DDxTutor: Clinical Reasoning Tutoring System with Differential Diagnosis-Based Structured Reasoning

Anonymous ACL submission

Abstract

Clinical diagnosis education requires students to master both systematic reasoning processes and comprehensive medical knowledge. While recent advances in Large Language Models (LLMs) have enabled various medical educational applications, these systems often provide direct answers that reduce students' cognitive engagement and lead to fragmented learning. We propose DDxTutor, a framework that follows differential diagnosis principles to decompose clinical reasoning into teachable components, consisting of (1) a structured reasoning module that analyzes clinical clues and synthesizes diagnostic conclusions, and (2) an interactive dialogue system that guides students through this process. To enable such tutoring, we construct *DDxReasoning*, a dataset of 933 clinical cases with fine-grained diagnostic steps verified by doctors. Our experiments demonstrate that fine-tuned LLMs achieve strong performance in both generating structured teaching references and conducting interactive diagnostic tutoring dialogues. Human evaluation by medical educators and students validates the framework's effectiveness for clinical diagnosis education. Code and data will be available.

1 Introduction

Clinical diagnosis, a core task in medical practice, involves synthesizing clinical information to reach a conclusion. As this process demands both depth and precision, systematic and scientific reasoning becomes indispensable (Fauci et al., 2008). For these demands, modern medical education has placed great emphasis on fostering clinical reasoning skills in students, aiming to strengthen the rigor and scientific foundation of diagnostic thinking (Schmidt and Mamede, 2015). Recent advances in Natural Language Processing (NLP), particularly Large Language Models (LLMs) like ChatGPT (Achiam et al., 2023) and DeepSeek (Liu et al., 2024a), have enabled new possibilities for interactive clinical education. These models have been applied to various medical educational tasks, from answering medical questions (Gilson et al., 2023) to providing feedback on junior doctors' diagnostic decisions (Huang et al., 2024).

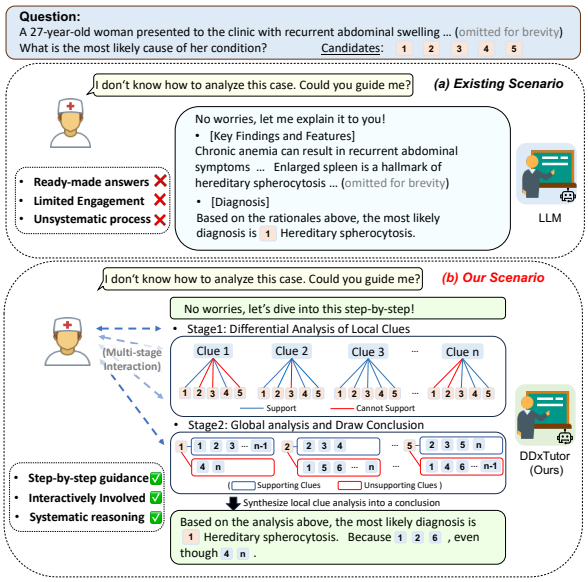


Figure 1: Comparison of clinical diagnostic training scenarios. (a) The existing scenario usually provides direct feedback with immediate answers and explanations, where students could become passive learners with an unsystematic knowledge intaking process. (b) Our proposed framework follows the thought of Differential Diagnosis (DDx), where medical students interactively engage in two steps: (1) independent analysis of individual clinical clues, and (2) global synthesis of findings to reach a diagnostic conclusion. Our structured approach enables systematic diagnostic reasoning while maintaining active student engagement with step-by-step guidance throughout the learning process.

However, current medical dialogue systems and diagnostic support tools still face significant limitations when applied to clinical reasoning education. Firstly, to the best of our knowledge, many existing medical LLMs are designed with a focus on providing direct diagnoses accompanied by accurate

and reasonable explanations (Sviridova et al., 2024; Tian et al., 2024). While these professional explanations serve as valuable references, they often present reasoning as a complete, ready-made output rather than a step-by-step process that students can actively participate in. From an educational perspective (Sweller, 1988), this approach reduces students' cognitive engagement, turning them into passive recipients rather than active participants in the diagnostic process. Although multi-turn dialogue systems allow students to ask follow-up questions about unclear explanations, this question-driven approach often leads to an unsystematic learning process (Barrows, 1986). As novice learners are still developing their clinical knowledge framework, their questions tend to address immediate comprehension gaps rather than following a structured diagnostic reasoning path, potentially missing critical diagnostic elements and relationships that are essential for building comprehensive clinical thinking skills (Schmidt and Mamede, 2015).

Facing these limitations in current medical education systems, we aim to devise a new framework that leverages LLMs to promote systematic clinical reasoning education with active student engagement across multiple diagnostic stages. To structure such an educational experience, we draw inspiration from Differential Diagnosis (DDx) (Fauci et al., 2008; First, 2013), a methodical process of weighing clinical evidence to identify the most likely diagnosis among various possibilities. In the first phase, individual patient clues are analyzed sequentially, examining their meaning, indication, specificity, and any supporting or unsupporting relationships with potential diagnoses. In the second phase, each analysis is synthesized to comprehensively evaluate diagnostic possibilities, arrive at a final conclusion, and offer explanations based on core evidence.

To operationalize this framework, we curated *DDxReasoning*, a dataset containing 933 doctor-verified clinical reasoning chains that capture these systematic diagnostic steps. This dataset serves dual purposes: it provides both fine-tuning data to enhance LLMs' ability to generate educational reasoning chains as tutoring reference. Moreover, our aims extend beyond having LLMs that merely follow these reasoning patterns. To enable interactive teaching, we further construct a dataset of simulated teacher-student dialogues. These dialogues demonstrate how to effectively guide students through the DDx reasoning process, enabling LLMs to conduct interactive clinical teaching while maintaining alignment with the structured reasoning objectives in DDxReasoning. Our main contributions are as follows:

(1) We propose *DDxTutor*, the first clinical diagnostic tutoring framework that leverages modern LLMs to implement systematic differential diagnosis teaching. By decomposing the diagnostic process into sequential clue analysis and global synthesis phases, our framework enables structured and interactive medical reasoning education.

(2) We construct *DDxReasoning*, a comprehensive dataset containing 933 expert-verified clinical cases with fine-grained diagnostic reasoning chains. This dataset not only serves as a benchmark for evaluating LLMs' diagnostic reasoning capabilities but also provides high-quality fine-tuning data for enhancing LLMs' tutoring value. Building upon this structured knowledge base, we further develop a dialogue generation approach that simulates teacher-student interactions aligned with the DDx reasoning process.

(3) Extensive experiments demonstrate the effectiveness of our approach from two complementary aspects: (a) LLMs fine-tuned on DDxReasoning demonstrate strong capabilities in comprehensive clinical scene analysis, providing structured analytical processes that serve as reliable teaching references; (b) The derivative dialogue tutoring system effectively performs student response tracking and analysis, providing adaptive explanations that guide students through the clinical reasoning process. These results validate our framework's ability to support systematic clinical reasoning education through both structured knowledge representation and interactive guidance.

2 Related Works

2.1 NLP for Medical Education

Medical education has emerged as a prominent research direction in Natural Language Processing (NLP). A notable application is the development of virtual patients powered by language models (Danforth et al., 2009; Menendez et al., 2015; Campillos-Llanos et al., 2020; Ali et al., 2021), which simulate authentic patient interactions to facilitate clinical training. These systems have demonstrated significant potential in enhancing medical education (Shi et al., 2024; Li et al., 2024b). In parallel, researchers have explored ways to improve pa-

tients' comprehension of medical instructions (Cai et al., 2023; Yao et al., 2024a). For instance, ChatCoach (Huang et al., 2024) functions as an AI copilot, helping healthcare providers refine their patient communication skills. Beyond these specialized educational tools, general-purpose medical question-answering systems (Yao et al., 2024b; Li et al., 2024a; Liu et al., 2024b) and applications designed for specific clinical tasks, such as medical history taking (Saley et al., 2024) and clinical note interpretation (Wang et al., 2024a), have also proven valuable as supplementary learning resources for medical students.

While these educational tools have shown promising results, they primarily focus on answering students' questions with direct explanations (as shown in Figure 1). Building upon this foundation, we propose to engage students through multiple steps of the DDx diagnostic process while tracking their fine-grained knowledge mastery.

2.2 Clinical Diagnosis Datasets (Benchmarks)

Clinical diagnosis datasets have emerged as crucial benchmarks for evaluating language models in healthcare applications. Traditional medical question-answering (QA) datasets, such as MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), and CMEXAM (Liu et al., 2024c), have laid the foundation for assessing medical knowledge comprehension. Recent advances have focused on developing more sophisticated diagnostic datasets (Li et al., 2024a; Wang et al.; Hou et al., 2024) that incorporate structured reasoning steps and support dynamic diagnostic processes. DDXPlus (Fansi Tchango et al., 2022) addresses the scarcity of training data for Automatic Diagnosis (AD) and Automatic Symptom Detection (ASD) by introducing a comprehensive synthetic dataset with differential diagnoses. Dual-Inf (Zhou et al., 2024) enhances diagnostic interpretability by combining expert annotations with Large Language Models (LLMs) to create an more explainable differential diagnosis framework.

These datasets have made valuable contributions to improving LLMs' diagnostic and explanatory capabilities. Complementing these efforts, our work emphasizes the educational value in clinical process dialogues by providing more fine-grained, comprehensive annotations and explanations to impart more systematic knowledge during student guidance.

3 Problem Formulation

Our proposed *DDxTutor* is designed following the core philosophy behind differential diagnosis (DDx): to convey fine-grained, intermediate reasoning steps throughout the diagnostic process. By transparently presenting the step-by-step thought process, our framework aims to empower students to develop a comprehensive understanding of clinical reasoning and sharpen their decision-making skills. To this end, *DDxTutor* is built around two pivotal tasks: (1) a structured differential diagnosis reasoning component that generates detailed teaching references, and (2) an interactive teacher-student module that provides dynamic guidance and feedback. We detail these two components below. Our framework is shown in Figure 2.

3.1 Structured DDx Reasoning

Following clinical differential diagnosis workflows (Elstein and Schwarz, 2002; Graber et al., 2005; First, 2013), we formulate a structured reasoning framework for clinical cases. Given a clinical diagnostic instance $\mathcal{I} = \{q, \mathbf{A}\}$, where q represents the clinical presentation and $\mathbf{A} = \{a_1, a_2, ..., a_m\}$ represents candidate diagnoses, the LLM generates a structured teaching reference $\mathcal{T} = \{\mathbf{L}, \mathbf{G}\}$ through a two-stage reasoning process, where $\mathbf{L}$ represents the local analysis for each clinical clue and $\mathbf{G}$ represents the global diagnostic synthesis respectively, which are described in detail as follows.

In the first stage, the LLM performs local analysis to generate $\mathbf{L}$. It begins by decomposing the clinical presentation q into a set of self-contained and relatively independent clinical clues $\mathbf{C} = \{c_1, c_2, ..., c_n\}$, where each c_i represents a complete and standalone clinical observation. For each clue c_i, the LLM conducts a detailed analysis $l_i \in \mathbf{L}$ that derives four new components: a specificity assessment of c_i ($spec_i$), an indication (ind_i) revealing the underlying clinical conditions, and two clue-candidate relation sets $\mathbf{P}_i$ and $\mathbf{N}_i$:

$$\begin{aligned} \mathbf{P}_i &= \left\{ (a_j, r_j^{(i)}) \mid c_i \rightarrow a_j, a_j \in \mathbf{A} \right\} \\ \mathbf{N}_i &= \left\{ (a_j, r_j^{(i)}) \mid c_i \not\rightarrow a_j, a_j \in \mathbf{A} \right\} \end{aligned} \quad (1)$$

where $r_j^{(i)}$ represents the corresponding reasoning, and $\rightarrow$ and $\not\rightarrow$ means whether the clue could support the candidate.

