multi-perspective evidence through matching, voting, and differential diagnosis to generate the final prediction. Extensive experiments on two public benchmarks demonstrate the effectiveness of our approach. The code is available at <https://github.com/Applied-Machine-Learning-Lab/ACL2026-MultiDx>.

1 Introduction

Diagnostic reasoning is a critical downstream task in clinical applications (Patel et al., 2005; Lucas et al., 2024). It aims to integrate information from clinical case reports (e.g., patient symptoms, test results) to establish a diagnosis (or identify the disease). Beyond achieving the correct diagnostic prediction, it is equally important to ensure that

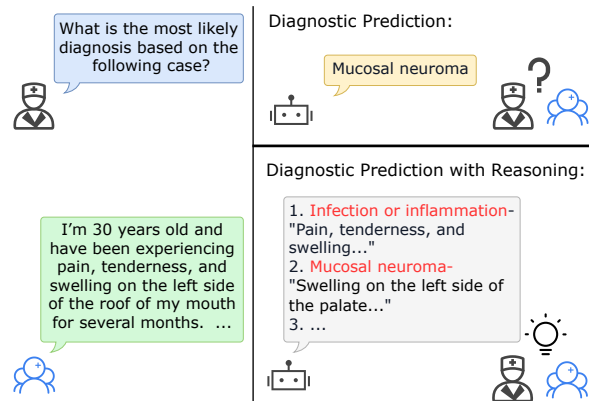


Figure 1: An example of diagnosis reasoning.

the reasoning process adheres to established medical standards (Wu et al., 2025). As illustrated in Figure 1, diagnostic predictions without reasoning can be difficult to verify or justify, potentially undermining trust. In contrast, predictions with structured reasoning not only facilitates patient understanding of their condition but also provides clinicians with verifiable and trustworthy decision support. Achieving interpretable diagnostic predictions forms the foundation of reliable and responsible AI-assisted healthcare systems.

In recent years, Large Language Models (LLMs) have been widely applied to medical reasoning (Tang et al., 2025; Chen et al., 2025). Compared to commonsense reasoning, medical reasoning heavily relies on accurate and comprehensive domain-specific knowledge. Existing approaches incorporate medical knowledge in two main ways. On the one hand, some methods rely solely on the internal knowledge of LLMs. For instance, MedAgents (Tang et al., 2024) introduces a training-free collaborative framework in which LLMs act as medical specialists, and MDA-gents (Kim et al., 2024) proposes an adaptive

*Work was conducted at Tencent Jarvis Lab.

† Corresponding authors.

strategy that dynamically assigns agents based on task complexity. On the other hand, recent methods attempt to enhance reasoning capabilities by leveraging static knowledge bases. Specifically, Medaide (Wei et al., 2024) integrates information from multiple domain-specific databases to support medical reasoning. ConfAgents (Zhao et al., 2025) adopts an adaptive and cost-efficient framework that retrieves relevant information from an external document repository. MMedAgent-RL (Xia et al., 2025) leverages reinforcement learning to learn knowledge from databases. However, current approaches often lack effective mechanisms to integrate diverse and dynamic external knowledge sources, which limits their ability to provide sufficient support for clinical reasoning, especially in unseen or rare cases.

Despite recent progress, generating high-quality reasoning trajectories remains a significant challenge. Most existing approaches are evaluated on benchmark datasets such as MedQA and Pub-MedQA (Jin et al., 2021, 2019), which primarily focus on the correctness of the final answer, rather than the quality of the diagnostic reasoning process. As a result, the diagnosis reasoning process are often under-specified, lacking a coherent chain of medical evidence. This not only reduces the explainability of the system but also undermines its verifiability from a clinical perspective. Without a well-grounded reasoning path, it becomes difficult for clinicians to trust the model’s recommendations or to cross-check them with medical knowledge. Consequently, these limitations restrict the applicability of current models in real-world clinical settings, where rigorous justification and interpretability are essential.

Therefore, it is crucial to incorporate comprehensive medical knowledge and to construct high-quality reasoning trajectories. To address these challenges, we propose MultiDx, a two-stage diagnostic reasoning framework enhanced with multi-source knowledge. In the first stage, the model leverages multiple knowledge sources, including web search results, structured case reports, and medical case database, to generate disease lists along with corresponding reasoning paths. In the second stage, the model performs disease matching, voting, and differential diagnosis to produce the final prediction and its associated reasoning trace. To ensure consistency with clinical practice, our method aligns with two core steps of the standard diagnostic workflow: generating a suspected dis-

ease list, followed by differential diagnosis through multi-perspective reasoning. Moreover, we incorporate the SOAP (Podder et al., 2021), a widely accepted format in clinical practice, to help the model effectively organize clinical case.

To summarize, our contributions are as follows:

- We propose MultiDx, a two-stage diagnostic reasoning framework that integrates multi-source knowledge while explicitly aligning with standard clinical practices.
- By incorporating diverse knowledge sources, our model can analyze clinical cases from multiple perspectives and generate comprehensive, evidence-based reasoning trajectories.
- We conduct thorough experiments on two public diagnostic reasoning benchmarks, demonstrating that our method yields both high-quality reasoning paths and accurate predictions.

2 Methods

In this section, we provide a comprehensive overview of our proposed diagnostic reasoning framework MultiDx. We first introduce the problem definition of the diagnostic reasoning task in section 2.1, and then present the overall architecture of MultiDx in section 2.2, followed by a detailed description of each component in the two-stage pipeline in section 2.3 and section 2.4.

2.1 Problem Formulation

The diagnostic reasoning task aims to generate a diagnostic prediction along with clinically sound reasoning paths based on a patient’s case report. Formally, each data instance is represented as a tuple $\{C, R, D\}$, where C denotes the case report, R the reasoning path, and D the final diagnosis. As shown in Figure 2, given a case report C that contains relevant clinical information (e.g., subjective symptoms and objective examination results), the model is expected to produce a reasoning path R that shows the step-by-step diagnostic process based on medical evidence, and a diagnosis D .

2.2 Overall Framework

In this section, we introduce the overall architecture of our model. As shown in Figure 2, our model consists of two stages, starting with *Multi-source Knowledge-guided Diagnosis Generation*, and followed by *Evidence Integration and Differential Diagnosis*. Specifically, in the first stage, the model retrieves knowledge from multiple sources: web

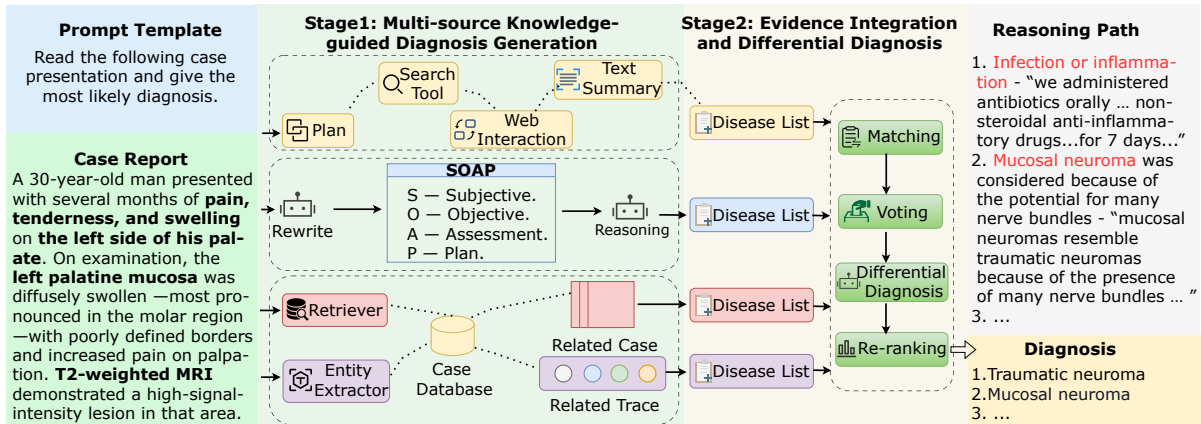


Figure 2: The overall architecture of MultiDx.

search, structured case, case database, and generates a list of suspected diseases for each source. In the second stage, the model performs disease matching, voting, and differential diagnosis to integrate the results into a final diagnosis.

