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

Zijie Liu[1], **Xinyu Zhao**[1], **Jie Peng**[2], **Jinhao Duan**[1], **Zhuangdi Zhu**[3]
Qingyu Chen[4], **Kaidi Xu**[5], **Xia Hu**[6], **Tianlong Chen**[1]

[1]UNC-CH [2]USTC [3]George Mason University
[4]Yale University [5]City University of Hong Kong [6]Rice University

Correspondence: tianlong@cs.unc.edu

Abstract

In real clinical practice, clinicians must sift through noisy and often conflicting information, progressively gathering and sequencing evidence before reaching conclusions. However, existing tuning methods for medical AI models are typically monologue-based — that is, models are fine-tuned on static question-answering (QA) tasks or medical articles, which fail to reflect the interactive and iterative nature of clinical reasoning. To bridge this gap, we introduce MuddyMaze, a benchmark designed to expose the limitations of current monologue-based tuning, and construct a large dialogue dataset of 22.2k doctor–patient interactions that capture stepwise diagnostic reasoning validated by medical experts. Building on those, we propose dialogue-tuning, a new fine-tuning paradigm that captures the internal reasoning dynamics unfolding across interactions.

To assess the effectiveness of our approach, we evaluated *dialogue-tuned* models on MuddyMaze, where they outperform *monologue-tuned* baselines (e.g., MedQA) by +16.1% in one-round and +4.1% in multi-round evidence ranking, while maintaining or even improving accuracy on standard medical QA benchmarks (e.g., PubMedQA). These results indicate that dialogue-tuning not only enhances reasoning robustness and evidence integration but also preserves the factual precision of traditional QA performance.

1 Introduction

Large language models (LLMs) have achieved remarkable progress in the medical domain, contributing to applications such as disease analysis, diagnostic support, and clinical decision assistance (Singhal et al., 2023; Li et al., 2023b,a; Chen et al., 2023; Peng et al., 2023; Kwon et al., 2024). Despite these advances, the evaluation of medical LLMs remains an open challenge. Most existing benchmarks are derived from medical licensing exams or research articles and are framed as multiple-choice QA or long-form reasoning tasks (Jin et al., 2021; Pal et al., 2022). These benchmarks have provided valuable insights into model knowledge and improved QA performance, but they remain highly structured and artificially clean, diverging from the complexities of real-world diagnostic practice (Chen et al., 2024a; Yao et al., 2024).

Limitations of current benchmarks. Current medical QA datasets neglect two essential properties of clinical reasoning. (i) *Stepwise and iterative reasoning.* In real practice, clinicians rarely have all information available upfront. Instead, they iteratively refine hypotheses by actively gathering information, asking targeted questions, and updating decisions as new evidence arrives. By contrast, medical QA benchmarks present complete information in a single turn, testing only the final interpretation of facts while ignoring the process of strategically acquiring and organizing evidence under uncertainty. (ii) *Noise and uncertainty.* Clinical data are often incomplete, ambiguous, or even conflicting—for instance, vague symptoms or false-positive test results. Physicians must weigh and prioritize evidence while resisting misleading signals. Current benchmarks, however, use well-structured case descriptions and offer no mechanism for assessing how models handle distractors, uncertainty, or noisy clinical contexts.

Our Benchmark: MuddyMaze. We address these gaps with MuddyMaze, an answer-conditioned benchmark that reframes medical QA as evidence ranking under noise. Each instance comprises background context, a diagnostic QA pair, and an evidence pool containing both relevant and distracting findings. Moreover, MuddyMaze evaluates reasoning through one-round and multi-round evidence ranking, simulating stepwise information gathering under varying noise and difficulty levels (see Section 4 for details). This design enables interpretable

Findings of the Association for Computational Linguistics: EACL 2026, pages 2858–2872
March 24-29, 2026 ©2026 Association for Computational Linguistics