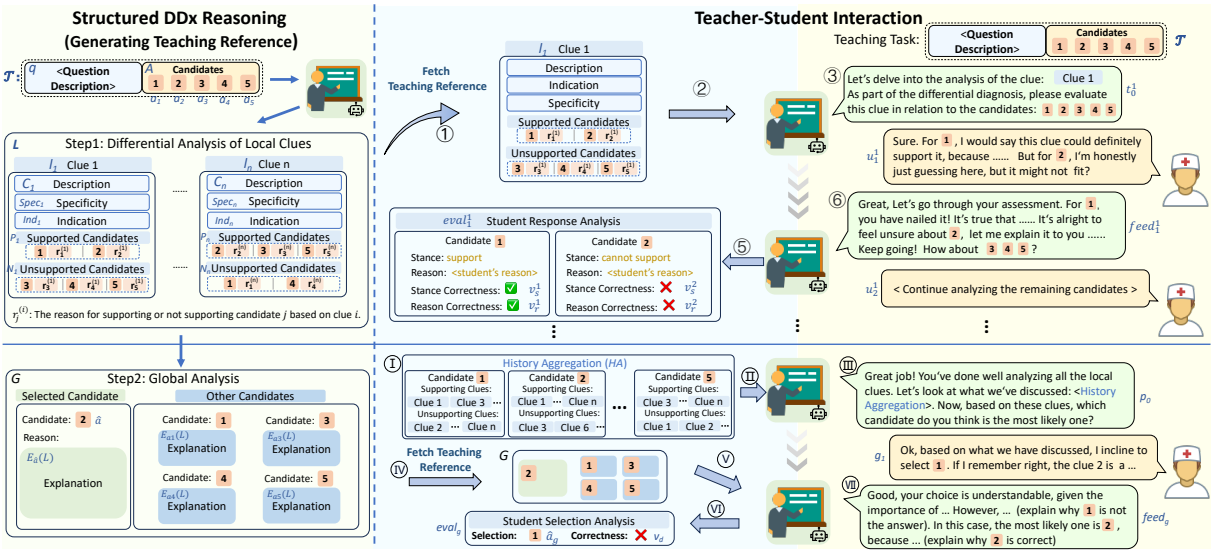


Figure 2: Our *DDxTutor* framework. Left: Teaching reference generation through structured differential diagnosis reasoning. Mid: Backend dialogue management, including teaching reference retrieval, student response analysis and record. Right: Interactive tutoring pipeline incorporating step-by-step clinical reasoning guidance and feedback.

In the Second stage, the LLM synthesizes all local analyses to generate a global synthesis $\mathbf{G}$ conditioning on $\mathbf{L}$:

$$\mathbf{G} = \{(\hat{a}, E_{\hat{a}}(\mathbf{L}))\} \cup \{(a_j, E_j(\mathbf{L})) \mid a_j \in \mathbf{A} \setminus \hat{a}\}$$

where $\hat{a} \in \mathbf{A}$ represents the selected diagnosis, and $E_j(\mathbf{L})$ denotes the explanation for candidate j based on evidence from $\mathbf{L}$.

3.2 Teacher-Student Interaction

Leveraging the teaching reference $\mathcal{T}$ generated from the *Structured DDx Reasoning* process, we design an interactive learning framework where the LLM serves as a medical educator guiding students through a systematic diagnostic reasoning process. The interaction follows a two-stage approach mirroring the structured DDx reasoning: (1) local clue analysis discussions and (2) global diagnostic synthesis discussions.

In the first stage, the LLM conducts detailed analytical dialogues for each clinical clue. Given a clinical clue analysis $l_i \in \mathbf{L}$, as defined in Section 3.1, the LLM initiates a structured discussion dialogue with student D^i_{local}:

$$D^i_{local,0:K} = \{(t^i_0), (u^i_1, t^i_1), (u^i_2, t^i_2), ..., (u^i_K, t^i_K)\}$$

where t^i_0 represents the initial teaching inquiry, u^i_k denotes the student's response in the k-th turn, and t^i_k represents the LLM's teaching response. Each teaching response t^i_k consists of two components: $t^i_k = (eval^i_k, feed^i_k)$, where $eval^i_k$ analyzes the student's response into a structured format: $eval^i_k = \{(a_j, stance_j, reason_j, v^j_s, v^j_r) \mid j \in J_k \subseteq \{1, ..., m\}\}$. Here, J_k represents the set of indices for candidates actually discussed by the student in the k-th turn, a_j represents the discussed candidate, $stance_j$ and $reason_j$ capture the student's supporting stance and reasoning of a_j respectively, while v^j_s and v^j_r are binary indicators denoting the correctness of the student's stance and reasoning as compared to l_i. Based on this evaluation, $feed^i_k$ provides targeted feedback by referencing the teaching reference l_i.

The complete local analysis phase comprises dialogues for all clinical clues:

$$D_{local} = \{D^i_{local}\}^n_{i=1}$$

In the second stage, the LLM facilitates a global diagnostic synthesis discussion D_{global} that follows a three-turn structure:

$$D_{global} = \{(\mathbf{L}), (p_1, g_1)\}$$

where p_1 denotes the student's final diagnostic decision with reasoning, and $g_1 = (eval_g, feed_g)$ represents the LLM's evaluation and feedback. The evaluation component $eval_g$ analyzes the student's final diagnosis: $eval_g = \{\hat{a}_g, v_d\}$ where $\hat{a}_g$ captures student's selected diagnosis and v_d is a binary indicator denoting the correctness of the student's selection. Based on this evaluation, $feed_g$ provides explanatory feedback incorporating the teaching reference $\mathbf{G}$.

4 DDx Reasoning Dataset

To support the structured DDx reasoning task described above, we developed a dataset based on clinical diagnosis questions from MedQA (Jin et al., 2021). Following the formulation in Section 3.1, each case in our dataset contains a clinical scenario $\mathcal{I} = \{q, \mathbf{A}\}$ and its corresponding teaching reference $\mathcal{T} = \{\mathbf{L}, \mathbf{G}\}$. The dataset consists of 933 cases (755 for training and 178 for testing), each presenting a complex clinical scenario with multiple symptoms and signs. Following the original setting of MedQA, each question has 5 candidate diagnoses.

We developed a two-stage dataset creation pipeline that aligns with our structured reasoning process, as shown in Appendix Figure 5.

Stage 1: Local Analysis Generation and Verification

First, for each clinical case $\mathcal{I}$, We use the OpenAI-o1 (Jaech et al., 2024) model to the local analysis component $\mathbf{L}$ by decomposing the question description q into independent clinical clues $\mathbf{C}$. For each clue c_i, it performs the structured analysis l_i as defined in Section 3.1, comprising a specificity assessment $spec_i$, clinical indication ind_i, supporting relationships $\mathbf{P}_i$, and unsupporting relationships $\mathbf{U}_i$.

Three experienced doctors then independently verify these local analyses, examining: (1) Clue decomposition - ensuring each clue represents a meaningful, independent clinical observation; (2) Clinical indications - validating the correct interpretation of symptoms and signs; (3) Specificity assessments - evaluating how uniquely each clue points to specific diagnoses; (4) Supporting/unsupporting relationships - verifying the correctness of candidate categorization and the validity of medical reasoning. For each sample, doctors independently review and mark problematic aspects of the local analyses. After all three doctors complete their reviews, we merge their annotations to identify overlapping concerns and unique issues. This merged feedback serves as the basis for a focused discussion to reach consensus on necessary modifications. This process continues until the merged feedback shows no remaining issues from any doctor.

Stage 2: Global Synthesis Generation and Verification

After verification of the local analyses, we feed the refined $\mathbf{L}$ back to OpenAI-o1 to generate the global synthesis component $\mathbf{G}$. This includes generating a final diagnostic decision $\hat{a}$, providing comprehensive reasoning for the chosen candidate, and explaining why other candidates were not selected.

The same doctor panel applies an identical review-merge-consensus process for the global analysis, focusing on (1) The logical coherence of the final decision, (2) The completeness of evidence integration, and (3) The validity of reasoning for both selected and rejected candidates. Similar to Stage 1, the annotations from all doctors are merged to identify common concerns and unique insights, followed by targeted discussions to resolve any remaining issues.

Statistics of DDxReasoning dataset, prompts used to build the data, examples, and more details are presented in Appendix A.

5 Knowledge-grounded Clinical Tutoring Dialogue Generation

Following the formulation of Teacher-Student Interaction in Section 3.2, we propose to simulate teacher-student dialogues based on the teaching reference $\mathcal{T}$. Simulated dialogue generation has been shown to be a cost-efficient and effective approach for developing educational systems (Wang et al., 2024b; Liu et al., 2024d). By creating diverse, high-quality simulated dialogues, we can effectively train the LLM to handle various student responses and teaching scenarios without the need for extensive real-world data collection.

Our dialogue generation process consists of two stages that mirror the structured DDx reasoning framework: local clue analysis dialogues and global diagnostic synthesis dialogues.

Local Analysis Dialogue Generation. For each clinical clue c_i, the dialogue begins with the teacher presenting the clue through an initial inquiry t_0^i and requesting analysis of its relationship to the candidates. In each turn, the student randomly selects k candidates to analyze ($1 \leq k \leq |\mathcal{A}_t|$), where $\mathcal{A}_t$ represents the remaining unanalyzed candidates.

For each selected candidate, the student provides a response that exhibits one of the following patterns: (1) correct stance with aligned reasoning, (2) correct stance with misaligned reasoning, (3) incorrect stance with misaligned reasoning, (4) random guessing, or (5) complete uncertainty. The teacher then evaluates this response against the teaching reference l_i and provides targeted feedback addressing both the correctness of stance and reasoning.

Global Synthesis Dialogue Generation. After

completing all local analysis dialogues, the teacher initiates a global discussion by presenting the previously analyzed clinical clues and requesting a final diagnostic decision. The student then synthesizes the evidence from all clues to select a final diagnosis and provide comprehensive reasoning for their choice. The teacher evaluates this final decision against the teaching reference $\mathbf{G}$ and provides detailed feedback that addresses both the diagnosis selection and the quality of synthesized reasoning.

We leverage GPT-4o (Hurst et al., 2024) to generate student responses and teacher feedback through carefully crafted prompts that ensure both dialogue coherence and pedagogical effectiveness. This approach creates authentic teacher-student interactions spanning diverse learning scenarios while remaining firmly grounded in the verified clinical knowledge from the DDx Reasoning Dataset. The resulting dataset comprises a Local Analysis Tutoring dialogue collection with 16,132 training and 3,760 testing turns, and a Global Synthesis dialogue collection with 1,506 training and 356 testing turns. A detailed illustration of this dialogue generation process is in Appendix B.

6 Experiments

We evaluate state-of-the-art LLMs on two core aspects of the DDxTutor framework: (1) the ability to generate comprehensive DDx reasoning chains that can serve as teaching references, and (2) the capacity to conduct tutoring dialogues that guide students through systematic diagnostic reasoning. For evaluation, we fine-tune three specialized groups of LLMs:

(1) Teaching Reference Generator: These models f_1 are trained to generate complete teaching references $\hat{\mathcal{T}}$ given the clinical case input $\mathcal{I}$: $\hat{\mathcal{T}} = f_1(\mathcal{I})$.

(2) Local Analysis Stage Dialogue Tutor: These models f_2 are trained to generate teaching responses $\hat{t}_k^i$ based on the clinical clue c_i, history dialogue $D^i_{local,0:(k-1)}$, student's last utterance u_k^i, and local teaching reference l_i: $\hat{t}_k^i = f_2(c_i, D_{local,0:(k-1)}, u_k^i, l_i)$.

(3) Global Synthesis Stage Dialogue Tutor: These models f_3 generate final feedback g_1 based on all local analyses with aggregated information HA, student's final diagnosis p_1, and global teaching reference $\mathbf{G}$: $\hat{g_1} = f_3(\mathbf{L}, HA, p_1, \mathbf{G})$.

We select models with parameters ranging from 3B to 32B, including 7 general open source LLMs: Qwen2.5-3B, Qwen2.5-7B, Qwen 2.5-14B, Qwen 2.5-32B (Qwen et al., 2025), LLama3.2-3B, LLama3.1-8B (Grattafiori et al., 2024), Phi4-14B (Abdin et al., 2024) and 2 Medical LLMs: Meditron-8B (Chen et al., 2023, 2024), and Med42-8B (Christophe et al., 2024).

6.1 Implementation Details

All the models are finetuned with LoRA (Hu et al., 2021) using the Unsloth framework (Daniel Han and team, 2023). We set the LoRA rank r=16 and α=16, targeting key transformer modules including attention layers and feed-forward networks. The models were trained for 3 epochs using AdamW optimizer, a learning rate of 2e-4, and a batch size of 8. A linear learning rate scheduler with 20 warmup steps was employed. For f_1 models, we further finetuned two 70B-level model: Qwen2.5-72B and Med42-70B with 4-bit Qlora (Dettmers et al., 2023). All the models are trained on a server with 2 A100 GPUs.

6.2 Evaluation Setting

We evaluate the f_1 models from the following aspects: **(1) Local Clue Analysis Capability**: the alignment of the decoupled local clues with teaching references using metrics $Bert_C$ and $Meteor_C$; the accuracy (Acc_{PN}) between predicted and ground-truth stances for each clue; the semantic and textual quality of reasoning explanations measured by $Bert_{PN}$ and $Meteor_{PN}$; the quality of clue specificity descriptions using $Bert_{Spec}$ and $Meteor_{Spec}$; the quality of indication description of the clue using $Bert_{Ind}$ and $Meteor_{Ind}$; **(2) Global Synthesis and Conclusion:** the final diagnosis selection accuracy $Acc_{\hat{a}}$; and the quality of final explanations using $Bert_E$ and $Meteor_E$. Here, $Bert$ and $Meteor$ refer to BertScore (Zhang et al., 2019) and Meteor Score (Banerjee and Lavie, 2005) respectively.