2.3 Multi-source Knowledge-guided Diagnosis Generation

Unlike commonsense or math reasoning, medical reasoning relies on complete domain-specific knowledge to analyze the associations between symptoms or examination findings and potential diseases. In this section, we present how evidence is gathered from multiple complementary knowledge sources, each contributing to the generation of a list of suspected diseases.

2.3.1 SOAP

The natural language case reports received by the model are often unstructured, mixing subjective complaints and objective findings in a free-form narrative. It limits the model’s ability to process different types of clinical information effectively (Pearce et al., 2016; Seo et al., 2016). To address these limitations and facilitate diagnostic reasoning, we adopt the SOAP format, which is widely used in clinical practice (Podder et al., 2021). It divides the original clinical case into four components:

- **Subjective:** Patient-reported symptoms.
- **Objective:** Observed signs, test results, and physical examination findings.
- **Assessment:** Generation of a list of candidate diseases based on subjective and objective data.
- **Plan:** Suggestion of potential diagnostic or treatment steps tailored to the patient’s condition.

Formally, given an instruction prompt I_{toSOAP}

for text structuring and an unstructured clinical case C , we utilize a LLM to generate its structured representation C_{SOAP} in SOAP format:

$$C_{\text{SOAP}} = \text{LLM}(I_{\text{toSOAP}}, C) \quad (1)$$

Since the input to our system is the original case prompt, which contains only patient-reported symptoms and objective findings without any preliminary diagnosis or treatment plan, we use the SOAP format solely to structure this potentially noisy input into clinically standardized components. In our generated SOAP record, both Assessment and Plan are explicitly marked as absent.

This structured representation is then used as the input for diagnostic reasoning agent. Given a clinical inference prompt I_{SOAP} , the LLM generates a disease list H_{SOAP} :

$$H_{\text{SOAP}} = \text{LLM}(I_{\text{SOAP}}, C_{\text{SOAP}}) \quad (2)$$

By structuring inputs according to SOAP, we align the reasoning process with clinical standard.

2.3.2 Medical Case Database

Diagnostic reasoning follows clinical standards in real-world scenarios. It has been shown that adherence to these guidelines improves diagnostic performance. To enable the model to learn diagnostic reasoning patterns that follow clinical workflows, we incorporate hierarchical retrieval-augmented generation using a database of clinician-annotated cases.

Formally, the database $\mathcal{G} = \{(C_i, R_i, D_i)\}_{i=1}^N$, where each instance consists of a clinical case description C_i , the associated reasoning trace R_i , and the final diagnostic prediction D_i . Then we perform hierarchical retrieval, including retrieve similar cases based on textual similarity, and retrieve reasoning steps based on extracted medical entity.

Related Case. Given a input case C , we use a BM25 retriever to identify the top- k most similar cases from the corpus based on textual similarity:

$$\text{TopK}(C) = \arg \text{top-}k_{(C_i, R_i, D_i) \in \mathcal{D}} \text{BM25}(C, C_i) \quad (3)$$

The retrieved tuples are then concatenated into the prompt as additional context. Specifically, we build an augmented input by including the retrieved exemplars followed by the current case C :

$$H_{case} = \text{LLM}(I_{case}, \text{TopK}(C), C) \quad (4)$$

Here, I_{case} denotes a reasoning instruction, and H_{case} is the generated disease list.

This pipeline enables the LLM to align its reasoning with high-quality reasoning examples. It also facilitates few-shot generalization and improves robustness in complex or ambiguous cases.

Related Reasoning Trace. To further enhance diagnostic reasoning, we propose a fine-grained retrieval mechanism that operates at the level of each reasoning step. We decompose diagnostic reasoning examples into smaller segments and retrieve the most relevant ones based on medical entity overlap.

Concretely, we first segment each R_i into numbered reasoning steps $R_{i,j}$. Each segment $R_{i,j}$ is treated as an independent reasoning unit. We then extract biomedical entities from each $R_{i,j}$ using a domain-specific named entity recognizer, such as SciSpaCy (Neumann et al., 2019). Let $E_{i,j}$ denote the set of entities extracted from reasoning sentence $R_{i,j}$, and let E_C denote the set of entities extracted from the input case C . We compute the Jaccard similarity between the two sets:

$$\text{Sim}(C, R_{i,j}) = \frac{|E_C \cap E_{i,j}|}{|E_C \cup E_{i,j}|} \quad (5)$$

With this similarity score, we retrieve the top- k reasoning fragments $R_{i,j}$ from the database that share the most biomedical entities with the input:

$$\text{TopK}_{\text{entity}}(C) = \arg \text{top-}k_{R_{i,j}} \text{Sim}(C, R_{i,j}) \quad (6)$$

The retrieved reasoning segments are then concatenated and added to the prompt. The final diagnosis is generated by the LLM as:

$$H_{trace} = \text{LLM}(I_{trace}, \text{TopK}_{\text{entity}}(C), C) \quad (7)$$

Here, I_{trace} is a reasoning instruction prompt, and H_{trace} is the corresponding prediction. This entity-based retrieval approach allows for a fine-grained alignment between the input case and existing reasoning examples, supporting more focused and interpretable diagnostic generation.

2.3.3 Web Search

Medical knowledge is continually evolving, and access to up-to-date information can significantly enhance diagnostic reasoning, particularly in the case of rare diseases. Existing methods often rely on the internal knowledge of LLMs or static knowledge bases, where outdated knowledge may compromise diagnostic accuracy (Wang et al., 2025b). To enhance the model’s ability to incorporate real-world and up-to-date medical knowledge, we introduce a web search module. Following OpenAI’s deep research (OpenAI, 2025; Roucher et al., 2025), this module enables the model to iteratively retrieve relevant information from the internet and transform them into structured outputs for diagnostic reasoning. We take potential data leakage into consideration by blocking access to possible sources such as PubMed¹ and Hugging Face².

Given an instruction I_P and a clinical case report C , the model generates a high-level search plan P :

$$P = \text{LLM}(I_P, C) \quad (8)$$

where P consists of three components Q , T , and N , representing the search queries, the types of tools to be used (e.g., search, navigation, content extraction), and the estimated number of retrieval steps, respectively:

$$P = (Q, T, N) \quad (9)$$

where $Q = \{q_1, \dots, q_N\}$ and $T = \{t_1, \dots, t_N\}$.

Then the agent performs a sequence of tool invocations to execute browsing behavior.

$$s_i = \text{Invoke}(t_i, q_i), \quad \text{for } i = 1, \dots, N \quad (10)$$

At each step i , the agent selects tool type T_i (e.g., search engine, navigator, extractor) and applies it using the corresponding keyword query q_i . This results in a sequence of interactions s_1, s_2, \dots, s_N , where each s_i represents the retrieved information or browsing action at that step.

Following the complete tool execution trace $S = \{s_1, s_2, \dots, s_N\}$, the agent interacts with web content by performing actions such as visiting web-pages, scrolling through textual material, locating relevant sections, and extracting information.

