Dialogue is Better Than Monologue: Instructing Medical LLMs via Strategical Conversation

Anonymous ACL submission

Abstract

Current medical AI models are trained primarily on static articles and question-answering (QA) tasks, and then evaluated on similar QA benchmarks. However, previous approaches fail to capture the dynamic real-world nature of clinical reasoning, particularly in handling ambiguous inputs (e.g., conflicting symptoms) and multi-step decision-making. To address this, we: ❶ introduce a comprehensive diagnostic benchmark, **MuddyMaze**, evaluating clinical reasoning with controlled noise and USMLE-aligned difficulty levels; ❷ curate a new dialogue dataset by converting 10.2k medical QA pairs and 12k PubMed articles into clinician-patient interactions; and ❸ develop dialogue-based fine-tuning that enhances reasoning capabilities. Experiments demonstrate significant improvements over traditional methods (+16.10% in one-round accuracy and +4.06% in multi-round reasoning), validating that dialogue-based training better aligns AI systems with real clinical workflows.

1 Introduction

Large language models (LLMs) have achieved significant advancements in the medical domain, especially in disease analysis and diagnostic assessment (Singhal et al., 2023; Li et al., 2023b,a; Chen et al., 2023; Peng et al., 2023; Kwon et al., 2024), while evaluating these medical LLMs remains a significant challenge. Many medical LLM benchmarks are derived from medical examinations or research articles, framed as multi-choice question-answering and long-answer reasoning tasks, (Jin et al., 2021; Pal et al., 2022). Although these benchmarks have provided valuable insight into the capabilities of medical LLMs which enhanced AI performance, they primarily work in articulated environments that diverge from the complexities of actual diagnostic procedures (Chen et al., 2024a; Yao et al., 2024).

The highly structured nature of such existing benchmarks and environments fail to reflect the reasoning required in real-world clinical scenarios, where systematic diagnostic investigation and step-by-step analysis are essential: (i) *Real-World Diagnosis is Stepwise*: Real-world diagnostic reasoning is an iterative, dynamic process where clinicians progressively refine hypotheses by actively seeking information—such as asking targeted follow-up questions or ordering specific tests—based on evolving evidence. In contrast, medical QA tasks provide all information upfront, testing only final interpretation—not the critical skill of deciding what to ask next. This omits the core challenge of real-world medicine: acting under uncertainty to strategically acquire missing information.

(ii) *Real-World Diagnosis Embraces Noise*: Clinical practice routinely deals with incomplete or conflicting data (e.g., vague symptoms, false-positive tests) requiring probabilistic reasoning. Current benchmarks present artificially clean cases, testing recall of textbook knowledge rather than the essential skill of uncertainty management through evidence weighting and differential diagnosis.

To address these gaps, we first analyze existing benchmarks (Jin et al., 2019, 2020) and observe that clinical information typically follows a natural sequence: from the patient background through examination findings to the final diagnosis. the sequential order mirrors real-world diagnostic processes. Based on these insights, we reframe question-answering tasks into a stepwise diagnostic pipeline: retrieving relevant clinical evidence and ranking it to mirror physician reasoning. In this new benchmark, models are required to integrate relevant evidence to support their answers under two tasks: ❶ the one-round ranking task that requires the model to rank all evidence once, and ❷ the multi-round ranking employing an iterative approach that retrieves and ranks evidence over several steps, gradually constructing the necessary context for decision-making. This reformulation reflects the reasoning process that doctors use when combining and synthesizing clinical evidence to reach a diagnosis or treatment decision. To reflect real-world challenging cases and noisy clinical contexts, we also introduce varying levels of difficulty—basic, advanced, and challenging—aligned with United States Medical Licensing Examination (USMLE) frameworks and irrelevant information sampled from other documents to mimic noise.

To address the limitations of static training, we explore dialogue-based fine-tuning as a method for training medical LLMs, converting multiple-choice datasets and PubMed articles into dialogue structures to enhance reasoning in diagnosis decisions. Our experiments show that this approach outperforms traditional methods on evidence retrieval and ranking benchmarks, demonstrating its effectiveness in improving the model's ability to capture intricate relationships and critical insights essential for medical reasoning. Our contributions are summarized below.

- We introduced an innovative benchmark, **MuddyMaze**, that transforms traditional document-based multi-choice tasks into step-by-step evidence ranking challenges, reflecting real-world diagnostic reasoning with tiered difficulty and noise levels.
- We developed dialogue datasets that imitate real-world doctor-patient reasoning processes using publicly available medical information, enabling models to train on clinically relevant scenarios while preserving patient confidentiality.
- We proposed a dialogue-based fine-tuning framework that enhances models' ability to capture logical relationships in medical evidence. Compared to standard approaches, ours improves clinical evidence retrieval, enhances differential diagnosis prioritization, and better mirrors real-world physician reasoning patterns.

2 Related Work

Medical Large Language Models. LLMs have demonstrated considerable progress in the medical field, (Singhal et al., 2023; Chen et al., 2023; Wu et al., 2023; Zhang et al., 2024b; Gema et al., 2024; Han et al., 2023; Xie et al., 2024), achieving state-of-the-art performance in medical question-answering and summarization task. Medical LLMs are typically fine-tuned on medical examinations, scientific literature, clinical guidelines, and clinical notes from EHRs to enable them to excel in a wide range of medical text analysis and real-world clinical tasks. Moreover, bio-focused large language models are being tailored specifically to address the unique challenges of biomedical research and healthcare applications (Luo et al., 2022; Bannur et al., 2023). For instance, BioBERT (Lee et al., 2020) and PubMedBERT (Gu et al., 2021) are foundational models pre-trained on PubMed articles, enabling them to excel in natural language understanding tasks such as named entity recognition (NER) and relation extraction.

Conversation Datasets in the Medical Field. Previously, medical benchmarks primarily focused on assessing knowledge using static question formats, such as multiple-choice tasks or extractive question answering. Examples include MedQA (Yao et al., 2024), MedMCQA (Pal et al., 2022), and PubMedQA (Jin et al., 2019), which are constructed from medical articles and examinations. Recent efforts have shifted towards collecting conversational datasets from real-world doctor-patient dialogues. Examples include MedDialog and ReMeDi (Yan et al., 2022) (Zeng et al., 2020), which introduces conversational benchmarks aimed at reflecting real-world medical scenarios through doctor-patient interactions. Additionally, there are some dialogue datasets target specific domains within healthcare, such as mental health counseling (Chen et al., 2024b), pediatrics (Zhang et al., 2024a), and Covid-19 (Ju et al., 2020).