For f_2, we evaluate their: **(1) Student Response Extraction capability**: measuring the intersection over union (IOU_a) between extracted student-analyzed candidates and ground truth, the accuracy of extracting students' supporting/non-supporting stances (Acc_{stance}), and the quality of extracted students' reasons ($Bert_{reason}$ and $Meteor_{reason}$); **(2) Student Response Judgment capability**: evaluating the accuracy of tutors' judgment on students' stances (Acc_{vs}) and the consistency between students' reasoning and teaching references (Acc_{vc}); **(3) Teacher Feedback Generation capability**:

Model	Bert_C	Meteor_C	Acc_{PN}	Bert_{PN}	Meteor_{PN}	Bert_{Ind}	Meteor_{Ind}	Bert_{Spec}	Meteor_{Spec}	Bert_E	Meteor_E	$\text{Acc}_{\hat{a}}$
Qwen2.5-3B	0.836	0.695	0.664	0.600	0.248	0.891	0.242	0.884	0.320	0.757	0.387	0.600
Llama3.2-3B	0.884	0.760	0.681	0.620	0.284	0.896	0.290	0.888	0.357	0.803	0.434	0.706
Qwen2.5-7B	0.842	0.706	0.707	0.639	0.269	0.894	0.274	0.822	0.316	0.823	0.442	0.777
Llama3.1-8B	0.887	0.769	0.739	0.677	0.334	0.899	0.309	0.896	0.389	0.849	0.466	0.832
Meditron-8B	0.901	0.779	0.730	0.669	0.337	0.900	0.316	0.898	0.398	0.828	0.451	0.773
Med42-8B	0.895	**0.790**	0.729	0.668	0.331	0.901	0.315	0.899	0.414	0.849	0.476	0.836
Qwen2.5-14B	0.881	0.720	0.734	0.665	0.294	0.893	0.265	0.895	0.359	0.825	0.441	0.776
Phi4-14B	0.896	0.787	0.767	**0.705**	**0.371**	**0.905**	**0.350**	0.903	**0.432**	0.863	**0.500**	0.858
Qwen2.5-32B	0.884	0.724	0.754	0.686	0.319	0.897	0.288	0.898	0.379	0.872	0.486	0.892
Qwen2.5-72B	0.888	0.741	**0.773**	0.703	0.328	0.897	0.296	0.897	0.383	0.867	0.498	0.874
Med42-70B	**0.912**	0.789	0.767	0.705	0.360	0.902	0.339	**0.904**	0.428	**0.877**	0.492	**0.901**
Phi4-14B (2shot)	0.759	0.611	0.718	0.654	0.300	0.895	0.264	0.894	0.368	0.798	0.345	0.738
GPT-4o-mini (2shot)	0.600	0.377	0.585	0.520	0.161	0.885	0.189	0.883	0.243	0.780	0.181	0.792
GPT-4o (2shot)	0.540	0.312	0.741	0.666	0.251	0.886	0.233	0.890	0.294	0.861	0.271	0.959

Table 1: Performance of Teaching Reference Generator f_1 on the *DDxReasoning Dataset*.

Model	IOU_a	Acc_{stance}	Bert_{reason}	Acc_{vs}	Acc_{vr}	Bert_{feed}	Meteor_{feed}	$\text{Acc}_{\hat{a}}$	Acc_{vd}	Bert_{feed_g}	Meteor_{feed_g}
Qwen2.5-3B	0.994	0.987	0.949	0.985	0.867	0.895	0.415	0.978	0.997	0.915	0.514
Llama3.2-3B	0.996	0.992	0.950	0.991	0.875	0.900	0.454	0.983	0.997	0.918	0.527
Qwen2.5-7B	0.995	0.975	0.938	0.977	0.868	0.896	0.420	0.922	0.997	0.913	0.512
Llama3.1-8B	0.997	0.992	0.954	0.992	0.898	0.901	0.462	0.992	0.994	0.918	0.526
Meditron-8B	0.996	0.991	0.954	0.992	0.900	0.901	0.456	0.994	0.997	0.918	0.527
Med42-8B	0.997	0.993	0.956	0.993	0.902	0.903	0.471	0.992	0.994	0.918	0.525
Qwen2.5-14B	0.991	0.989	0.952	0.989	0.885	0.898	0.437	0.961	0.997	0.915	0.518
Phi4-14B	**0.998**	**0.994**	**0.957**	**0.994**	**0.914**	**0.905**	**0.481**	**0.997**	**1.000**	**0.920**	**0.539**
Qwen2.5-32B	0.997	0.993	0.953	0.993	0.903	0.900	0.451	0.994	0.997	0.916	0.523
Phi4-14B (2shot)	0.890	0.965	0.874	0.848	0.740	0.878	0.312	0.553	0.991	0.910	0.506
GPT-4o-mini (2shot)	0.886	0.893	0.880	0.878	0.704	0.881	0.360	0.919	0.997	0.901	0.476
GPT-4o (2shot)	0.981	0.955	0.881	0.919	0.789	0.880	0.335	0.800	1.000	0.896	0.469

Table 2: Performance of Local Analysis Tutor f_2 and Global Synthesis Tutor f_3 models on the simulated dialogues.

assessing feedback quality using $Bert_{feed}$ and $Meteor_{feed}$.

For f_3, following a similar evaluation logic as f_2, we assess its capability to extract students' final diagnosis selection ($Acc_{\hat{a}g}$), the accuracy of tutors' judgment (Acc_{vd}), and the quality of teacher feedback ($Bert_{feed\_g}$ and $Meteor_{feed\_g}$).

More detailed explanation of our evaluation protocol is presented in Appendix C.

6.3 Results of Teaching Reference Generators

The quantitative evaluation of teaching reference generators (f_1 models) reveals several key findings. As shown in Table 1 and Figure 3, model performance generally improves with increasing parameter count, following established scaling laws (Kaplan et al., 2020; Zhang et al.). This trend is particularly evident in the Qwen model family, which demonstrates consistent performance gains across parameter scales.

Phi4-14B achieves optimal efficiency with the best balance between performance and computational cost among evaluated models. Notably, medical domain-specialized models like Med42-8B and Meditron-8B fail to demonstrate their expected advantages in general medical QA (Chen et al., 2023; Christophe et al., 2024), despite medical pretraining. This underperformance may be attributed to the limited representation of complex, multi-step clinical reasoning chains in their pre-training datasets, which drag back their performance to near or even their parent model, Llama3.1-8b.

Further analysis through correlation mapping (Figure 4) reveals that the final diagnosis accuracy ($Acc_{\hat{a}}$) is most strongly correlated with three local analysis metrics: stance accuracy (Acc_{PN}) and reasoning quality measures ($Bert_{PN}$ and $Meteor_{PN}$). This phenomenon aligns with clinical intuition that accurate stance identification and sound reasoning at each local analysis step are fundamental to reaching correct final diagnoses.

Few (two) shot results of Phi4-14b, GPT-4o-mini and GPT-4o demonstrate that few-shot models seem to take a different path to reach solutions, or to some extent, they are 'slacking off'. This is evidenced by their poor clue decoupling scores ($Bert_C$ and $Meteor_C$) while maintaining high final selection accuracy ($Acc_{\hat{a}}$). This could indicate that few-shot models are leveraging their pre-trained knowledge to make direct diagnostic leaps rather than following the systematic reasoning process we aim to teach. While this may achieve cor-

rect final diagnoses, it bypasses the step-by-step analytical process that is crucial for medical education. This observation highlights the importance of fine-tuning in ensuring models adhere to structured diagnostic reasoning patterns rather than relying on shortcuts that, while potentially effective for diagnosis, may be less valuable for teaching purposes.

6.4 Results of Dialogue Tutors

The evaluation of dialogue tutors (f_2 and f_3 models) reveals promising capabilities across multiple dimensions (Table 2). All fine-tuned models demonstrate robust performance in recognizing candidate diagnoses from student responses—a critical ability for maintaining effective dialogue flow and monitoring student progress. The models also exhibit high precision in assessing student response correctness, as evidenced by strong Acc_{vs} and Acc_{vr} scores.

Significantly, even models with relatively modest parameters (3B) achieve satisfactory performance in workflow control, suggesting the framework's viability across various model scales. The minimal performance differential between models of different sizes indicates that dialogue tutoring, when supported by teaching references, may not demand the same model capacity as reference generation. To validate this hypothesis, we conducted an ablation study without teaching references (detailed in Appendix Table 8). The results align with our expectations, showing a predictable performance decline when Teaching material input is removed.

7 Human Evaluation

To validate our framework's effectiveness, we conducted rigorous human evaluations following the methodology established in NoteChat (Wang et al., 2024a), which focused on simulating patient-doctor dialogues. Our evaluation encompassed both expert assessment and user experience from two perspectives.

For evaluating the pedagogical value of Teaching References, we enlisted three experienced educators, each with over a decade of medical teaching experience at government-funded health institutions. These experts evaluated 60 Teaching References, comprising 20 samples each from Phi4-14b, GPT-4o (2-shot), and the ground truth from *DDxReasoning* (GT). Their assessment focused on the materials' potential as both immediate teaching resources and structured draft materials.

From the learning perspective, we engaged five medical students to assess the quality of local and global dialogues generated by our f_2 and f_3 models respectively. The evaluation covered 20 cases per dialogue type, comparing outputs from LLaMA-3.2-3B, Phi4-14B, and GPT-4o (2-shot).

Participants in both evaluations ranked outputs on a 1-to-3 scale, with 1 being most preferred. We quantified preferences using Mean Reciprocal Rank (MRR), as detailed in Table 3. The results demonstrate that our fine-tuned Phi4-14B model achieves comparable or superior performance to GPT-4o (2-shot) mode within our DDx Tutoring framework, highlighting its potential for practical applications.

Teaching Reference f_1		
GT	Phi4-14B	GPT4o (2shot)
0.850	0.600	0.383
Local Analysis Dialogue f_2		
llama3.2-3B	Phi4-14B	GPT4o (2shot)
0.425	**0.758**	0.650
Global Synthesis Dialogue $f3$		
llama3.2-3B	Phi4-14B	GPT4o (2shot)
0.483	**0.708**	0.642

Table 3: MRR scores for human evaluation.

8 Conclusion

This paper presents DDxTutor, a novel framework that leverages LLMs to provide systematic clinical diagnostic education through structured reasoning and interactive tutoring. By decomposing the diagnostic process into sequential clue analysis and global synthesis phases, our approach enables active student engagement while maintaining scientific rigor in clinical reasoning education. The framework is supported by DDxReasoning, a comprehensive dataset of expert-verified clinical reasoning chains that serves both as fine-tuning data and evaluation benchmark.

Our experimental results demonstrate that fine-tuned LLMs can effectively generate structured teaching references and conduct pedagogically sound tutoring dialogues. The strong performance of mid-sized models like Phi4-14B suggests that effective clinical tutoring systems are achievable with reasonable computational resources. Human evaluation from both medical educators and students validates the educational value of our framework, with experts highlighting the quality of generated teaching materials and students confirming the effectiveness of the interactive tutoring approach.

9 Limitation

While DDxTutor demonstrates promising potential in clinical education, several important limitations need to be acknowledged:

First, DDxTutor's educational scope is relatively focused. Although differential diagnosis represents a classical approach to clinical reasoning, it is not the sole objective of medical education. Traditional medical teaching encompasses various pedagogical formats, including didactic lectures, Problem-Based Learning (PBL), case-based discussions, and bedside teaching. Each format serves distinct educational goals and develops different aspects of clinical competency. In this context, DDxTutor primarily addresses structured diagnostic reasoning training and may need adaptation to support broader educational objectives and teaching modalities.

Second, our framework makes several assumptions about student-teacher interactions, particularly in dialogue construction. For instance, during local clue analysis, we assume students will provide structured responses containing both their analysis and supporting rationale. While this assumption aligns with conventional medical discussions and proved effective during human evaluation, it may not fully capture the diversity of student responses in real educational settings. As a pioneering work in this domain, DDxTutor currently has limited capability to handle off-pattern or unexpected student responses. Future iterations could potentially integrate agent workflow frameworks to enhance robustness in real-world applications, allowing more flexible and adaptive responses to diverse student interaction patterns.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mohammad Rafayet Ali, Taylan Sen, Benjamin Kane, Shagun Bose, Thomas M Carroll, Ronald Epstein, Lenhart Schubert, and Ehsan Hoque. 2021. Novel computational linguistic measures, dialogue system and the development of sophie: Standardized online patient for healthcare interaction education. *IEEE Transactions on Affective Computing*, 14(1):223–235.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Howard S Barrows. 1986. A taxonomy of problem-based learning methods. *Medical education*, 20(6):481–486.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Pengshan Cai, Zonghai Yao, Fei Liu, Dakuo Wang, Meghan Reilly, Huixue Zhou, Lingxi Li, Yi Cao, Alok Kapoor, Adarsha Bajracharya, et al. 2023. Paniniqa: Enhancing patient education through interactive question answering. *Transactions of the Association for Computational Linguistics*, 11:1518–1536.