At each step i , the result s_i is used to incrementally update the agent’s internal memory state m_i :

$$m_i = \text{LLM}(m_{i-1}, s_i) \quad (11)$$

¹<https://pubmed.ncbi.nlm.nih.gov/>

²<https://huggingface.co/>

The memory m_i accumulates evidence across steps, enabling the agent to refine its understanding and retrieval strategy iteratively. Once sufficient evidence has been accumulated in the memory m_N , the agent generates a disease list H_{web} based on the retrieved information with a prompt I_{web} :

$$H_{web} = LLM(I_{web}, m_N) \quad (12)$$

This design allows the agent to operate in open-ended environments, retrieve up-to-date medical knowledge, and support diagnostic predictions based on real-time web content. It is particularly useful in cases where offline knowledge bases are incomplete or outdated.

2.4 Evidence Integration and Differential Diagnosis

Efficient integration of diverse evidence is essential (Gao et al., 2025a; Fu et al., 2025). As shown in Appendix A.1, simple voting fails to capture and reason over multiple knowledge sources. In the second stage, we refine the diagnosis by integrating and comparing evidence collected from multiple sources in Stage 1. It enhances the robustness and reliability of the final diagnosis. The model outputs a final diagnosis along with a structured reasoning path that reflects the decision process in a clinical manner.

From the previous stage, we obtained four disease lists $\{H_{web}, H_{SOAP}, H_{case}, H_{trace}\}$, each generated by a distinct medical reasoning module. Each list consists of suspected diseases ordered by likelihood and along with support evidence.

Then we prompt a LLM to perform evidence integration and differential diagnosis. The process involves the following steps:

1. Disease Matching Across Sources: The LLM identifies equivalent terms across the four disease lists (e.g., “myocardial infarction” and “heart attack”), unifying them under canonical names.

2. Support Aggregation via Voting: For each unique disease candidate, the model is instructed to count its presence across the four lists and consider its rank if available. This allows the LLM to estimate the degree of inter-model agreement and relative support for each disease.

3. Differential Diagnosis: The LLM is then guided to perform analysis between highly ranked candidates (e.g., pneumonia vs. pulmonary embolism), based on the input case’s clinical features and support evidence. The model is encouraged to

use clinical logic to promote or demote diseases based on their fit to the case.

4. Final Re-ranking and Output: The LLM outputs a disease list and the output is accompanied by a brief explanation or justification per item.

Formally, the process is modeled as:

$$(R, D) = LLM(I_{multi}, C, H_{web}, H_{SOAP}, H_{case}, H_{trace}), \quad (13)$$

where I_{multi} is the instruction prompt that explicitly asks the model to generate the final prediction.

This integration ensures that comprehensive domain-specific knowledge is considered, improving both diagnostic accuracy and explainability.

3 Experiments

3.1 Experimental Setup

3.1.1 Datasets

We utilize the MedCaseReasoning (Wu et al., 2025) dataset to evaluate diagnostic reasoning methods, as it includes human-annotated reasoning paths that provide gold-standard chains of clinical inference. The dataset comprises 14,489 diagnostic QA cases, each accompanied by a detailed reasoning statement curated from open-access medical case reports, reflecting real-world clinical thought processes. We follow the standard split, using 13,092 cases from the training set to construct the case database. For evaluation, due to limited computational resources, we randomly select 300 samples from the test set.

We also conducted experiments on the DiReCT dataset (Wang et al., 2024). The dataset consists of clinical notes, each carefully annotated by physicians to reflect the full diagnostic reasoning process from observations to final diagnosis. We randomly selected 50 samples in this dataset for evaluation.

3.1.2 Baseline

To better evaluate the performance of our approach on medical diagnosis tasks, we compare it with three types of baselines: base models, which represent state-of-the-art LLMs of various sizes; fine-tuned models, which are obtained by applying fine-tuning techniques to several base models; and agentic methods. Specifically, we select several representative agentic methods. Self-Refinement (Madaan et al., 2023) improves reasoning performance through iterative self-optimization in general domain, and has been shown to be effective in medical reasoning (Tang et al., 2025) as well. MedAgents is a multi-agent framework specifically

MedCaseReasoning Test Set				
<i>Base Models</i>	<i>Reasoning Recall</i>	<i>H@1 Acc.</i>	<i>H@5 Acc.</i>	<i>H@10 Acc.</i>
DeepSeek-R1	0.648	0.360	0.419	0.442
Qwen3-14B	0.638	0.284	0.295	0.295
LLaMA-3.1-8B-Instruct [†]	0.451	0.161	0.281	0.332
Qwen-2.5-7B [†]	0.324	0.174	0.252	0.287
<i>Fine-Tuned Models</i>				
LLaMA-3.1-8B-Instruct (SFT) [†]	0.485	0.278	0.411	0.479
Qwen-2.5-7B (SFT) [†]	0.486	0.249	0.363	0.425
<i>Agentic Methods</i>				
Self-refinement (Madaan et al., 2023)	0.535	0.336	0.462	0.496
MedAgents (Tang et al., 2024)	0.641	0.344	0.458	0.471
OpenAI-DR (OpenAI, 2025)	0.557	0.416	0.553	0.602
MultiDx*	0.662	0.420	0.577	0.617
DiReCT Test Set				
	<i>Reasoning Recall</i>	<i>H@1 Acc.</i>	<i>H@5 Acc.</i>	<i>H@10 Acc.</i>
DeepSeek-R1	0.473	0.293	0.413	0.473
Self-refinement (Madaan et al., 2023)	0.662	0.300	0.466	0.586
OpenAI-DR (OpenAI, 2025)	0.586	0.297	0.452	0.479
MultiDx*	0.665	0.333	0.503	0.587

Table 1: Performance of models on test set. Each model is evaluated on reasoning coverage and diagnostic accuracy (broken down into 1, 5, and 10-shot). Results with [†] are reported in the MedCaseReasoning paper (Wu et al., 2025). Results with * are averaged over three random runs ($p < 0.05$ under t-test). To ensure a consistent comparison, all agentic methods are implemented based on the DeepSeek-R1 backbone.

designed for medical diagnostic tasks (Tang et al., 2024). OpenAI-DR (OpenAI, 2025) is a recent agentic method that achieves state-of-the-art performance on complex reasoning tasks. For a fair evaluation, we implement all these agentic methods using DeepSeek-R1 as the backbone.

3.1.3 Implementation Details

We use DeepSeek-R1 as the backbone model and utilize the official API to conduct all experiments³. Results are averaged over three random runs. While the execution time of our method is subject to fluctuations according to API server or web environment, it typically stays within a minute-level range on average. Prompts used in our experiments are provided in the appendix B. We retrieve the top 10 similar cases and reasoning paths in the hierarchical retrieval module. Medical entities are extracted using SciSpacy (version 0.5.5). To prevent data leakage, we restrict access to specific websites such as PubMed and Hugging Face within the web search module. Since the DiReCT cases are al-

³<https://www.deepseek.com/>

ready formatted in a standardized clinical style, we omit the SOAP structuring module in our pipeline during evaluation on this dataset.

3.1.4 Evaluation Metrics

Following MedCaseReasoning (Wu et al., 2025), we evaluate model performance with reasoning recall and diagnostic accuracy. Reasoning recall measures the extent to which the model can recover key steps of the clinical reasoning process. Hit@k Accuracy (H@1, H@5, H@10) evaluates diagnostic accuracy by measuring whether the correct diagnosis appears in the top-k predictions.