3 Methodology

This section presents our approach to improving medical reasoning in AI. We first distinguish between monologue and dialogue training formats. Next, we detail our method for converting standard medical datasets into dialogue form using an LLM. Finally, we present dialogue-tuning, which leverages this format to improve reasoning by simulating real-world doctor-patient interactions.

3.1 Preliminary

Monologue Tuning. Monologue-style datasets primarily involve static, non-interactive formats such as multi-choice question answering and article-based tasks. Medical models often rely on biomedical literature for pretraining or fine-tuning, leveraging either the entire article or just the abstract. Moreover, fine-tuning on multi-choice question-answering datasets is a standard practice for many medical AI models. However, this Monologue-style format, while useful for acquiring medical knowledge, often fails to mimic the complexity of real-world diagnostic reasoning, which requires iterative problem-solving and logical synthesis.

Dialogue Tuning. Dialogue datasets, in contrast, simulate real-world interactions by focusing on dynamic exchanges between patients and doctors. These datasets emphasize the iterative nature of diagnostic reasoning, where information unfolds progressively over multiple conversational turns. Notably, doctor-patient dialogues showcase how doctors determine the next steps based on the information provided. This process highlights how doctors gather details and arrive at a final diagnosis. Compared to monologues, dialogues inherently capture the dynamic reasoning process characteristic of real-world diagnostic scenarios.

A detailed comparison between monologue tuning and dialogue tuning is provided in Figure 1.

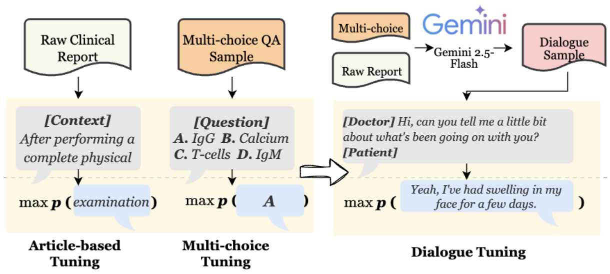

Figure 1: Previous medical LLMs are trained on next token prediction with medical text (Article-based tuning) or medical Question-Answer pair (Multi-choice tuning). For our dialogue tuning, we convert raw article and multi-choice QA samples into dialogue samples with Gemini-2.5 flash.

3.2 Data Reformation for Dialogue Tuning

To better simulate real-world doctor–patient interactions, we reformulated existing multiple-choice datasets (10.2k question–answer pairs) and approximately 12k article-based datasets—into a dialogue format, as shown in Figure 2.

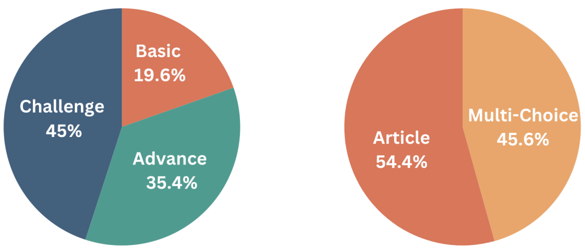

Figure 2: The left pie chart represents the ratio of difficulty levels in our benchmark. While the right pie chart represents the proportion of multiple-choice question-answering sets and articles used during the tuning stage, the dialogues generated from these sources are equal in quantity to them.

Multi-Choice to Dialogue. For multiple-choice tasks, we transformed each question and its context into a doctor-patient dialogue format. This approach aligns with the inherent structure of medical QA, which typically includes key clinical components such as patient demographics, symptom descriptions, physical examination findings, and diagnostic outcomes. The dialogue framework was designed to simulate natural clinical interactions: patients present their medical history and current symptoms, while physicians systematically elicit information and guide the conversation toward an accurate diagnosis and appropriate treatment plan.

Article-Based Tasks to Dialogue. For article-based tasks, we adapted a parallel dialogue conversion methodology. This process began by extracting structured case reports from medical literature, including key clinical elements such as patient histories, physical examination findings, diagnostic workups, and treatment outcomes. The reports were systematically converted into naturalistic physician-patient dialogues, maintaining all essential medical information while translating technical details into comprehensible clinical talks.

Implementation. We employed the Gemini-2.5 Flash to perform the task reformulation, as shown in Figure 1. For each dataset, we designed specific prompts tailored to guide the model in generating the desired dialogue format while adhering to the constraints. These prompts ensured that the original task's essential details were faithfully preserved in the reformulated dialogue. The number of dialogue rounds and their length varied according to the complexity and amount of information in the source context. The complete prompts used for both multiple-choice and article-based tasks are provided in Appendix A. Additionally, we conducted human evaluations to validate the quality of the reformulated dialogues; these results are presented in the subsection 5.2.

3.3 Implementation of Dialogue Tuning.

Our dialogue-tuning approach trains the LLM to generate physician responses conditioned on the previous conversational context. This differs fundamentally from conventional next-token prediction by operating at the level of complete dialogue acts rather than individual tokens.

Training Objective. Given a dialogue sequence $D = \{u_1, ..., u_T\}$ where each u_t is an utterance alternating between patient (P) and doctor (D) roles, we optimize the model to predict doctor responses conditioned on preceding dialogue history. Formally:

Let $x_t = \{u_1, ..., u_{t-1}\}$ be the context preceding doctor's turn u_t (where u_{t-1} must be patient utterance). The training objective maximizes:

$$\mathcal{L} = -\sum_{t \in \mathcal{T}_D} \sum_{i=1}^{|u_t|} \log P(w_i^t | w_{<i}^t, x_t) \quad (1)$$

where $\mathcal{T}_D$ denotes doctor turn indices, w_i^t represents the i-th token in doctor utterance u_t, and $x_t = \{u_1, ..., u_{t-1}\}$ contains all preceding dialogue (ending with patient input). The loss is computed only over doctor response tokens while conditioning on the full conversational context.