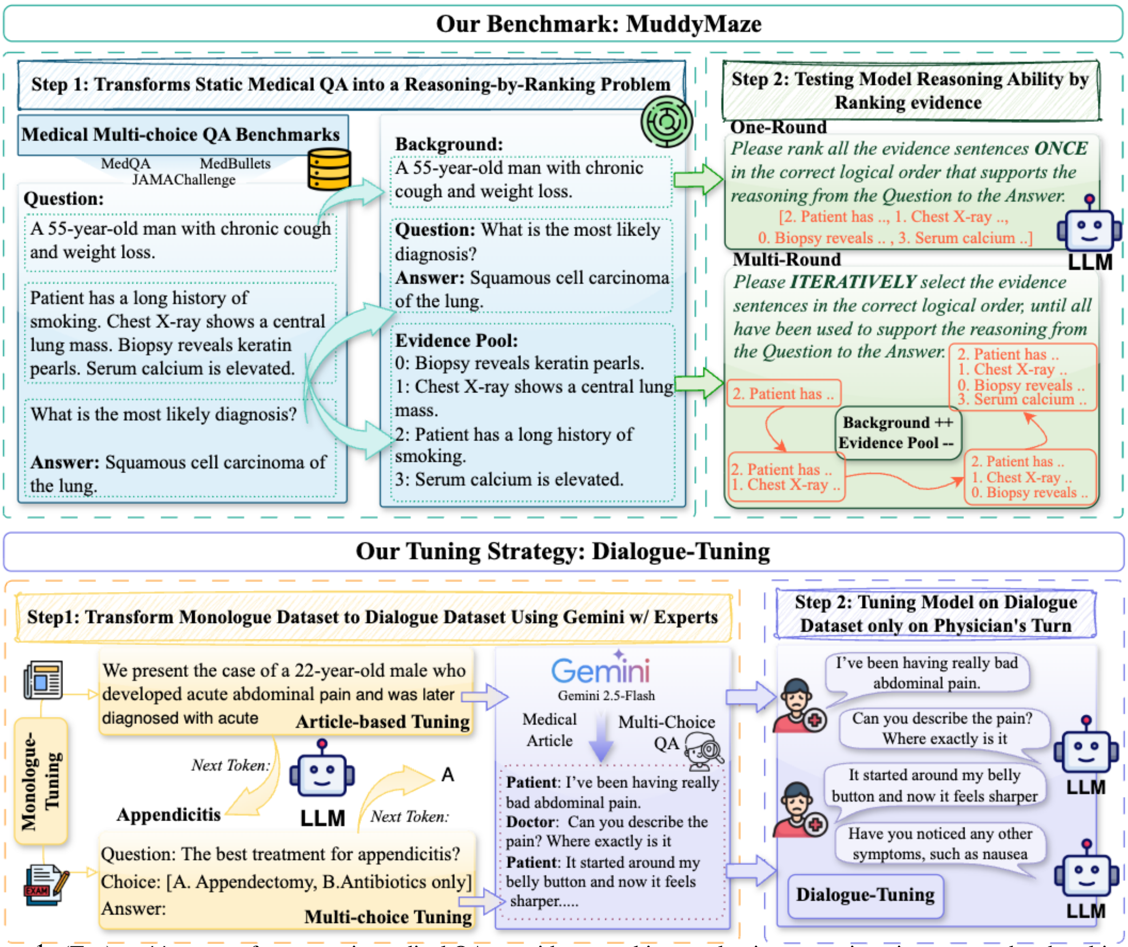
Our Benchmark: MuddyMaze
Step 1: Transforms Static Medical QA into a Reasoning-by-Ranking Problem
Medical Multi-choice QA Benchmarks
MedQA
MedBullets
JAMAChallenge
Question:
A 55-year-old man with chronic cough and weight loss.
Patient has a long history of smoking. Chest X-ray shows a central lung mass. Biopsy reveals keratin pearls. Serum calcium is elevated.
What is the most likely diagnosis?
Answer: Squamous cell carcinoma of the lung.
Background:
A 55-year-old man with chronic cough and weight loss.
Question: What is the most likely diagnosis?
Answer: Squamous cell carcinoma of the lung.
Evidence Pool:
0: Biopsy reveals keratin pearls.
1: Chest X-ray shows a central lung mass.
2: Patient has a long history of smoking.
3: Serum calcium is elevated.
Step 2: Testing Model Reasoning Ability by Ranking evidence
One-Round
Please rank all the evidence sentences ONCE in the correct logical order that supports the reasoning from the Question to the Answer.
[2. Patient has .., 1. Chest X-ray ..,
0. Biopsy reveals .. , 3. Serum calcium ..]
LLM
Multi-Round
Please ITERATIVELY select the evidence sentences in the correct logical order, until all have been used to support the reasoning from the Question to the Answer.
2. Patient has ..
1. Chest X-ray ..
0. Biopsy reveals ..
3. Serum calcium ..
2. Patient has ..
Background ++
Evidence Pool --
2. Patient has ..
1. Chest X-ray ..
2. Patient has ..
1. Chest X-ray ..
0. Biopsy reveals ..
Our Tuning Strategy: Dialogue-Tuning
Step1: Transform Monologue Dataset to Dialogue Dataset Using Gemini w/ Experts
Monologue-Tuning
We present the case of a 22-year-old male who developed acute abdominal pain and was later diagnosed with acute
Article-based Tuning
Next Token:
Appendicitis
LLM
A
Next Token:
Question: The best treatment for appendicitis?
Choice: [A. Appendectomy, B.Antibiotics only]
Answer:
Multi-choice Tuning
Gemini
Gemini 2.5-Flash
Medical Article
Multi-Choice QA
Patient: I've been having really bad abdominal pain.
Doctor: Can you describe the pain? Where exactly is it
Patient: It started around my belly button and now it feels sharper.....
Step 2: Tuning Model on Dialogue Dataset only on Physician's Turn
I've been having really bad abdominal pain.
Can you describe the pain? Where exactly is it
LLM
It started around my belly button and now it feels sharper
Have you noticed any other symptoms, such as nausea
Dialogue-Tuning
LLM