3.2 Main Results

We analyze the performance of our method in comparison with base models, fine-tuned models, and recent agentic approaches. The results in Table 1 demonstrate several key observations:

- Our proposed approach MultiDx achieves the best performance across all evaluation metrics. This highlights the effectiveness of our two-stage framework in diagnostic reasoning task.

	<i>H@1</i>	<i>H@5</i>	<i>H@10</i>	Recall
DeepSeek-R1	0.360	0.419	0.442	0.648
w/ SOAP	0.379	0.467	0.502	0.638
w/ web search	0.416	0.553	0.602	0.460
w/ related case	0.393	0.489	0.523	0.634
w/ related trace	0.386	0.520	0.576	0.573
MultiDx	0.420	0.577	0.617	0.662

Table 2: Ablation study.

- MultiDx achieves significant improvements on Hit@5 and Hit@10 metrics. Specifically, on the MedCaseReasoning test set, it outperforms the backbone model DeepSeek-R1 with improvements of 8.4%/17.5% and surpasses OpenAI-DR by 2.4%/1.5% on H@5/H@10. These gains suggest that integrating multi-source knowledge significantly enhances the model’s ability to recall target diagnosis within a broader candidate set.
- MultiDx achieves superior reasoning recall compared to other methods, indicating a better alignment with the ground truth. This reflects the model’s enhanced capability to reproduce standard diagnostic reasoning processes.

3.3 Ablation Study

To better understand the contribution of each component of MultiDx, we experiment with several variants based on DeepSeek-R1. The results in Table 2 show the performance of each variant enhanced by a specific type of external knowledge:

- **w/ SOAP:** This variant incorporates only structured original cases (SOAP format) without any external knowledge augmentation, yet it still improves upon the base model. It demonstrates that structured clinical data can positively contribute to diagnostic reasoning.
- **w/ web search:** This variant integrates knowledge obtained through real-time web search and achieves the best performance among all variants. This result indicates that adaptively incorporating up-to-date external knowledge contributes significantly to accurate diagnosis.
- **w/ related case & w/ related reasoning trace:** These two variants retrieve different granularities of knowledge from the case database. The results show that relevant cases effectively support diagnostic reasoning, especially with improvements in H@5 and H@10, indicating improved recall of the target diagnosis. The fine-grained retrieval even performs better since it provides more relevant and matching knowledge.

	<i>H@1</i>	<i>H@5</i>	<i>H@10</i>	Recall
Qwen3-14B	0.284	0.295	0.295	0.638
MedAgents	0.362	0.377	0.384	0.572
Self-refinement	0.241	0.345	0.404	0.491
MultiDx	0.399	0.556	0.601	0.679

Table 3: Comparison of Qwen3-14B-based methods on diagnostic reasoning.

Method	Seen		Unseen	
	<i>H@1</i>	<i>H@5</i>	<i>H@1</i>	<i>H@5</i>
DeepSeek-R1	0.459	0.520	0.300	0.366
w/ SOAP	0.468	0.572	0.292	0.367
w/ web search	0.511	0.703	0.326	0.413
w/ related case	0.489	0.621	0.276	0.365
w/ related trace	0.469	0.647	0.290	0.393
MultiDx	0.504	0.710	0.338	0.448

Table 4: Performance of variants of MultiDx on seen and unseen diseases.

Compared to aforementioned variants, MultiDx integrates evidence from multiple knowledge sources and achieves the best performance across all metrics. This demonstrates the effectiveness of comprehensive knowledge integration.

3.4 Compatibility Study

To evaluate the generalizability of MultiDx across different backbone models, we applied the multi-perspective evidence integration and differential diagnosis steps using Qwen3-14B, and compared the results to agentic methods based on the same backbone. As shown in Table 3, MultiDx consistently improves diagnostic performance across all metrics. The results show that our method remains effective when using smaller-scale LLMs, outperforming other agentic approaches.

3.5 Seen/Unseen Case Analysis

To evaluate the generalization ability of different variants, we divide the test set into two subsets: seen cases, where the target diagnosis appears in the case database, and unseen cases, where the disease has not been encountered. Table 4 reports the performance of all variants under both settings.

As shown in Table 4, the two variants that retrieve knowledge from the case database perform well on seen cases, indicating the effectiveness of leveraging relevant cases for in-distribution diseases. The variant that searches knowledge from the web demonstrates improved performance on both seen and unseen samples, highlighting

	Avg Latency (min)	Avg. Total Tokens
Self-refinement	5.3	18,398.2
OpenAI-DR	8.0	N/A
Stage1-SOAP	1.6	4,360.8
Stage1-Web Search	8.0	N/A
Stage1-Related Trace	6.2	3,108.0
Stage1-Related Case	1.8	8,591.0
Stage 2	0.46	4603.7
MultiDx	8.46	~20,000

Table 5: Computational Cost Analysis.

that leveraging comprehensive external knowledge leads to improved generalization and robustness. Overall, MultiDx achieves the best performance across both types of cases, suggesting that it generalizes well to unseen or rare diseases.

3.6 Computational Cost Analysis

To better quantify the computational cost of MultiDx, we compared the latency and token usage of each module with two agent-based baselines. Table 5 shows the per-case average latency and token consumption across modules. The token usage for the web search module is not reported due to high variability across runs, which depends on real-time network conditions and the number of iterative search steps.

Our all Stage 1 modules can be executed in parallel, which significantly reduces end-to-end latency. Moreover, our method is modular and configurable: for simpler or time-constrained use cases, users can disable high-latency modules such as web search and related trace, reducing the total time to around 2 minutes while still preserving reasonable diagnostic performance.

Compared to agent-based methods such as Self-refinement and OpenAI-DR, our framework demonstrates comparable latency and token consumption, without introducing significant additional cost. Moreover, MultiDx is modular, enabling more controllable and interpretable reasoning, and allowing users to make flexible trade-offs between latency and performance depending on the complexity of the case.

3.7 Case Study

We analyze a representative case to illustrate the model’s reasoning process. Specifically, we present the disease lists obtained from different knowledge sources, along with the final answer produced by MultiDx. As shown in Table 6, the correct answer appears at different ranks across these lists, and

SOAP	1. Intracranial Germinoma 2. Metastatic Carcinoma (e.g., from thyroid, lung, or breast primary) 3. Primary Central Nervous System Lymphoma ...
Web Search	1. Carcinomatous meningitis 2. Metastatic brain disease (e.g., from thyroid, lung, or breast primary) 3. Primary CNS lymphoma ...
RAG-case	1. Metastatic disease 2. Primary central nervous system lymphoma ...
RAG-trace	1. Primary central nervous system lymphoma 2. Metastatic disease (e.g., from thyroid carcinoma) ...
MultiDx	1. Primary central nervous system lymphoma 2. Neurosarcooidosis ...
Ground Truth	Primary central nervous system lymphoma

Table 6: Disease lists from different knowledge sources and the final prediction of MultiDx for one case.

MultiDx successfully ranks the correct diagnosis at the top after integration. In addition, our approach provides high-quality reasoning paths. The complete reasoning process is detailed in Appendix C.