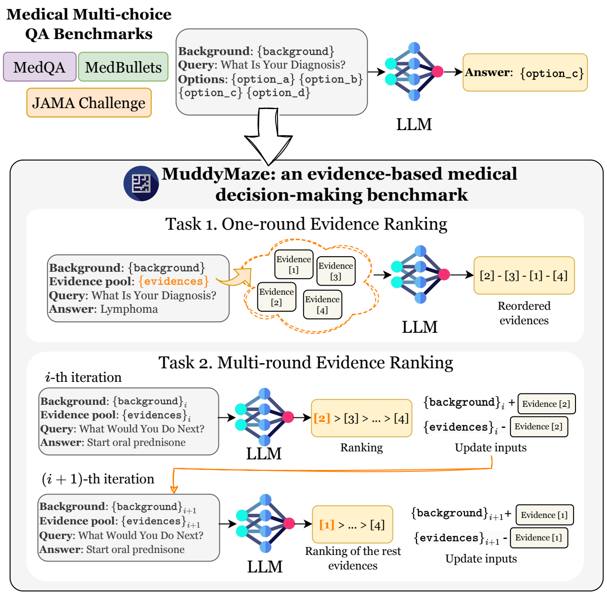

Figure 3: The MuddyMaze benchmarks encompass two settings: one-round evidence ranking and multi-round evidence ranking. In the one-round evidence ranking, the model is required to identify the correct evidence and output it in order. In the multi-round evidence ranking, the model must update the current information with each selection, iterating via several rounds to reach the endpoint.

4 Our Benchmark: Muddy Maze

To evaluate the reasoning abilities of AI models under realistic medical scenarios, we introduce **Muddy Maze**, a benchmark designed to test logical reasoning and evidence-based decision-making through a structured framework, in Figure 3. It includes: ❶ background information (e.g., patient history and presenting symptoms), ❷ an evidence pool containing both relevant clinical findings and distracting information, and ❸ a diagnostic question requiring selection of appropriate supporting evidence. By using varying difficulty levels and realistic noise, Muddy Maze mimics the uncertain, step-by-step nature of real medical decision-making. This provides a clear way to measure how well models perform in complex clinical scenarios.

4.1 Dataset Sources

Muddy Maze is built using the following question answering datasets: ❶ **MedQA**: Includes questions from USMLE Step 1, Step 2, and Step 3 exams. Step 1 questions requiring foundational medical knowledge and straightforward reasoning. Step 2 and Step 3 questions focusing on clinical reasoning and decision-making. ❷ **MedBullets Step 2/3**: Contains high-quality questions for USMLE Step 2 and Step 3 preparation. ❸ **JAMA Challenge**: Includes complex, real-world clinical cases published in the Journal of the American Medical Association (JAMA). These cases designed to simulate the most difficult and ambiguous medical reasoning scenarios. The preprocessing steps for MedBullets and the JAMA Challenge follow the methodology outlined in (Chen et al., 2024a).

4.2 Difficulty Levels Aligned with USMLE

The benchmark is divided into three difficulty levels to reflect the progressive challenges of medical training which aligned with United States Medical Licensing Examination:

- **Basic Level.** Based on MedQA Step 1 questions. Focuses on foundational medical knowledge with straightforward evidence identification.
- **Advanced Level.** Combines MedQA Step 2 & Step 3 and MedBullets questions. Requires the integration of multiple pieces of evidence and reasoning through complex information.
- **Challenge Level.** Derived from JAMA Challenge cases. Simulates challenging real-world clinical cases from a range of medical domains.

4.3 Noise Levels Aligned with the Real World

The benchmark can randomly select irrelevant evidence from the dataset and add it to the current evidence set to challenge the LLM's judgment, the basis shown in Appendix B. If the noise level is set to 0, no additional information is included, and all evidence contributes directly to the final diagnosis. Otherwise, irrelevant evidence, essentially random noise, gets added to the evidence pool. The purpose of this functionality is to evaluate whether the model can maintain its performance when faced with noisy information, simulating real-world scenarios. Similarly, doctors must identify and rely

Algorithm 1 Multi Round Muddy Maze

Input: Background Information BI, Question Q, Answer A, Evidence Sentences $E = \{e_1, e_2, \ldots, e_n\}$, Total Attempts T
Initialize $i \leftarrow \emptyset$ {No sentence selected yet}
for $t = 1$ **to** T **do**
  if $t = 1$ **then**
    Display current BI, Q, A, and E
    Prompt model to select a single sentence index $i_t \in \{1, 2, \ldots, n\}$
  else
    Update $BI \leftarrow BI \cup e_{i_{t-1}}$ {Add previous sentence to Background Information}
    Display updated BI, Q, A, and E
    Prompt model to select a single sentence index $i_t \in \{1, 2, \ldots, n\}$
  end if
  if $t = T$ **then**
    break {Final attempt reached}
  end if
end for
Return i_T {Final selected sentence index}

only on relevant information for the final diagnosis rather than treating all information as equally important.

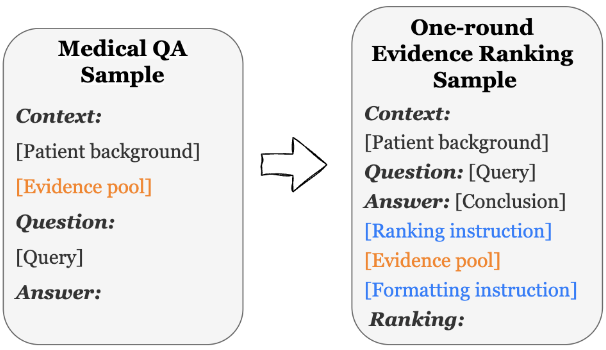

Figure 4: Format document Question-Answering sample to our One-Round evidence ranking sample

4.4 Details of One-Round and Multi-Round Tasks

We reformatted the traditional QA tasks into *evidence ranking tasks* to emphasize reasoning:

- **One-Round Evidence Ranking.** This task involves determining the correct sequence of evidence sentences that logically connects given background information to a final diagnosis or suggestion. The format shows in Figure 4.

  The model is provided with a background, a conclusion, and a pool of unordered evidence sentences. The objective is to arrange these evidence sentences in a single, coherent order that forms a logical reasoning chain, ensuring that each sentence builds upon the previous one and collectively supports the conclusion. The sequence must reflect clear dependencies—introducing foundational facts before statements that rely on them—and demonstrate holistic reasoning from start to finish. The final ordered list should provide a smooth and complete transition from the background to the diagnosis or suggestion without redundancy or logical gaps.

- **Multi-Round Evidence Ranking.** This task presents the same background, final diagnosis or suggestion, and pool of evidence as the one-round setting, but instead of ordering all evidence at once, the model selects one sentence at a time. After each selection, the chosen evidence is added to the background, updating the context for the next decision. This step-by-step approach requires the model to iteratively identify the most informative evidence at each stage, gradually narrowing down diagnostic possibilities and constructing a coherent reasoning chain leading to the final conclusion.

