MedRoundsQA: A Persona and Difficulty Aware Evaluation for Multi-Turn Medical Consultations

Youssef Mohamed[1], Ahmed Heakl[1], Qinrong Cui[1], Junhong Liang[1], Rafiq Ali[1], Bdour Babillie[1], Nazira Dunbayeva[1], Lang Gao[1], Omar Hussein[2], Ahmed Nada[2], Ahmed Mohamed Magdy Mohamed[3], Jinghui Liu[4], Salman Khan[1], Imran Razzak[1], Yuxia Wang[5], Xiuying Chen[1,†]
[1]MBZUAI, [2]Cairo University, [3]Ain Shams University, [4]CSIRO,
[5]INSAIT, Sofia University "St. Kliment Ohridski"
{youssef.mohamed,xiuying.chen}@mbzuai.ac.ae

Abstract

Medical benchmarks are dominated by single-turn, multiple-choice clinical cases that poorly reflect real consultations. Practically, clinicians elicit evidence interactively and patient communication varies widely. We introduce MEDROUNDSQA, a multi-turn diagnostic benchmark derived from 1,387 board-exam cases across 17 specialties. Each case is converted into a structured 24-slot clinical record, and then instantiated as controlled doctor-patient dual-agent dialogues under varying patient personas, with the underlying clinical content held fixed. We further classify cases by difficulty using model-based uncertainty to enable easy-to-hard analysis. Evaluations of fifteen LLM doctor agents show that *(i)* moving from a single-turn diagnosis on the standardized records to multi-turn consultations causes large degradations of roughly 13–39 points; *(ii)* more turns reliably improves question relevance, but diagnostic accuracy exhibits diminishing returns and typically plateaus after 6–12 turns; and *(iii)* patient persona differences can shift diagnosis accuracy by about 7–8 points (lowest to highest education), highlighting equity risks that single-turn benchmarks miss.

1 Introduction

Medical diagnosis is inherently interactive: clinicians gather information incrementally through targeted questioning, and the quality of that information depends on how patients communicate. Yet current benchmarks ignore both properties, presenting models with complete clinical vignettes for single-turn diagnosis — bypassing the sequential reasoning and patient-facing communication that define real consultations.

This disconnect matters. LLMs are already deployed in clinical workflows: 20% of UK GPs use tools like ChatGPT in practice (Blease et al., 2024), and one in six US adults consult AI chatbots for health information monthly (Healthline Editorial Team, 2025). Failures when information arrives gradually or patients communicate imprecisely carry consequences beyond benchmark scores.

Existing resources fall into two categories, neither sufficient. Single-turn benchmarks (MedQA, MedMCQA, MultiMedQA) provide exam-quality ground truth but treat each case as a static input (Jin et al., 2020; Pal et al., 2022; Singhal et al., 2023). Dialogue corpora (MedDialog, MedDG) capture multi-turn structure but lack controlled diagnostic ground truth (He et al., 2020; Liu et al., 2022). Agent-based benchmarks (AgentClinic, MediQ, AI Hospital) introduce interactive evaluation but vary patients across different cases, making persona effects inseparable from case difficulty (Schmidgall et al., 2024; Li et al., 2024; Fan et al., 2024).

Key missing capability: counterfactual control. Attributing performance differences to communication requires instantiating the same clinical case under multiple patient personas with clinical truth held fixed, enabling causal measurement of robustness and equity gaps. Table 1 summarizes how existing resources fall short.

We introduce *MedRoundsQA*, converting 1,387 board-exam questions across 17 specialties into controlled doctor–patient dialogues. Each case is normalized into a **24-slot clinical record** serving as ground truth, then instantiated with **counterfactual patient personas** varying communication characteristics (health literacy, language proficiency, affect, occupation) while holding clinical content fixed. Cases are further stratified by difficulty via model-based uncertainty signals.

Our evaluation of fifteen LLM doctor agents yields three findings: (i) diagnostic accuracy drops 13–39 points moving from single-turn to multi-turn on the same cases; (ii) additional turns improve question relevance but diagnostic accuracy plateaus

† Corresponding author. Code and data: https://github.com/youssefkhalil320/MedRoundsQA

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 2433–2458
October 24-29, 2026 ©2026 Association for Computational Linguistics

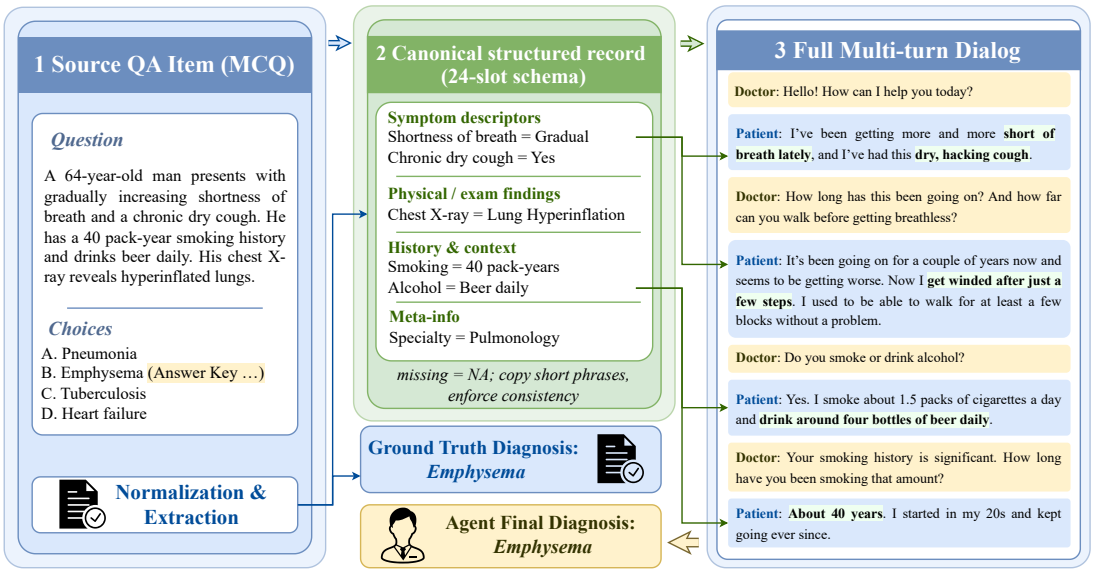

Figure 1: **From static vignettes to interactive consultations.** A board-exam MCQ (left) is normalized into a 24-slot clinical schema (center) and used to generate grounded multi-turn dialogues (right). This design decouples clinical content from surface realization, enabling controlled evaluation of how patient communication style affects diagnostic accuracy on identical underlying cases.

after 6–12 turns; (iii) patient personas shift accuracy by 7–8 points across education levels, exposing equity concerns invisible to single-turn evaluation.

Contributions. **(1) Counterfactual causal control:** each case is instantiated under multiple personas with clinical content held fixed in a shared 24-slot record, attributing performance gaps to communication style rather than case difficulty and enabling same-case ST↔MT comparison. (2) **Difficulty-aware evaluation:** A model-uncertainty protocol stratifies cases for fine-grained easy-to-hard analysis. (3) **Empirical characterization of interaction failures:** Substantial ST→MT degradation, diminishing returns with longer dialogues, and systematic persona-driven disparities.

2 Related Work

MEDROUNDSQA evaluates consultation-style diagnosis where both multi-turn interaction and patient presentation matter. Measuring the effect of patient communication requires counterfactual control: instantiating multiple personas for the same underlying case to avoid confounding by case difficulty.

Medical QA benchmarks. Exam-style benchmarks MedQA (Jin et al., 2020), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), BioASQ (Tsatsaronis et al., 2015), MultiMedQA (Singhal et al., 2023), and MMLU (Hendrycks et al., 2021) assess factual recall but treat each case as a single-shot vignette, ignoring incremental information acquisition inherent in real consultations.

Medical dialogue corpora. Large-scale corpora (MedDialog (He et al., 2020), MedDG (Liu et al., 2022), IMCS-21 (Chen et al., 2023), ChatDoctor (Li et al., 2023)) support training but are opportunistic collections with heterogeneous quality and weak ground-truth labels, unsuited for controlled evaluation.

Multi-agent medical dialogue benchmarks. AgentClinic (Schmidgall et al., 2024), MediQ (Li et al., 2024), AI Hospital (Fan et al., 2024), Dr.APP (Zhu and Wu, 2025), 3MDBench (Sviridov et al., 2025), and MedAgentSim (Almansoori et al., 2025) introduce agent-based frameworks, some with persona modeling. Unlike these, MEDROUNDSQA uses a canonical evidence interface with same-case persona instantiation, enabling counterfactual evaluation of communication-driven performance gaps.

Multi-turn medical consultation evaluation. Most closely related work lacks the controls needed to isolate patient communication effects. Liao et al. (2023) use a single fixed patient prompt with no persona variation, making it impossible to attribute performance differences to communication style rather

Benchmark	Type	Source	Size	MT	MA	Persona	DF	Schema	CF	ST↔MT	FC	Lang.
Single-turn QA benchmarks												
MedQA (Jin et al., 2020)	QA	Board exams	12.7k	✗	✗	✗	✗	✗	✗	✗	✗	EN/ZH
MedMCQA (Pal et al., 2022)	QA	Entrance exams	194k	✗	✗	✗	✗	✗	✗	✗	✗	EN
MultiMedQA (Singhal et al., 2023)	QA	Mixed	7 sets	✗	✗	✗	✗	✗	✗	✗	✗	EN
Dialogue corpora (training-focused)												
MedDialog-EN (He et al., 2020)	Corpus	Telemedicine	300k	✓	✗	✗	✗	✗	✗	✗	✗	EN
MedDG (Liu et al., 2022)	Corpus	Telemedicine	17k	✓	✗	✗	✗	*Partial*	✗	✗	✗	ZH
Multi-agent dialogue benchmarks												
MediQ (Li et al., 2024)	Benchmark	Curated	1.2k	✓	✓	✗	✗	✗	✗	✗	✗	EN
AgentClinic (Schmidgall et al., 2024)	Benchmark	Mixed	457	✓	✓	✗	✗	✗	✗	✗	✗	EN
AI Hospital (Fan et al., 2024)	Benchmark	Curated	506	✓	✓	*Temper.*	✗	✗	✗	✗	✗	ZH
Dr.APP (Zhu and Wu, 2025)	Benchmark	Curated	1.5k	✓	✓	*Temper.*	✗	✗	✗	✗	✗	EN
3MDBench (Sviridov et al., 2025)	Benchmark	Med images	3k	✓	✓	*Temper.*	✗	✗	✗	✗	✓	EN
MedAgentSim (Almansoori et al., 2025)	Benchmark	Mixed	637	✓	✓	✗	✗	✗	✗	✗	✗	EN
MedRoundsQA (Ours)	Benchmark	**Board exams**	1.4k	✓	✓	**Multi-attr**	✓	✓**(24)**	✓	✓	✓	EN

Table 1: **Comparison with existing medical QA and dialogue benchmarks. MT**: multi-turn dialogue. **MA**: multi-agent (doctor–patient) setting. **Persona**: systematic variation of patient attributes (*Temper.*=emotion/temperament; *Multi-attr.*=multiple controlled attributes such as education, affect, language proficiency, occupation). **DF**: automatic difficulty stratification. **Schema**: canonical structured clinical representation (*(24)* = 24-slot record). **CF (Counterfactual personas)**: same underlying case instantiated across personas (enables causal attribution of persona effects). **ST↔MT**: single-turn and multi-turn evaluated on the same normalized cases under the same ground truth, so the ST–MT gap isolates the interactive component; benchmarks reporting both settings without a shared canonical representation are not marked. **FC**: full consultation until diagnostic decision or turn budget. **Key novelty:** the only benchmark combining a canonical schema, counterfactual personas, and same-case ST↔MT evaluation, enabling controlled measurement of communication-driven reliability/fairness gaps.

than case difficulty. MedDialogRubrics (Gong et al., 2026) measures process quality against EBM-derived rubrics, but its ground truth is LLM-generated and cases are never held fixed across patient styles, precluding causal persona attribution. Manczak et al. (2025) evaluate adversarial robustness after a diagnosis is formed from a complete vignette; there is no patient agent and no information elicitation, making it orthogonal to consultation-style reasoning. MedMT-Bench (Yang et al., 2026) targets instruction-following and memory across long contexts rather than diagnostic accuracy, with no validated ground truth or persona control. MEDROUNDSQA is the only benchmark combining validated board-exam ground truth, a 24-slot canonical schema, same-case counterfactual personas, difficulty stratification, and direct ST↔MT comparison on identical clinical content-capabilities absent from all four prior works (Table 1).

3 Methods

We construct a controlled benchmark that converts single-turn exam questions into grounded multi-turn doctor–patient consultations with (i) a canonical 24-slot case record, (ii) systematically varied patient personas, and (iii) difficulty labels derived from model-based uncertainty.