Figure 1: (Top) MuddyMaze reframes static medical QA as evidence ranking, evaluating reasoning via one-round and multi-round modes. (Bottom) Dialogue-Tuning converts QA and article data into doctor–patient dialogues and fine-tunes models only on physician turns, encouraging stepwise clinical reasoning.

measurement of how models reason under uncertainty rather than in perfectly structured settings.

Dialogue-based training. While MuddyMaze evaluates reasoning, training remains limited if models continue to learn only from static QA objectives that provide final-answer supervision. To approximate clinicians' stepwise reasoning, we introduce a dialogue-based fine-tuning paradigm that reformulates medical QA and article cases into doctor–patient interactions. During training, the model generates doctor responses conditioned on evolving conversational context, thereby learning to reason iteratively rather than memorizing direct QA mappings. Dialogue-based training encourages models to organize and prioritize evidence more effectively in noisy scenarios, leading to consistent improvements on MuddyMaze and stable performance on QA benchmarks.

- We introduce MuddyMaze, a benchmark that transforms static QA into stepwise evidence ranking under noise, with both one-round and multi-round evaluation and tiered difficulty levels.
- We curate a large-scale dialogue dataset simulating doctor–patient interactions, enabling models to learn iterative reasoning strategies.
- We propose a dialogue-based fine-tuning framework that enhances evidence-ranking performance as a proxy for reasoning quality, while maintaining strong results on conventional QA benchmarks.

2 Related Work

Medical Large Language Models. LLMs have shown rapid progress in the medical domain (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 results on medical QA and summarization tasks. These models are typically fine-tuned on medical exams, scientific literature, clinical guidelines, and EHR notes to adapt to real-world clinical text understanding. In parallel, bio-focused language models have been developed for biomedical research and healthcare applications (Luo et al., 2022; Bannur et al., 2023). For example, BioBERT (Lee et al.,

2020) and PubMedBERT (Gu et al., 2021), both pre-trained on PubMed, serve as foundational models for tasks such as named entity recognition and relation extraction.

Conversation Datasets in the Medical Field. Early medical benchmarks mainly assessed factual knowledge through static QA formats such as multiple-choice or extractive tasks, exemplified by MedQA (Yao et al., 2024), MedMCQA (Pal et al., 2022), and PubMedQA (Jin et al., 2019). Recent work has shifted toward dialogue-based settings that reflect real doctor–patient communication, including MedDialog (Zeng et al., 2020b) and ReMeDi (Yan et al., 2022). Several domain-specific datasets further extend this paradigm, covering mental health counseling (Chen et al., 2024b), pediatrics (Zhang et al., 2024a), and COVID-19 consultations (Ju et al., 2020). More recently, studies such as MediQ (Li et al., 2024) and UoT (Hu et al., 2024) explored interactive information-seeking dialogues, where models actively query for missing clinical details—highlighting a growing focus on reasoning-oriented medical conversations.

3 Methodology

This section presents our methodology for enhancing clinical reasoning in medical LLMs through `dialogue-tuning`. We reformulate standard medical QA and article-based data into synthetic doctor–patient dialogues, where the model learns to generate physician responses conditioned on evolving conversational context. By fine-tuning on the physician's turns, `dialogue-tuning` encourages stepwise reasoning behaviors that better reflect real clinical decision-making processes.

3.1 From Monologue to Dialogue Tuning

Monologue Tuning. Most medical training datasets are `monologue-style`, such as multi-choice QA or article-based tasks, where all relevant information is given upfront and the model is trained to output a final answer. While effective for knowledge recall, this format oversimplifies clinical reasoning, which in reality is iterative, uncertain, and evidence-driven.

Dialogue Tuning. Dialogue-style supervision better mirrors real diagnostic workflows. In doctor–patient interactions, information is revealed progressively rather than all at once. At each step, the clinician integrates known facts, asks targeted questions, and updates diagnostic hypotheses. To approximate this process, we reformulate training data into multi-turn dialogues where the model generates physician responses conditioned on prior context. Each doctor turn requires the model to (i) identify relevant findings from the patient's utterances and (ii) determine the next appropriate inquiry or interpretation. This setup provides structured supervision for iterative evidence integration and hypothesis refinement—key aspects of clinical reasoning absent in standard `monologue-style` training. A detailed comparison between `monologue-` and `dialogue-tuning` is provided in Figure 1.