  This iterative process mirrors the way doctors incrementally gather and synthesize evidence in clinical reasoning, ensuring that each piece of information contributes to the final diagnosis or decision-making.

The benchmark is explicitly designed to reflect real-world medical practice, and the multi-round process can be described with the (Algorithm 1):
• **Step-by-Step Reasoning.** Multi-round evidence ranking mirrors how doctors iteratively gather and synthesize evidence during diagnosis and treatment planning.
• **Handling Distractors.** The inclusion of irrelevant evidence simulates the noisy and complex environment of real-world clinical data.
• **Complex Scenarios.** The progressive difficulty levels (basic, advanced, challenge) mirror the stages of medical training and ensure comprehensive testing.

4.5 Evaluation Metric

Multi-Hop Accuracy. A metric evaluates the model's ability to both identify the correct evidence sentences and place them in the correct sequential order to support the reasoning chain. The metric is defined as:

$$\text{Multi-Hop Accuracy} = \frac{\sum_{i=1}^{N} \mathbb{I}(e_i = \hat{e}_i \wedge p_i = \hat{p}_i)}{N}$$

where, N represents the total number of evidence sentences in the reasoning chain. Each e_i is the i-th evidence sentence in the ground-truth reasoning chain, while $\hat{e}_i$ is the corresponding evidence sentence selected by the model. Similarly, p_i denotes the position of the i-th evidence sentence in the ground truth, and $\hat{p}_i$ is the position assigned by the model. The indicator function $\mathbb{I}$ returns 1 if both $e_i = \hat{e}_i$ and $p_i = \hat{p}_i$, and 0 otherwise.

This formula calculates the proportion of evidence sentences where the model correctly identifies both the content and its position in the reasoning chain, providing a comprehensive measure of reasoning accuracy.

Single-Wise Accuracy. metric evaluates the model's ability to predict correct pairs of consecutive evidence sentences in the reasoning chain. This metric measures how well the model captures the sequential relationship between evidence elements, ensuring that not only the individual evidence sentences but also their relationship in the chain is accurate. The metric is defined as:

$$\begin{array}{c}\text{Single-Wise}\\ \text{Accuracy}\end{array} = \frac{1}{N-1} \sum_{i=1}^{N-1} \mathbb{I} \begin{pmatrix} (e_i, e_{i+1}) = (\hat{e}_i, \hat{e}_{i+1}) \vee \\ (e_i, e_{i+1}) = (\hat{e}_{i+1}, \hat{e}_i), \end{pmatrix}$$

where N represents the total number of evidence sentences in the reasoning chain. The pair (e_i, e_{i+1}) denotes the ground-truth pair of consecutive evidence sentences in the reasoning chain, while $(\hat{e}_i, \hat{e}_{i+1})$ represents the corresponding pair predicted by the model. The accuracy is calculated using an indicator function $\mathbb{I}$, which returns 1 if $(e_i, e_{i+1}) = (\hat{e}_i, \hat{e}_{i+1})$ and 0 otherwise.

This metric highlights the model's ability to maintain the correct sequential relationships between evidence elements, ensuring consistency in multi-round and one-round reasoning tasks. However, the metric is relatively loose, as it accommodates bidirectional consistency to account for reasoning paths that may flow in either direction.

5 Experiment

In this section, we want to validate the effectiveness of our proposed dialogue tuning strategy.

5.1 Experiments Setup

We designed three comparison groups to show our methods advantages: (1) For multiple-choice data, we compared the raw model with models fine-tuned on either the original multiple-choice format (Multi-Choice) and its dialogue-converted version (Dialogue(MC)); (2) For medical articles, we similarly compared the raw model with models fine-tuned on either the original articles (Article) and their dialogue-converted versions (Dialogue(Article)); and (3) For combined training, we tested the raw model against both a baseline model trained on original multiple-choice and article data together (Baseline) and a model trained on both dialogue datasets combined (Combined Dialogue).

***Q1*: What is the effect of dialogue tuning on reasoning across all difficulty levels in single round setting? *A1*: Improves reasoning across basic, advanced, and challenge levels.**

The results (Table 1, Table 2, Table 3) demonstrate that dialogue tuning significantly enhances the reasoning performance of models compared to both multi-choice, article-based tuning and baseline strategies across varying levels of task difficulty (basic, advanced, and challenge). We incorporate three noisy level, which means that we add different amounts of irrelevant evidence to the evidence pool, to distract the model's choice, which act as the noise in the real world.

Compared to multi-choice tuning, dialogue tuning shows significant improvements, with a 19.46% higher Multi-Hop Accuracy and an 18.86% increase in Single-Wise Accuracy at the basic level (see Table 1). This advantage persists in other difficulty levels, where dialogue tuning consistently outperforms multi-choice tuning. When tuning with case reports, dialogue tuning does not exhibit as large of an improvement but still achieves a 2.54% higher Multi-Hop Accuracy in the advanced setting and a 1.45% increase in the challenge setting. Additionally, our baseline strategies—which combine case reports and multiple-choice questions for tuning—perform notably worse than our combined dialogue tuning approach. For example, in basic settings, the baseline achieves only 28.82% Single-Wise Accuracy, whereas dialogue tuning reaches 43.40%. Even in the more challenging task, combined dialogue tuning maintains a clear advantage, scoring 47.58% (Single-Wise Accuracy) compared to the baseline's 39.23%(Single-Wise Accuracy), and 30.02% versus 22.04% in another metric.

Table 1: Performance of Llama-3.2-3B-Instruct and Qwen2.5-3B-Instruct under three tuning strategies across noise levels, evaluated on Basic tasks in One-Round setting.