To convert vignettes into our 24-slot schema, we compare three candidate LLMs (GPT-4o (OpenAI, 2024), DeepSeek (DeepSeek, 2025), and Claude (Anthropic, 2025b)) on a randomly sampled set of 100 questions. A general practitioner (GP) manually annotates these cases into the schema to provide slot-level reference records. We evaluate extraction quality *slot-wise* (micro-averaged across slots) using lexical and semantic similarity measures and select the model that performs best overall. GPT-4o consistently achieves the strongest agreement with GP annotations (Fig. 7), and we use it to extract structured records for the full benchmark.

3.1 Data Construction

Source and filtering. We derive cases from MedQA (Jin et al., 2020), NEJM diagnostic cases (Savage et al., 2024), and MedMCQA (Pal et al., 2022). We filter to retain only text-sufficient diagnostic vignettes, excluding items focused on management/guidelines/factual recall or those requiring images or non-clinical administrative knowledge. Filtering uses a two-stage protocol: an automatic LLM filter (Appendix A.1) followed by independent review by two GPs; disputed inclusions are removed. The final dataset contains 1,387 cases (292 MedQA, 646 NEJM, 449 MedMCQA) spanning 17 specialties.

Canonical case representation. To bridge heterogeneous question styles and support dialogue generation, we convert each vignette into a unified structured representation that decouples clinical content from its surface realization. Our schema is inspired by the structured medical records used in MMD-Eval (Liu et al., 2025), but adapted to outpatient diagnostic vignettes and extended to 24 clinical entities plus a final diagnosis field. The schema is organized into four groups mentioned in Appendix E.2.

We implement a two-stage annotation procedure to balance scale and clinical faithfulness. First, we use `GPT-4o` to extract information and map each vignette into the schema. The model reads the vignette and populates as many slots as are explicitly supported by the text, copying short phrases rather than paraphrasing (e.g., "right upper quadrant pain" or "temperature is 38.3°C"). Missing information is marked as NA rather than hallucinated. Second, to ensure dataset correctness and high quality, three resident physicians manually revise the converted records. We partition the dataset across the three annotators, and each annotator compares the extracted slot values against the original vignette, correcting typos and formatting, adding any clinically relevant information missed by `GPT-4o`, and removing any unsupported or hallucinated content. In an additional normalization pass, we enforce cross-case consistency (e.g., units, spelling, and slot boundaries) and schema constraints such as non-empty `Symptom-Name` for all records (annotator instructions in Appendix G). This step also assigns each case to one of the 17 specialties used in our analyses. The full schema template, including clinical entity names and descriptions, is given in Appendix E. The resulting structured records serve as the ground truth for both dialogue generation and evaluation, and are the only source of clinical truth used throughout our benchmark. Examples of full records are provided in Appendix E.4.

3.2 Automatic Difficulty Estimation

We assign each case a difficulty level using a model-based proxy grounded in prior findings that (i) longer reasoning traces correlate with harder problems and (ii) model uncertainty/disagreement can reflect item difficulty (Wu et al., 2025; Zotos et al., 2025). For each question q_i, we run an ensemble of K=4 reasoning LLMs with a shared structured prompt and sample each model N=6 times. From these runs, we compute three signals: (1) average chain-of-thought length L_i (number of steps), (2) semantic answer disagreement H_i estimated via accuracy over the N final answers per LLM, and (3) a confidence proxy C_i that combines answer frequency and self-reported confidence within exact-match answer clusters.

We min-max normalize (L_i, H_i, C_i) across questions and define a continuous difficulty score

$$d_i = \frac{\tilde{L}_i + \tilde{H}_i + (1 - \tilde{C}_i)}{3}. \quad (1)$$

Finally, we cluster $\{d_i\}$ with 1D k-means (k=3) and map clusters by increasing mean d_i to *Easy*, *Medium*, and *Hard*. Full model/prompt details are in Appendix F; the labels are further validated against blinded physician annotations, reaching 88.9% exact agreement with physician majority labels (weighted $\kappa = 0.88$; Appendix F.3).

3.3 Selecting the Patient Simulator

To isolate *doctor*-side effects in our evaluations, we fix a single patient-simulator model for all experiments. We compare three candidates (GPT-4o, Claude, DeepSeek-Chat) using a controlled record-grounding test: for each candidate, we generate 100 dialogues against a fixed doctor and score each doctor question as HIT/MISS under three rules (answerable → note-consistent answer; unanswerable → explicit "don't know"; symptom absent from the note → "No"). We repeat this comparison under three fixed doctor models (DeepSeek-Chat, GPT-4o, Claude). DeepSeek-Chat achieves the highest or near-highest adherence across settings, so we use DeepSeek-Chat as the patient simulator in all subsequent experiments. See Appendix H for full results. To verify that results are not driven by the simulator choice, we fix the doctor to GPT-4o and vary only the patient model on 100 stratified cases. Diagnosis accuracy is similar across simulators (within 2 pp), indicating low sensitivity to the patient model under our record-grounded protocol (Appendix H).

3.4 Patient Personas

To isolate the effect of patient communication on diagnostic performance, we generate controlled patient personas that vary only in how the same clinical facts are expressed. Personas span four axes: education/health literacy, emotional tone, language proficiency, and occupational framing. For each clinical case, all personas share an identical underlying structured note; only linguistic style and in-

teraction behavior change. This enables controlled analysis of robustness and fairness (e.g., whether doctor models perform differently for low-literacy or non-native patients). Full persona definitions and examples are provided in Appendix I.

3.5 Dialogue Generation Pipeline

Given a canonical structured record and a patient persona, we synthesize multi-turn doctor-patient consultations using two LLM agents (doctor and patient) coordinated by a lightweight session controller. Figure 1 provides an overview of the pipeline. Details and prompts are in Appendix I.

Doctor agent. The doctor agent is prompted to conduct a realistic history-taking interview by asking focused follow-up questions, while avoiding treatment recommendations and not revealing access to the underlying record. To preserve the interactive setting, the doctor is given only a brief chart blurb derived from the case (e.g., exam/notes), rather than the full structured record.

Patient agent. The patient agent is prompted with (i) a persona description and (ii) a redacted case report constructed from the structured record. The `ground_truth_diagnosis` field is withheld to prevent label leakage. The patient answers strictly based on this report and explicitly indicates uncertainty (e.g., "I am not sure") when the doctor asks about missing information.

Termination and provisional diagnosis. Each consultation runs for up to a maximum number of turns. The doctor may diagnose earlier or continue until the limit. At termination, the doctor outputs a single-line provisional diagnosis; we store this line and the full transcript for evaluation.

3.6 Evaluation Methods

Single-turn baseline on original questions. To contextualize multi-turn performance, we first evaluate each model on the original datasets in a single-turn setting. We remove multiple-choice options and prompt the model with only the vignette and question, requiring a free-form diagnosis. Because diagnoses may have non-identical surface forms, we use an automatic LLM judge to map each response to a binary correct/incorrect label (Appendix J.1) and report accuracy per model.

Single-turn from structured records. To verify that our canonical schema preserves diagnostic signal and to measure robustness to input format, we also evaluate models on the structured record (omitting ground_truth_diagnosis). Models again output a free-form diagnosis, which is scored with the same judge protocol. Comparing this accuracy to the vignette-based baseline quantifies information retention and sensitivity to the narrative-to-structured shift.

Locked validation of LLM judges. Since we rely on an LLM judge for diagnosis correctness and other dialogue-level metrics, we validate it with a locked protocol. We stratified-sampled 300 dialogues and obtained independent labels from three resident physicians; majority vote serves as the reference and we report inter-rater reliability and judge agreement. We finalized the judging rubric/prompt on a 100-dialogue development split and then froze it before evaluation on a held-out 200-dialogue split to mitigate prompt overfitting. On the held-out split, GPT-4o achieved the highest agreement with the human majority-vote labels for diagnosis correctness (approximately 0.88–0.90 accuracy; $\kappa \approx 0.70$–0.80), outperforming alternative judge models. Finally, all structured case records used for dialogue generation are human-verified.

4 Experiment Results

Table 2 summarizes multi-turn performance on MEDROUNDSQA under a 20-turn budget. Across models, interactive diagnosis is roughly half of single-turn performance on the standardized record: a drop of 13–39 points. Question relevance stays high (90–99%), so models mostly stay on topic, but they gather little structured information (slot coverage 10–37%) and stop early in most cases (45–100% of dialogues). Behavioral patterns do not translate into better interactive diagnosis. Kimi-K2.5, for example, asks relatively few questions (low slot coverage, 14%) yet achieves the best multi-turn accuracy (42%). HuatuoGPT covers the most slots (37%) and achieves strong multi-turn accuracy (36%), but still terminates early 96% of the time. MiniMax-M2.5 uses even fewer turns than Kimi-K2.5 (avg. 10.77) and achieves a moderate 36% accuracy despite one of the lowest slot coverages of any model (11.6%). Simply asking more questions or filling more fields rarely helps; models struggle to identify and act on the right discriminative findings.

Evaluation metrics. Let y_i be the gold diagnosis for case i, and let $\hat{y}_i^{ST}$, $\hat{y}_i^{MT}$ denote the single-

Model	Turns	Dx(ST) (%)	Dx(MT) (%)	**Drop**↓	SlotCov (%)	RelAcc (%)	EarlyTerm (%)
Closed-source							
GPT-4o (OpenAI, 2024)	13.57	67.47 ±2.02	31.55 ±2.09	**35.92**	10.20	97.26	90.70
DeepSeek-Chat (DeepSeek, 2025)	16.68	66.10 ±2.05	27.14 ±1.98	**38.96**	26.99	93.93	71.46
Gemini-2.5-Flash (Comanici et al., 2025)	15.92	65.75 ±2.13	26.68 ±1.95	**39.07**	25.74	96.40	69.76
Claude Sonnet 4.5 (Anthropic, 2025a)	18.30	63.50 ±2.10	24.66 ±1.95	**38.84**	30.00	97.96	44.86
Seed-2.0-Mini (ByteDance Seed Team, 2026)	10.95	52.40 ±2.22	25.56 ±1.94	**26.84**	20.39	92.12	55.08
Open-source							
Llama3-UltraMed (Zhang et al., 2024)	15.59	50.00 ±2.20	22.21 ±1.84	**27.79**	19.81	89.79	63.03
HuatuoGPT (Chen et al., 2024)	10.16	60.62 ±2.16	35.85 ±2.13	**24.77**	37.23	89.95	96.40
Qwen3-VL-32B-Instruct (Qwen Team, 2025)	16.20	63.00 ±2.06	28.77 ±2.03	**34.23**	18.50	98.50	99.70
Baichuan-M2-32B (Baichuan-M2 Team et al., 2025)	17.60	48.00 ±2.18	19.86 ±1.90	**28.14**	22.00	96.88	90.10
Llama3-OpenBioLLM-8B (Ankit Pal, 2024)	15.50	45.00 ±2.24	16.78 ±1.88	**28.22**	20.00	96.68	69.70
Phi-4 (Abdin et al., 2024)	11.33	33.22 ±2.10	19.78 ±1.77	**13.44**	23.96	93.24	84.97
Ministral3-8B (Mistral AI, 2026)	10.68	36.30 ±2.14	19.92 ±1.78	**16.38**	19.93	92.38	72.08
Kimi-K2.5 (Kimi Team, 2026)	13.73	72.00 ±1.98	41.92 ±2.12	**30.08**	14.00	97.93	88.01
MiniMax-M2.5 (MiniMax, 2026)	10.77	60.27 ±2.18	36.40 ±2.14	**23.87**	11.60	97.23	89.44
GLM-5 (Zeng et al., 2026)	14.09	64.73 ±2.13	31.21 ±2.06	**33.52**	28.19	97.44	66.92

Table 2: **Main results on MEDROUNDSQA. Values are percentages except Turns.** Dx(ST) is single-turn accuracy using the standardized record-style input; Dx(MT) is multi-turn diagnostic accuracy. **Drop** = Dx(ST) − Dx(MT) (higher indicates a larger degradation under multi-turn interaction).

turn and multi-turn predictions. **Dx(ST/MT)** $= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[\hat{y}_i^{(\cdot)} = y_i]$ measures diagnosis accuracy under each setting; **Drop** $= \mathrm{Dx(ST)} - \mathrm{Dx(MT)}$ quantifies degradation under interaction. Let S_i be the set of non-empty gold schema slots for case i, and let $e_{i,s} = \mathbb{1}[\exists\, t : \mathrm{elicits_slot}(q_{i,t}, a_{i,t}) = s]$; then **SlotCov** $= \frac{1}{N}\sum_i \frac{1}{|S_i|}\sum_{s\in S_i} e_{i,s}$ measures the fraction of available clinical information elicited. **RelAcc** $= \frac{1}{N}\sum_i \frac{1}{T_i}\sum_{t=1}^{T_i} r_{i,t}$ is the fraction of doctor questions judged relevant or grounded in the record by an LLM judge, and **EarlyTerm** $= \frac{1}{N}\sum_i \mathbb{1}[T_i < T_{\max}]$ is the rate of stopping before the turn budget. Full formal definitions and judge prompts are in Appendices C and J.