Leonardo Campillos-Llanos, Catherine Thomas, Éric Bilinski, Pierre Zweigenbaum, and Sophie Rosset. 2020. Designing a virtual patient dialogue system based on terminology-rich resources: Challenges and evaluation. *Natural Language Engineering*, 26(2):183–220.

Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Zeming Chen, Alejandro Hernández-Cano, et al. 2024. Meditron-70b: Scaling medical pretraining for large language models. https://huggingface.co/OpenMeditron/Meditron3-8B.

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical llms.

Douglas R Danforth, Mike Procter, Richard Chen, Mary Johnson, and Robert Heller. 2009. Development of virtual patient simulations for medical education. *Journal For Virtual Worlds Research*, 2(2).

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Arthur S Elstein and Alan Schwarz. 2002. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *Bmj*, 324(7339):729–732.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. Ddxplus: A new dataset for automatic medical diagnosis. *Advances in neural information processing systems*, 35:31306–31318.

Anthony S Fauci, Eugene Braunwald, Dennis L Kasper, Stephen L Hauser, Dan L Longo, J Larry Jameson, and Joseph Loscalzo. 2008. Harrison's principles of internal medicine. In *Harrison's principles of internal medicine*, pages 2754–2754.

Michael B First. 2013. *DSM-5-TR® Handbook of Differential Diagnosis*. American Psychiatric Pub.

Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. 2023. How does chatgpt perform on the united states medical licensing examination (usmle)? the implications of large language models for medical education and knowledge assessment. *JMIR medical education*, 9(1):e45312.

Mark L Graber, Nancy Franklin, and Ruthanna Gordon. 2005. Diagnostic error in internal medicine. *Archives of internal medicine*, 165(13):1493–1499.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ruihui Hou, Shencheng Chen, Yongqi Fan, Lifeng Zhu, Jing Sun, Jingping Liu, and Tong Ruan. 2024. Msdiagnosis: An emr-based dataset for clinical multi-step diagnosis. *arXiv preprint arXiv:2408.10039*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hengguan Huang, Songtao Wang, Hongfu Liu, Hao Wang, and Ye Wang. 2024. Benchmarking large language models on communicative medical coaching: A dataset and a novel system. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1624–1637.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

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*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

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

Yanzeng Li, Cheng Zeng, Jialun Zhong, Ruoyu Zhang, Minhao Zhang, and Lei Zou. 2024b. Leveraging large language model as simulated patients for clinical education. *arXiv preprint arXiv:2404.13066*.

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

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. 2024b. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. 2024c. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Zhengyuan Liu, Stella Xin Yin, Geyu Lin, and Nancy F. Chen. 2024d. Personality-aware student simulation for conversational intelligent tutoring systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 626–642, Miami, Florida, USA. Association for Computational Linguistics.

Elisa Menendez, Blície Balisa-Rocha, Monique Jabbur-Lopes, Wanderson Costa, José Rafael Nascimento, Marcos Dósea, Leila Silva, and Divaldo Lyra Junior. 2015. Using a virtual patient system for the teaching of pharmaceutical care. *International journal of medical informatics*, 84(9):640–646.

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.

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, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Vishal Saley, Goonjan Saha, Rocktim Das, Dinesh Raghu, et al. 2024. Meditod: An english dialogue dataset for medical history taking with comprehensive annotations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16843–16877.

Henk G Schmidt and Sílvia Mamede. 2015. How to improve the teaching of clinical reasoning: a narrative review and a proposal. *Medical education*, 49(10):961–973.

Xiaoming Shi, Zeming Liu, Li Du, Yuxuan Wang, Hongru Wang, Yuhang Guo, Tong Ruan, Jie Xu, Xiaofan Zhang, and Shaoting Zhang. 2024. Medical dialogue system: A survey of categories, methods, evaluation and challenges. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2840–2861, Bangkok, Thailand. Association for Computational Linguistics.

Ekaterina Sviridova, Anar Yeginbergen, Ainara Estarrona, Elena Cabrio, Serena Villata, and Rodrigo Agerri. 2024. Casimedicos-arg: A medical question answering dataset annotated with explanatory argumentative structures. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18463–18475.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285.

Yuanhe Tian, Ruyi Gan, Yan Song, Jiaxing Zhang, and Yongdong Zhang. 2024. ChiMed-GPT: A Chinese medical large language model with full training regime and better alignment to human preferences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7156–7173, Bangkok, Thailand. Association for Computational Linguistics.

Mina Valizadeh and Natalie Parde. 2022. The ai doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6638–6660.

Bowen Wang, Jiuyang Chang, Yiming Qian, Guoxin Chen, Junhao Chen, Zhouqiang Jiang, Jiahao Zhang, Yuta Nakashima, and Hajime Nagahara. Direct: Diagnostic reasoning for clinical notes via large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2024a. Notechat: a dataset of synthetic patient-physician conversations conditioned on clinical notes. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15183–15201.

Junling Wang, Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2024b. Book2Dial: Generating teacher student interactions from textbooks for cost-effective development of educational chatbots. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9707–9731, Bangkok, Thailand. Association for Computational Linguistics.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207.

Zonghai Yao, Nandyala Siddharth Kantu, Guanghao Wei, Hieu Tran, Zhangqi Duan, Sunjae Kwon, Zhichao Yang, and Hong Yu. 2024a. README: Bridging medical jargon and lay understanding for patient education through data-centric NLP. In *Findings of the Association for Computational Linguistics:*

EMNLP 2024, pages 12609–12629, Miami, Florida, USA. Association for Computational Linguistics.

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

Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. When scaling meets llm finetuning: The effect of data, model and finetuning method. In *The Twelfth International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Shuang Zhou, Mingquan Lin, Sirui Ding, Jiashuo Wang, Genevieve B Melton, James Zou, and Rui Zhang. 2024. Interpretable differential diagnosis with dual-inference large language models. *arXiv preprint arXiv:2407.07330*.

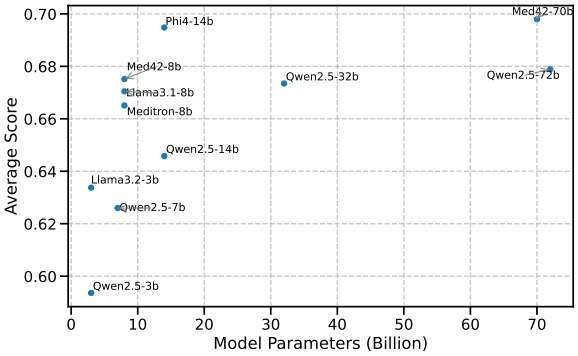


Figure 3: Performance Vs Parameters of Teaching Reference Generator f_1 models.

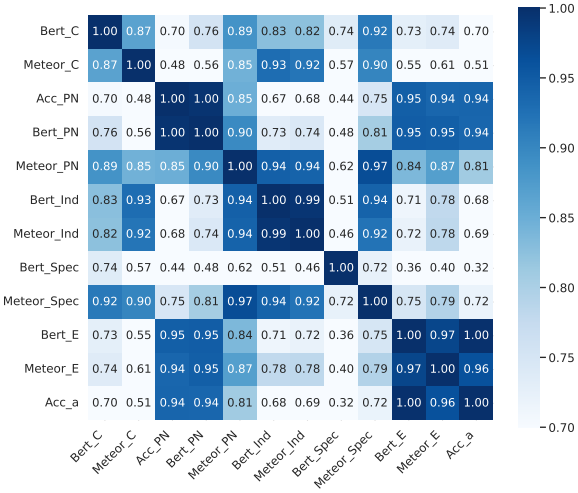


Figure 4: The correlation matrix of the performance Table 1 of Teaching Reference Generators.

A More Details on DDxReasoning Dataset

Our DDxReasoning dataset was developed through a systematic two-stage annotation process combining LLM generation with expert medical verification. Initially, we selected 1,000 clinical cases from the MedQA dataset (Jin et al., 2021), with an intended split of 800 training and 200 testing cases. However, through our rigorous verification process, some cases were eliminated due to quality concerns, resulting in a final dataset of 933 cases (755 training and 178 testing cases).

The first stage focuses on generating and validating detailed local analyses for individual clinical clues. We utilize OpenAI O1 with Local Analysis Prompt ($PromptL$, presented in Table 4) to generate initial structured analyses for each clinical presentation, breaking down complex cases into discrete, analyzable components. Each generated analysis comprises a specificity description, clini-

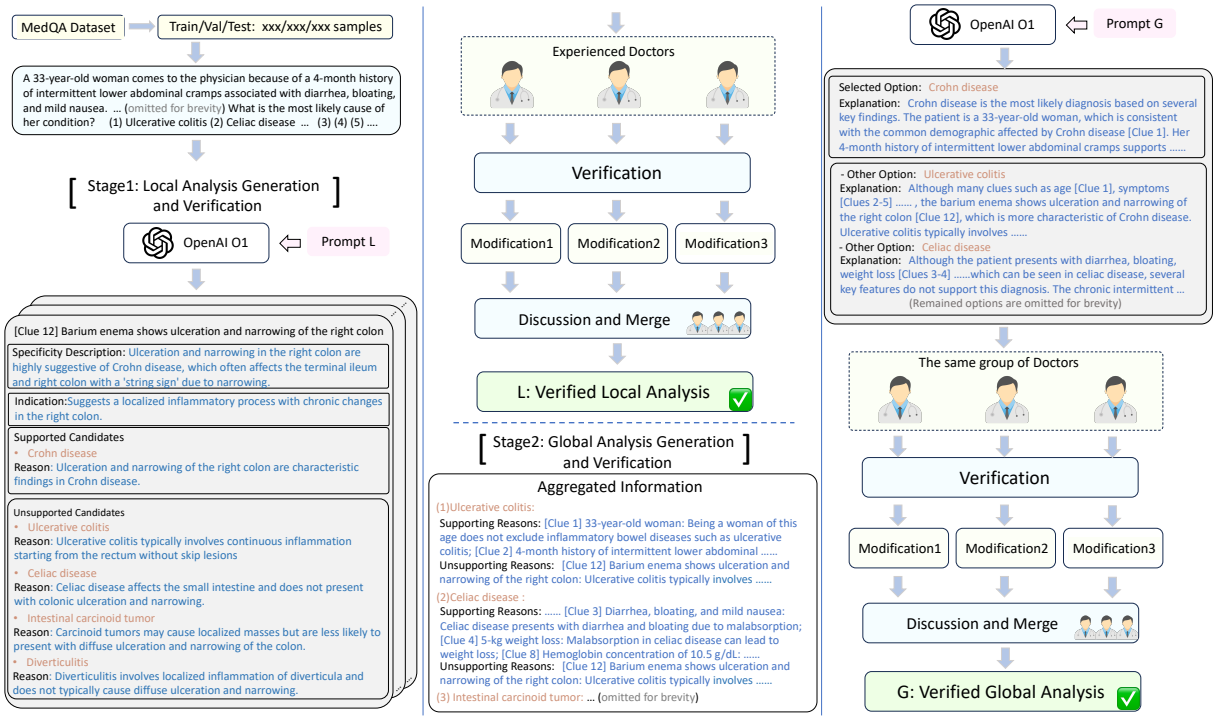


Figure 5: The two-stage generation and verification pipeline of the proposed *DDxReasoning* dataset

cal indication, and detailed relationships with candidate diagnoses, classifying them into supported and unsupported categories. During this stage, approximately 49 cases were removed due to unclear or ambiguous clinical presentations that could not be effectively decomposed into distinct clues.

These local analyses then undergo thorough verification by a panel of three experienced doctors. Each doctor independently proposes modifications (Modification 1, Modification 2, Modification 3), followed by a collaborative discussion and merge phase. This verification process continues iteratively until all three doctors reach consensus, resulting in the Verified Local Analysis L.

In the second stage, we aggregate the verified local analyses and use OpenAI O1 again with Global Analysis Prompt (Prompt G, presented in Table 4) to generate comprehensive diagnostic reasoning, integrating all verified clues to form final diagnostic conclusions with detailed supporting and unsupporting reasons for each candidate. This global analysis undergoes the same rigorous verification process by the same group of doctors, leading to the final Verified Global Analysis G. Through this two-stage process, an additional 18 cases were eliminated due to inconsistencies in clinical reasoning or lack of consensus among experts.

The **statisics** of the DDxReasoning Dataset is presented in Table 5. An sample of the dataset is presented in Figure 6.

All our annotators are experienced doctors with PhD degrees, and they are compensated at rates satisfying local market guidelines.

B Knowledge-grounded Clinical Tutoring Dialogue Generation

B.1 Dialogue Generation of Local Analysis

To create a knowledge-grounded dialogue dataset that enables LLMs to effectively perform clinical tutoring, we propose a structured dialogue generation framework (Figure 7). The framework aims to develop three key capabilities in the fine-tuned LLM: (1) student response awareness - understanding what clinical elements the student analyzed in their last utterance, (2) stance recognition - identifying whether the student believes a discussed clue supports specific candidates, and (3) judgment capability - assessing the correctness of student stances and reasoning.