Basic (Multi-Hop Acc)										
Model	Noise Level	Raw	Multi-Choice	Dialogue(MC)	Raw	Article	Dialogue(Article)	Raw	Baseline	Combined Dialogue
Llama-3.2-3B-Instruct	0	0.2707	0.1473	**0.3253**	0.2707	0.3221	**0.3317**	0.2707	0.2464	**0.3335**
	1	0.3526	0.1114	**0.4779**	0.3526	0.4477	**0.5224**	0.3526	0.302	**0.4882**
	3	0.2482	0.1232	**0.2842**	0.2482	0.3269	**0.3754**	0.2482	0.2545	**0.3919**
Qwen2.5-3B-Instruct	0	0.2804	0.2101	**0.3289**	0.2804	**0.2993**	0.271	0.2804	0.2183	**0.3045**
	1	0.3715	0.2198	**0.4249**	0.3715	0.3697	**0.3779**	0.3715	0.254	**0.4007**
	3	0.3337	0.2023	**0.3404**	**0.3337**	0.3262	0.3259	0.3337	0.2746	**0.343**
Average		0.3095	0.1690	**0.3636**	0.3095	0.3487	**0.3674**	0.3095	0.2583	**0.3770**
Basic (Single-Wise Acc)										
Model	Noise Level	Raw	Multi-Choice	Dialogue(MC)	Raw	Article	Dialogue(Article)	Raw	Baseline	Combined Dialogue
Llama-3.2-3B-Instruct	0	0.4104	0.1991	**0.4528**	0.4104	0.4624	**0.4635**	0.4104	0.2397	**0.4629**
	1	0.4171	0.15	**0.5104**	0.4171	0.4891	**0.561**	0.4171	0.2421	**0.5044**
	3	0.2856	0.1303	**0.3096**	0.2856	0.3389	**0.3587**	0.2856	0.1951	**0.3552**
Qwen2.5-3B-Instruct	0	0.3845	0.298	**0.4239**	0.3845	**0.3836**	0.339	**0.3845**	0.2642	0.3735
	1	0.5434	0.4427	**0.5568**	0.5434	**0.5364**	0.5137	0.5434	0.4569	**0.5347**
	3	**0.4023**	0.2731	0.3717	0.4023	**0.3797**	0.3671	**0.4023**	0.3314	0.3733
Average		0.4072	0.2489	**0.4375**	0.4072	0.4317	**0.4338**	0.4072	0.2882	**0.4340**

Table 2: Performance of Llama-3.2-3B-Instruct and Qwen2.5-3B-Instruct under three tuning strategies across noise levels, evaluated on Adavance tasks in One-Round setting.

Advance (Multi-Hop Acc)										
Model	Noise Level	Raw	Multi-Choice	Dialogue(MC)	Raw	Article	Dialogue(Article)	Raw	Baseline	Combined Dialogue
Llama-3.2-3B-Instruct	0	0.1092	0.052	**0.1578**	0.1092	0.1574	**0.1588**	0.1092	0.0867	**0.1683**
	1	0.1815	0.0522	**0.424**	0.1815	0.3965	**0.4804**	0.1815	0.1995	**0.4308**
	3	0.1508	0.0487	**0.2198**	0.1508	0.2474	**0.3014**	0.1508	0.1539	**0.2925**
Qwen2.5-3B-Instruct	0	0.1449	0.1374	**0.1667**	0.1449	0.154	**0.1603**	0.1449	0.1207	**0.1668**
	1	0.339	0.2375	**0.479**	0.339	**0.3873**	0.3765	0.339	0.2888	**0.4496**
	3	**0.2995**	0.2108	0.2965	0.2995	0.2822	**0.3**	0.2995	0.2391	**0.315**
Average		0.2042	0.1231	**0.2906**	0.2042	0.2708	**0.2962**	0.2042	0.1815	**0.3038**
Advance (Single-Wise Acc)										
Model	Noise Level	Raw	Multi-Choice	Dialogue(MC)	Raw	Article	Dialogue(Article)	Raw	Baseline	Combined Dialogue
Llama-3.2-3B-Instruct	0	0.1926	0.0823	**0.2897**	0.1926	**0.3154**	0.2939	0.1926	0.1274	**0.2937**
	1	0.3178	0.1031	**0.6159**	0.3178	0.5859	**0.6676**	0.3178	0.2319	**0.6092**
	3	0.2604	0.0852	**0.3754**	0.2604	0.4021	**0.4405**	0.2604	0.1837	**0.4232**
Qwen2.5-3B-Instruct	0	0.2557	0.218	**0.2733**	0.2557	**0.2714**	0.2506	0.2557	0.1973	**0.2713**
	1	0.5579	0.561	**0.6919**	0.5579	**0.6681**	0.6521	0.5579	0.6469	**0.6777**
	3	0.4131	0.3552	**0.4368**	0.4131	0.4309	**0.4363**	0.4131	0.4332	**0.4511**
Average		0.3329	0.2341	**0.4472**	0.3329	0.4456	**0.4568**	0.3329	0.3034	**0.4544**

Table 3: Performance of Llama-3.2-3B-Instruct and Qwen2.5-3B-Instruct under three tuning strategies across noise levels, evaluated on Challenge tasks in One-Round setting.

Challenge (Multi-Hop Acc)										
Model	Noise Level	Raw	Multi-Choice	Dialogue(MC)	Raw	Article	Dialogue(Article)	Raw	Baseline	Combined Dialogue
Llama-3.2-3B-Instruct	0	0.1019	0.0697	**0.1154**	0.1019	0.1187	**0.121**	0.1019	0.0888	**0.1254**
	1	0.2931	0.0698	**0.4288**	0.2931	0.4449	**0.5142**	0.2931	0.301	**0.4725**
	3	**0.227**	0.0767	0.2144	0.227	0.2577	**0.267**	0.227	0.1935	**0.2834**
Qwen2.5-3B-Instruct	0	0.1192	0.1031	**0.1266**	0.1192	**0.1232**	0.1217	0.1192	0.1087	**0.125**
	1	0.4037	0.347	**0.4997**	0.4037	**0.456**	0.4478	0.4037	0.3594	**0.4935**
	3	0.2802	0.2062	**0.2865**	0.2802	0.2769	**0.2927**	0.2802	0.2712	**0.3013**
Average		0.2375	0.1454	**0.2786**	0.2375	0.2796	**0.2941**	0.2375	0.2204	**0.3002**
Challenge (Single-Wise Acc)										
Model	Noise Level	Raw	Multi-Choice	Dialogue(MC)	Raw	Article	Dialogue(Article)	Raw	Baseline	Combined Dialogue
Llama-3.2-3B-Instruct	0	0.2026	0.118	**0.2316**	0.2026	0.2314	**0.2318**	0.2026	0.1588	**0.2392**
	1	0.5586	0.2013	**0.6779**	0.5586	0.6848	**0.7384**	0.5586	0.4599	**0.6967**
	3	0.4357	0.1728	**0.4549**	0.4357	0.4854	**0.4898**	0.4357	0.3336	**0.4945**
Qwen2.5-3B-Instruct	0	0.2115	0.1937	**0.2217**	0.2115	**0.2229**	0.2218	0.2115	0.1965	**0.2154**
	1	0.6758	0.6564	**0.7356**	0.6758	**0.7284**	0.7076	0.6758	0.7156	**0.725**
	3	0.4629	0.4198	**0.4824**	0.4629	0.4804	**0.485**	0.4629	**0.4895**	0.4839
Average		0.4245	0.2937	**0.4674**	0.4245	0.4722	**0.4791**	0.4245	0.3923	**0.4758**