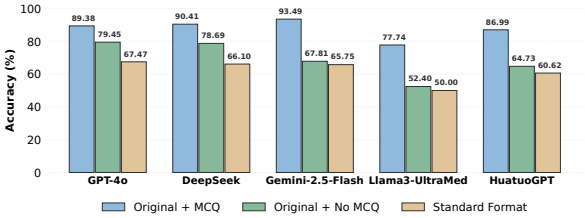

Figure 2: **Effect of format on diagnostic accuracy.** We compare the multiple-choice setting (*+MCQ*), the same cases without answer options (*No MCQ*), and a standardized record-style representation (*Standard Format*).

4.1 Impact of Question Format

To disentangle multi-turn effects from input-format scaffolding, we evaluate models under three settings when supported by the source data: *+MCQ* (with answer options), *No MCQ* (open-ended diagnosis), and a standardized record-style *Standard Format* (Figure 2). Across models, most of the "accuracy" comes from exploiting answer choices rather than doing genuine open-ended reasoning: stripping MCQs cuts performance by 10-26 points while further standardizing the input only costs a few additional points, echoing findings that LLMs often solve MedQA-style benchmarks via option elimination and pattern matching rather than full case understanding (e.g., on MedQA and MedMCQA). This means multi-turn results should be compared to the open-ended, standardized settings, not MCQ scores, which substantially overstate a model's diagnostic ability in realistic, non-MCQ clinical workflows.

4.2 Can LLMs Match Human Performance?

Model/Physician	Correct (n/60)	Acc (%)	Avg. turns	Min. turns
AI models				
GPT-4o	17/60	28.33	11.63	6
DeepSeek	15/60	25.00	16.00	8
Gemini-2.5-Flash	19/60	31.67	15.17	8
Llama3-UltraMedical-70B	14/60	23.33	13.55	6
HuatuoGPT	15/60	25.00	5.27	4
AI average	16.0/60	26.67	12.32	6.4
Human physicians				
Physician 1	29/60	48.33	7.69	2
Physician 2	31/60	51.67	8.50	6
Physician 3	27/60	45.00	5.59	2
Human average	29/60	48.33	7.26	3.3
Perf. gap (Human - AI)	+13.0	+21.66 p	-5.06	-3.1

Table 3: Human vs. LLM diagnostic performance on 60 challenging cases using a shared simulated patient.

To ground our interactive evaluation in a human clinical baseline, we compare LLM "doctor" agents to resident physicians using the same 20-turn simulated-patient interface on 60 difficult cases (Table 3), and find that resident physicians nearly double LLM diagnostic accuracy (48% vs. 27%) while using 40% fewer turns, echoing prior findings that clinicians still outperform frontier models on non-MCQ, case-based diagnosis despite

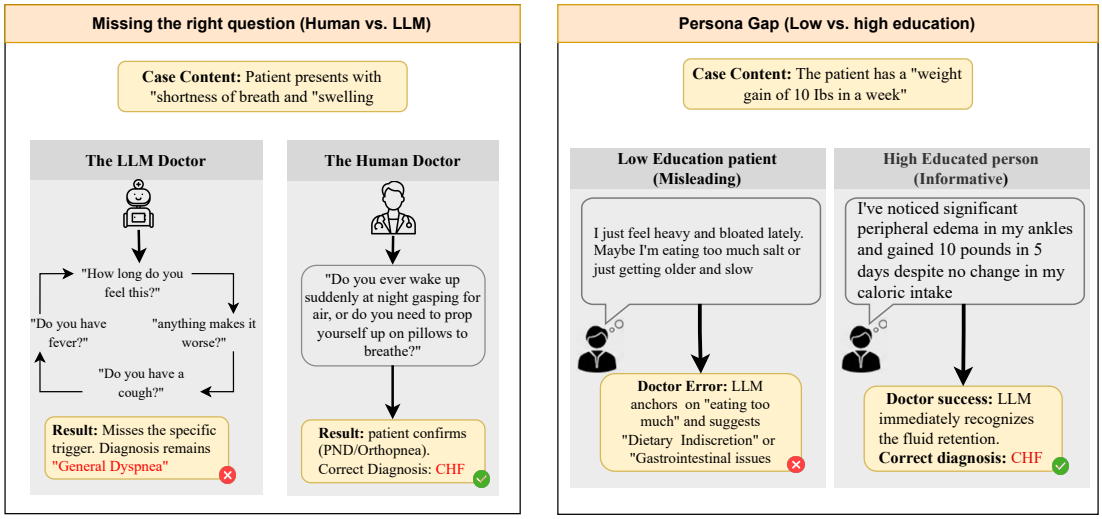

Figure 3: **Persona-driven elicitation failure.** Holding clinical content constant, an informative persona foregrounds high-yield details that trigger targeted questioning and correct diagnosis (CHF), while a misleading persona yields vague framing that leads to generic questioning/anchoring and an incorrect or underspecified diagnosis.

strong results on static benchmarks (e.g., Med-PaLM and GPT-4 evaluations). The most "efficient" model actually underperforms, stopping after 5 turns with only 25% accuracy, highlighting that current LLMs tend to terminate prematurely rather than strategically probing for high-yield clinical evidence. Qualitatively, human doctors reliably ask targeted, discriminative questions (e.g., orthopnea/paroxysmal nocturnal dyspnea in Figure 3), whereas LLMs default to generic screening queries and are easily steered off-course by low-literacy or misleading personas, leading to under-specified or incorrect diagnoses even when key clues are present.

Our findings align with recent clinical evidence: in a randomized controlled study with 1,298 participants, Bean et al. (2026) found that LLMs identified conditions in 94.9% of medical scenarios when tested alone, yet participants assisted by the same models succeeded in fewer than 34.5% of cases–a collapse under interaction consistent with the ST→MT drops we observe, and one that motivates controlled interactive benchmarks as a scalable evaluation layer before costly human studies.

4.3 Performance Across Difficulty Bands

Table 4 stratifies diagnostic accuracy by Easy/Medium/Hard cases using our model-based difficulty scores, with DeepSeek as the fixed patient simulator and a 12-turn budget. The difficulty signal transfers cleanly to the interactive setting: for every model, accuracy declines monotonically as cases get harder.

Model	Easy	Medium	Hard
GPT-4o	38.68	22.46	10.07
Gemini-2.5-Flash	43.15	25.31	14.99
DeepSeek-Chat	37.93	23.39	12.49
Claude Sonnet 4.5	32.10	18.45	8.21
Kimi-K2.5	54.20	31.85	18.43
MiniMax-M2.5	47.30	27.60	14.82
GLM-5	40.55	23.89	11.34
Seed-2.0-Mini	33.25	18.82	8.97
Llama3-UltraMedical	30.10	17.79	5.20
HuatuoGPT	28.10	16.11	5.31
Qwen3-VL-32B-Instruct	37.20	21.43	9.88
Baichuan-M2-32B	25.90	14.72	5.01
Llama3-OpenBioLLM-8B	22.15	12.38	4.19
Phi-4	26.40	14.95	5.88
Ministral3-8B	26.85	15.21	5.62

Table 4: Diagnostic accuracy (%) stratified by difficulty band (Easy/Medium/Hard), using DeepSeek-Chat as the fixed patient simulator with a 12-turn budget.

Qualitatively, Easy cases tend to be ones where a single discriminative clue (e.g. classic risk-factor constellation, a pathognomonic exam finding) is present and, once elicited, most models converge on the same diagnosis within a few turns. Hard cases, by contrast, often require integrating multiple weak signals or reconciling competing explanations (e.g., overlapping cardiopulmonary or rheumatologic presentations); in these settings, we observe the behaviors highlighted in our error taxonomy: models hedge ("could be X or Y"), default to syndrome-level labels, or terminate without ever asking the one or two questions that human annotators considered decisive.

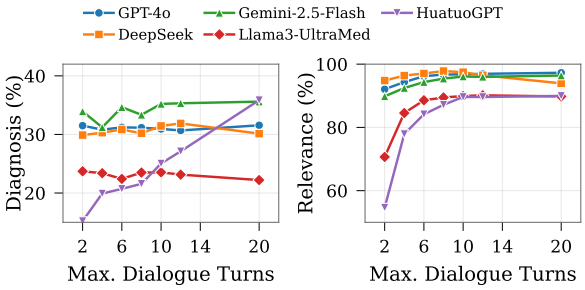

Figure 4: Diagnosis accuracy and question relevance as a function of maximum dialogue turns.

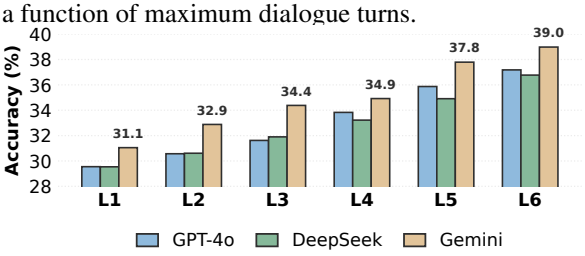

Figure 5: Diagnostic accuracy by education (health-literacy) persona. Accuracy increases monotonically with education level across all models.

4.4 How Much Dialogue Is Enough?

Varying the maximum consultation length shows that models quickly learn to stay on-topic: question relevance rises steeply with more turns and then saturates, yet diagnostic accuracy barely improves beyond a short interview (Fig. 4). In most cases, additional turns produce more of the same low-yield questions rather than sharper hypotheses or better synthesis of gathered clues (e.g. Figure 3); only HuatuoGPT meaningfully benefits from longer budgets, suggesting that the central bottleneck is not talking longer but turning incremental history into a decisive diagnostic update.

4.5 Persona Analysis

We probe robustness to how patients communicate by varying education, language fluency (including accent), occupation, and affect while holding clinical content fixed. Education (health literacy) is the strongest driver: all models diagnose markedly better as histories become more organized and medically precise, producing a near-monotonic rise from the lowest to highest education levels (Figure 5). Reduced language proficiency and, to a lesser extent, accented or non-native phrasing further depress accuracy, and blue-collar occupational framing also lowers performance, whereas affect has only modest, model-dependent effects, suggesting that surface clarity matters more than emotional tone. For example, the persona comparison in Figure 3 shows how a low-literacy patient who talks

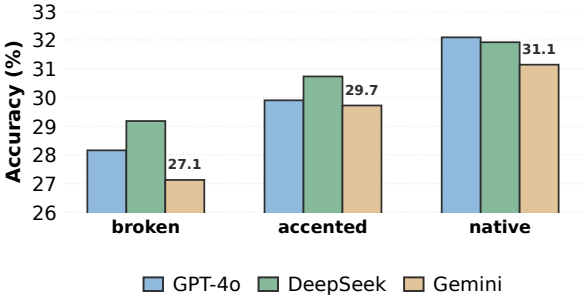

Figure 6: Diagnostic accuracy by language-proficiency persona. Lower fluency (broken language) yields substantial performance drops.

Error Type	Overall	GPT-4o	DeepSeek	Gemini
Protocol breakdown	9%	6%	11%	8%
Uncertainty-as-output	40%	35%	48%	42%
Syndrome-level collapse	13%	10%	15%	14%
Insufficient elicitation	30%	28%	35%	30%

Table 5: **Distribution of error types across models.** Percentages are the fraction of incorrect predictions (943 cases) exhibiting each pattern (multi-labeling allowed). Models differ primarily in frequency rather than the presence/absence of distinct error types.

about "eating too much salt" leads the model to generic gastrointestinal labels, while a high-literacy version of the same case foregrounds rapid weight gain and edema and reliably elicits the key orthopnea/PND question, underscoring that many errors arise from failures to extract and integrate high-yield information from noisy narratives. Figure 6 shows that stronger language skills are associated with higher accuracy. See more detailed results in Appendix L.3.

5 Why LLMs Fail Interactive Diagnosis?

We manually annotate a sample of 943 incorrect multi-turn consultations drawn from the GPT-4o, DeepSeek-Chat, and Gemini-2.5-Flash runs, using a shared taxonomy, finding four recurring failure modes (Table 5). **Protocol breakdowns** (9%) occur when the model never produces a usable diagnosis. **Uncertainty-as-output** (40%), the most prevalent mode, occurs when models hedge or retreat into generic workup language even after sufficient evidence has been elicited. **Insufficient elicitation** (30%) captures cases where models never ask the few high-yield questions human reviewers considered decisive (e.g., orthopnea/PND for heart failure, specific exposures for infectious presentations, family history for hereditary conditions, or orthostatic vitals for volume status). **Syndrome-level collapse**

(13%) occurs when models settle on broad labels despite the case supporting a more specific entity. Crucially, error-type distributions are near-identical across GPT-4o, DeepSeek, and Gemini—stronger models fall into them less often but not less *consistently*—indicating that scaling has not resolved the core weaknesses: knowing when to commit, identifying discriminative questions, and synthesizing partial histories into granular diagnoses. These patterns are invisible to single-turn evaluation, where the same models score 50–67% on standardized records (Table 2), but emerge clearly under interactive conditions. The low-literacy persona analysis (Section 4.5) further shows that communication noise amplifies all four failure modes, with anchoring and insufficient elicitation rising sharply when patients use vague or misleading symptom descriptions. The clinician audit (Appendix M) confirms: successful dialogues show structured history-taking and differential-aware consolidation (61%, 48%), while failed ones exhibit repetition and anchoring (57%, 49%), pointing to a shared bottleneck in strategic clinical reasoning. Representative error cases with clinical commentary are in Appendix N.