Response State Design

For each clinical clue $c_i \in \mathbf{C}$, the student response should analyze its relationship with candidate diagnoses. Each analysis contains two key components:

- A stance $stance_j$ indicating whether the clue supports or does not support the candidate a_j

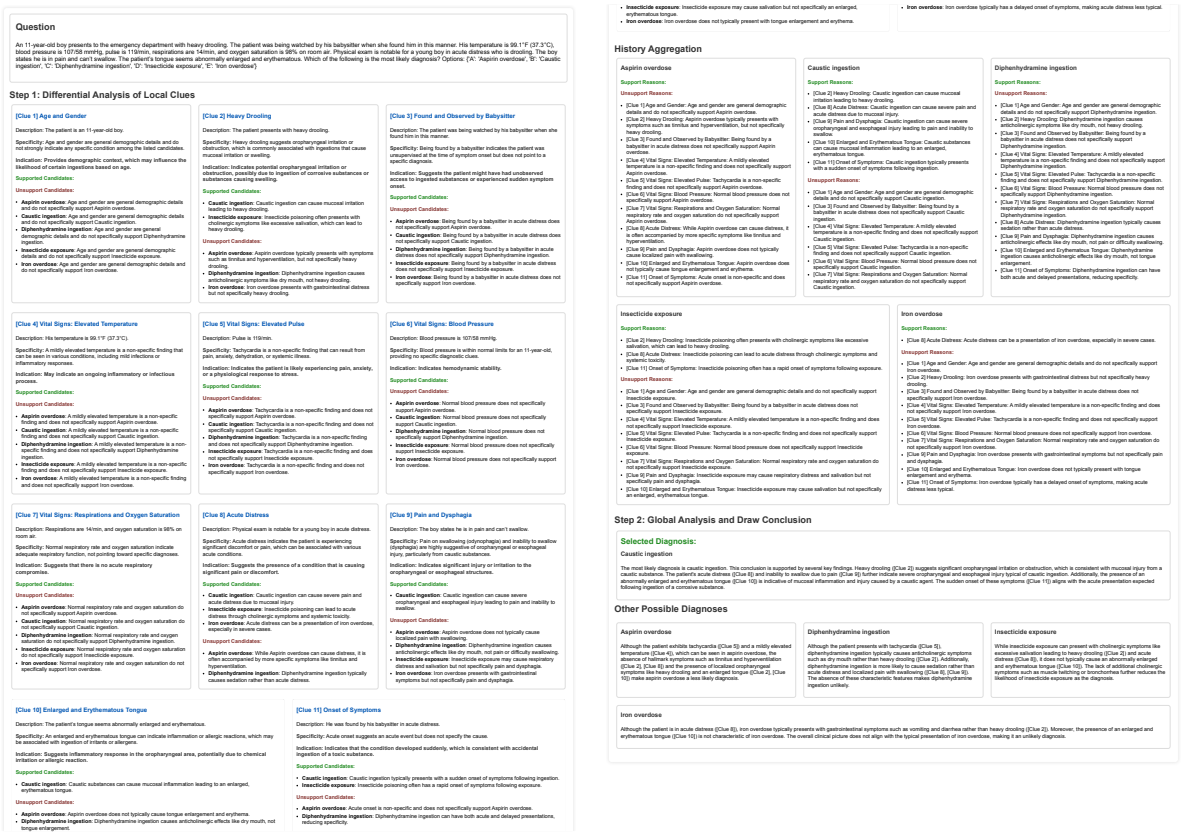


Figure 6: A sample of our **DDxReasoning** Dataset (Best viewed when zoomed in).

- A reasoning text $reason_j$ explaining the stance

The teacher evaluates these responses against the teaching reference $l_i \in \mathbf{L}$, assessing both stance correctness (v_s^j) and reasoning quality (v_r^j).

Generation Process

For a given clinical case with question q, candidates set $\mathbf{A}$, and a specific clue c_i, our dialogue generation follows a structured process that simulates a teacher-student dialogue analyzing how the clue supports or cannot support each candidate.

First, the teacher simulator combines q, c_i, and $\mathbf{A}$ to generate the initial teaching inquiry t_0^i.

Then we enter the student-teacher dialogue loop. At the beginning of each loop k, the student obtains unanswered candidates from the system state:

$$A_{remaining} = \mathbf{A} \setminus \bigcup_{k=1}^{t} J_k \tag{2}$$

where J_k represents the set of candidates discussed in turn k.

The student's response state is simulated by randomly choosing from five possible scenarios for each candidate:

1. Correct stance with correct reasoning
2. Correct stance with flawed reasoning
3. Incorrect stance with incorrect reasoning
4. Random guess (stance correct/incorrect, with planned reason as None)
5. Complete uncertainty

For each candidate to be addressed, we randomly select one response state to form the student's response plan. We then use GPT4o to generate simulated reasoning based on the response state, grounded in the teaching reference l_i.

For example, consider a scenario where the student's response plan includes: - For candidate a_1: correct stance and correct reasoning (scenario 1) - For candidate a_2: correct stance but flawed reasoning (scenario 2).

In this case, GPT4o will: (1) Generate $reason_1$ that closely aligns with the correct reasoning from l_i. (2) Generate $reason_2$ that intentionally deviates from the correct reasoning in l_i.

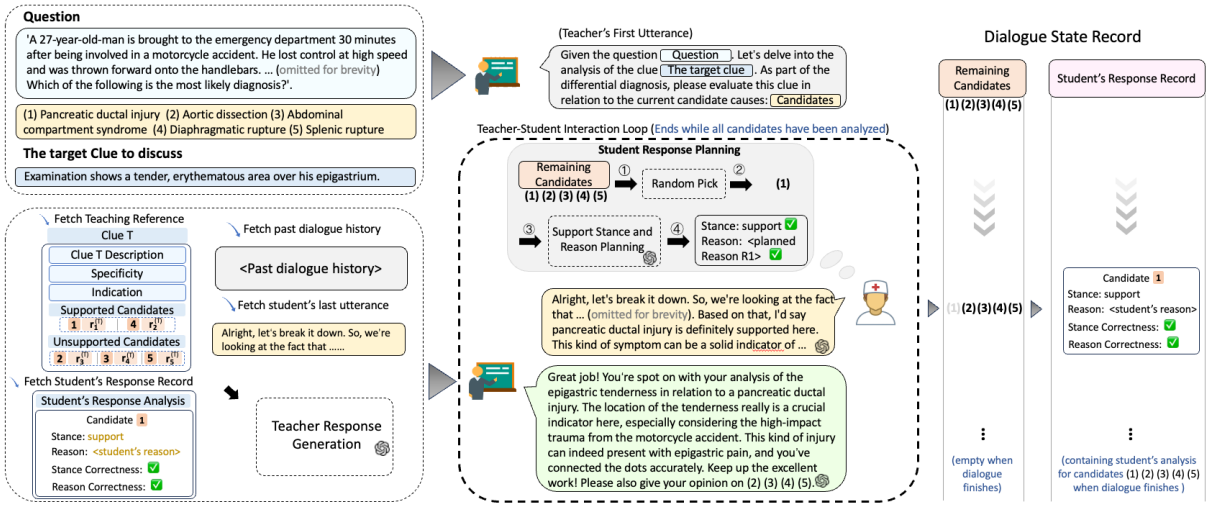


Figure 7: The Generation pipeline of Local Analysis Dialogues

This process yields a structured response evaluation:

$$eval_k^i = \{(a_j, stance_j, reason_j, v_s^j, v_r^j) \mid j \in J_k\} \quad (3)$$

We then prompt GPT4o to generate a complete student response u_k^i following this structured plan. The generated response is expected to closely align with the evaluation structure, allowing us to use the components as ground truth labels for the dialogue output.

Teacher Response Framework

The teacher's response t_k^i includes two components: evaluation ($eval_k^i$) and feedback ($feed_k^i$). The teacher's feedback strategy F is determined by comparing student responses against the teaching reference l_i:

$$F(a_j) = \begin{cases} \text{Confirm} & \text{if } v_s^j \wedge v_r^j \\ \text{Partial} & \text{if } v_s^j \wedge \neg v_r^j \\ \text{Correct} & \text{if } \neg v_s^j \end{cases} \quad (4)$$

In real scenarios, we expect the LLM to generate $eval_k^i$ and $feed_k^i$ end-to-end, with $feed_k^i$ being conditioned on $eval_k^i$. The feedback component is generated based on the known evaluation of the student's response to ensure appropriate adaptive guidance.

The dialogue D_{local}^i progresses through turns until all candidates are analyzed. Each turn consists of:

(1) Teacher requesting analysis of unaddressed candidates (2) Student providing stance and reasoning for selected candidates (3) Teacher evaluating responses and providing appropriate feedback (4) System updating J_k with newly discussed candidates

This structured approach ensures systematic coverage of clinical reasoning while maintaining natural dialogue flow. The final dialogue dataset captures both correct and incorrect reasoning patterns, enabling the LLM to learn appropriate response evaluation and feedback generation strategies.

B.2 Global Synthesis Dialogue Generation

After completing all local clue analyses $\{D_{local}^i\}_{i=1}^n$, we generate the global diagnostic synthesis dialogue D_{global}. For each case, we generate two variants of D_{global}: one with correct diagnosis selection ($v_d = 1$) and another with incorrect selection ($v_d = 0$). This approach enables the model to learn appropriate feedback strategies for both successful and unsuccessful diagnostic reasoning.

The dialogue generation process consists of three main components:

Teacher's First Utterance: The LLM generates the initial prompt by combining:

- A summary of all analyzed clinical clues $\mathbf{C}$
- A request for final diagnostic decision based on $\mathbf{L}$

Student Response Generation: For each dialogue variant, we simulate the student's response p_1 through:

1. Diagnosis selection:

$$\hat{a}_g = \begin{cases} \hat{a} & \text{if } v_d{=}{=}1 \\ a_j \in \mathbf{A} \setminus \{\hat{a}\} & \text{if } v_d{=}{=}0 \end{cases} \quad (5)$$

PROMPT L

Question: {question}
Please extract the clues from the question for differential analysis. For each clue, analyze it systematically using the following structure:
1. Clue Name: Clearly label the clue using concise and descriptive language.
2. Description: Provide a detailed description of the clue, retaining as much of the original wording as possible.
3. Specificity Description: Assess whether the clue is diagnostically specific and explain its relevance to particular conditions (e.g., "The presence of xxx is highly specific for the xxx disease"). 4. Indication: Analyze what the clue suggests based on medical knowledge and its implications (e.g., "This finding could suggest xxx because of xxx").
5. Supported Candidates: Identify which candidate diagnoses the clue could support and explain why (e.g., "This clue supports xxx because xxx"). If there are no supported diagnoses, leave this empty.
6. Unsupported Candidates: Identify which candidate diagnoses the clue could not support and explain why (e.g., "This clue refutes xxx because xxx" or "This clue is unrelated to xxx because xxx"). If all the candidates are supported, leave this empty.
Note: Please always ensure that the sum of supported and unsupported diagnoses is equal to the number of all candidates.
Here is a JSON format of the expected output: {Local Clue Analysis Json}.

PROMPT G

{question}
Below is a clue-by-clue analysis for each candidate in the options: Clue_Differential_Analysis
Please generate a final decision with a detailed differential diagnosis explanation based on the given clues. Your task is to determine the most likely diagnosis and explain why each candidate is selected or rejected.

Requirements:
Citation Format: Cite specific clues using the format [Clue X] (e.g., [Clue 1], [Clue 3]). Avoid combining multiple clue citations (e.g., [Clue 1,2] or [Clue 3,5,10]). When discussing multiple related clues, clearly state each finding with its corresponding clue index.

Example citation: "Although scleroderma is more common in middle-aged women [Clue 1] and presents with fatigue [Clue 2] and difficulty swallowing [Clue 4], it typically does not present with upper esophageal webs [Clue 9], koilonychia [Clue 7], or iron deficiency anemia [Clue 10]. Additionally, the presence of neck pain [Clue 3] and leukocytosis [Clue 11] is inconsistent with typical scleroderma features, making it a less likely diagnosis."

Evidence Evaluation: Prioritize strong, decisive clues over weaker or peripheral evidence. Focus on clues that are sufficient to support diagnostic decisions. Avoid overemphasis of non-specific or irrelevant findings.

Rejected Candidates: Provide clear, evidence-based explanations for why each rejected candidate is less likely. Base rejections on specific clue citations. Explain any missing key features expected for these diagnoses.

Balanced Analysis: Use concessive reasoning (although-style) when analyzing candidates with mixed evidence. Explicitly weigh supporting versus conflicting evidence. Clearly justify final decisions when evidence is mixed.