***Q2*: Does dialogue tuning still show the advantage in the multi-round setting? *A2*: Yes, it still outperforms traditional tuning methods.**

In Figure 6, we clearly demonstrate the perfor-

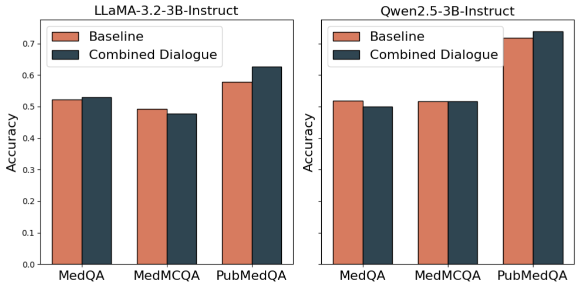

Figure 5: Comparison of scores between baseline and combined dialogue approaches for LLaMA 3.2-3B Instruct and Qwen2.5-3B-Instruct across MedQA, MedMCQA, and PubMedQA datasets. The combined dialogue approach consistently improves performance.

mance of fine-tuned model in the multi-round setting across confusion levels 0 to 5. For the LLaMA 3.2-3B Instruct, our dialogue tuning exhibits clear advantages across all three experimental settings, and further enhances performance in multi-round tasks—achieving a 4.06% improvement even in the noisiest environment (level 5). As for the Qwen 2.5-3B Instruct model, while it does not show as significant a gap compared to LLaMA, it still maintains an advantage, particularly in high-noise environments. Specifically, it achieves a 3.16% improvement over multi-choice tuning strategies at confusion level 5.

Beyond synthetic multi-turn tasks, our dialogue-tuned models also demonstrate comparable or better performance on widely-used medical QA datasets including MedQA, MedMCQA, and PubMedQA (Figure 5), further confirming the practical generalize ability of our approach.

5.2 Human Evaluation

To assess the fidelity of our LLM-generated dialogues to the original context, we conducted a human evaluation study with 32 participants in U.S., including medical students. Participants rated each dialogue on a 5-point scale ranging from Fully Covered (4: all essential information preserved) to Not Covered (0: core information missing or distorted).

The results demonstrated strong performance, shown in Figure 7: 79% of the MC dialogues and 74% of the Report dialogues were rated as Fully Covered (score 4), indicating that the generated dialogues preserved all key information from the original context. Notably, only 3–4% of outputs fell into the Minimally/Not Covered categories (scores 0–1), suggesting rare failures in coherence. These findings confirm that our LLM-generated dialogues are highly faithful to the source material, achieving the primary goal of dependency on and comprehensive coverage of the raw context. The human evaluation thus validates the reliability of our approach for producing trustworthy dialogue outputs.

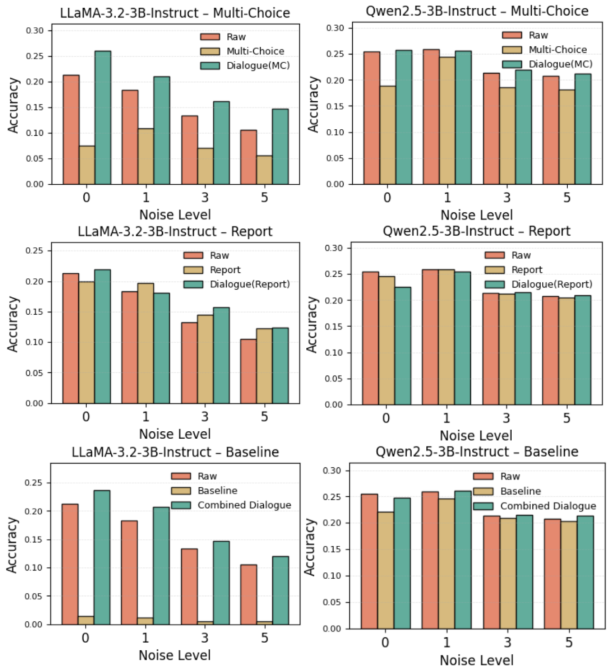

Figure 6: Performance of Llama-3.2-3B-Instruct and Qwen2.5-3B-Instruct under three tuning strategies across four noise levels, average all difficulty levels, evaluated on Multi-Round setting.

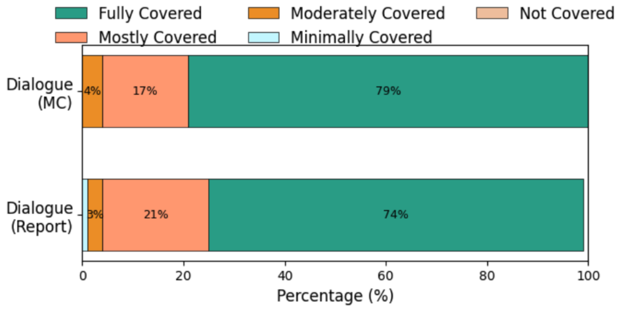

Figure 7: Human Evaluation Performance.

6 Conclusion

In this work, we introduce a novel benchmark, **Muddy Maze**, designed to evaluate the logical reasoning and evidence-based decision-making capabilities of language models in realistic medical scenarios. Moreover, we demonstrate that **dialogue tuning** significantly improves model performance compared to multi-choice and article-based tuning, through extensive experiments.

This work highlights the importance of a dynamical approach to advancing reasoning in medical AI systems. Dialogue tuning aligns training with the step-by-step cognitive processes required for diagnostic decision-making, providing a framework for developing more reliable models.

Limitations

The dialogue generation process, which relies solely on Gemini-2.5 Flash, may introduce certain biases. Even though we randomly sample some of the generated dialogues for human evaluation, relying on a single large language model for dialogue generation could lead to model-specific biases—particularly in how it structures conversations and prioritizes certain types of medical information.