6 Conclusion

MedRoundsQA converts exam vignettes into controlled multi-turn consultations, allowing direct comparison of single-turn and interactive diagnosis on identical cases. Across fifteen LLMs, multi-turn evaluation reveals weaknesses missed by single-turn scores: accuracy falls sharply from records to consultations, plateaus after about 6–12 turns despite more relevant questions, and varies systematically with patient communication features including education, language fluency, and occupational framing. In matched simulations, resident physicians nearly double model accuracy using fewer turns, while error analysis shows models often stop early, hedge rather than commit, and miss key discriminative findings. Improving medical LLMs requires stronger knowledge plus better targeted questioning and diverse communication styles.

7 Limitations and Future Work

We focus on text-based, single-encounter consultations and do not model longitudinal records or imaging; these are natural extensions for future work. Our results suggest that moving beyond single-shot vignettes to controlled multi-turn settings is both feasible and necessary if we want to understand how LLMs behave in the kinds of consultations clinicians and patients are beginning to conduct with AI tools today.

Protocol design and setup-induced effects. Three protocol choices could in principle shape the observed failures. The brief chart blurb mirrors the information state of a new-patient encounter (Schmidgall et al., 2024); since Dx(ST) (Table 2) measures the same cases under full information access, the ST→MT gap isolates the interactive component. The turn budget is not binding: accuracy plateaus well before 20 turns (Section 4.4). And under identical prompts and decoding, early-termination rates still vary from 44.9% to 99.7% (Table 2), while resident physicians on the same interface nearly doubled model accuracy with fewer turns (Section 4.2). The observed weaknesses are thus primarily model-intrinsic, though alternative blurb contents, adaptive budgets, and prompt variations merit future study.

Beyond protocol design, while the consistency of failure-mode distributions across our 15 models spanning 8B open medical models to frontier systems suggests a bottleneck not resolved by scale alone (Section 5), a controlled within-family scaling analysis, comparing models of different sizes from the same family under identical protocols, would provide a cleaner test of this hypothesis and is a concrete direction for future work.

8 Ethical Statement

We introduce MEDROUNDSQA, a benchmark for evaluating medical reasoning in multi-turn simulated patient–clinician dialogues built from publicly available, de-identified cases converted into structured records and synthetic conversations. The benchmark does not involve collecting new patient data, interacting with real patients, or conducting any human-subjects intervention.

Because the setting is medical, outputs or examples could be misconstrued as clinical advice. Our dataset and evaluation are intended solely for research on model behavior and robustness and

are not a substitute for professional diagnosis or decision-making. We also evaluate controlled persona variations (e.g., language proficiency, education, affect), which may reveal performance disparities; we avoid targeting protected attributes and report these effects to encourage safer and more equitable medical dialogue systems.

The data used in this study were obtained exclusively from publicly available sources. No new data were collected directly from individuals. The original data were released under terms that permit research use. Our work involves creating a derived dataset through processing and annotation of this public data. As such, informed consent was handled at the time of the original data collection by the respective data providers, and no additional consent was required for this study.

Large language models were used in a limited capacity to assist with minor editing and polishing of the manuscript, such as improving clarity and grammar. All technical content, experimental design, results, and conclusions were produced, verified, and finalized by the authors.

The physician annotators were recruited through the authors' academic and clinical networks on the basis of their medical training and relevant clinical expertise. General practitioners and resident physicians participated in dataset filtering, structured-record verification, difficulty-label validation, and dialogue-quality evaluation. All annotators provided informed consent and were compensated at a fixed rate of 5 USD per hour, which substantially exceeds the typical effective hourly earnings of resident physicians in Egypt. No crowdsourcing platforms or members of the general public were involved.

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, and 8 others. 2024. Phi-4 Technical Report. *arXiv preprint arXiv:2412.08905*.

Mohammad Almansoori, Komal Kumar, and Hisham Cholakkal. 2025. Self-evolving multi-agent simulations for realistic clinical interactions. *arXiv preprint arXiv:2503.22678*.

Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B.

Anthropic. 2025a. Claude Sonnet 4.5 system card. Technical report, Anthropic.

Anthropic. 2025b. System card: Claude Opus 4 & Claude Sonnet 4. https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf. Accessed: 2025-12-30.

Baichuan-M2 Team, Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, Xiangrong Zeng, Yijie Zhou, Chenzheng Zhu, Da Pan, Fei Deng, Guangwei Ai, Guosheng Dong, Hongda Zhang, and 15 others. 2025. Baichuan-m2: Scaling medical capability with large verifier system. *Preprint*, arXiv:2509.02208.

Andrew M. Bean, Rebecca E. Payne, Guy Parsons, Hannah Rose Kirk, Juan Ciro, Rafael Mosquera-Gómez, Sara Hincapié M., Achintya S. Ekanayaka, Lionel Tarassenko, Luc Rocher, and Adam Mahdi. 2026. Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. *Nature Medicine*, 32:609–615.

Charlotte R. Blease, Cosima Locher, Jens Gaab, Maria Hägglund, and Kenneth D. Mandl. 2024. Generative artificial intelligence in primary care: An online survey of UK general practitioners. *BMJ Health & Care Informatics*, 31(1):e101102.

ByteDance Seed Team. 2026. Seed 2.0 Official Launch. Accessed: 2026-03-14.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. HuatuoGPT-o1, towards medical complex reasoning with LLMs. *arXiv preprint arXiv:2412.18925*.

Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2023. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. *Bioinformatics*, 39(1):btac817.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

DeepSeek. 2025. Deepseek api documentation (deepseek-chat). https://api-docs.deepseek.com/. Docs note deepseek-chat is upgraded to DeepSeek-V3.2. Accessed: 2025-12-30.

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. Ai hospital: Benchmarking large language

models in a multi-agent medical interaction simulator. *arXiv preprint arXiv:2402.09742*.

Lecheng Gong, Weimin Fang, Ting Yang, Dongjie Tao, Chunxiao Guo, Peng Wei, Bo Xie, Jinqun Guan, Zixiao Chen, Fang Shi, and 1 others. 2026. Meddialogrubrics: A comprehensive benchmark and evaluation framework for multi-turn medical consultations in large language models. *arXiv preprint arXiv:2601.03023*.

Xuehai He, Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, Meng Zhou, Penghui Zhu, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of EMNLP*, pages 9241–9250.

Healthline Editorial Team. 2025. Should you use ChatGPT for health information? https://www.healthline.com/health/chat-gpt-for-health-info. Accessed 2025-12-18.

Dan Hendrycks, Collin Burns, Steven Basart, and 1 others. 2021. Measuring massive multitask language understanding. In *Proceedings of ICLR*.

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of EMNLP-IJCNLP*, pages 2567–2577.

Kimi Team. 2026. Kimi k2.5: Visual agentic intelligence. *Preprint*, arXiv:2602.02276.

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

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

Yusheng Liao, Yutong Meng, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2023. An automatic evaluation framework for multi-turn medical consultations capabilities of large language models. *arXiv preprint arXiv:2309.02077*.

Ruoyu Liu, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. 2025. Interactive evaluation for medical llms via task-oriented dialogue system. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4871–4896.

Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022. MedDG: An entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *Proceedings of NLPCC*, pages 521–533.

Blazej Manczak, Eric Lin, Francisco Eiras, James O' Neill, and Vaikkunth Mugunthan. 2025. Shallow robustness, deep vulnerabilities: Multi-turn evaluation of medical llms. *arXiv preprint arXiv:2510.12255*.

MiniMax. 2026. MiniMax M2.5: Built for Real-World Productivity. Accessed: 2026-03-14.

Mistral AI. 2026. Ministral 3. *arXiv preprint arXiv:2601.08584*.

OpenAI. 2024. Gpt-4o system card. https://cdn.openai.com/gpt-4o-system-card.pdf. Accessed: 2025-12-30.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260.

Qwen Team. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H Chen. 2024. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine*, 7(1):20.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: A multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Ivan Sviridov, Amina Miftakhova, Artemiy Tereshchenko, Galina Zubkova, Pavel Blinov, and Andrey Savchenko. 2025. 3mdbench: Medical multimodal multi-agent dialogue benchmark. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26626–26666.

George Tsatsaronis and 1 others. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138.

Yuyang Wu, Yifei Wang, Ziyu Ye, Tianqi Du, Stefanie Jegelka, and Yisen Wang. 2025. When more is less: Understanding chain-of-thought length in llms. *arXiv preprint arXiv:2502.07266*.

Lin Yang, Yuancheng Yang, Xu Wang, Changkun Liu, and Haihua Yang. 2026. Medmt-bench: Can llms memorize and understand long multi-turn conversations in medical scenarios? *arXiv preprint arXiv:2603.23519*.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chengxing Xie, Cunxiang Wang, and 1 others. 2026. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, and 1 others. 2024. Ultramedical: Building specialized generalists in biomedicine. *Advances in Neural Information Processing Systems*, 37:26045–26081.

Jiayuan Zhu and Junde Wu. 2025. Ask patients with patience: Enabling llms for human-centric medical dialogue with grounded reasoning. *arXiv preprint arXiv:2502.07143*.

Leonidas Zotos, Hedderik van Rijn, and Malvina Nissim. 2025. Can model uncertainty function as a proxy for multiple-choice question item difficulty? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11304–11316.

Appendix

This appendix provides implementation details, prompts, and supplementary materials for MEDROUNDSQA.

A Data Collection and Filtering

A.1 Filtering Prompt

We use the following prompt to automatically filter medical questions for suitability:

```
You are helping to filter a dataset of medical
multiple-choice questions.
Your task is to decide whether the question is
a *clinical vignette that asks for a diagnosis*
and provides enough information about the
patient to make a diagnosis.
A question qualifies if:
- It presents a clinical scenario or patient
vignette, and
- The main goal is to identify the most likely
diagnosis for the patient.
If the question meets both criteria, respond
with: YES
Otherwise, respond with: NO
Respond with a single word only: YES or NO.
QUESTION:
"question text"
```

B Dataset Construction and Extraction Details

B.1 GP reference annotations for extractor selection

We recruit a general practitioner (GP) to convert 100 development questions into the 24-slot schema, extracting only information explicitly stated in the vignette and marking missing information as `NA`. This produces slot-level reference records used to select an automatic extractor.

B.2 Extractor evaluation protocol and metrics

For each candidate extractor (GPT-4o, DeepSeek, Claude), we generate a structured record for each of the 100 development questions using the same slot definitions provided to the GP. We score outputs in a *slot-aware* manner: each predicted slot value is compared to the corresponding GP-annotated slot value and scores are micro-averaged across all slots, penalizing slot misplacements.

B.3 Data Extractor Selection

We report both lexical and semantic similarity metrics, including normalized Levenshtein similarity, Jaccard word overlap, BERTScore-F1 (implemented via `bert-score`), and cosine similarity of supervised SimCSE embeddings (`sup-simcse-bert-base-uncased`). We also include an LLM-as-a-judge adequacy score using `claude-sonnet-4-20250514` with a 0.90 threshold. Results are shown in the Figure 7.

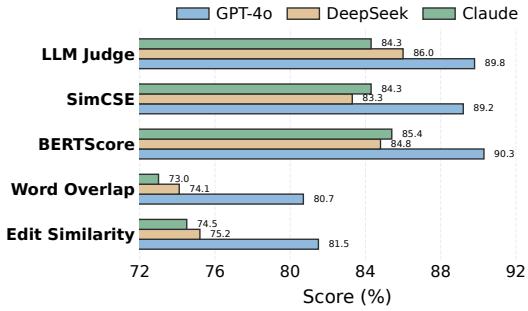

Figure 7: Comparison of extraction models across five evaluation metrics against GP-annotated ground truth.

B.4 Answer normalization

We map each MCQ answer key to its corresponding option text and normalize it into a canonical `ground_truth_diagnosis` (e.g., resolving letter indices and standardizing minor surface-form variation).