Example reasoning: "Although Candidate A shows feature X [Clue 3], which supports the diagnosis, the absence of feature Y [Clue 5], a hallmark finding, makes it less likely."
Important Note: Your analysis will serve as a teaching resource for identifying strong and sufficient evidence in diagnostic reasoning. Therefore: Ensure all cited clues are accurate and directly relevant. Select only strong, definitive clues to support or reject diagnoses. Exclude weak or ambiguous evidence from core supporting arguments.
Here is a JSON format of the expected output: {Global Analysis Json}

Table 4: The $PromptL$ and $PromptG$ we used to utilized OPENAI o1 to extract a initial local clues analysis and global analysis when we curated the DDxReasoning dataset.

Split	Num	Num of Clues	Max Tokens	Mean Tokens
Train	755	7047	12325	5223
Test	178	1643	9353	5043

Table 5: The statistics of the DDxReasoning dataset are summarized as follows. The "Num of Clues" refers to the total number of clues analyzed during the dataset's creation. "Max Tokens" and "Mean Tokens" indicate the maximum and average token counts within the dataset, respectively.

where $\hat{a}$ is the correct diagnosis from $\mathbf{G}$.

2. Reasoning generation: GPT4o generates student-like reasoning that:
   - References relevant analyses from $\mathbf{L}$
   - Employs natural language patterns
   - Demonstrates diagnostic conviction while maintaining educational tone

Teacher's Response Generation: The teacher's response $g_1 = (eval_g, feed_g)$ is generated conditionally:

For correct diagnosis ($v_d = 1$):
- Confirmation of correct selection
- Validation of key reasoning points
- Supplementary supporting evidence from $\mathbf{G}$
- Integration of clinical elements from $\mathbf{L}$

For incorrect diagnosis ($v_d = 0$):
- Analysis of reasoning flaws
- Presentation of correct diagnosis $\hat{a}$
- Comprehensive justification based on $\mathbf{G}$
- References to relevant evidence from $\mathbf{L}$

The resulting global synthesis dialogue follows this structure:

$$D_{global} = \{(\mathbf{L}), (p_1, g_1)\} \quad (6)$$

This structured approach generates diverse training examples for both successful and unsuccessful diagnostic scenarios, enabling the LLM to learn appropriate response strategies while maintaining educational effectiveness.

B.3 The Statistics of the Simulated Dialogues

Following the dataset design of Task-Oriented Dialogue systems (Valizadeh and Parde, 2022; Wei et al., 2018), we convert our generated dialogues into Supervised Finetuning datasets (SFT). The Statistics of the dataset is presented in Table 6

Split	Num	Mean Total Tokens	Mean Gen Tokens
Local Clue Analysis Dialogue			
Train	16132	2064	418
Test	3760	2507	641
Global Synthesis Dialogue			
Train	1506	6764	428
Test	356	6574	434

Table 6: The statistics of the simulated tutoring dialogue datasets, Local Clue Analysis Dialogue and Global Synthesis Dialogue. Here the Mean Gen Tokens means the average number of tokens in generated responses, while Mean Total Tokens represents the average length of the entire dialogue including both input and output.

C More details of the Evaluation Protocol

This section provides detailed explanations of our evaluation metrics for assessing the three types of fine-tuned models: Teaching Reference Generator (f_1), Local Analysis Dialogue Tutor (f_2), and Global Synthesis Dialogue Tutor (f_3).

C.1 Content Alignment Strategy

When evaluating structured outputs from LLMs against ground truth references, a key challenge is that the order of analyses (e.g., candidate diagnoses, clinical clues) may differ between the prediction and reference, while the content remains semantically equivalent. To address this, we employ the Hungarian algorithm for optimal content matching before computing evaluation metrics.

Specifically, for any two sets of content that need alignment (e.g., predicted vs. reference candidates), we:

1. Construct a cost matrix M where $M_{ij} = 1 - sim(p_i, r_j)$
2. $sim(p_i, r_j)$ computes the textual similarity ratio between prediction p_i and reference r_j
3. Apply the Hungarian algorithm to find the optimal matching that minimizes total matching cost

This matching process ensures accurate evaluation by properly aligning corresponding contents regardless of their order in the structured output. The matched pairs are then used for computing various evaluation metrics detailed below.

C.2 Teaching Reference Generator (f_1) Evaluation

The evaluation of f_1 models focuses on two main aspects:

Local Clue Analysis Capability:

- **Clue Decomposition Quality:**
  - $Bert_C(\hat{c}, c)$: BertScore measuring semantic similarity between predicted clues $\hat{c}$ and ground-truth clues c
  - $Meteor_C(\hat{c}, c)$: Meteor score assessing textual alignment between predicted and ground-truth clues
- **Clue-Candidate Relationship Analysis:**
  - Acc_{PN}: Accuracy of predicted support/non-support stances

$$Acc_{PN} = \frac{|\{(c_i, a_j)|\hat{s}_{ij} = s_{ij}\}|}{|\{(c_i, a_j)\}|} \quad (7)$$

    where $\hat{s}_{ij}$ and s_{ij} are predicted and ground-truth stances
  - $Bert_{PN}(\hat{r}, r)$ and $Meteor_{PN}(\hat{r}, r)$: Quality metrics comparing predicted reasoning explanations $\hat{r}$ with ground-truth reasoning r in $\mathbf{P}_i$ and $\mathbf{N}_i$
- **Clue Property Analysis:**
  - $Bert_{Spec}(\hat{spec}, spec)$ and $Meteor_{Spec}(\hat{spec}, spec)$: Quality metrics comparing predicted specificity assessment $\hat{spec}_i$ with ground truth $spec_i$
  - $Bert_{Ind}(\hat{ind}, ind)$ and $Meteor_{Ind}(\hat{ind}, ind)$: Quality metrics comparing predicted indication description $\hat{ind}_i$ with ground truth ind_i

Global Synthesis Capability:

- **Diagnostic Accuracy:**

$$Acc_{\hat{a}} = \frac{|\{\mathcal{I}|\hat{a} = a^*\}|}{|\{\mathcal{I}\}|} \quad (8)$$

  where a^* is the ground-truth diagnosis
- **Explanation Quality:**
  - $Bert_E(\hat{E}, E)$ and $Meteor_E(\hat{E}, E)$: Quality metrics comparing predicted diagnosis explanations $\hat{E}_j(\mathbf{L})$ with ground truth $E_j(\mathbf{L})$

C.3 Local Analysis Dialogue Tutor (f_2) Evaluation

We evaluate f_2 models on three aspects:
Student Response Extraction:

- **Candidate Coverage:**

$$IOU_a = \frac{|\hat{J}_k \cap J_k|}{|\hat{J}_k \cup J_k|} \tag{9}$$

  where $\hat{J}_k$ and J_k are predicted and ground-truth discussed candidate sets

- **Stance Extraction:**

$$Acc_{stance} = \frac{|\{j \in J_k | \hat{stance}_j = stance_j\}|}{|J_k|} \tag{10}$$

- **Reasoning Extraction:** $Bert_{reason}(\hat{r}, r)$ and $Meteor_{reason}(\hat{r}, r)$ measuring quality between predicted reasoning $\hat{r}$ and ground-truth reasoning r

Response Judgment:

- **Stance Judgment:**

$$Acc_{vs} = \frac{|\{j \in J_k | \hat{v_s^j} = v_s^j\}|}{|J_k|} \tag{11}$$

  where $\hat{v_s^j}$ and v_s^j are predicted and ground-truth judgments on student's stance

- **Reasoning Judgment:**

$$Acc_{vr} = \frac{|\{j \in J_k | \hat{v_r^j} = v_r^j\}|}{|J_k|} \tag{12}$$

  where $\hat{v_r^j}$ and v_r^j are predicted and ground-truth judgments on student's reasoning

Feedback Generation: $Bert_{feed}(\hat{f}, f)$ and $Meteor_{feed}(\hat{f}, f)$ measuring quality between predicted feedback $\hat{feed}_k^i$ and ground-truth feedback $feed_k^i$

C.4 Global Synthesis Dialogue Tutor (f_3) Evaluation

The evaluation of f_3 models follows similar principles:
Diagnosis Extraction:

$$Acc_{\hat{a}g} = \frac{|\{\mathcal{I} | \hat{a}_g = a_g\}|}{|\{\mathcal{I}\}|} \tag{13}$$

where $\hat{a}_g$ and a_g are predicted and ground-truth student's diagnosis selections
Diagnosis Judgment:

$$Acc_{vd} = \frac{|\{\mathcal{I} | \hat{v_d} = v_d\}|}{|\{\mathcal{I}\}|} \tag{14}$$

where $\hat{v_d}$ and v_d are predicted and ground-truth judgments on student's final diagnosis
Global Feedback Quality: $Bert_{feed\_g}(\hat{f}_g, f_g)$ and $Meteor_{feed\_g}(\hat{f}_g, f_g)$ measuring quality between predicted global feedback $\hat{feed}_g$ and ground-truth feedback $feed_g$

For f_1 model's BertScore calculation, we use RoBERTa-large (Liu, 2019) as the base model. For f_2 and f_3 models, we employ Longformer (Beltagy et al., 2020) as the base model for BertScore computation. All scores are computed by comparing the generated content against expert-annotated ground truth references.

D More Experimental Studies

D.1 Finetuning models to perform single clue analysis.

Model	Acc_{PN}	$Bert_{PN}$	$Meteor_{PN}$	$Bert_{Spec}$	$Meteor_{Spec}$	$Bert_{Ind}$	$Meteor_{Ind}$
Qwen2.5-3b	0.700	0.636	0.279	0.899	0.353	0.896	0.281
Llama3.2-3b	0.722	0.661	0.329	0.906	0.406	0.900	0.314
Qwen2.5-7b	0.747	0.68	0.311	0.903	0.376	0.897	0.296
Llama3.1-8b	0.754	0.692	0.351	0.908	0.426	0.902	0.328
Med42-8b	0.759	0.698	0.363	0.909	0.428	0.903	0.33
Qwen2.5-14b	0.753	0.684	0.313	0.905	0.388	0.899	0.305
Phi4-14b	**0.785**	**0.724**	**0.391**	**0.916**	**0.467**	**0.907**	**0.362**
Qwen2.5-32b	0.768	0.702	0.341	0.909	0.410	0.902	0.321

Table 7: Experiental results of finetuned models to perform single clue analysis.

In this section, we investigate models' capability to perform focused analysis on individual clinical clues. For this purpose, we further fine-tuned a specialized model f_4: $\hat{l}_i = f_4(c_i)$, where f_4 takes a single clue description c_i as input and yields its comprehensive analysis l_i. This capability is valuable for "just-in-time" clinical teaching scenarios - when students encounter unfamiliar symptoms during case discussions, instructors can instantly query f_4 to generate focused mini-lectures about specific clinical manifestations, maintaining the natural flow of case-based discussions while addressing knowledge gaps in real-time.

As shown in Table 7, models demonstrate strong performance in analyzing individual clinical clues, particularly in stance identification (Acc_{PN}) and specificity description ($Bert_{Spec}$). The Phi4-14B model achieves the best overall performance with notably high scores in stance accuracy (0.724) and

clinical indication analysis ($Bert_{Ind}$ = 0.907). Interestingly, while larger models generally perform better, the improvement margin narrows in single-clue analysis tasks compared to the comprehensive teaching reference generation task discussed earlier. This suggests that accurate analysis of individual clinical manifestations may have a lower parameter requirement threshold than integrating multiple pieces of evidence for final diagnosis.

This observation complements our previous findings regarding teaching reference generators (f_1), where models showed stronger scaling effects in multi-clue reasoning tasks. The relatively stable performance across model sizes in single-clue analysis further justifies the practical value of f_4 in educational settings, as even smaller models can provide reliable focused analysis for immediate teaching needs.

D.2 Dialogue Tutoring without Teaching Reference

We also study the effect of removing Teaching Reference for f_2 and f_3 models, the result are presented in Table 8.

This study reveals nuanced impacts of removing teaching references across different components of the DDxTutor framework. For local analysis dialogue tutors (f_2), the absence of teaching references leads to a notable decline in student judgment capabilities, as evidenced by decreased Acc_{vs} and Acc_{vr} scores. This degradation clearly demonstrates the crucial role of teaching references in supporting effective dialogue-based instruction at the individual clue analysis stage.

Interestingly, when removing the global teaching reference $\mathbf{G}$ from the global dialogue tutor (f_3), the impact on its judgment capability remains relatively minimal. We hypothesize that this robustness stems from the presence of complete Local Clue Analysis $\mathbf{L}$ in f_3's context, which continues to serve as a structured foundation for information integration. This suggests that f_3 can inherently derive accurate diagnostic reasoning by leveraging precise local analyses, even without explicit global teaching guidance.

This observation underscores an insight about our framework: the accuracy of initial local clue analysis significantly influences the quality of subsequent global synthesis. The relative stability of f_3's performance, contingent on accurate $\mathbf{L}$, validates our framework's emphasis on building strong foundations through precise local analysis before proceeding to global integration.