References

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, and 1 others. 2023. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2024a. Benchmarking large language models on answering and explaining challenging medical questions. *arXiv preprint arXiv:2402.18060*.

Po-Chaun Chen, Mahdin Rohmatillah, You-Teng Lin, and Jen-Tzung Chien. 2024b. Convcounsel: A conversational dataset for student counseling. In *2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, and 1 others. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Aryo Pradipta Gema, Pasquale Minervini, Luke Daines, Tom Hope, and Beatrice Alex. 2024. Parameter-efficient fine-tuning of llama for the clinical domain. *Preprint*, arXiv:2307.03042.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2023. Medalpaca – an open-source collection of medical conversational ai models and training data. *Preprint*, arXiv:2304.08247.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

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*.

Zeqian Ju, Subrato Chakravorty, Xuehai He, Shu Chen, Xingyi Yang, and Pengtao Xie. 2020. Coviddialog: Medical dialogue datasets about covid-19.

Taeyoon Kwon, Kai Tzu-iunn Ong, Dongjin Kang, Seungjun Moon, Jeong Ryong Lee, Dosik Hwang, Beomseok Sohn, Yongsik Sim, Dongha Lee, and Jinyoung Yeo. 2024. Large language models are clinical reasoners: Reasoning-aware diagnosis framework with prompt-generated rationales. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18417–18425.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Wenqiang Li, Lina Yu, Min Wu, Jingyi Liu, Meilan Hao, and Yanjie Li. 2023a. Doctorgpt: A large language model with chinese medical question-answering capabilities. In *2023 International Conference on High Performance Big Data and Intelligent Systems (HDIS)*, pages 186–193. IEEE.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023b. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, and 1 others. 2023. A study of generative large language model for medical research and healthcare. *NPJ digital medicine*, 6(1):210.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine. *Preprint*, arXiv:2304.14454.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, Xinyu Zhou, Lingfei Qian, Huan He, Dennis Shung, Lucila Ohno-Machado, Yonghui Wu, Hua Xu, and Jiang Bian. 2024. Me llama: Foundation large language models for medical applications. *Preprint*, arXiv:2402.12749.

Guojun Yan, Jiahuan Pei, Pengjie Ren, Zhaochun Ren, Xin Xin, Huasheng Liang, Maarten de Rijke, and Zhumin Chen. 2022. Remedi: Resources for multi-domain, multi-service, medical dialogues. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3013–3024.

Zonghai Yao, Zihao Zhang, Chaolong Tang, Xingyu Bian, Youxia Zhao, Zhichao Yang, Junda Wang, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, and 1 others. 2024. Medqa-cs: Benchmarking large language models clinical skills using an ai-sce framework. *arXiv preprint arXiv:2410.01553*.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, and 1 others. 2020. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9241–9250.

Qian Zhang, Panfeng Chen, Jiali Li, Linkun Feng, Shuyu Liu, Mei Chen, Hui Li, and Yanhao Wang. 2024a. Pediabench: A comprehensive chinese pediatric dataset for benchmarking large language models. *arXiv preprint arXiv:2412.06287*.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. 2024b. Alpacare:instruction-tuned large language models for medical application. *Preprint*, arXiv:2310.14558.

A Dialogue Generation

A.1 Why We Need Dialogue Tuning?

Dialogue tuning is proposed as a more effective approach for capturing logical relationships, as the dialogue format inherently mirrors the reasoning process found in human doctor-patient interactions.

Interactive Nature of Dialogue. The dialogue format enables iterative, question-and-answer reasoning that mimics real diagnostic processes. Clinicians progressively gather information through targeted questioning, with each response reducing diagnostic uncertainty. This stepwise approach organizes clinical data logically, making the reasoning chain more transparent and tractable for AI systems.

Empirical Support for Dialogue Tuning. Our experiments in section 5 demonstrate that dialogue tuning significantly outperforms multi-choice and article-based tuning in tasks designed with our benchmark. Key findings include: ❶ Dialogue-tuned models exhibit higher accuracy in reasoning tasks, including those requiring multi-step integration of evidence. ❷ Dialogue-tuned models remain robust even when tackling challenging tasks or integrating noisy inputs, maintaining performance where other approaches falter. ❸ They also have competitive capabilities in general medical question-answering tasks, as the model is fine-tuned with multiple-choice QA and medical articles.

A.2 Prompt Templates

Multi-choice QA to Dialogue: Please convert the following paragraph into a doctor-patient dialogue. Ensure that all the information provided, including personal details, symptoms, examination findings, diagnosis, and treatment, is included. Most important is the final answer, "{{Answer}}", which must be included in the dialogue without any changes. Use natural conversational language to connect the details, but do not introduce any new information. The dialogue should not be too redundant:
{{Document}}

Article to Dialogue: Please convert the following paragraph into a doctor-patient dialogue. Ensure that all the information provided, like personal details, symptoms, examination findings, diagnosis, and treatment, is included. Use natural conversational language to connect the details, but do not introduce any new information. The dialogue should not be too redundant:
{{Document}}

A.3 Task Reformulation Examples

This example demonstrates how we use the above prompt to guide Gemini 2.5-flash in transitioning from multiple-choice question answering to a doctor-patient dialogue format.

```
A 23-year-old pregnant woman at 22 weeks
gestation presents with burning upon
urination. She states it started 1 day
ago and has been worsening despite drinking
more water and taking cranberry extract.
She otherwise feels well and is followed by
a doctor for her pregnancy. Her temperature
is 97.7°F (36.5°C), blood pressure is
122/77 mmHg, pulse is 80/min, respirations
are 19/min, and oxygen saturation is 98% on
room air. Physical exam is notable for an
absence of costovertebral angle tenderness
and a gravid uterus. Which of the following
is the best treatment for this patient?
```

```
Doctor: Hi there, I understand you're
having some trouble. What brings you in
today?
Patient: Hi Doctor. I'm 22 weeks pregnant,
and for the past day, I've had burning when
I pee. It's really uncomfortable.
Doctor: I see. And how long has this been
going on?
Patient: Just since yesterday, and it's
actually getting worse, even though I've
been drinking a lot of water and taking
cranberry extract.
Doctor: I understand. Other than the
burning, how are you feeling?
Patient: I feel fine otherwise. My
pregnancy is being followed regularly by
my OB.
Doctor: Okay, let's take a look. I'm going
to check your vitals and do a quick exam.
Your temperature is 97.7, blood pressure
is 122 over 77, pulse is 80, respirations
are 19, and your oxygen is 98%. Everything
looks good there. I'm also checking for any
tenderness in your back near your kidneys...
that's good, no tenderness. And I can feel
your uterus is growing as expected. Given
your symptoms and the examination, the best
treatment for this is Nitrofurantoin.
```