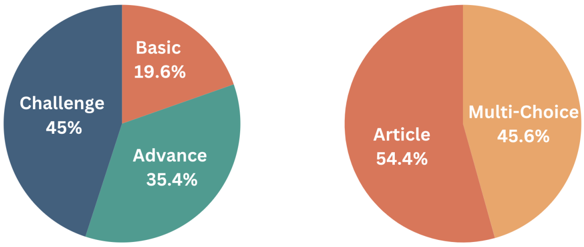

Figure 2: The left pie chart represents the ratio of difficulty levels in our benchmark. While the right pie chart represents the proportion of multiple-choice question-answering sets and articles used during the tuning stage, the dialogues generated from these sources are equal in quantity to them.

3.2 Dialogue Tuning Objective

To better align training with clinical reasoning, we convert existing datasets into synthetic doctor–patient dialogues (see Figure 2). Two sources are used: 10.2k QA pairs from the MedQA (Jin et al., 2021) training set and 12k articles from PubMed. ❶ *Multi-Choice to Dialogue*: Each QA item is reformulated into a dialogue where the patient introduces demographics, symptoms, or test results, and the physician probes further and provides interpretations. ❷ *Article to Dialogue*: Article Cases are similarly decomposed into exchanges: patients describe their history and findings in accessible language, while doctors interpret, ask follow-up questions, and reason about possible diagnoses. In both cases, technical details are preserved but revealed turn by turn, allowing the dialogue to unfold progressively and mirror real diagnostic encounters where physicians refine hypotheses step by step.

Implementation. We employ Gemini-2.5 Flash (Li et al., 2024) to perform the conversion, using carefully designed prompts to ensure faithfulness to the source material (Appendix A). Dialogue length and number of turns vary with sample complexity. To assess quality, we conduct human evaluation on a stratified sample, confirming that essential information is preserved (subsection 5.2).

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 Setup. In dialogue-tuning, the model plays the role of the physician. Given a dialogue prefix

$$H_t = \{u_1, d_1, \ldots, u_{t-1}, u_t\},$$

where u_i and d_i denote patient and physician's turns, the model predicts the next physician utterance d_t. The objective is standard token-level cross-entropy:

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

where $\mathcal{T}_D$ indexes physician turns. The loss is computed only on physician responses, conditioning on the full conversational context. Although the optimization objective remains unchanged, dialogue-tuning differs conceptually from monologue-tuning by training models to generate reasoning steps rather than final answers.

4 MuddyMaze: Benchmark for Evidence Ranking

To evaluate reasoning under realistic clinical uncertainty, we introduce MuddyMaze, a benchmark that reformulates QA into structured evidence-ranking tasks (Figure 1). Each instance comprises: (i) *background information* (e.g., patient history and symptoms), (ii) a *diagnostic QA pair* specifying the target problem, and (iii) an *evidence pool* containing gold supporting sentences mixed with clinically plausible distractors. The model must output an *ordered evidence sequence* that connects the background to the answer while filtering out irrelevant information. This formulation transforms static QA into a reasoning-by-ranking problem, directly testing whether models can identify, prioritize, and organize evidence in noisy clinical contexts.

4.1 Dataset Sources

MuddyMaze is constructed by reformulating multiple publicly available medical QA benchmarks into an evidence-ranking format. For each instance, we decompose the original QA item into background, QA-pair and evidence sentences. Moreover, distractors are introduced by sampling clinically plausible but irrelevant content to ensure robustness against noise. Specifically: ❶ **MedQA Test Set (USMLE Step 1–3)** (**Jin et al., 2021**): Provides foundational and advanced clinical questions. ❷ **MedBullets Step 2/3** (**Chen et al., 2025**): Offers high-quality preparation questions that require integrating multiple evidence pieces. These items enrich the benchmark's advanced-level reasoning scenarios. ❸ **JAMA Challenge** (**Chen et al., 2025**): Contains complex, real-world case vignettes from the *Journal of the American Medical Association*, representing the most challenging and ambiguous diagnostic reasoning tasks.

4.2 Difficulty Levels Aligned with USMLE

To mirror the progressive complexity of medical training, MuddyMaze is organized into three tiers aligned with USMLE stages:

- **Basic Level.** Derived from MedQA Step 1 items focusing on single-fact biomedical reasoning.
- **Advanced Level.** Based on Step 2/3 and MedBullets questions requiring multi-evidence synthesis across history, labs, and guidelines.
- **Challenge Level.** Formed from JAMA Challenge cases with ambiguous findings and subtle distractors, requiring prioritization under uncertainty.