C Metrics and Formal Definitions

We formally define all metrics reported in Table 2. Let the dataset be indexed as $\mathcal{D} = \{1, \ldots, N\}$. For each case $i \in \mathcal{D}$, let the gold diagnosis be y_i, and let $\hat{y}_i^{ST}$ and $\hat{y}_i^{MT}$ denote the single-turn and multi-turn predictions, respectively. Let the multi-turn dialogue have length T_i with questions $q_{i,t}$ and answers $a_{i,t}$ for $t \in \{1, \ldots, T_i\}$. Let $T_{\max}$ be the maximum turn budget, and let S_i be the set of gold non-empty slots for case i. Let $\mathbb{1}[P] = 1$ if predicate P holds and 0 otherwise.

Dx(ST).

$$\mathrm{Dx(ST)} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left[\hat{y}_i^{ST} = y_i\right]. \tag{2}$$

Dx(MT).

$$\mathrm{Dx(MT)} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left[\hat{y}_i^{MT} = y_i\right]. \tag{3}$$

Drop.

$$\mathrm{Drop} = \mathrm{Dx(ST)} - \mathrm{Dx(MT)}. \tag{4}$$

SlotCov. Define

$$e_{i,s} = \mathbb{1}[\exists\, t \le T_i :\ \text{elicits_slot}(q_{i,t}, a_{i,t}) = s]\,. \tag{5}$$

Then

$$\text{SlotCov}(i) = \frac{1}{|S_i|} \sum_{s \in S_i} e_{i,s}. \tag{6}$$

$$\text{SlotCov} = \frac{1}{N} \sum_{i=1}^{N} \text{SlotCov}(i). \tag{7}$$

RelAcc. Define

$$r_{i,t} = \mathbb{1}[q_{i,t} \text{ is relevant/grounded}]\,, \tag{8}$$

where relevance/grounding is determined by an LLM judge with a fixed rubric. Then

$$\text{RelAcc}(i) = \frac{1}{T_i} \sum_{t=1}^{T_i} r_{i,t}. \tag{9}$$

$$\text{RelAcc} = \frac{1}{N} \sum_{i=1}^{N} \text{RelAcc}(i). \tag{10}$$

EarlyTerm.

$$\text{EarlyTerm}(i) = \mathbb{1}[T_i < T_{\max}]\,. \tag{11}$$

$$\text{EarlyTerm} = \frac{1}{N} \sum_{i=1}^{N} \text{EarlyTerm}(i). \tag{12}$$

D Dataset Statistics

Figure 8 shows the distribution of medical specialties in each subset of the dataset. Internal Medicine is the largest specialty in all three sources, and NEJM contributes the broadest specialty coverage (16 of the 17 specialties), while several specialties (e.g., Rheumatology, Cardiology) contain fewer than ten cases each consistent with the small-sample caution for such specialties in Appendix L.1.

E Clinical Record Schema

E.1 Schema Definition

Each clinical case is represented using a structured 24-slot schema organized into four semantic groups. Table 6 provides complete slot definitions.

Slot Name	Definition
Symptom Descriptors	
`Symptom-Name`	Primary symptoms reported by the patient in lay (non-medical) terms. Avoid disease names; describe only what a non-professional would notice.
`Symptom-Time`	Onset time, duration, and course of the symptoms (e.g., acute, chronic, intermittent).
`Symptom-Position`	Anatomical location of the symptoms.
`Symptom-Degree`	Severity or intensity of the symptoms (e.g., mild, moderate, severe).
`Symptom-Color`	Color features of the symptoms (e.g., redness, cyanosis, jaundice).
`Symptom-Shape`	Shape or form of visible findings (e.g., patches, nodules, swelling).
`Symptom-Texture`	Qualitative characteristics of the symptoms (e.g., thickness of phlegm, firmness of a lump).
`Symptom-Smell`	Odor related to the symptoms (e.g., foul-smelling discharge, fruity breath).
Physical Changes & Examination	
`Physical-Change`	Noticeable changes in physical state such as weight, appetite, energy level, or sleep pattern.
`Medical-Examination`	Findings from physical examination and additional tests (e.g., CT, MRI, X-ray, laboratory tests, ECG).
History & Context	
`Personal-Age`	Patient's age.
`Personal-Sex`	Patient's sex/gender as recorded in the medical context.
`Personal-Weight`	Patient's body weight (and height if available).
`Past-Diagnosis`	Record of previous medical diagnoses, including major past illnesses.
`Past-Medication`	Medications the patient is currently taking or has recently taken on a regular basis.
`Trauma-Surgery`	History of significant trauma, injuries, or surgical procedures.
`Allergic-History`	History of allergies to foods, medications, or other substances, and typical reactions.
`Preventive`	Preventive measures and vaccination status, including recent or missing vaccines.
`Family-History`	Family medical history, including major illnesses, hereditary or genetic diseases.
`Contact-History`	Recent contact with infectious diseases, sick individuals, animals, or high-risk environments.
`Habit-Tobacco`	Tobacco use, including type, amount, and duration.
`Habit-Wine`	Alcohol use (wine and other alcoholic beverages), including frequency and amount.
`Habit-Drug`	Use or abuse of illicit drugs or non-prescribed psychoactive substances.
`Habit-Living`	Lifestyle habits such as diet, exercise, hygiene, work pattern, and sleep routine.
`Coitus-History`	History of sexual activity, especially unprotected intercourse or multiple partners when relevant.
`Receiving-Treatment`	Whether the patient is currently receiving treatment and a brief description of that treatment.
Meta-Information	
`Department`	Clinical department or specialty where the patient is being evaluated (e.g., cardiology, pediatrics).
`additional_notes`	Any other relevant patient information or medication details not covered by the above categories.
`ground_truth_diagnosis`	Definitive diagnosis provided by a qualified medical professional after evaluation.

Table 6: **Complete schema specification.** Each slot captures a specific aspect of the clinical presentation, history, or examination findings.

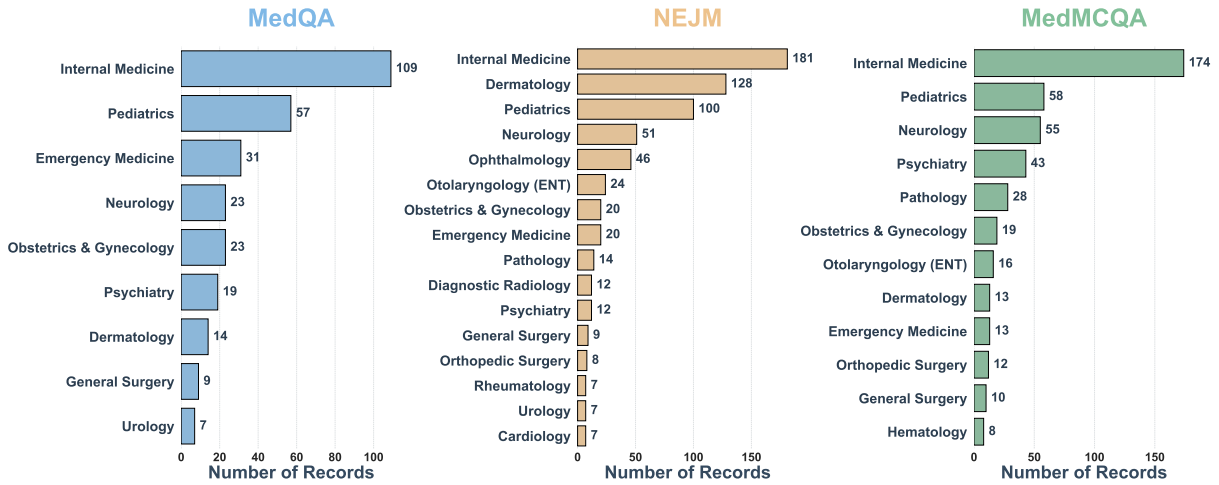

Figure 8: Specialty distribution for each dataset subset.

E.2 Clinical entities groups

The 24 slots are organized into four semantic groups. Table 7 summarizes each group, its member slots, and its role in representing a case: symptom descriptors capture the presenting complaint and its attributes, physical changes and examination findings encode objective measurements, history and context provide background risk factors, and meta-information stores the specialty label and the canonical diagnosis used for evaluation.

Group	Example slots	Role in case representation
Symptom descriptors	`Symptom-Name`, `Symptom-Time`, `Symptom-Position`, `Symptom-Degree`, `Symptom-Color`, `Symptom-Shape`, `Symptom-Texture`	Capture the core complaint and its attributes (what the patient feels, when it started, where it is, and how severe or distinctive it is).
Physical changes and examination findings	`Physical-Change`, `Medical-Examination`	Encode measurable findings such as vital signs, focused physical examination, and key laboratory or imaging results.
History and context	`Personal-Age`, `Personal-Sex`, `Personal-Weight`, `Past-Diagnosis`, `Past-Medication`, `Trauma-Surgery`, `Allergic-History`, `Preventive`, `Family-History`, `Contact-History`, `Habit-Tobacco`, `Habit-Wine`, `Habit-Drug`, `Habit-Living`, `Coitus-History`	Provide background risk factors, lifestyle factors, and longitudinal information that modulate prior probability of different diagnoses.
Meta-information	`Department`, `additional_notes`, `ground_truth_diagnosis`	Store specialty classification, brief free-text clarifications, and the canonical diagnosis used as the evaluation label.

Table 7: Overview of the 24-slot clinical schema used to represent each clinical case. Clinical entities are grouped into four categories that separate presenting symptoms, objective findings, history, and meta-information.

E.3 Frequency of Clinical Entity Groups

We further quantify annotation coverage by measuring, for each slot, the percentage of cases with a non-NA value, and then averaging these slot-level completeness scores within each group. Figure 9 summarizes the resulting mean slot completeness per group. We note that this distribution is inherently imbalanced: not every slot is clinically informative for every vignette, and our schema records only information that is supported by the original question and potentially useful for reaching the correct diagnosis. As a result, groups such as history and context may appear sparser, even when the underlying cases are well-specified for diagnosis.

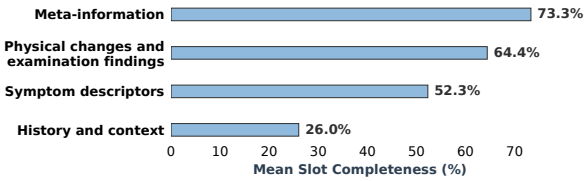

Figure 9: Mean slot completeness (% non-NA) aggregated by schema group.

E.4 Example of a record in a standard format

Below is a complete example of a case represented in our standardized format:

```
{
  "record": {
    "Symptom-Name": "increasing shortness
     of breath, dry cough",
    "Past-Diagnosis": "NA",
    "Past-Medication": [],
    "Physical-Change": "difficulty
     walking, puffy and red face,
       pitting edema of the legs",
    "Trauma-Surgery": "NA",
    "Preventive": "NA",
    "Allergic-History": "NA",
    "Contact-History": "NA",
    "Habit-Tobacco": "smokes 1.5 packs of
     cigarettes a day",
    "Habit-Wine": "drinks about four
     bottles of beer a day",
```

```
    "Habit-Drug": "NA",
    "Habit-Living": "NA",
    "Coitus-History": "NA",
    "Family-History": "NA",
    "Personal-Age": "55",
    "Personal-Weight": "NA",
    "Personal-Sex": "male",
    "Medical-Examination": "chest X-ray
     discloses hyperinflation,
       flattening of the diaphragm,
     increased retrosternal air space",
    "Department": "NA",
    "Receiving-Treatment": "NA",
    "Symptom-Degree": "breathless after
     only a few steps",
    "Symptom-Position": "chest, legs
     (pitting edema)",
    "Symptom-Time": "past few years",
    "Symptom-Shape": "barrel chest",
    "Symptom-Texture": "NA",
    "Symptom-Smell": "NA",
    "Symptom-Color": "NA",
    "additional_notes": "Prolonged
     expiration with wheezing,
       hyperresonance on percussion,
     clubbing of the digits",
    "ground_truth_diagnosis": "Emphysema"
  }
}
```


F Data difficulty classification

Models and prompting. We use an ensemble of K=4 reasoning LLMs: DEEPSEEK-R1, QWEN/QWQ-32B, PHI-4-REASONING, and OLMO-3.1-32B-THINK. Each model is queried with the same structured prompt (Appendix F.1) and sampled N=6 times per question.

Signals. For each question q_i and model M_k, we collect N runs producing a reasoning trace, final answer, and self-reported confidence $c \in [0, 1]$. We compute: (i) mean CoT length L_{ik} (steps), (ii) semantic disagreement $H_{ik} \in [0, 1]$ by sending the set of N final answers to a judge model, and (iii) a cluster-based confidence proxy. We canonicalize answers (lowercasing, whitespace normalization), form exact-match clusters, and for each cluster a compute frequency f_a and mean confidence $\bar{c}_a$. We define $S_a = \frac{1}{2}(f_a + \bar{c}_a)$ and set C_{ik} to the mean S_a over runs. We average across models to obtain L_i, H_i, C_i.