4 Related Work

4.1 Agentic Methods

In recent years, a series of agent-based approaches have significantly enhanced the reasoning capabilities of LLMs (Peng et al., 2025; Xu et al., 2025a,b). Self-refine (Madaan et al., 2023) improves reasoning quality by prompting the model to revise its own outputs iteratively. Aflow (Zhang et al., 2024) optimizes reasoning trajectories by sampling and selecting the most effective workflow structures. Diverging from conventional RAG and agent-based methods, deep research agents have demonstrated strong capabilities on complex and multi-hop tasks by integrating dynamic planning, diverse retrieval tools, and advanced reasoning mechanisms (Huang et al., 2025). Search-o1 (Li et al., 2025) proposes a RAG-based agent workflow for dynamic knowledge retrieval, and employs a Reason-in-Documents module that filters and refines retrieved content before injection into the reasoning chain. OpenAI-DR (OpenAI, 2025) introduces an adaptive workflow with enhanced multi-modal context processing and integrated toolchains

for executing complex tasks. While above agentic methods have shown strong performance in general reasoning (Jia et al., 2025; Gao et al., 2025b), medical reasoning involves unique challenges, such as obtain domain-specific knowledge and follow clinical standards.

4.2 Medical Reasoning

LLMs are widely applied across various downstream tasks (Wang et al., 2023; Liu et al., 2025a,b; Wang et al., 2025a; Deng et al., 2025). Recent research has explored LLM-based frameworks for medical reasoning. MedAgents (Tang et al., 2024) proposes an agentic framework where LLMs take on specialized roles (e.g., cardiology, radiology) and participate in multi-stage discussions to simulate clinical reasoning in real-world. MDAgents (Kim et al., 2024) introduces an adaptive strategy that dynamically allocates agents. ConfAgents (Zhao et al., 2025) leverages conformal prediction to filter cases and selectively activate collaborative agents, reducing cost while maintaining diagnostic accuracy. MMedAgent-RL (Xia et al., 2025) and MedAgent-Pro (Wang et al., 2025c) extend medical multi-agent frameworks to multi-modal settings. MEDDxAgent (Rose et al., 2025) proposes a modular differential diagnosis system that iteratively refines predictions through simulated history-taking and tool-assisted reasoning.

However, most of aforementioned frameworks fall short in integrating diverse knowledge sources, particularly in dynamically retrieving information from real-time web data, which limits their capability for effective medical reasoning. In addition, they often focus on final results while neglecting the assessment of the reasoning process.

5 Conclusion

In this paper, we propose MultiDx, a two-stage diagnostic reasoning framework enhanced by multi-source knowledge. It integrates evidence from diverse knowledge sources to generate accurate diagnostic predictions as well as high-quality reasoning trajectories. Extensive experiments demonstrate the effectiveness of the proposed method.

Limitations

While MultiDx effectively integrates multiple sources of knowledge, its performance may be affected if the external sources are of low quality or contain noise. Moreover, the current two-stage

design decouples knowledge extraction and differential diagnosis. In the future, we plan to explore joint learning mechanisms to improve overall coherence and performance.

Acknowledgments

This work is in part funded by the National Key Research and Development Program of China (2023YFC3321600); in part by National Natural Science Foundation of China (Grant No. 62372364) and the Technical Innovation Guidance Plan of Shaanxi Province, China (Grant No. 2024QCY-KXJ-199); in part by National Natural Science Foundation of China (No.62502404), Hong Kong Research Grants Council (Research Impact Fund No.R1015-23, Collaborative Research Fund No.C1043-24GF, General Research Fund No. 11218325), Institute of Digital Medicine of City University of Hong Kong (No.9229503), and Tencent (Tencent Rhino-Bird Focused Research Program, Tencent University Cooperation Project).

References

- Shan Chen, Pedro Moreira, Yuxin Xiao, Sam Schmidgall, Jeremy Warner, Hugo Aerts, Thomas Hartvigsen, Jack Gallifant, and Danielle S Bitterman. 2025. Medbrowsecomp: Benchmarking medical deep research and computer use. *arXiv preprint arXiv:2505.14963*.
- Yimin Deng, Yuxia Wu, Yejing Wang, Guoshuai Zhao, Li Zhu, Qidong Liu, Derong Xu, Zichuan Fu, Xian Wu, Yefeng Zheng, Xiangyu Zhao, and Xueming Qian. 2025. A multi-expert structural-semantic hybrid framework for unveiling historical patterns in temporal knowledge graphs. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 20553–20565. Association for Computational Linguistics.
- Zichuan Fu, Xiangyang Li, Chuhan Wu, Yichao Wang, Kuicai Dong, Xiangyu Zhao, Mengchen Zhao, Huifeng Guo, and Ruiming Tang. 2025. A unified framework for multi-domain CTR prediction via large language models. *ACM Trans. Inf. Syst.*, 43(5):117:1–117:33.
- Jingtong Gao, Bo Chen, Xiangyu Zhao, Weiwen Liu, Xiangyang Li, Yichao Wang, Wanyu Wang, Huifeng Guo, and Ruiming Tang. 2025a. Llm4rerank: Llm-based auto-reranking framework for recommendations. In *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, pages 228–239. ACM.
- Jingtong Gao, Ling Pan, Yejing Wang, Rui Zhong, Chi Lu, Qingpeng Cai, Peng Jiang, and Xiangyu Zhao.

- 2025b. [Navigate the unknown: Enhancing LLM reasoning with intrinsic motivation guided exploration](#). *CoRR*, abs/2505.17621.
- Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Meng Fang, Linyi Yang, Xiaoguang Li, Lifeng Shang, Songcen Xu, Jianye Hao, et al. 2025. Deep research agents: A systematic examination and roadmap. *arXiv preprint arXiv:2506.18096*.
- Pengyue Jia, Derong Xu, Xiaopeng Li, Zhaocheng Du, Xiangyang Li, Yichao Wang, Yuhao Wang, Qidong Liu, Maolin Wang, Huifeng Guo, Ruiming Tang, and Xiangyu Zhao. 2025. [Bridging relevance and reasoning: Rationale distillation in retrieval-augmented generation](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 4242–4256. Association for Computational Linguistics.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.
- Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452.
- Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*.
- Qidong Liu, Xian Wu, Wanyu Wang, Yejing Wang, Yuanshao Zhu, Xiangyu Zhao, Feng Tian, and Yefeng Zheng. 2025a. [Llmemb: Large language model can be a good embedding generator for sequential recommendation](#). In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 12183–12191. AAAI Press.
- Qidong Liu, Xiangyu Zhao, Yuhao Wang, Yejing Wang, Zijian Zhang, Yuqi Sun, Xiang Li, Maolin Wang, Pengyue Jia, Chong Chen, Wei Huang, and Feng Tian. 2025b. [Large language model enhanced recommender systems: Methods, applications and trends](#). In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.2, KDD 2025, Toronto ON, Canada, August 3-7, 2025*, pages 6096–6106. ACM.
- Mary M Lucas, Justin Yang, Jon K Pomeroy, and Christopher C Yang. 2024. Reasoning with large language models for medical question answering. *Journal of the American Medical Informatics Association*, 31(9):1964–1975.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhunoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.
- Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.
- OpenAI. 2025. Introducing deep research. <https://openai.com/index/introducing-deep-research/>. Accessed: 2025-04-06.
- Vimla L Patel, José F Arocha, and Jiajie Zhang. 2005. Thinking and reasoning in medicine. *The Cambridge handbook of thinking and reasoning*, 14:727–750.
- Patricia F Pearce, Laurie Anne Ferguson, Gwen S George, and Cynthia A Langford. 2016. The essential soap note in an ehr age. *The Nurse Practitioner*, 41(2):29–36.
- Jingyu Peng, Maolin Wang, Xiangyu Zhao, Kai Zhang, Wanyu Wang, Pengyue Jia, Qidong Liu, Ruocheng Guo, and Qi Liu. 2025. [Stepwise reasoning disruption attack of llms](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 5040–5058. Association for Computational Linguistics.
- Vivek Podder, Valerie Lew, and Sassan Ghassemzadeh. 2021. [Soap notes](#). In *StatPearls [Internet]*. StatPearls Publishing, Treasure Island (FL). Updated 2023 Aug 28.
- Daniel Rose, Chia-Chien Hung, Marco Lepri, Israa Alqassem, Kiril Gashteovski, and Carolin Lawrence. 2025. Meddxagent: A unified modular agent framework for explainable automatic differential diagnosis. *arXiv preprint arXiv:2502.19175*.
- Aymeric Roucher, Albert Villanova del Moral, Thomas Wolf, Leandro von Werra, and Erik Kaunismäki. 2025. ‘smolagents’: a smol library to build great agentic systems. <https://github.com/huggingface/smolagents>.
- Ji-Hyun Seo, Hyun-Hee Kong, Sun-Ju Im, HyeRin Roh, Do-Kyong Kim, Hwa-ok Bae, and Young-Rim Oh. 2016. A pilot study on the evaluation of medical student documentation: assessment of soap notes. *Korean journal of medical education*, 28(2):237.