4.3 Noise Levels Reflecting Real-World Uncertainty

In clinical practice, not all available information is reliable or relevant. To capture this, MuddyMaze introduces noise levels by injecting distractor sentences into the evidence pool.

- **Level 0 (Clean).** No distractors are added; all evidence is directly relevant to the diagnostic task.
- **Level $n > 0$ (Noisy).** A specified number of distractors are sampled from unrelated clinical contexts and inserted into the pool alongside the gold supporting evidence. Distractors may appear clinically plausible but do not contribute to the correct reasoning chain.

This design assesses a model's ability to sustain reasoning performance under uncertainty. By introducing controlled noise, the benchmark evaluates whether models can accurately identify and prioritize informative evidence while disregarding misleading or irrelevant inputs..

4.4 Task Settings

Two complementary settings probe distinct reasoning behaviors:

- **One-Round Evidence Ranking.** The model receives background, QA pair, and the full evidence

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 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.

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 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.

pool, and must output a complete ordered evidence chain in one step. This setting tests global reasoning organization without intermediate feedback.

- **Multi-Round Evidence Ranking.** At each round t, the model selects one evidence sentence, which is appended to the background before the next iteration. The process continues for T rounds, yielding a progressive evidence sequence $\{i_1, i_2, \ldots, i_T\}$. This iterative formulation evaluates stepwise evidence acquisition and integration.

Together, the two settings assess both holistic reasoning and incremental evidence gathering.

4.5 Evaluation Metrics

MuddyMaze evaluates model reasoning from both global and local perspectives through two complementary metrics. Each instance consists of background information, a diagnostic QA pair, and an evidence pool from which the model must output an ordered chain of supporting sentences.

- **Multi-Hop Accuracy.** This metric measures whether the model exactly reconstructs the annotated reasoning chain, considering both content and position. For a chain of length N, let e_i denote the i-th gold evidence sentence with position p_i. A prediction at position i is correct only if the selected sentence $\hat{e}_i = e_i$ and its assigned position $\hat{p}_i = p_i$. The score 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 $\mathbb{I}$ is the indicator function. This metric provides a strict assessment of whether the model fully recovers the global reasoning chain.

- **Single-Wise Accuracy.** This metric evaluates whether the model preserves the adjacency relations between consecutive evidence sentences. A ground-truth pair (e_i, e_{i+1}) is considered correct if the predicted sequence contains $(\hat{e}_i, \hat{e}_{i+1}) = (e_i, e_{i+1})$ or its reversed order $(\hat{e}_i, \hat{e}_{i+1}) = (e_{i+1}, e_i)$. Formally:

Algorithm 1 Multi Round `MuddyMaze`

```
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}
```

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

This metric reflects the model's ability to maintain coherent transitions between reasoning steps even when the full order is imperfect, thus highlighting local consistency under uncertainty.

Together, Multi-Hop and Single-Wise Accuracy form a comprehensive evaluation of reasoning quality. The former emphasizes exact recovery of the global reasoning structure, while the latter captures the preservation of clinically meaningful intermediate relations, offering a balanced view of both holistic and incremental reasoning performance.

5 Experiment

This section evaluates the effectiveness of our proposed `dialogue-tuning` strategy through three controlled comparison settings.

5.1 Experiments Setup

We conduct experiments to evaluate how dialogue-based training affects reasoning performance across different types of medical data. (i) **Multiple-Choice Setting.** Models are fine-tuned on standard multiple-choice QA data (*Multi-Choice*) and its dialogue-converted variant (*Dialogue(MC)*), compared against the zero-shot *Raw* model. This setup tests whether dialogue-style training helps models move beyond pattern recognition toward evidence-based reasoning within structured QA formats. (ii) **Article-Based Setting.** Using medical article data, we compare models fine-tuned directly on raw article text (*Article*) versus their dialogue-transformed versions (*Dialogue(Article)*). This evaluates whether conversational reformulation enhances reasoning when training on narrative, unstructured medical text. (iii) **Combined Setting.** We further integrate both modalities to test generalization. A *Baseline* model is trained on the mix of Multi-Choice QA and Article data, while the *Combined Dialogue* model uses their dialogue-converted counterparts. This setting examines whether `dialogue-tuning` benefits scale consistently across heterogeneous data sources.

All models are evaluated on the proposed `MuddyMaze` benchmark under both *One-Round* and *Multi-Round* reasoning modes. Each mode is tested across two noise levels (1, 3) and three difficulty tiers (Basic, Advanced and Challenge), allowing us to systematically assess reasoning robustness under varying uncertainty. We employ two backbone models—Llama-3.2-3B-Instruct (LLaMA) (Grattafiori et al., 2024) and Qwen2.5-3B-Instruct (Qwen) (Qwen et al., 2025)—and report two complementary metrics: *Multi-Hop Accuracy*, which measures global reconstruction of the reasoning chain, and *Single-Wise Accuracy*, which captures local coherence between consecutive evidence steps.