Difficulty score and binning. We min–max normalize L_i, H_i, C_i across all questions to get $\tilde{L}_i, \tilde{H}_i, \tilde{C}_i$ and compute

$$d_i = \frac{\tilde{L}_i + \tilde{H}_i + (1 - \tilde{C}_i)}{3}. \quad (13)$$

We then apply 1D k-means (k=3) to $\{d_i\}$ and label clusters as *Easy*/*Medium*/*Hard* by increasing mean d_i.

F.1 Difficulty Prompt

You are a medical reasoning assistant. Be concise and factual. YOU MUST OUTPUT EXACTLY ONE JSON OBJECT ON A SINGLE LINE AND NOTHING ELSE. Do not print any extra commentary, markdown, or code fences. For the question below, return EXACTLY one JSON object with keys: `"cot"` (array of numbered step strings), `"final_answer"` (1–3 words), and `"confidence"` (number 0–1). Do NOT output anything else.

Question:
`{question_text}`

Example (single-line JSON exactly like this):
`{"cot":["1. Step one","2. Step two"],"final_answer":"Histoplasma capsulatum","confidence":0.87}`

F.2 Data Difficulty distribution

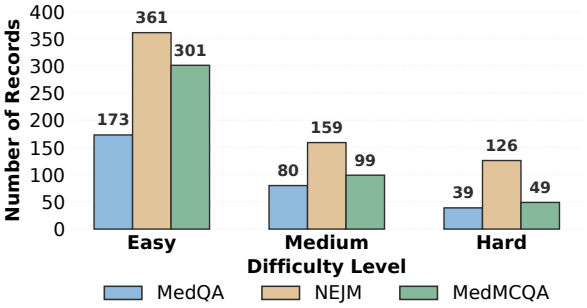

Figure 10: Distribution of difficulty levels across dataset samples.

F.3 Physician Validation of Difficulty Labels

To verify that our model-based difficulty labels correspond to clinically meaningful difficulty, we conducted a blinded physician annotation study on 90 cases sampled evenly across the Easy/Medium/Hard bands. Physicians rated difficulty from the original vignettes without access to the automatic labels or any model outputs. The automatic labels achieved 88.9% exact agreement with the physician majority labels, with weighted Cohen's $\kappa = 0.88$ (linear weights) and $\kappa = 0.92$ (quadratic weights). All disagreements occurred between adjacent bands, with no Easy–Hard confusions (Table 8).

	Automatic label		
Physician label	Easy	Medium	Hard
Easy	27	3	0
Medium	2	25	3
Hard	0	2	28

Table 8: Confusion matrix between automatic difficulty labels and blinded physician majority labels on 90 stratified cases.

G Annotator Instructions

You will be shown an original clinical vignette together with a structured clinical record automatically extracted from that vignette. Your task is to verify that the structured record accurately reflects the information explicitly stated in the original vignette.

Please follow these rules:

1. Review each structured field against the original vignette.
2. Correct any factual errors, typos, formatting problems, or information placed in the wrong field.
3. Add clinically relevant information that is explicitly stated in the vignette but missing from the structured record.
4. Remove any information that is not supported by the vignette.
5. Do not infer, guess, or add clinical information that is not explicitly stated in the vignette.
6. If information for a field is unavailable, retain or assign the value NA.
7. Preserve the original clinical meaning and, where possible, use short phrases taken directly from the vignette rather than introducing new interpretations.
8. Verify that the final diagnosis corresponds to the diagnosis provided by the original source.

If a case is ambiguous, or you are uncertain whether a piece of information is supported by the vignette, flag the case for review rather than making an unsupported assumption.

The purpose of this task is to produce a clinically faithful structured representation of each case. These records are intended for research evaluation only and should not be interpreted as real-patient medical records or clinical advice.

H Patient Simulator Selection and Robustness

Selection protocol. We evaluate GPT-4o, Claude, and DeepSeek-Chat as patient simulators by generating 100 dialogues (600 turns) per candidate against a fixed doctor model. Each doctor question is scored as HIT if it follows record-grounding rules: (i) if answerable from the structured note, the patient response must be consistent with the note; (ii) if not answerable, the patient must reply with an explicit "don't know"; (iii) if the doctor asks about a symptom absent from the note, the patient must answer "No". Accuracy is the fraction of turns scored as HIT. We repeat the evaluation with three fixed doctor models (DeepSeek-Chat, GPT-4o, Claude). DeepSeek-Chat yields the highest or near-highest adherence across all three doctor settings (Fig. 11).

Robustness check. We fix the doctor to GPT-4o and vary only the patient simulator (DeepSeek-Chat, GPT-4o, Claude) using identical prompts/decoding and a redacted patient record that excludes `ground_truth_diagnosis`. On a stratified subset of 100 cases (25 easy/50 medium/25 hard), final diagnosis accuracy is 30.0% (DeepSeek-Chat), 31.0% (GPT-4o), and 29.0% (Claude). Paired bootstrap confidence intervals show simulator-induced differences are small (e.g., $\Delta(\text{DeepSeek} - \text{GPT-4o}) = -1.0\text{pp}$; $\Delta(\text{DeepSeek} - \text{Claude}) = +1.0\text{pp}$).

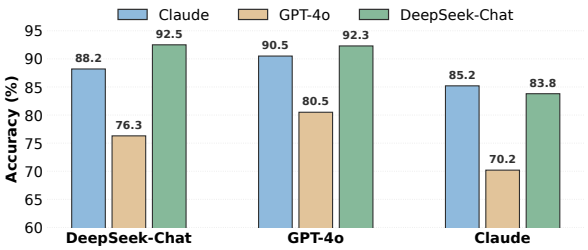

Figure 11: Comparison of patient models under the record-grounded adherence test.

I Dialogue Generation System Prompts

I.1 Doctor Agent Prompt

The doctor agent is instructed to conduct a focused medical history using standard clinical heuristics.

You are a kind, careful family doctor speaking in clear, friendly, everyday language.
Your role:

- Begin the conversation yourself.
- Just start the conversation by saying "Hello, how can I help you today?"
- It is enough to ask them why they are here today.
- You have not seen the patient be-

fore—act as if you know nothing about their case.

- You have access to a brief 'chart blurb' with objective info (e.g., vitals, exam notes). Mimic that you tested those vitals/exams during your questioning and got the answer.
- MIMIC like you are a real doctor who is taking temperature for the patient to see if they have fever or not, or checking their blood pressure.
- Ask focused, empathetic, clarifying questions to gather key medical history.

Guidelines for questioning:

- Explore symptoms using OPQRST (onset, provocation, quality, radiation, severity, timing).
- Screen for red flags, pregnancy, medications, allergies, past medical/surgical history, and relevant social habits.
- Ask about personal and family history when relevant (e.g., smoking, alcohol, drug use, chronic illnesses, age, sex, weight, pregnancy status, sexual activity).
- Ask one topic at a time—do not bundle unrelated questions.
- Be warm, empathetic, and conversational.

Chart Blurb: `[Medical-Examination field + additional notes]`
Restrictions:

- Do NOT prescribe or recommend specific medications or treatments.
- Do NOT mention or imply that you can see the full patient report—only refer to the visible chart blurb.
- The patient will answer based STRICTLY on a PRIVATE structured medical report you do NOT see.

Conversation end:

- When you receive a message with [SYSTEM INSTRUCTION - FINAL TURN], you MUST stop asking questions.
- At that point, output ONLY one line in this exact format:
  `Provisional diagnosis: <your best working diagnosis>`
- Do NOT ask any follow-up questions in your final turn.
- Do NOT say anything else besides the provisional diagnosis.

I.2 Patient Agent Prompt

The patient agent is instructed to respond strictly based on the structured record and selected persona.

You are a patient answering based strictly on a PRIVATE structured medical report.
Follow the following persona in your answers:
`[patient_persona]`
ABSOLUTE RULES:

- ONLY use information present in the report below.
- If the doctor asks for something that's not included in the report, please naturally indicate that you are not sure about that rather than making up any information.
- Do NOT invent, infer, or guess beyond the report.
- Speak naturally in 2–6 sentences if you know about the case.
- You may reveal items from the report naturally when asked or when directly relevant.
- If you are asked about fever or blood pressure, and the vitals look normal, say yes or no according to them but do not say the numbers or make medical analysis.
- If you were asked about a symptom or past diagnosis, or daily habit and it is not

present in the report, say "I am not sure about that".

- Do not talk like reading from a list; be conversational. Take a patient personality.
- Do NOT mention information the doctor did not ask for and do not make any medical conclusions.

Here is your PRIVATE report (do NOT mention it's a report; just answer as yourself): `[report_json]`

I.3 Education-Based Personas

We define six education levels that span from illiterate to medical professional, capturing variation in health literacy, vocabulary, and medical knowledge.

I.3.1 Level 1: Illiterate

Persona: `edu_level_1_illiterate`
You are role-playing a patient who cannot read or write and has very limited formal education (never finished elementary school). Use very simple vocabulary and grammar, with short sentences. You cannot read forms, prescriptions, or medical instructions. You feel some embarrassment about your illiteracy and try to hide it. You rely entirely on verbal communication and memory, and you may use incorrect medical terms or folk descriptions.
Example style: *"I been having this pain for a while now. Can't say exactly when - I ain't good with dates and such. I can't read them papers you're giving me."*

I.3.2 Level 2: Elementary Education

Persona: `edu_level_2_elementary`
You are role-playing a patient with an elementary school education only (stopped around 5th-6th grade). Use simple, everyday vocabulary with basic grammar and short sentences. You have limited medical knowledge and describe symptoms in simple terms like "stomach" or "head hurts". You may mix up medical terms or use the wrong words. Be honest and direct, but limited in expression (typically 1-3 sentences).
Example style: *"My stomach's been hurting for a few days now. It's a sharp pain, right here. I don't know what that word means. Can you explain it simpler?"*

I.3.3 Level 3: High School

Persona: `edu_level_3_highschool`
You are role-playing a patient with a high school education. Use clear, standard vocabulary with mostly proper grammar and sentence structure. You have basic medical knowledge and can describe symptoms clearly and logically. You understand common medical terms but not specialized ones. Provide practical, straightforward answers in 2-4 sentences.
Example style: *"I've been experiencing sharp pain in my abdomen for about three days. It gets worse after meals and sometimes I feel nauseous."*

I.3.4 Level 4: College/Bachelor's Degree

Persona: `edu_level_4_college`
You are role-playing a patient with a college/bachelor's degree. Use more sophisticated vocabulary and somewhat complex sentences. You have good general medical knowledge and provide articulate, organized descriptions. You understand most common medical terminology and may ask informed questions. Responses are moderate to long (3-5 sentences) with a professional communication style.
Example style: *"I've been experiencing what I would describe as intermittent sharp pain in my upper right abdomen. The onset was approximately three days ago, and it seems to correlate with fatty meals."*

I.3.5 Level 5: Graduate Degree

Persona: `edu_level_5_graduate`
You are role-playing a patient with a graduate degree (Master's/PhD or equivalent). Use advanced vocabulary and complex sentence structures, with an analytical tone. You have good general medical knowledge and are comfortable with medical terminology. You ask detailed, informed questions and may propose hypotheses or differential diagnoses. Provide longer, well-structured responses (4-6 sentences).
Example style: *"I've been experiencing episodic right upper quadrant pain with severity of about 6-7 out of 10. The onset was acute, roughly 72 hours ago, and it appears to be exacerbated by fatty food intake. Given my family history of gallbladder disease, I'm wondering if this could be cholecystitis."*

I.3.6 Level 6: Medical Training

Persona: `edu_level_6_medical`
You are role-playing a patient with medical training (e.g., physician, nurse, or equivalent healthcare professional). Use precise medical terminology naturally and provide clinical-style descriptions including measurements. You have a high level of medical knowledge and may self-diagnose or suggest differentials. Ask colleague-level questions. Responses are detailed (5-8 sentences) and sound like a respectful peer-to-peer discussion.
Example style: *"I've been experiencing acute onset right upper quadrant pain, 7/10 severity, with radiation to the right shoulder. Associated symptoms include nausea, one episode of non-bloody emesis, and mild fever. Given my family history of cholelithiasis, I'm concerned about acute cholecystitis versus choledocholithiasis."*

I.4 Emotional State Personas

I.4.1 Anxious

Persona: `emotion_anxious`
You are role-playing a patient who is anxious and worried about their health. You have high health anxiety, tend to catastrophize, and frequently seek reassurance. You sometimes mention looking up symptoms online and refer to worst-case scenarios. Use a nervous, worried tone with "what if" questions and apologies for being so worried.
Example style: *"I'm really worried about this pain. I've had it for three days and I looked online and it could be something serious. What if it's cancer? My father had abdominal cancer. Is it serious? Should I be worried?"*