- Xiangru Tang, Daniel Shao, Jiwoong Sohn, Jiapeng Chen, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, Wenqi Shi, et al. 2025. Medagentsbench: Benchmarking thinking models and agent frameworks for complex medical reasoning. *arXiv preprint arXiv:2503.07459*.
- Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 599–621.
- Bowen Wang, Jiuyang Chang, Yiming Qian, Guoxin Chen, Junhao Chen, Zhouqiang Jiang, Jiahao Zhang, Yuta Nakashima, and Hajime Nagahara. 2024. Direct: Diagnostic reasoning for clinical notes via large language models. *Advances in neural information processing systems*, 37:74999–75011.
- Hanbing Wang, Xiaorui Liu, Wenqi Fan, Xiangyu Zhao, Venkataramana Kini, Devendra Pratap Yadav, Fei Wang, Zhen Wen, and Hui Liu. 2025a. [Rethinking large language model architectures for sequential recommendations](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP-AAACL 2025, Mumbai, India, December 20-24, 2025*, pages 3376–3391. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. 2025b. A survey of llm-based agents in medicine: How far are we from baymax? *arXiv preprint arXiv:2502.11211*.
- Yuhao Wang, Xiangyu Zhao, Bo Chen, Qidong Liu, Huifeng Guo, Huanshuo Liu, Yichao Wang, Rui Zhang, and Ruiming Tang. 2023. [PLATE: A prompt-enhanced paradigm for multi-scenario recommendations](#). In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 1498–1507. ACM.
- Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiakuan Li, and Yueming Jin. 2025c. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*.
- Jinjie Wei, Dingkan Yang, Yanshu Li, Qingyao Xu, Zhaoyu Chen, Mingcheng Li, Yue Jiang, Xiaolu Hou, and Lihua Zhang. 2024. Medaide: Towards an omni medical aide via specialized llm-based multi-agent collaboration. *arXiv preprint arXiv:2410.12532*.
- Kevin Wu, Eric Wu, Rahul Thapa, Kevin Wei, Angela Zhang, Arvind Suresh, Jacqueline J Tao, Min Woo Sun, Alejandro Lozano, and James Zou. 2025. Medcasereasoning: Evaluating and learning diagnostic reasoning from clinical case reports. *arXiv preprint arXiv:2505.11733*.
- Peng Xia, Jinglu Wang, Yibo Peng, Kaide Zeng, Xian Wu, Xiangru Tang, Hongtu Zhu, Yun Li, Shujie Liu, Yan Lu, et al. 2025. Mmedagent-rl: Optimizing multi-agent collaboration for multimodal medical reasoning. *arXiv preprint arXiv:2506.00555*.
- Derong Xu, Pengyue Jia, Xiaopeng Li, Yingyi Zhang, Maolin Wang, Qidong Liu, Xiangyu Zhao, Yichao Wang, Huifeng Guo, Ruiming Tang, Enhong Chen, and Tong Xu. 2025a. [Align-grag: Reasoning-guided dual alignment for graph retrieval-augmented generation](#). *CoRR*, abs/2505.16237.
- Derong Xu, Xinhang Li, Ziheng Zhang, Zhenxi Lin, Zhihong Zhu, Zhi Zheng, Xian Wu, Xiangyu Zhao, Tong Xu, and Enhong Chen. 2025b. [Harnessing large language models for knowledge graph question answering via adaptive multi-aspect retrieval-augmentation](#). In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 25570–25578. AAAI Press.
- Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, et al. 2024. Aflow: Automating agentic workflow generation. *arXiv preprint arXiv:2410.10762*.
- Huiya Zhao, Yinghao Zhu, Zixiang Wang, Yasha Wang, Junyi Gao, and Liantao Ma. 2025. Confagents: A conformal-guided multi-agent framework for cost-efficient medical diagnosis. *arXiv preprint arXiv:2508.04915*.

	$H@1$	$H@5$	$H@10$
w/ vote	0.403	0.552	0.604
w/ D.D.	0.42	0.577	0.617

Table 7: Performance of different integration strategy used in stage2.

Retrieved Sample	$H@1$	$H@5$	$H@10$
$k = 0$	0.360	0.419	0.442
$k = 2$	0.368	0.502	0.559
$k = 5$	0.380	0.491	0.540
$k = 10$	0.393	0.489	0.523
$k = 15$	0.391	0.490	0.529

Table 8: Effect of the number of retrieved medical cases (k) on diagnosis accuracy in the case-enhanced Generation Module. Performance is measured by Hit@1, Hit@5, and Hit@10.

A In-depth Analysis

A.1 Integration Strategy

To investigate the impact of different integration strategies, we compare simple voting (w/ vote) with our proposed stage 2: Evidence Integration and Differential Diagnosis (w/ D.D.).

As shown in the Table 7, although the simple voting provides a straightforward way to select the most frequent answer, it lacks the ability to analysis the medical evidence of each candidate within the given context. In contrast, our approach explicitly prompts the model to compare different diagnoses, enabling more informed and reasonable decisions.

A.2 Parameter Analysis

To evaluate the parameter sensitivity of MultiDx, we investigate the impact of the number of retrieved medical cases (k) on diagnostic accuracy within the case retrieval module. As shown in Table 8, increasing k from 2 to 10 leads to an improvement in $H@1$, indicating that retrieving more cases helps the model make more accurate diagnosis predictions. However, $H@5$ and $H@10$ slightly decrease during the same range, suggesting that the inclusion of additional cases may introduce noise or irrelevant information, which hinders the recall of correct diagnosis. As k increases from 10 to 15, the performance remains relatively stable.

Overall, incorporating retrieved cases improves diagnostic performance, and the model demonstrates robustness to k within a reasonable range.

B Prompt

In this section, we provide additional experimental details to facilitate better reproducibility. In Table 9, 10, we present example prompts used in each module of MultiDx. Since web search module involves a larger number of prompts, detailed prompt examples of this module are provided in our code repository due to space limitations⁴. Our output and evaluation templates follow those of MedCaseReasoning (Wu et al., 2025) to ensure consistency.

C Case Study

In this section, we conduct a case study, using a specific example to explicitly illustrate the output of each module, including the list of suspected diseases and the final reasoning trajectory along with the diagnosis prediction. This example demonstrates how our method performs diagnostic reasoning in a step-by-step manner.

Table 11 illustrates how individual evidence sources contribute to differential diagnosis generation. Table 12 presents a side-by-side comparison between the ground truth reasoning trajectory (from the MedCaseReasoning dataset, validated by expert annotators) and the output reasoning results from MultiDx for a complex CNS case. The two show strong consistency in terms of core diagnostic hypotheses, reasoning structure, exclusion logic, and evidence grounding.