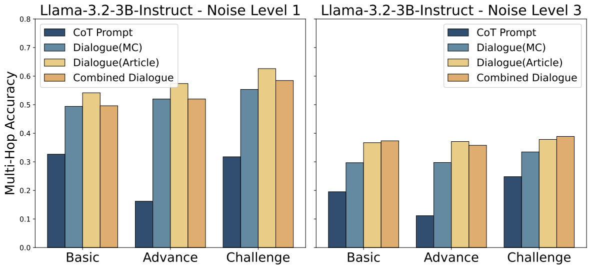

Figure 3: Comparison between `dialogue-tuning` and Chain-of-Thought (CoT) prompting in the One-Round setting across Basic, Advanced, and Challenge tasks under Noise Levels 1 and 3.

***Q1*: What is the effect of `dialogue-tuning` on reasoning across all difficulty levels in *One-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). 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 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.

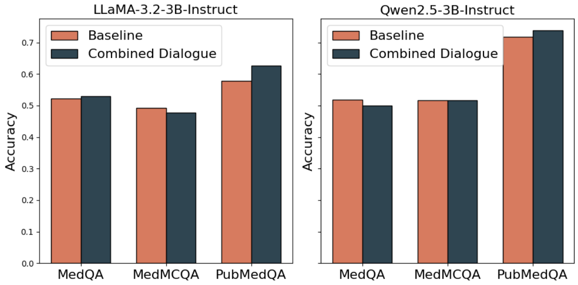

Figure 4: Comparison of scores between baseline and combined dialogue approaches for LLaMA 3.2-3B Instruct and Qwen2.5-3B-Instruct across MedQA, MedMCQA, and PubMedQA datasets. The combined dialogue approach consistently improves performance.

***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 performance 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.

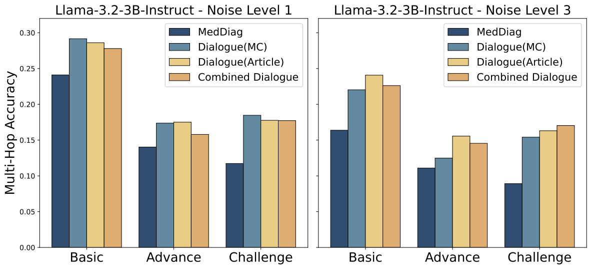

Figure 5: Comparison between `dialogue-tuning` and MedDiag-tuned models in the Multi-Round setting across three difficulty levels under Noise Levels 1 and 3.

***Q3*: How does `dialogue-tuning`compare to existing reasoning and dialogue-based training approaches? *A3*: It achieves more robust reasoning than CoT and MedDiag baselines while generalizing well to standard QA tasks.**

As shown in Figure 3, dialogue-tuned models consistently surpass CoT prompting across all difficulty levels and noise conditions, indicating that reasoning-oriented dialogue supervision provides stronger inductive bias than explicit step-by-step prompting. In the Multi-Round setting (Figure 5), our synthetic dialogue-tuned models also match or exceed the performance of models trained on MedDiag (Zeng et al., 2020a) conversations, suggesting that strategic dialogue reformulation effectively captures diagnostic reasoning structure. Finally, as shown in Figure 4, dialogue-tuned models generalize well to widely used general medical QA datasets (MedQA (Yao et al., 2024), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019)), confirming the robustness and transferability 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

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**

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.

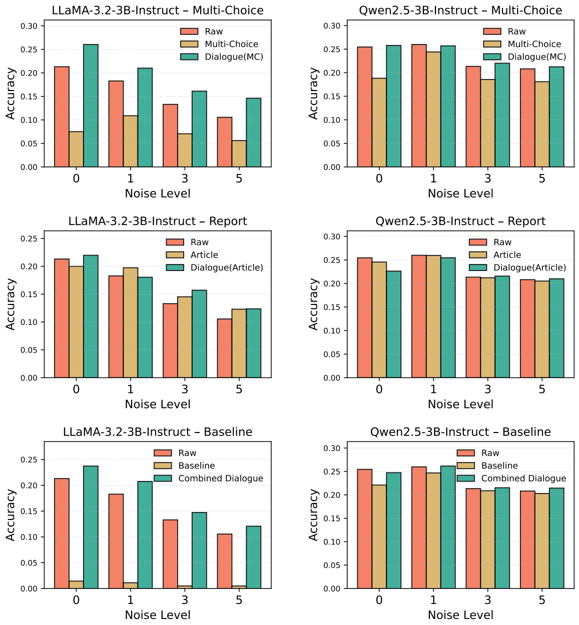

Figure 6: Performance of Llama-3.2-3B-Instruct and Qwen2.5-3B-Instruct under three tuning strategies across four noise levels.