A.4 Tuning Details

In our tuning strategies, we consistently use LoRA tuning with the following parameters: rank = 8, lora_alpha = 32, and lora_dropout = 0.1. In all three comparison studies, both the control group and the experimental group share the same hyperparameters. In the *Multi-Choice Comparison Study*, we set Qwen2.5-3B to 2 epochs, a learning rate of 5×10^{-4}, and a max length of 1024; LLaMA-3.2-3B to 2 epochs, a learning rate of 5×10^{-4}, and a max length of 1024. In the *Article Comparison Study*, we configure Qwen2.5-3B with 2 epochs, a learning rate of 1×10^{-5}, and a max length of 1024; LLaMA-3.2-3B with 2 epochs, a learning rate of 5×10^{-5}, and a max length of 2048. Finally, in the *Combined Dialogue and Baseline Study*, we assign Qwen2.5-3B to 2 epochs, a learning rate of 2×10^{-5}, and a max length of 2048; LLaMA-3.2-3B to 2 epochs, a learning rate of 5×10^{-5}, and a max length of 2048.

Testbed. We fine-tuned the Llama 3.2-3B Instruct and Qwen 2.5-3B Instruct models using 2 NVIDIA RTX 6000 GPUs, each with 48GB of memory. We running our benchmark also on the NVIDIA RTX 6000 48GB GPUs.

B Benchmark: Muddy Maze

Dataset. Our benchmark integrates the MedQA-USMLE Test Set, Medbullets, and JAMA Clinical Challenge. The fine-tuning is based on the MedQA-USMLE Train Set, which have around 10.2k question-answer pairs and approximately 12k PubMed articles. All dataset information shows in Figure 2.

B.1 Clinical and Examination Basis for MuddyMaze Benchmark

The design of MuddyMaze is rigorously aligned with established medical licensing exams and real-world diagnostic workflows. Below, we outline its foundations in several key areas:

- **USMLE Step 2 Clinical Skills (CS).** It required examinees would take a history, perform a physical exam, formulate differential diagnoses, and write a patient note.
- **USMLE Step 3 CCS.** It assesses clinical decision-making through Computer-based Case Simulations (CCS). These simulations require doctors to diagnose and manage patients by sequentially ordering tests, interpreting results, and initiating treatments—all while filtering out irrelevant information (like incidental findings or patient anecdotes) that could distract from critical decisions.
- **Medical Jeopardy competitions.** An answer-first format, where contestants hear a clinical "clue" (e.g., "This tumor causes episodic hypertension and headaches") and must respond with the correct question (e.g., "What is pheochromocytoma?"). It required clinicians compete to solve clinical puzzles by connecting fragmented clues—such as symptoms, labs, or imaging findings—into accurate diagnoses. Contestants must rapidly prioritize key evidence while ignoring distractors, mirroring real-world diagnostic reasoning.

The design of MuddyMaze integrates core principles from these real-world clinical assessments:

- USMLE Step 2 CS's iterative data gathering (history → exam → tests) inspired our multi-round evidence ranking, where models must simulate a clinician's stepwise reasoning.
- USMLE Step 3 CCS's emphasis on prioritizing actions amid distractions (e.g., ignoring incidental findings) directly aligns with MuddyMaze's noise injection and dynamic evidence selection.
- Medical Jeopardy's answer-first format—requiring contestants to reverse-engineer diagnoses from clues—parallels our one-round evidence chaining, where models reconstruct logical sequences (e.g., lab → imaging → diagnosis) from fragmented inputs.

Together, these connections validate MuddyMaze's clinical fidelity, ensuring it tests not just medical knowledge, but the decision-making workflows and noise resilience essential in practice.

B.2 Prompt Template

```
Here is the background information:
"{{prerequisit}}"
Question: {{question}}
Answer: {{answer}}
Below are several evidence sentences.
Identify the {{groundtruth zoo length}}
sentences that, if added to the background
information, would support inferring the
answer based on the given question-answer
pair. Please choose the sentence in logical
order!
{{tagged maze}}
Provide only the indices of the relevant
sentences in brackets formatted like this:
[ ], no more than {{groundtruth zoo length}}
sentences.
ANSWER:
```

```
Here is the background information:
"{{prerequisit}}"
Question: {{question}}
Answer: {{answer}}
Below are several evidence sentences. Based
on the given question-answer pair, please
select which sentence should be added to the
background information to support inference
of the answer.
{{tagged maze}}
You have {{groundtruth zoo length}}
attempts in total to make a selection; this
is your {{i_th}} attempt. Please choose the
sentence in logical order!
Provide only the indices of the relevant
sentences in brackets formatted like this:
[ ]
ANSWER:
```

B.3 Task Reformulation Examples

This is an example of transitioning from the traditional question-answering task to our benchmark. The results are generated using LLaMA 3.1-8B at the basic task level, with a noise level of 0, in a one-round setting.

```
A 67-year-old man with transitional cell
carcinoma of the bladder comes to the
physician because of a 2-day history
of ringing sensation in his ear. He
received this first course of neoadjuvant
chemotherapy 1 week ago. Pure tone
audiometry shows a sensorineural hearing
loss of 45 dB.
Question: The expected beneficial effect
of the drug that caused this patient's
symptoms is most likely due to which of
the following actions?
Answer: Cross-linking of DNA
```

```
Here is the background information: Ä
67-year-old man with transitional cell
carcinoma of the bladder comes to the
physician because of a 2-day history of
ringing sensation in his ear.¨
Question: The expected beneficial effect
of the drug that caused this patient's
symptoms is most likely due to which of
the following actions?
Answer: Cross-linking of DNA
Below are several evidence sentences.
Identify the 2 sentences that, if added to
the background information, would support
inferring the answer based on the given
question-answer pair. Please choose the
sentence in logical order!
0: Pure tone audiometry shows a
sensorineural hearing loss of 45 dB.
1: He received this first course of
neoadjuvant chemotherapy 1 week ago.
Provide only the indices of the relevant
sentences in brackets formatted like this:
[ ], no more than 2 sentences.
ANSWER: [1], [0]
```