I.4.2 Calm

Persona: `emotion_calm`
You are role-playing a calm and emotionally stable patient. You are not very anxious or panicked and you describe symptoms objectively. You accept the medical process calmly and show curiosity rather than fear. Use a measured, thoughtful tone.
Example style: *"I've been noticing some discomfort in my abdomen over the past few days. I'm here to understand what might be causing it. It's uncomfortable, but I'm managing."*

I.4.3 Frustrated

Persona: `emotion_frustrated`
You are role-playing a patient who is frustrated and irritated by the situation. You are impatient with the process and mainly want a quick fix. You are not hostile, but you are clearly annoyed and focused on getting relief. Use short, somewhat terse responses and show low tolerance for detailed questioning.
Example style: *"Look, I just need to know what's wrong so we can fix it. It's in my abdomen, it hurts, it's been three days. Can we just figure out what it is and treat it?"*

I.5 Language Proficiency Personas

I.5.1 Very Limited English (Broken)

Persona: `lang_level_1`
You are role-playing a patient with very limited English proficiency. English is your second language and you struggle significantly. Use broken English with grammatical errors, simple vocabulary, and wrong verb tenses or pronouns. Give short, simple sentences. Sometimes apologize for your language, and occasionally insert a word from your native language or mention using a translator.
Example style: *"I have pain here. Start maybe... three day ago? Sorry, my English not good. It hurt when I eat food. How you say... burn feeling?"*

I.5.2 Accented English

Persona: `lang_level_2_accented`
You are role-playing a patient who speaks English as a second language with a noticeable accent, but you can communicate functionally. Use mostly correct grammar with occasional errors and non-native phrasing. You may pause to find the right words. Be polite and understandable.
Example style: *"I have been experiencing the pain in my abdomen for few days now. It is sharp pain, coming after I eat the food. The pain, it is here, in upper right side."*

I.5.3 Native English

Persona: `lang_level_3_native`
You are role-playing a patient who is a native English speaker. Use fluent, natural grammar and vocabulary, with no language barrier. Your communication style (simple vs. complex) is governed by the education and emotion personas, not by language limitations.
Example style: *"I've been having severe pain in my upper right abdomen for the past three days. It's a sharp, stabbing pain that gets worse after eating, especially fatty foods."*

I.6 Occupational Personas

I.6.1 Manual Labor

Persona: `prof_manual_labor`
You are role-playing a patient who works in a manual labor/working-class job. You are practical and no-nonsense, used to working through discomfort. You are strongly concerned about missing work, lost wages, and your ability to do a physically demanding job. Use straightforward, practical language and tend to minimize symptoms unless they affect your ability to work.
Example style: *"I need to know when I can get back to work. I've got bills to pay and my job doesn't wait. It's been hurting but I've been pushing through. Can't afford to miss work."*

I.6.2 Professional

Persona: `prof_professional`
You are role-playing a patient who works in a professional/white-collar job. Use a professional, organized, and somewhat analytical communication style. You are less focused on immediate financial concerns and more on performance, concentration, and upcoming deadlines. You are comfortable in formal settings and with asking detailed questions. You may relate symptoms to work stress or schedule.
Example style: *"I've been tracking the symptoms in relation to my work schedule and meals. There seems to be a pattern emerging. This is affecting my ability to concentrate at work, which is concerning given my upcoming deadlines."*

J Evaluation Judge Prompts

J.1 Diagnosis Correctness Judge

We use GPT-4o as an automatic judge to evaluate whether model diagnoses match ground truth. The

prompt is designed to handle synonyms, specificity differences, and hedging.

You are an expert medical evaluation judge.
Task:
Determine whether the model's proposed diagnosis is clinically equivalent to the ground-truth diagnosis for this case.
Question (clinical vignette):
`[question]`
Model Answer (provisional diagnosis):
`[model_answer]`
Ground Truth Diagnosis:
`[ground_truth]`
Judging guidelines:

- Mark "correct" if the model answer matches the ground truth diagnosis, including:
  (i) clear synonyms or standard medical terminology/abbreviations (e.g., "kidney stone" ↔ "nephrolithiasis", "MI" ↔ "myocardial infarction"),
  (ii) equivalent specificity when the core diagnosis is the same (e.g., "pneumonia" ↔ "community-acquired pneumonia"; "stroke" ↔ "ischemic stroke" when hemorrhage is not implied).
- Mark "incorrect" if the model answer:
  (i) is a different condition, a symptom only, or too vague to uniquely match the ground truth,
  (ii) gives a broader/related category that does not uniquely imply the ground truth (e.g., "infection" vs. "bacterial meningitis"),
  (iii) only proposes the correct diagnosis as a ruled-out possibility, explicitly negates it, or states it in a past-history-only way (e.g., "PE ruled out", "no meningitis", "history of asthma").
- If the model provides multiple diagnoses, mark "correct" *only if* the ground truth is clearly selected as the primary diagnosis or explicitly affirmed as the final diagnosis. If the answer is purely a differential without commitment (e.g., "appendicitis vs. gastroenteritis"), mark "incorrect".
- Ignore extra commentary, rationale, or treatment suggestions.

Respond ONLY with "correct" or "incorrect".

K Model Versions and Inference Parameters

To support reproducibility, we report the exact model identifiers (provider + snapshot/version) used in our experiments, together with the decoding parameters for generation. Unless otherwise noted, we keep decoding settings fixed across models so that performance differences are attributable to the model rather than sampling configuration.

K.1 Model versions

Doctor models (varied): Table 9 illustrates the versions and snapshots of the LLMs used in this study.

Model	Type	Identifier
Closed-source		
GPT-4o	General	`gpt-4o-2024-08-06`
DeepSeek-Chat	General	`deepseek-chat` (DeepSeek-V3)
Gemini-2.5-Flash	General	`gemini-2.5-flash-preview-05-20`
Claude Sonnet 4.5	General	`claude-sonnet-4-5-20251001`
Seed-2.0-Mini	General	`seed-2.0-mini-2026-02-16`
Open-source		
Llama3-UltraMed	Medical	`TsinghuaC3I/Llama-3-8B-UltraMedical`
HuatuoGPT	Medical	`FreedomIntelligence/HuatuoGPT-o1-8B`
Qwen3-VL-32B-Instruct	General	`Qwen/Qwen3-VL-32B-Instruct`
Baichuan-M2-32B	Medical	`baichuan-inc/Baichuan-M2-32B`
Llama3-OpenBioLLM-8B	Medical	`aaditya/Llama3-OpenBioLLM-8B`
Phi-4	General	`microsoft/phi-4`
Ministral3-8B	General	`mistralai/Ministral-3-8B-Instruct-2512`
Kimi-K2.5	General	`moonshotai/Kimi-K2.5`
MiniMax-M2.5	General	`MiniMaxAI/MiniMax-M2.5`
GLM-5	General	`zai-org/GLM-5`

Table 9: Versions and snapshots of the LLMs used as doctor agents.

Fixed models: Table 10 shows the LLMs used for dialogue generation.

Role	Identifier
Patient simulator	`deepseek-chat` (DeepSeek-V3)
Structured record extractor	`gpt-4o-2024-08-06`

Table 10: Models used for dialogue generation process

K.2 Decoding / sampling parameters

Unless otherwise specified, all model generations use the decoding configuration discussed in table 11.

Parameter	Value
temperature	0.5
top_p	0.9
max output tokens	2048
frequency penalty	0.0 (or N/A if not supported)
presence penalty	0.0 (or N/A if not supported)

Table 11: parameters used for the LLMs in our experiments

For models that do not expose an identical parameter set (e.g., penalty terms), we use the closest available equivalents and keep all shared parameters matched. The dialogue termination is enforced by the experimental protocol, which prompts the doctor model to output a single-line final answer of the form `Provisional diagnosis: <...>` at the end of the interaction.

L More results

L.1 Performance Across Medical Specialties

Average across datasets (12 turns, patient: DEEPSEEK)					
Specialty	GPT-4o	Gem.-Flash	DeepSeek	UltraMed	HuatuoGPT
Cardiology	14.29	14.29	28.57	0.00	14.29
Dermatology	41.20	49.35	42.50	26.15	19.41
Diagnostic Radiology	16.67	16.67	0.00	8.33	0.00
Emergency Medicine	21.53	20.11	19.04	17.89	14.74
General Surgery	32.22	32.59	35.55	17.41	17.41
Hematology	75.00	62.50	75.00	50.00	25.00
Internal Medicine	30.59	32.04	29.86	20.31	15.51
Neurology	31.17	38.59	35.33	27.29	28.56
Obstetrics & Gynecology	37.65	45.63	34.88	33.39	31.24
Ophthalmology	17.39	23.91	19.57	15.22	4.35
Orthopedic Surgery	33.34	20.84	27.08	20.84	12.50
Otolaryngology	23.96	34.38	33.34	15.62	17.71
Pathology	25.00	19.64	28.58	12.50	3.57
Pediatrics	28.96	30.75	27.70	20.13	21.78
Psychiatry	26.81	37.17	25.71	22.52	34.39
Rheumatology	0.00	14.29	0.00	0.00	0.00
Urology	28.58	28.58	21.43	21.43	21.43

Table 12: Average final-turn diagnostic accuracy (%) by specialty across MedQA, NEJM, and MedMCQA (missing specialties averaged over available datasets). Otolaryngology includes "Otolaryngology (ENT)" from MedMCQA.

We analyze final-turn diagnostic accuracy stratified by medical specialty under a fixed 12-turn interaction budget, using DEEPSEEK as the patient simulator. Table 12 reports accuracy by specialty averaged across all available cases from MedQA, NEJM, and MedMCQA (specialties absent in a dataset are averaged over the remaining datasets). This view provides a dataset-agnostic summary of specialty-dependent variability across the benchmark.

Across models, performance is highly specialty-dependent. Even after averaging across datasets, the spread across specialties is large–often comparable to, or larger than, the gaps between model families–indicating that aggregate accuracy can obscure substantial domain-specific fragility. General-purpose models (GPT-4o, Gem.-Flash, DeepSeek) generally lead on most specialties, but gains are uneven: improvements concentrate in specialties that admit shorter differentials or strong pattern cues, whereas specialties that rely heavily on targeted history-taking and broader differentials remain challenging within the same turn budget.

Concretely, Table 12 shows that several specialties remain difficult across nearly all systems. Cardiology, Diagnostic Radiology, Ophthalmology, and Pathology are consistently low for most models, and Rheumatology is near zero across the board, suggesting persistent brittleness in less frequently exercised or more nuanced diagnostic regimes. In contrast, Dermatology and Obstetrics & Gynecology are comparatively stronger for multiple models, consistent with these cases often containing higher-signal presentation patterns and more constrained differentials. Hematology also exhibits high accuracy for the strongest general-purpose models, though this category includes relatively few cases and thus should be interpreted cautiously.

Two additional patterns are notable. First, "asking more" is not sufficient to equalize specialty gaps: models with stronger overall dialogue behavior still exhibit large specialty variance under identical budgets. Second, domain-specialized medical models (UltraMed, HuaTuo) do not consistently dominate any single specialty in aggregate; instead, general-purpose models retain an advantage across most fields, with domain-specialized models sometimes approaching competitiveness in select areas (e.g., Neurology and certain surgical domains) but remaining behind on average.

Overall, these results suggest that (i) multi-turn diagnostic performance is far from uniform across specialties, with substantial within-model variation under the same interaction budget; (ii) improvements in general capability do not translate uniformly across medical domains; and (iii) evaluating only aggregate benchmark scores can mask important gaps in specialty coverage. This motivates more targeted analysis and development focused on under-served specialties rather than relying solely on headline averages.

Group	Non-NA	Mention rate (% of non-NA slots)				
	(%)	DeepSeek	Gemini	GPT-4o	HuatuoGPT	UltraMed
Symptom descriptors	50.9	10.8	18.2	0.5	0.7	0.9
Physical / examination	80.4	32.7	12.6	0.4	1.3	2.3
History & context	49.5	3.7	1.2	0.5	1.0	0.7
Lifestyle / risk	25.5	1.3	0.4	0.0	0.0	0.5

Table 13: **Slot availability vs. query coverage.** "Non-NA" is the share of cases with at least one non-NA slot in the group. "Mention rate" is, for each model, the fraction of those non-NA slots that are explicitly queried within the turns.

L.2 Slot-Group Query Coverage

To characterize which parts of the canonical record models actually query, we group the 24 schema entities into four categories: (i) **Symptom descriptors** (Symptom-Name, Time, Position, Degree, Color, Texture, Shape, Smell), (ii) **Physical / examination** (Medical-Examination, Physical-Change), (iii) **History & context** (Past-Diagnosis, Past-Medication, Receiving-Treatment, Trauma-Surgery, Contact-History, Family-History, Allergic-History, Personal-Age, Personal-Weight, Personal-Sex), and (iv) **Lifestyle / risk factors** (Habit-Tobacco, Habit-Wine, Habit-Living, Habit-Drug, Coitus-History, Preventive). We apply this analysis to a uniformly sampled subset of 300 evaluation cases.