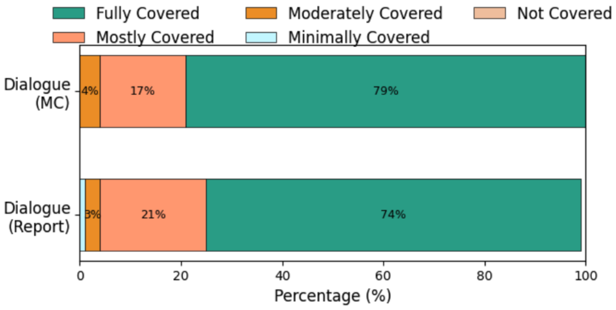

Figure 7: Human Evaluation Performance.

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.

6 Conclusion

In this work, we introduced MuddyMaze, a novel benchmark for evaluating the logical reasoning and evidence-based decision-making abilities of medical language models under realistic clinical uncertainty. By reframing medical QA as an evidence-ranking problem, MuddyMaze captures both holistic and stepwise reasoning behaviors that are critical in diagnostic practice. We further proposed dialogue-tuning, a training strategy that reformulates static QA and article data into simulated doctor–patient interactions, where the model is updated only on the physician's turns. Comprehensive experiments demonstrate that our method is both effective and efficient.

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. We appreciate this important point. To encourage diversity and robustness, our benchmark incorporates samples from multiple sources (MedQA, MedBullets, JAMA Challenge), each with varied disease coverage and linguistic styles.

Additionally, the noise injection and multi-step reasoning design of MuddyMaze introduce substantial unpredictability and complexity, mirroring unconventional and unpredictable situations that appear in real-world clinical decisions.

We acknowledge that more explicit evaluation of rare diseases or population shifts would strengthen our claims. As part of future work, we plan to extend our framework to clinical case simulations derived from local hospital EHR data, enabling evaluation on broader demographics and out-of-distribution cases.

Acknowledgement

Dr. Qingyu Chen is supported by the National Institutes of Health National Library of Medicine under Award Number R01LM014604.

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*.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025. Benchmarking large language models on answering and explaining challenging medical questions. *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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei Koh, and Bryan Hooi. 2024. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in large language models. *Preprint*, arXiv:2402.03271.

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.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Preprint*, arXiv:2406.00922.

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.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020a. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, and 1 others. 2020b. 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: MuddyMaze

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]
```

C Error Analysis

C.1 Case Analysis: Comparison of RAW, CoT-Prompt, and Dialogue-Tuned Models

1. **Raw Model:** The raw model often produces shallow, pattern-based outputs that loosely connect keywords like "chemotherapy" and "hearing loss" without deep reasoning. **Incorrect**

2. **COT-prompt Model:** Despite generating a coherent verbal explanation under the "Let's think step by step" prompt, the CoT model fails to select the correct evidence order. It narrates a cause → effect chain but ultimately chooses `[0, 1]`, reversing that logic. This illustrates a common CoT failure: the explanation and the action (evidence selection) are decoupled. The model focuses more on sounding logical than being structurally accurate. **Incorrect**

3. **Dialogue Tuning Model:** In contrast, the dialogue-tuned model demonstrates structured, clinically grounded reasoning: it identifies the treatment (chemotherapy) as the initiating factor, links it to a known adverse effect (sensorineural hearing loss), and leverages the timing of symptom onset to strengthen the causal inference. The selected evidence—[1], [0]—is minimal, correctly ordered, and directly supports the answer. **Correct**

Conclusion In summary, while raw and CoT-prompted models may produce plausible answers or fluent explanations, they often lack structured, causal reasoning. Dialogue-tuned models, by contrast, learn to reason in a clinically faithful, step-wise manner—aligning explanation, evidence selection, and causal order. This leads to more accurate, interpretable, and reliable multi-step inference in medical tasks.

C.2 Supplementary Outputs

Question and GroundTruth

Here is the background information: "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." 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. Answer: [1,0]

Raw Model Output

[0], [1]
These sentences support inferring that the drug caused the patient's symptoms because they provide evidence of a sensorineural hearing loss, which is consistent with ototoxicity, a known side effect of certain chemotherapy agents.

COT-prompt Model Output

To answer this question, let's analyze the situation step by step:

1. The patient has transitional cell carcinoma of the bladder and is experiencing a ringing sensation in his ear (tinnitus), which is a common side effect of certain chemotherapy drugs.
2. The patient recently received a course of neoadjuvant chemotherapy, which suggests that the tinnitus is likely related to this treatment.
3. The question asks about the expected beneficial effect of the drug that caused the patient's symptoms, which is tinnitus.
4. The answer is "Cross-linking of DNA", which is a mechanism of action associated with certain chemotherapy drugs, such as platinum-based agents (e.g., cisplatin).

Now, let's examine the evidence sentences:

0: Pure tone audiometry shows a sensorineural hearing loss of 45 dB.
This sentence suggests that the patient's hearing loss is related to the chemotherapy, but it doesn't directly support the answer.

1: He received this first course of neoadjuvant chemotherapy 1 week ago.
This sentence confirms that the patient recently received chemotherapy, which is consistent with the patient's symptoms.

To support inferring the answer, we need another sentence that connects the chemotherapy to the mechanism of action mentioned in the answer. Sentence 0 provides this connection by mentioning the hearing loss, which is a known side effect of certain chemotherapy drugs that cause cross-linking of DNA.

The relevant sentences are:
[0], [1]

Dialogue-Tuning Model Output

[1], [0]
These sentences support the inference that the drug causing the patient's symptoms is likely cisplatin, a common chemotherapeutic agent used to treat transitional cell carcinoma of the bladder. Cisplatin is known to cause ototoxicity, leading to sensorineural hearing loss, and its ototoxic effects can be detected through pure tone audiometry. The timing of the chemotherapy (1 week ago) also supports this inference, as cisplatin's ototoxic effects typically manifest within a week of treatment.

C.3 Error Analysis

Here is an example:

```
Here is the background information: "A
46-year-old man is brought to the emergency
department for evaluation of altered
mental status." Question: Which of the
following is the most likely cause of this
patient's symptoms? Answer: Benzodiazepine
intoxication.
Below are several evidence sentences.
Identify the 8 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: On physical examination, an alcoholic
smell and slurred speech are noted.
1: An ECG shows no abnormalities.
2: His pupils are normal.
3: He is somnolent but responsive when
aroused.
4: Neurological exam shows diminished deep
tendon reflexes bilaterally and an ataxic
gait.
5: Blood alcohol concentration is 0.04%.
6: He was found on the floor in front of
his apartment.
7: His pulse is 64/min, respiratory rate
is 15/min, and blood pressure is 120/75 mm
Hg.
Then provide only the indices of the
relevant sentences in brackets formatted
like this: [ ], no more than 8 sentences.
ANSWER:
```

- **Model Answer:** [0, 3, 4, 7, 6, 5, 1, 2]
- **Ground Truth:** [6, 3, 7, 0, 4, 2, 5, 1]

In this example, although our model is trained to capture clinical reasoning patterns, the nature of the evidence set poses a particular challenge. Many of the supporting sentences (e.g., 2, 5, 1) are clinically relevant but less explicit in their reasoning cues, requiring subtle inferences rather than obvious trigger words. This makes it harder for a smaller 3B-parameter model to consistently identify and order such evidence correctly. The model performs well on more straightforward cases but may struggle when clinical logic is embedded in finer-grained or seemingly trivial details. Enhancing the model's ability to handle such nuanced clinical reasoning remains a key direction for our future work.

D Details on Human evaluation and Human Annotation

Here is annotator agreement and survey introduction:

> Welcome and thank you for participating! In this questionnaire, we’re evaluating how well our system, which uses a LLM, transforms original paragraphs into dialogue format. Your feedback will help us assess the quality of the generated dialogues and improve future outputs. Please rate each dialogue based on how well it preserves the essential information from the original paragraph, using the following scale: 4 – Fully Covered: All essential information is present and clearly conveyed. 3 – Mostly Covered: Most key ideas are included, with minor omissions. 2 – Moderately Covered: About half of the vital information is included. 1 – Minimally Covered: Only a few important points are mentioned. 0 – Not Covered: The dialogue misses or distorts the core information. A score of 4 means the dialogue is excellent; 0 means it’s far off the mark. At the end of the questionnaire, there will be space for you to leave any suggestions or thoughts on how we can improve the prompts or the dialogue generation process. We truly appreciate your time and insights! If you have any questions, feel free to reach out to us anytime.

One error case analysis is shown below: *Doctor: “Unfortunately, you passed away three months after the onset of the right-shoulder swelling.*

Here the patient is simultaneously alive (able to converse) and deceased, revealing a temporal-logic breakdown in the LLM’s generation. Such errors slipped through the coverage rubric because all factual points from the source paragraph were indeed “present,” yet their combination was nonsensical.

Moreover, we conducted an additional annotation on a subset of 500 dialogues with the help of 20 domain experts across five specialties—Infectious Disease, Pediatrics, Oncology, Ophthalmology, and Neurology. Our previous evaluation covered 300 samples. The results reflect average agreement among annotators rather than single-annotator judgments, ensuring high reliability. In this larger set, 76% of sampled dialogues were rated as “fully covered” with clinically coherent reasoning.