D.3 qualitative results

We present some qualitative results for Teaching Reference Generators f_1, Local Analysis Dialogue Tutor f_2, and Global Synthesis Dialogue Tutor f_3.

$f1$ models: Two Teaching Reference result from GPT-4o (2 shot) results are presented in Figure 8 and Figure 9. At the same time, the comparsion results on the same cases, which are generated by finetuned Phi4-14b model, are presented in Figure 10 and 11.

f_2 and f_3 models: A local clue analysis dialogue case and a global synthesize dialogue case, with results from Llama3.2-3b (finetuned), Phi4-14b (finetuned) and GPT-4o (2 shot In-Context Learning), are presented in Table 9 and 10.

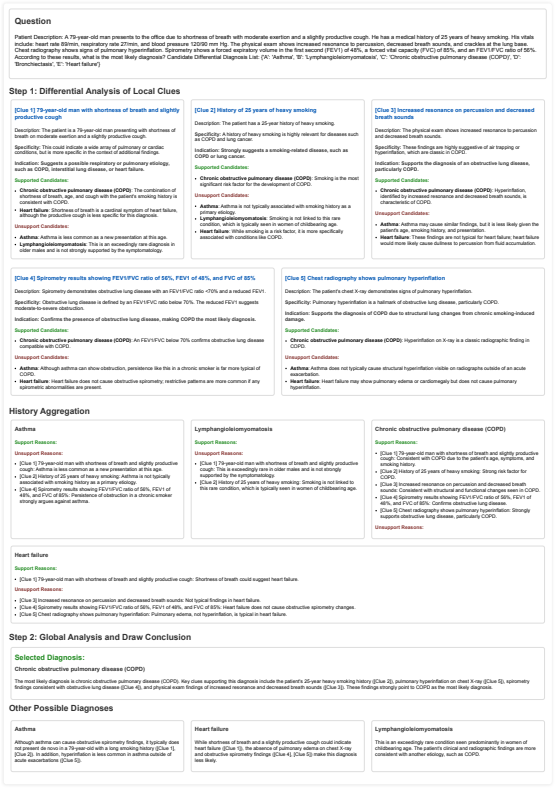


Figure 8: Teaching Reference Generation Case #1 (Model: GPT-4o 2-shot In-context Learning)

E Implementation and writing

Due to computational constraints, we report results from a single training run, which we believe demonstrates the model's potential while acknowledging the need for future studies on performance stability. We mildly used AI assistants to polish paper writing.

Model	IOU$_a$	Acc$_{stance}$	Bert$_{reason}$	Acc$_{vs}$	Acc$_{vr}$	Bert$_{feed}$	Meteor$_{feed}$	Acc$_{\hat{a}}$	Acc$_{vd}$	Bert$_{feed_g}$	Meteor$_{feed_g}$
Qwen2.5-3b	0.995	0.989	0.947	0.828	0.804	0.891	0.397	0.994	0.963	0.907	0.482
Llama3.2-3b	0.996	0.990	0.951	0.842	0.819	0.896	0.430	0.989	0.949	0.908	0.467
Qwen2.5-7b	0.994	0.990	0.947	0.844	0.812	0.892	0.408	0.889	0.969	0.906	0.477
Llama3.1-8b	0.997	0.992	0.953	0.863	0.835	0.898	0.446	**0.997**	0.975	0.909	0.473
Meditron-8b	0.996	0.991	0.953	0.853	0.830	0.898	0.442	**0.997**	0.972	0.909	0.475
Med42-8b	0.997	0.993	0.955	0.868	0.838	0.899	0.456	0.989	0.966	0.909	0.475
Qwen2.5-14b	0.982	0.991	0.948	0.866	0.836	0.894	0.421	0.966	0.980	0.909	0.490
Phi4-14b	**0.997**	**0.993**	**0.955**	**0.888**	**0.857**	**0.901**	**0.468**	**0.997**	**0.989**	**0.912**	0.484
Qwen2.5-32b	0.996	0.993	0.950	0.878	0.846	0.897	0.437	**0.997**	0.980	0.911	**0.498**

Table 8: Performance of Local Dialogue Mode f_2 and Global Dialogue Model f_3 after removing Teaching Reference L and G on the simulated dialogue datasets.

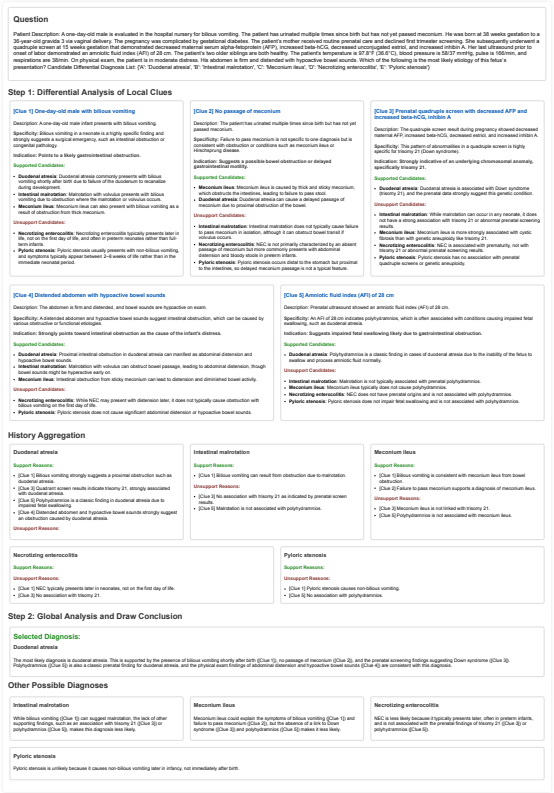
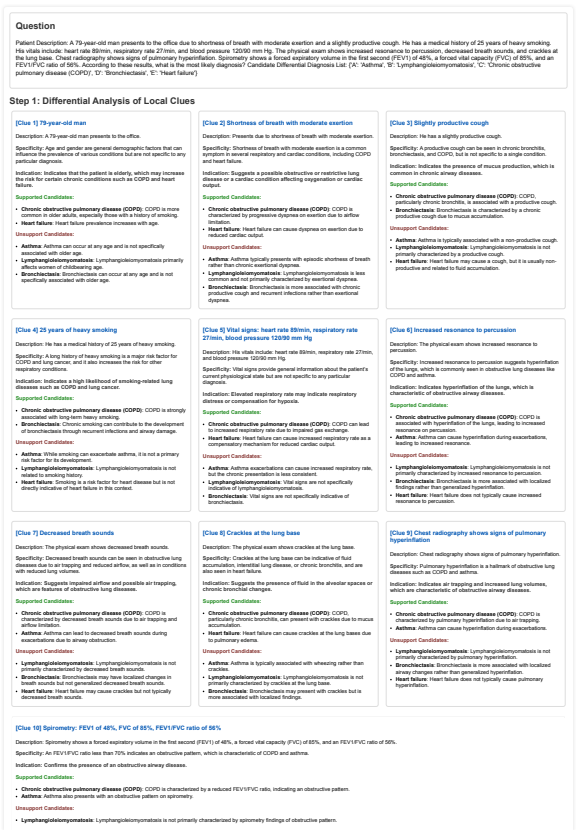


Figure 9: Teaching Reference Generation Case #2 (Model: GPT-4o 2-shot In-context Learning)

F Potential Risks

First, despite high performance metrics, no AI system can guarantee perfect accuracy in medical reasoning. Students might occasionally be exposed to incorrect or incomplete teaching references, which could affect their learning trajectory. To mitigate this risk, we recommend that DDxTutor be used under proper supervision from experienced medical educators who can identify and correct any potential errors. Second, there is a risk that students might become overly reliant on AI-based tutoring systems, potentially developing diagnostic habits that prioritize pattern matching over critical thinking. Medical educators should emphasize that DDxTutor is a supplementary tool to enhance learning, not a replacement for developing independent clinical reasoning skills.



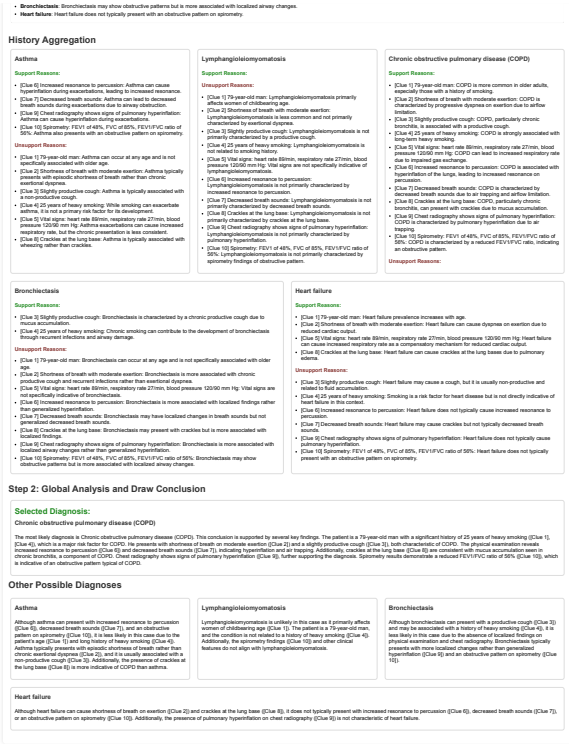
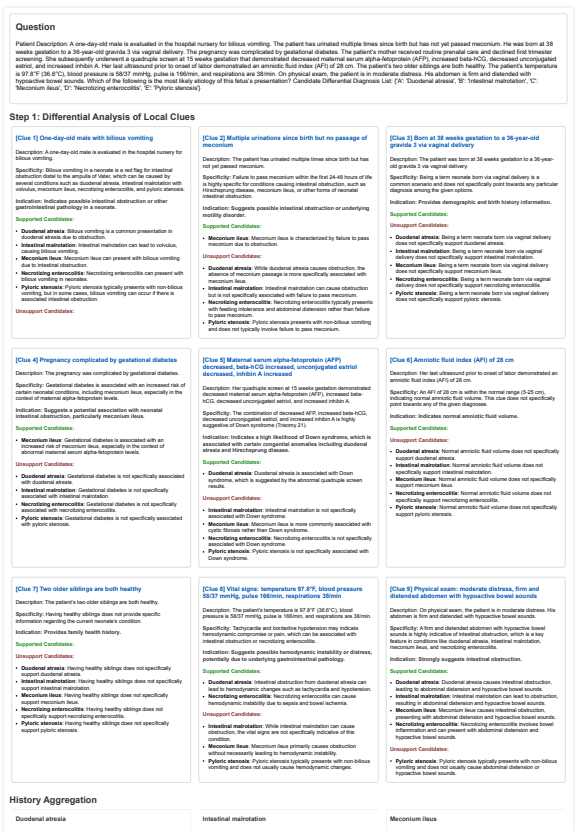


Figure 10: Teaching Reference Generation Case #1 (Model: Phi4-14B finetuned)



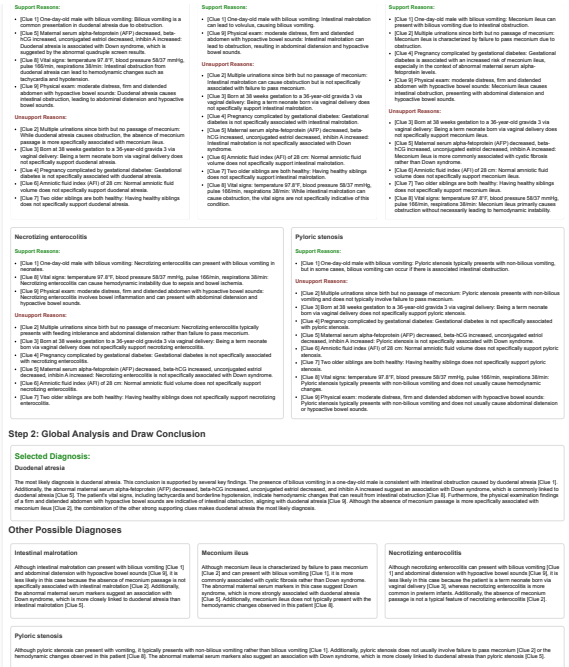


Figure 11: Teaching Reference Generation Case #2 (Model: Phi4-14B finetuned)