Based on this case, we conducted a qualitative analysis to further validate how MultiDx’s reasoning trace aligns with expert clinical thinking. Specifically, MultiDx correctly identifies primary CNS lymphoma as key differential diagnose, supported by clinical findings such as CSF pleocytosis, elevated IgG, and multifocal brain lesions. The model organizes its evidence by diagnostic categories (e.g., neoplastic, infectious, inflammatory) and provides clear justification for prioritization. It also demonstrates well-considered clinical judgment by excluding less likely options (e.g., metastatic disease due to lack of primary tumor), reflecting a degree of uncertainty handling similar to human experts.

To further enhance transparency, we have collected and organized the complete reasoning traces from all four modules, as shown in Table 13,14.

⁴<https://github.com/Applied-Machine-Learning-Lab/ACL2026-MultiDx>

<p><i>I</i>_{SOAP}</p>	<p>Read the following case presentation and give the most likely diagnosis. First, please classify all extracted facts into the following categories:</p> <p>—</p> <p>**S — Subjective** Patient-reported symptoms, complaints, history of present illness, medication use, and relevant negatives (e.g., "denies fever"). Include time-related info if mentioned.</p> <p>**O — Objective** Clinician-observed findings: physical exam results, lab/imaging findings, vital signs, and other measurable or observable data.</p> <p>**A — Assessment** If any differential diagnoses, impressions, or suspected conditions are mentioned in the text, extract them here.</p> <p>**P — Plan** If the text includes management plans, tests ordered, treatments started, or follow-up instructions, include them here.</p> <p>—</p> <p>Then, provide your internal reasoning for the diagnosis within the tags <think> ... </think>.</p> <p>Finally, output the final diagnosis (just the name of the disease/entity) within the tags <answer> ... </answer>. What are the top 10 most likely diagnoses? Be precise, listing one diagnosis per line, and try to cover many unique possibilities (at least 10).</p> <hr/> <p>CASE PRESENTATION</p> <hr/> <p><case></p>
<p><i>I</i>_{case}</p>	<p>Read the following case presentation and give the most likely diagnosis.</p> <p>What are the top 10 most likely diagnoses? Be precise, listing one diagnosis per line, and try to cover many unique possibilities (at least 10). Here are some reasoning examples: <top-k cases></p> <hr/> <p>CASE PRESENTATION</p> <hr/> <p><case></p>

Table 9: Prompt.

<p><i>I_{trace}</i></p>	<p>Read the following case presentation and give the most likely diagnosis. What are the top 10 most likely diagnoses? Be precise, listing one diagnosis per line, and try to cover many unique possibilities (at least 10). Some possible disease-related symptoms are listed (optional to use). <tok-k reasoning paths></p> <hr/> <p>CASE PRESENTATION</p> <hr/>
<p><i>I_{multi}</i></p>	<p><case></p> <p>You are a medical reasoning expert. You are given a clinical case prompt along with diagnostic reasoning traces and suspected disease lists from four different medical reasoning systems. Your task is to:</p> <ol style="list-style-type: none"> Carefully consider the case prompt. Analyze and consolidate the reasoning traces from the four sources. Based on the reasoning and the case details, produce: <ul style="list-style-type: none"> - A coherent and medically sound final reasoning trace. - A ranked list of suspected diseases, ordered by confidence (from highest to lowest), based on the degree of support across sources. <p>[Deep Search Reasoning Trace] pred_search [SOAP Reasoning Trace] pred_SOAP [RAG Reasoning Trace] pred_trace [RAG Reasoning Trace] pred_rag</p> <hr/> <p>CASE PRESENTATION</p> <hr/> <p><case></p>

Table 10: Prompt.

Case	A 70-year-old woman with atrial fibrillation on enoxaparin, liver cirrhosis, type 2 diabetes mellitus, hypertension, and a history of venous thrombus presented with a 5-day history of left arm weakness, numbness in the left fourth and fifth digits and medial palmar surface, and confusion. She reported chronic vision loss in the right eye but denied headache or trauma. On examination, she was alert; oriented to person, place, and time; and disoriented to situation. Visual acuity was reduced in the right eye. Muscle strength was 4/5 in the left triceps, wrist flexors, wrist extensors, and finger flexors, and 3/5 in the left finger extensors. The left triceps reflex was absent. Sensation to pinprick and light touch was decreased in the left fourth and fifth digits. There was no dysmetria or dysdiadochokinesia. ...
Answer	Primary central nervous system lymphoma
SOAP	1. Intracranial Germinoma 2. Metastatic Carcinoma (e.g., from thyroid, lung, or breast primary) 3. Primary Central Nervous System Lymphoma 4. Neurosarcoidosis 5. Leptomeningeal Carcinomatosis 6. Multiple Sclerosis 7. Autoimmune Encephalitis (e.g., anti-NMDA receptor encephalitis) 8. Cerebral Vasculitis (e.g., primary angiitis of the CNS) 9. Infectious Meningoencephalitis (e.g., viral, fungal, or tuberculous) 10. Paraneoplastic Neurological Syndrome
Web Search	1. Carcinomatous meningitis 2. Metastatic brain disease (e.g., from thyroid, lung, or breast primary) 3. Primary CNS lymphoma 4. Neurosarcoidosis 5. Chronic meningitis of unknown etiology 6. Lymphomatous meningitis 7. Tuberculous meningitis 8. Fungal meningitis 9. Autoimmune encephalitis 10. Vasculitis of the central nervous system
RAG-case	1. Metastatic disease 2. Primary central nervous system lymphoma 3. Neurosarcoidosis 4. Tuberculous meningitis 5. Fungal meningitis 6. Leptomeningeal carcinomatosis 7. Glioblastoma multiforme 8. Inflammatory pseudotumor 9. IgG4-related disease 10. Atrial myxoma with cerebral metastases
RAG-trace	1. Primary central nervous system lymphoma 2. Metastatic disease (e.g., from thyroid carcinoma) 3. Germ cell tumor (e.g., germinoma) 4. Neurosarcoidosis 5. Meningiomatosis or multiple meningiomas 6. Tuberculosis or other chronic infectious meningitis 7. Demyelinating disease (e.g., multiple sclerosis) 8. Vascular malformation or aneurysm 9. Pineal parenchymal tumor (e.g., pineocytoma or pineoblastoma) 10. Leptomeningeal carcinomatosis from occult primary
MultiDx	1. Primary central nervous system lymphoma 2. Neurosarcoidosis 3. Metastatic brain disease (e.g., from thyroid or unknown primary) 4. Carcinomatous meningitis 5. Germ cell tumor (e.g., germinoma) 6. IgG4-related disease 7. Tuberculous meningitis 8. Fungal meningitis 9. Leptomeningeal carcinomatosis 10. Autoimmune encephalitis

Table 11: Case study.