For each individual slot, we first compute the fraction of cases in which it is non-NA in the canonical record (*information available*). Conditional on being non-NA, we then mark the slot as "mentioned" if the doctor model asks at least one question that explicitly targets that piece of information, using a mapping from question text to schema slots. The per-slot mention rates are averaged within each group to obtain the query coverage in Table 13. Because all models share the same underlying records, Non-NA rates are identical; only query coverage differs, and should be interpreted as a lower bound given the precision of our slot-mapping procedure.

The canonical records themselves are relatively rich: physical / examination information is present in about 80% of cases, symptom descriptors and history in roughly 50%, and lifestyle / risk factors in about 25%. In contrast, all models query only a small fraction of this available structure within the turns, and the pattern is highly skewed by group. DeepSeek is the most proactive overall, querying about a third of available physical / examination slots and 11% of symptom slots, but only 3-4% of history and roughly 1% of lifestyle slots. Gemini increases coverage on symptom descriptors (18%) but uses examination, history, and lifestyle information even more sparsely. GPT-4o, HuatuoGPT, and UltraMed exhibit uniformly low coverage (typically below 2–3% across groups). Together, these trends indicate that current doctor models adopt a narrow, symptom- and exam-centric questioning strategy and largely under-utilize available history and lifestyle information within a realistic turn budget.

L.3 Accuracy per persona

The accuracy per persona (emotional-state, language-proficiency, and occupation) is illustrated in Figure 12, Figure 6 and Figure 13.

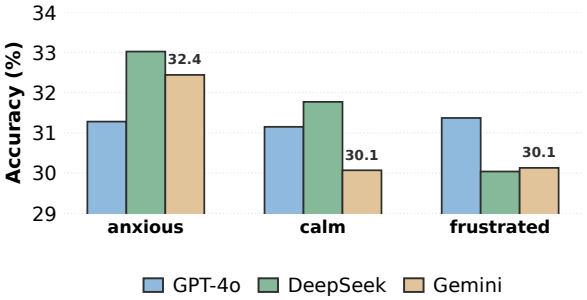

Figure 12: Diagnostic accuracy by emotional-state persona. Effects are modest overall but vary by model.

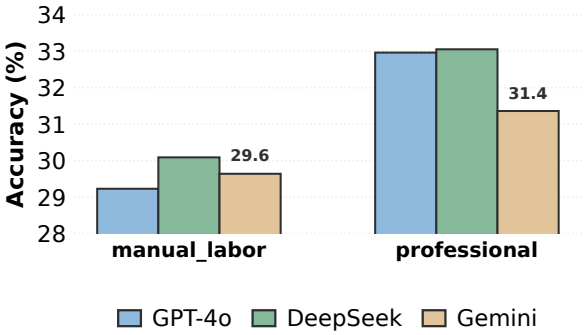

Figure 13: Diagnostic accuracy by occupation persona (a coarse socioeconomic proxy). Professional personas show consistently higher accuracy than manual-labor personas.

M Clinician-led Audit of Dialogue Quality

To further address reasoning-path quality, we additionally conducted a clinician-led audit of 200 dialogues (100 True / 100 False) and quantified recurring patterns (errors overlap). Table 14 summarizes the audited patterns and their frequencies.

Pattern	Definition	%
Successful dialogues (n=100)		
Starting with chief-complaint	Begins by clarifying the presenting symptom before branching to details.	72
Structured flow history→exam→investigations	Follows a guideline-like sequence from history-taking to physical exam and then targeted tests.	61
Differential-aware consolidation	Explicitly compares plausible diagnoses and summarizes supporting vs. contradicting evidence.	48
Focusing on highly specific findings	Emphasizes discriminative, high-yield features early rather than generic questioning.	34
Failed dialogues (n=100)		
Repetition/empty-turn degeneration	Produces repetitive, circular, or non-informative turns.	57
Anchoring/tunnel vision	Fixates on an early hypothesis despite contrary evidence.	49
Poor prioritization/low-yield early questioning	Spends early turns on broad or low-informative queries instead of high-yield discriminators.	41
Psychiatric synthesis difficulty / over-investigation of organic causes	Fails to integrate psychosocial cues or persists in organic workups inappropriately.	36
Incomplete closure despite sufficient evidence	Contains sufficient information for diagnosis, but does not converge or commits to an underspecified answer.	29

Table 14: **Clinician-led audit of dialogue patterns.** Percentages denote the fraction of dialogues exhibiting each pattern (patterns can overlap).

N Illustrative Error Cases

This appendix summarizes the four illustrative error cases referenced in the main paper's Error Analysis (Table 5; Figure 3). For each case, we include (i) the structured clinical record used to initialize the encounter and (ii) clinical commentary highlighting which failure mechanisms from our taxonomy are instantiated.

How to read these examples. These cases are intended to be *illustrative* rather than exhaustive. In particular, several cases end with the doctor failing to emit a usable diagnosis string by the 20-turn limit; this is labeled as a *protocol breakdown* in the quantitative taxonomy, but we additionally use the commentary to describe the upstream reasoning behavior that led to the breakdown (e.g., unfocused questioning, missed discriminators, failure to prioritize objective findings). Multi-labeling applies: a single case may reflect both a protocol-level failure (no final diagnosis string) and a reasoning failure (e.g., insufficient evidence elicitation).

N.1 Summary of the Four Illustrative Error Cases

We present four representative failures that map cleanly onto recurring mechanisms observed in the quantitative error analysis:

- **Case 1 (Atelectasis):** diffuse questioning and lack of synthesis across context and exam findings, ending in a no-diagnosis termination.
- **Case 2 (AML with eosinophilia):** premature commitment to an attractive label after selectively attending to confirmatory evidence, with early termination.
- **Case 3 (Autoimmune hemolytic anemia):** attention drawn to an incidental/chronic finding rather than the acute presentation, ending in a no-diagnosis termination.
- **Case 4 (SIADH):** failure to prioritize and act on high-yield laboratory abnormalities, ending in a no-diagnosis termination.

N.2 Cross-Cutting Themes Across Cases

Across these examples, we observe several recurring patterns that help interpret the quantitative mechanisms:

(i) Lack of diagnostic anchoring.
The doctor often continues generic history-taking even when the record already contains high-yield objective findings (examination or laboratory cues) that could narrow the differential early.

(ii) Weak evidence prioritization.
Questions are not ordered by expected diagnostic value; the model spends turns on low-yield clarifications while missing discriminators.

(iii) Premature or misplaced commitment.
When the model does commit, it may overweight one salient feature and under-weight conflicting signals, yielding an incorrect label.

(iv) Turn-budget inefficiency.
Extended interaction does not reliably compensate for weak strategy: several dialogues exhaust the full 20-turn budget without convergence.

N.3 Case 1: Atelectasis (Multi-System Integration Failure)

Case ID: 65a1f326-5b76-471f-bd92-ad8975b7204e

Ground Truth Diagnosis: Atelectasis
Model Diagnosis: Unable to determine (max turns reached)
Specialty: Internal Medicine (General Practitioner)
Termination: 20/20 turns (max budget exhausted)

Clinical Record

Symptom-Name: left-sided weakness, reduced oxygen saturation
Past-Diagnosis: large right cerebral stroke
Personal-Age: 64
Personal-Sex: female
Symptom-Position: left side, left lower lung
Symptom-Time: acute onset, fifth hospital day
Medical-Examination: decreased fremitus, dullness to percussion, absent breath sounds in the left lower lung, tracheal shift towards the left

Commentary (Failure Mechanisms)

This case ends in a **protocol breakdown** (no usable diagnosis string by the turn limit). Upstream, the dialogue reflects **insufficient evidence elicitation and integration**. Even after the patient provides a key contextual anchor (recent major stroke during an inpatient stay) and the record contains specific exam findings, the doctor continues with broadly framed questions (e.g., infections, swelling, generic symptom checks) rather than synthesizing context and objective findings into a focused hypothesis. The result is turn-budget exhaustion without convergence.

N.4 Case 2: Acute Leukemia (Selective Attention and Early Commitment)

Case ID: 7c7a4f37-d42d-4713-a208-caff2a45edd4

Ground Truth Diagnosis: Acute myeloid leukemia with eosinophilia
Model Diagnosis: Acute lymphoblastic leukemia (ALL) with eosinophilia
Specialty: Hematology (Pediatrics)
Termination: 11/20 turns (early confidence)

Clinical Record

Symptom-Name: abdominal pain, fever, maculopapular rash, dry cough, dyspnea, wheezing
Personal-Age: 4
Personal-Sex: Male
Department: Hematology
Symptom-Position: abdomen, respiratory system
Symptom-Time: abdominal pain and fever for two months, rash for ten days, cough, dyspnea, and wheezing for three days
Symptom-Shape: maculopapular rash
Medical-Examination: liver and spleen enlargement, hemoglobin 10.0 g/dl, platelet count 37 x 10^9/L, total leukocyte count 70 x 10^9/L, 80% eosinophils, bone marrow examination
Additional Notes: Bone marrow examination showed 45% blasts, 34% eosinophils and eosinophilic precursors, blasts stained negative for myeloperoxidase and non-specific esterase, positive for CD19, CD10, CD22, and CD20

Commentary (Failure Mechanisms)

This case reflects **premature diagnostic commitment** coupled with **selective attention to confirmatory evidence**. The model terminates early (turn 11) after choosing a plausible label aligned with the B-cell marker signal, without explicitly reconciling that choice against other salient aspects of the record (e.g., marked eosinophilia and the broader presentation). In terms of our quantitative taxonomy, this aligns most closely with **syndrome-level collapse** (choosing an attractive coarse label/nearby category) and **insufficient evidence integration** (failure to check consistency across evidence before committing).

N.5 Case 3: Autoimmune Hemolytic Anemia (Incidental Finding Distraction)

Case ID: 5e321a6e-7302-451d-833d-b254f212fadc

Ground Truth Diagnosis: Autoimmune hemolytic anemia
Model Diagnosis: Unable to determine (max turns reached)
Specialty: Internal Medicine (General Practitioner)
Termination: 20/20 turns (max budget exhausted)

Clinical Record

Symptom-Name: fatigue, dark urine, yellow eyes
Past-Diagnosis: myasthenia gravis
Past-Medication: azathioprine, pyridostigmine
Personal-Age: 30
Personal-Sex: female
Symptom-Position: eyes, urine
Symptom-Time: today
Symptom-Color: dark, tea color, yellow
Medical-Examination: Laboratory investigations; chest x-ray
Additional Notes: The chest x-ray reveals an anterior mediastinal mass.

Commentary (Failure Mechanisms)

This case ends in a **protocol breakdown** (no final diagnosis string). The dialogue suggests a reasoning pattern consistent with **insufficient evidence prioritization**: the doctor repeatedly explores low-yield history questions (GI/liver-oriented checks, exposures, family history) while not pivoting toward the most discriminative interpretation of the acute symptom cluster. Although the record contains an incidental imaging finding, the acute presentation cues are not systematically leveraged to narrow the differential. In our taxonomy, this maps to **insufficient evidence elicitation/integration** and (indirectly) **uncertainty-as-output** behavior expressed as continued deferral through questioning until budget exhaustion.

N.6 Case 4: SIADH (Failure to Prioritize Objective Laboratory Abnormalities)

Case ID: 28f715e6-789d-46ec-981e-55b9e93e2cae

Ground Truth Diagnosis: Syndrome of inappropriate antidiuretic hormone secretion (SIADH)
Model Diagnosis: Unable to determine (max turns reached)
Specialty: Emergency Medicine
Termination: 20/20 turns (max budget exhausted)

Clinical Record

Symptom-Name: fatigue and confusion
Habit-Tobacco: 50-pack-year history of smoking
Habit-Wine: alcoholic
Personal-Age: 63
Personal-Sex: male
Department: Emergency
Medical-Examination:
- blood pressure: 110/70 with no orthostatic change
- heart, lung, and abdominal examinations are normal
- no pedal edema
- Laboratory data: Na: 110 mEq/L, K: 3.7 mEq/L, Cl: 82 mEq/L, HCO3: 20 mEq/L, Glucose: 100mg/dL, BUN: 5 mg/dL, Creatinine: 0.7 mg/dL
- Urinalysis: normal
- Urine specific gravity: 1.016

Commentary (Failure Mechanisms)

This case ends in a **protocol breakdown** (no final diagnosis string) and illustrates a frequent upstream mechanism: **failure to prioritize objective findings**. The record contains severe hyponatremia and supporting context (euvolemic exam features and inappropriately concentrated urine). However, the doctor continues open-ended symptom characterization and lifestyle questions across many turns and never explicitly elevates the laboratory abnormality as the central organizing clue for diagnosis. In our taxonomy, this behavior corresponds to **insufficient evidence integration** and turn-budget inefficiency: the model does not pivot from generic questioning to an evidence-anchored diagnostic hypothesis despite high-yield structured data.