TEACHER
Given the question: 'A 44-year-old man presents to the family medicine clinic for some small bumps on his left thigh. The lesions are non-pruritic and have been present for the last 3 weeks. He reports feeling fatigued and malaise for the past few months. The patient has no known medical problems and takes no medications. He smokes one pack of cigarettes per day and uses intravenous street drugs. His heart rate is 82/min, the respiratory rate is 14/min, the temperature is 36.7°C (98.1°F), and the blood pressure is 126/80 mm Hg. Auscultation of the heart is without murmurs. Lungs are clear to auscultation bilaterally. Three 2-3 mm, dome-shaped, hardened papules are noted on the left thigh. Central umbilication can be observed in each papule. There is a non-tender cervical lymphadenopathy present. Which of the following is the most likely diagnosis?'.Let's delve into the analysis of the clue: 'Three 2-3 mm, dome-shaped, hardened papules are noted on the left thigh. Central umbilication can be observed in each papule.'. As part of the differential diagnosis, please evaluate this clue in relation to the current candidate causes: 'A: Kaposi sarcoma, B: Bed bug bite, C: Molluscum contagiosum, D: Atopic dermatitis, E: Acute urticaria', determining whether it supports or cannot support each of them, and provide a clear explanation of your reasoning."
STUDENT
Hmm, okay, so starting with Kaposi sarcoma... honestly, I'm not sure. The description of the bumps and the other clues don't immediately scream Kaposi sarcoma to me, but I kind of feel like I don't know enough about how it might show up to totally rule it out either. So, yeah, I'd have to say "I don't know" for now.bed bug bites... again, I can't say I'm confident here either. The timing and the appearance of the bumps don't immediately make me think of bed bugs, but at the same time, I feel like I don't know their exact presentation well enough to make a clear call. So I'd stick with "I don't know" on this one too.'
LOCAL DIALOGUE TUTOR 1 (LLAMA3.2 3B FINETUNED)
[Student's Response Extraction and Analysis] ["candidate_name": "(A) Kaposi sarcoma", "student's support stance": "N.A.", "student's reasoning": "I don't know.", "student's support stance correctness": false, "student's reasoning correctness": false, "candidate_name": "(B) Bed bug bite", "student's support stance": "N.A.", "student's reasoning": "I don't know.", "student's support stance correctness": false, "student's reasoning correctness": false] [/Student's Response Extraction and Analysis] [Teacher's Response] Sure, let's go through your responses together and see how they match up with the ground truth. Here's my feedback: — **Candidate: Kaposi sarcoma** - **Your stance:** You said you didn't know whether the clue supports or opposes Kaposi sarcoma. - **Ground truth:** The clue does not support Kaposi sarcoma. - **Feedback:** That's totally okay—sometimes it's better to admit when you're not sure than to guess incorrectly. That said, here's the deal: the clue doesn't support Kaposi sarcoma because the specific lesion characteristics described (dome-shaped, hardened papules with central umbilication) aren't consistent with how Kaposi sarcoma presents. Kaposi sarcoma typically shows up as violaceous, irregular plaques or nodules rather than the dome-shaped, umbilicated bumps we're seeing here. So, even though you didn't take a stance, you're right to feel uncertain because this clue doesn't fit Kaposi sarcoma. — **Candidate: Bed bug bite** - **Your stance:** You also said you didn't know whether the clue supports or opposes bed bug bites. - **Ground truth:** The clue does not support bed bug bites. - **Feedback:** Again, it's totally fine to feel unsure here. But just to clarify, the specific lesion characteristics described (dome-shaped, hardened papules with central umbilication) aren't consistent with bed bug bites. Bed bug bites typically appear as erythematous, itchy papules or bumps, often in a linear or clustered pattern, rather than the dome-shaped, umbilicated bumps we're seeing here. So, this clue doesn't fit bed bug bites. — Final Thoughts: You're doing great—recognizing when you're unsure is a key part of clinical reasoning, and it's better to admit uncertainty than to guess incorrectly. Keep practicing, and these patterns will become more intuitive over time! [/Teacher's Response]
LOCAL DIALOGUE TUTOR 2 (PHI4-14B FINETUNED)
[Student's Response Extraction and Analysis] ["candidate_name": "(A) Kaposi sarcoma", "student's support stance": "N.A.", "student's reasoning": "I don't know.", "student's support stance correctness": false, "student's reasoning correctness": false, "candidate_name": "(B) Bed bug bite", "student's support stance": "N.A.", "student's reasoning": "I don't know.", "student's support stance correctness": false, "student's reasoning correctness": false] [/Student's Response Extraction and Analysis] [Teacher's Response] Alright, let's go through your responses one by one and see how they stack up. I'll give you feedback based on what you said and the ground truth reasoning. — **1. Kaposi sarcoma** You said: *"I'm not sure. The description of the bumps and the other clues don't immediately scream Kaposi sarcoma to me, but I kind of feel like I don't know enough about how it might show up to totally rule it out either. So, yeah, I'd have to say 'I don't know' for now."* **Feedback:** No worries about not being sure here—this is a tricky one! Let me clarify: the clue about the "3 2-3 mm, dome-shaped, hardened papules with central umbilication" does **not** support Kaposi sarcoma. Here's why: Kaposi sarcoma lesions are typically **flat or slightly elevated**, **purplish or violaceous**, and **not umbilicated**. So, this clue doesn't fit with Kaposi sarcoma at all. You're absolutely right to feel unsure, and it's okay to admit when you don't know something. The important thing is to keep learning, and now you know how to rule this one out based on the lesion description! — **2. Bed bug bite** You said: *"Again, I can't say I'm confident here either. The timing and the appearance of the bumps don't immediately make me think of bed bugs, but at the same time, I feel like I don't know their exact presentation well enough to make a clear call. So I'd stick with 'I don't know' on this one too."* **Feedback:** You're on the right track here! The clue about the "3 2-3 mm, dome-shaped, hardened papules with central umbilication" does **not** support bed bug bites. Bed bug bites typically present as **red, itchy welts** or **swollen bumps**, and they don't have the **dome shape or central umbilication** described in this case. So, while you didn't take a firm stance, your instinct to question this as a possibility is spot on. And now you know that bed bug bites don't match the lesion description here. Great job! — Final Thoughts: You're doing a great job working through these! It's okay to feel unsure sometimes—clinical reasoning takes practice, and you're building the skills to analyze clues more confidently. Keep it up! [/Teacher's Response]
LOCAL DIALOGUE TUTOR 3 (GPT4O (2-SHOT))
[Student's Response Extraction and Analysis] "'json ["candidate_name": "Kaposi sarcoma", "student's support stance": "N.A.", "student's reasoning": "The description of the bumps and the other clues don't immediately scream Kaposi sarcoma to me, but I kind of feel like I don't know enough about how it might show up to totally rule it out either.", "student's support stance correctness": false, "student's reasoning correctness": false , "candidate_name": "Bed bug bite", "student's support stance": "N.A.", "student's reasoning": "The timing and the appearance of the bumps don't immediately make me think of bed bugs, but at the same time, I feel like I don't know their exact presentation well enough to make a clear call.", "student's support stance correctness": false, "student's reasoning correctness": false] "' [/Student's Response Extraction and Analysis] [Teacher's Response] Let's review these candidates you've mentioned. Kaposi Sarcoma: You expressed uncertainty about this condition and acknowledged a lack of knowledge to either support or rule it out. The truth is, the appearance of the lesions you've described—dome-shaped papules with central umbilication—are not characteristic of Kaposi sarcoma. Kaposi sarcoma typically presents as purple, red, or brown lesions, often in plaques or patches, not dome-shaped papules with central umbilication. So, it's understandable to be unsure, but knowing the hallmark differences can guide you more accurately. Bed Bug Bite: Similar to Kaposi Sarcoma, you felt unsure about bed bug bites due to a lack of detailed knowledge on presentation. Bed bug bites often appear as small, itchy wheals and usually don't have a central umbilication, which is a characteristic clue for Molluscum contagiosum, not bed bug bites. This understanding could help you differentiate in practice, even when feeling uncertain. Considering you've now evaluated Kaposi Sarcoma and Bed Bug Bite, let's move forward. Could you analyze the same clue in relation to the remaining candidates: "Molluscum contagiosum," "Atopic dermatitis," and "Acute urticaria"?

Table 9: A Local Analysis Tutoring Dialogue Case.

QUESTION
An 8-year-old girl presents to her pediatrician with intensely pruritic lesions over her buttocks, as shown. These lesions occur intermittently but have worsened over time. Her medical history is remarkable for iron deficiency, for which she is on ferrous sulfate. The patient has also experienced intermittent episodes of mild diarrhea, previously diagnosed as lactose intolerance. Her height is at the 30th percentile and weight is at the 25th percentile for age and sex. An immunoglobulin A (IgA) tissue transglutaminase titer is 5 times the upper limit of normal. Which of the following is the most likely cause of this patient's condition? 'A': 'Crohn's disease', 'B': 'Giardiasis', 'C': 'Celiac disease', 'D': 'Common variable immune deficiency', 'E': 'Henoch-Schonlein purpura' TEACHER Great job! You've done well analyzing all the local clues. Let's look at what we've discussed: <An Aggregation of Local Clue Analysis> Now, based on these clues, which candidate do you think is the most likely one?
STUDENT
Based on what we discussed, the clues seem to really point towards it. First, there are the intensely pruritic lesions over her buttocks that could be dermatitis herpetiformis, which is like, strongly linked to celiac disease. Then, there's the iron deficiency; since celiac disease leads to malabsorption, it makes sense it would cause iron deficiency anemia. Also, the patient has intermittent mild diarrhea and growth at lower percentiles, which can also be explained by the malabsorption happening in celiac disease. But I think what really seals it for me is the elevated IgA tissue transglutaminase titer, which is pretty specific for celiac disease. From what I understand, that clue is pretty significant. So yeah, I think Celiac disease fits best here!
GLOBAL DIALOGUE TUTOR 1 (LLAMA3.2 3B FINETUNED)
[Students' Response Analysis]'s Selection: (C) Celiac disease; Selection Correctness: True[/Students' Response Analysis] [Teacher's Response]Great job! You've correctly identified celiac disease as the most likely cause of this patient's condition. Let's go over why this diagnosis fits so well with the clinical clues we have.Firstly, the intensely pruritic lesions over the buttocks are a significant clue. These are characteristic of dermatitis herpetiformis, which is closely linked to celiac disease. This is a key point that helps us narrow down the diagnosis.Additionally, the elevated IgA tissue transglutaminase titer is a very strong indicator of celiac disease. This test is highly specific for the condition, and its elevation is a crucial finding in this case.The history of iron deficiency anemia is another important clue. Celiac disease leads to malabsorption, which can cause iron deficiency. This aligns perfectly with the patient's medical history.Moreover, the intermittent mild diarrhea and the growth percentiles at the 30th percentile for height and 25th percentile for weight further support the diagnosis of celiac disease. These symptoms suggest malabsorption, which can affect nutrient absorption, leading to these clinical manifestations.You've done an excellent job in piecing together these clues to arrive at the correct diagnosis. Keep up the great work in synthesizing clinical information![/Teacher's Response]
GLOBAL DIALOGUE TUTOR 2 (PHI4-14B FINETUNED)
[Students' Response Analysis]S tudent's Selection: (C) Celiac disease; Selection Correctness: True[/Students' Response Analysis] [Teacher's Response]Great job! You've correctly identified celiac disease as the most likely diagnosis for this patient. Let's reinforce why this is the correct choice by reviewing the key clinical features that support it.Firstly, the intensely pruritic lesions over the buttocks are highly suggestive of dermatitis herpetiformis, which is a skin manifestation strongly associated with celiac disease. This is a significant clue that points us in the direction of celiac disease.Secondly, the elevated IgA tissue transglutaminase titer is a crucial finding. This test is highly specific for celiac disease, making it a strong indicator that the patient has this condition.Additionally, the history of iron deficiency anemia and intermittent mild diarrhea suggests malabsorption, which is a hallmark of celiac disease. This malabsorption can also explain the patient's growth percentiles being at the lower end, as celiac disease can impair nutrient absorption and affect growth.You've done an excellent job piecing together these clues to arrive at the correct diagnosis. Keep up the great work in analyzing clinical scenarios![/Teacher's Response]"
GLOBAL DIALOGUE TUTOR 3 (GPT4O (2-SHOT))
[Students' Response Analysis]Student's Selection: (C) Celiac disease; Selection Correctness: True[/Students' Response Analysis][Teacher's Response]Excellent analysis! You've accurately identified ¨Celiac disease¨ as the most likely diagnosis, and your reasoning is well-founded.Let's reinforce the key diagnostic elements that led you to this conclusion. The intensely pruritic lesions over the buttocks you mentioned [Clue 2] are indeed characteristic of dermatitis herpetiformis, which is closely associated with celiac disease. Furthermore, the elevated IgA tissue transglutaminase titer [Clue 6] is highly specific for celiac disease, making this clue particularly significant.Your observation regarding iron deficiency [Clue 3] is critical, as malabsorption due to celiac disease can lead to iron deficiency anemia. The patient's intermittent mild diarrhea [Clue 4] and suboptimal growth percentiles [Clue 5] further support malabsorption, a hallmark of celiac disease.This diagnostic conclusion is a great synthesis of understanding the implications of IgA tissue transglutaminase test results in conjunction with recognizing the dermatological and systemic manifestations of celiac disease in children. Excellent work in putting together these clinical pieces to arrive at the correct diagnosis. Keep honing your diagnostic reasoning skills![/Teacher's Response]

Table 10: A Global Synthesis Tutoring Dialogue Case.