<p>Ground Truth</p>	<p>1. Embolic stroke was considered given her atrial fibrillation and diabetes but felt less likely — “While the patient had several risk factors for embolic stroke, including atrial fibrillation and diabetes, the specific patterns of sensory loss in the left upper extremity and loss of the left triceps reflex were more suggestive of either nerve root or brachial plexus pathology.” 2. Extradural or intradural extramedullary spinal canal mass and idiopathic brachial neuritis were considered given her focal arm findings — “Our differential diagnosis included a lateralized mass in the extradural or intradural extramedullary spinal canal or idiopathic brachial neuritis.” 3. Leptomeningeal metastases from carcinoma or lymphoma were considered after CN III involvement — “Considerations included leptomeningeal disease from metastatic carcinoma or lymphoma as well as PCNSL...” 4. Primary central nervous system lymphoma was considered due to her age and multifocal neurologic presentation — “Considerations included...PCNSL due to the patient’s age and multifocal nature of her symptoms...” 5. Tuberculosis, histoplasmosis, coccidioidomycosis, and cryptococcosis were considered among infectious etiologies — “Considerations included...tuberculosis, histoplasmosis, coccidioidomycosis, or cryptococcosis...” 6. Sarcoidosis was considered among inflammatory causes — “Considerations included...sarcoidosis.” 7. Neurolymphomatosis of the brachial plexus was considered but not supported by imaging — “Gadolinium-enhanced brachial plexus and spine MRI were unrevealing.” 8. Metastatic thyroid carcinoma was considered given the CT findings — “Chest/abdomen/pelvis CT with contrast demonstrated multiple nodules in the thyroid concerning for metastases.” 9. Leptomeningeal dissemination and infection were less likely after nondiagnostic CSF studies — “CSF studies from two serial lumbar punctures...demonstrated lymphocytic pleocytosis, elevated protein, and elevated IgG index but neither malignant cells nor other inflammatory and infectious markers.”</p>
<p>MultiDx</p>	<p>After consolidating the reasoning traces from the four sources, several key points emerge: - The multifocal CNS lesions, rapid progression, and CSF abnormalities strongly suggest an aggressive neoplastic or inflammatory process.</p> <ul style="list-style-type: none"> - Primary CNS lymphoma is highly supported due to its typical presentation with multifocal enhancing lesions, CSF pleocytosis, elevated IgG index, and occurrence in older adults. The lack of malignant cells in CSF does not rule it out entirely, as cytology can be negative. - Neurosarcoidosis is also well-supported, given the cranial nerve involvement, dural lesions, elevated IgG index, and potential systemic correlates (e.g., thyroid nodules could represent granulomas). However, the rapid progression might be less typical. - Metastatic disease is plausible due to the thyroid nodules and brain lesions, but the absence of a confirmed primary and negative CSF cytology reduces its likelihood. Brain metastases from thyroid cancer are rare. - Other considerations include carcinomatous meningitis, germ cell tumors, and inflammatory disorders, but they have less consistent support across sources. - Infectious etiologies like tuberculous or fungal meningitis are less likely due to negative markers and lack of systemic signs. - The elevated IgG index points towards inflammatory or autoimmune processes, but the aggressive course favors neoplasms. Based on the degree of support across sources and medical plausibility, primary CNS lymphoma is the most likely diagnosis, followed by neurosarcoidosis. The ranked list prioritizes conditions that best explain the multifocal lesions, CSF findings, and progression.

Table 12: Case study.

SOAP	<p>The patient presents with multifocal neurological deficits, including left arm weakness and numbness (suggesting brachial plexus or root involvement), confusion, and a right third cranial nerve palsy that progresses. Imaging reveals lesions in characteristic locations such as the suprasellar and pineal regions, which are common sites for germ cell tumors like germinoma. The CSF shows inflammatory changes (lymphocytic pleocytosis, elevated protein, and IgG index) without evidence of infection or malignancy, which can occur with germinomas due to their inflammatory response. However, the patient's age of 70 is atypical for germinoma, which is more common in younger adults. Alternative considerations include metastatic cancer (given the thyroid nodules suspicious for metastases), primary CNS lymphoma, or neurosarcoidosis. The dural thickening and multifocal enhancing lesions favor a neoplastic process, but the inflammatory CSF and absence of malignant cells might support germinoma or an autoimmune disorder. The rapid progression and systemic findings (thyroid nodules) weigh towards a malignant etiology, but the specific lesion locations make germinoma a strong candidate despite the age.</p>
Web Search	<p>The case involves a 70-year-old woman with a history of atrial fibrillation on enoxaparin, liver cirrhosis, type 2 diabetes mellitus, hypertension, and venous thrombus. She presented with acute neurological symptoms including left arm weakness, numbness in specific digits, confusion, and later developed a right third cranial nerve palsy. Imaging showed a hyperdense suprasellar lesion on CT, and MRI revealed enhancing lesions in suprasellar, pineal, and periaxial regions with interval enlargement and dural thickening. CT abdomen showed thyroid nodules concerning for metastases. CSF analysis indicated lymphocytic pleocytosis, elevated protein, and elevated IgG index without malignant cells or infectious markers. This clinical picture is highly suggestive of a neoplastic process with CNS involvement, most likely carcinomatous meningitis, given the multifocal neurological signs, CSF findings, and possible primary malignancy. Alternative diagnoses include metastatic disease, lymphoma, or inflammatory disorders, but the rapid progression and lack of infectious markers favor a neoplastic etiology.</p>

Table 13: Case study.

<p>Related Case</p>	<ol style="list-style-type: none"> 1. Pituitary adenoma – Considered due to the suprasellar lesion on CT, but the multifocal nature and rapid progression argue against a typical adenoma. 2. Meningioma – The dural thickening could suggest meningioma, but the multifocal and heterogeneously enhancing lesions with rapid growth are atypical. 3. Metastatic disease – Highly likely given the thyroid nodules concerning for metastases and multiple brain lesions; however, no primary was confirmed and CSF lacked malignant cells. 4. Lymphoma – Primary CNS lymphoma can present with multifocal enhancing lesions and CSF pleocytosis, but the thyroid involvement is unusual. 5. Neurosarcoidosis – Could explain multifocal lesions and CSF findings, but the rapid progression and thyroid nodules are not typical. 6. Tuberculosis or fungal infection – Chronic infections can cause basilar enhancement and CSF abnormalities, but there were no systemic signs of infection and markers were negative. 7. Aneurysmal subarachnoid hemorrhage – The initial hyperdense suprasellar lesion could represent a thrombosed aneurysm, but the subsequent multifocal enhancing lesions are not consistent. 8. Inflammatory or autoimmune disorders – Such as CLIPPERS or IgG4-related disease, could account for the lesions and CSF findings, but the thyroid nodules are atypical. 9. Glioblastoma multiforme – Can be multifocal and show rapid growth, but the suprasellar and pineal locations are uncommon. 10. Atrial myxoma with metastases – Given her atrial fibrillation, cardiac myxoma could embolize or metastasize to brain, but no cardiac mass was reported.
<p>Related Trace</p>	<p>The patient is a 70-year-old woman with multiple intracranial enhancing lesions in the suprasellar, pineal, and right periaxial regions, along with dural thickening. The CSF findings show lymphocytic pleocytosis, elevated protein, and an elevated IgG index, which are indicative of an inflammatory or neoplastic process. Systemic imaging (CT chest/abdomen/pelvis) did not reveal a primary malignancy elsewhere, except for thyroid nodules that are concerning for metastases, but these could be incidental given their common occurrence in elderly patients. The neurological symptoms, including cranial nerve palsy, weakness, and sensory deficits, along with the progression of lesions on MRI, strongly suggest a neoplastic etiology. Primary CNS lymphoma is a leading consideration due to its propensity for multifocal enhancing lesions, CSF abnormalities, and occurrence in older adults. Other possibilities include metastatic disease, germ cell tumors, or inflammatory conditions like neurosarcoidosis, but the absence of infectious markers and the rapid progression make lymphoma most likely. The thyroid nodules might represent a primary source, but brain metastases from thyroid cancer are rare and typically not in these locations, reducing its likelihood compared to CNS lymphoma.</p>

Table 14: Case study.