Human or *LLM* as Standardized Patients? A Comparative Study for Medical Education

Anonymous ACL submission

Abstract

Standardized Patients (SP) are indispensable for clinical skills training but remain expensive, inflexible, and difficult to scale. Existing large-language-model (LLM)based SP simulators promise lower cost yet show inconsistent behavior and lack rigorous comparison with human SP. We present EasyMED, a multi-agent framework combining a Patient Agent for realistic dialogue, an Auxiliary Agent for factual consistency, and an Evaluation Agent that delivers actionable feedback. To support systematic assessment, we introduce SPBench, a benchmark of real SPdoctor interactions spanning 14 specialties and eight expert-defined evaluation criteria. Experiments demonstrate that EasyMED matches human SP learning outcomes while producing greater skill gains for lower-baseline students and offering improved flexibility, psychological safety, and cost efficiency.

1 Introduction

Medical education underpins professional healthcare training, fostering essential competencies in clinical reasoning, communication, and decision-making (Cleland and Durning, 2022). Deficiencies in these skills contribute to diagnostic errors and preventable harm, with communication failures alone responsible for thousands of in-hospital deaths annually (Ball et al., 2015). To improve learning efficiency and scalability, recent studies have explored artificial intelligence (AI)particularly LLMfor clinical education. LLM enable interactive simulation and automated feedback at low cost, offering new possibilities for scalable, practice-oriented training (Thirunavukarasu et al., 2023).

Patient simulation is central to such experiential learning, allowing students to practice reasoning and communication in realistic but safe settings. The Standardized Patientactors trained to portray clinical casesis widely recognized as the gold standard (Sayers et al., 2024; Ma et al., 2023), yet remains costly, labor-intensive, and difficult to scale, limiting case diversity and accessibility (Zendejas et al., 2013).

To mitigate these constraints, previous researchers have explored LLM-based virtual SP simulators as low-cost and scalable alternatives (Ye and Tang, 2025; Du et al., 2024; Khandwaha and Law, 2025). Most systems adopt a single-agent, prompt-driven architecture, where a pretrained model performs patient simulation, intent understanding, and dialogue control within one process. This unified design couples persona control, information disclosure, and evaluation, leading to unstable role behavior, factual errors, and uncontrolled information flow (Sirdeshmukh et al., 2025; Laban et al., 2025). Evaluation methods are usually ad hoc, based on coarse rubrics or single-turn scoring, which limits analysis of interactive reasoning and learning outcomes (Burke et al., 2024; Wang et al., 2025). As a result, whether virtual SP can match the performance and educational value of human SP remains unclear (Simzine, 2025).

In this research, we introduce ***EasyMED***, a multi-agent framework that transforms virtual SP training into a structured and interactive process. Unlike traditional single-agent models (Li et al., 2023), EasyMED decomposes patient simulation, intent recognition, and evaluation into coordinated agents with distinct roles. Each agent aligns with the workflow of real clinical practice: the **Patient Agent** simulates realistic symptom expression and emotional behavior, the **Auxiliary Agent** interprets the doctors conversational intent to guide contextually appropriate responses, and the **Evaluation Agent** monitors and assesses interactions to ensure consistency and clinical validity. This design establishes an **interpretable framework** that enhances controllability and error traceability

through clearly defined, task-specific agent roles.

To benchmark interactive behaviors that matter pedagogically, we introduce ***SPBench***, a dataset derived from authentic SP-student dialogues covering 14 specialties and eight expert-defined assessment criteria. It standardizes prompts, information disclosure rules, and scoring procedures, enabling consistent comparison of models and prompting strategies under realistic clinical scenarios.

We further validate ***EasyMED*** through a four-week controlled study comparing it with human SP training under identical content and scoring rubrics. The results demonstrate that *EasyMED* achieves comparable or even superior learning outcomesparticularly among novice learnerswhile providing greater flexibility, psychological safety, and cost efficiency.

In summary, our contributions are threefold: **(1) Multi-Agent Framework for Virtual SP Training**: We propose *EasyMED*, effectively simulates the structured interactive process of the realistic SPs and training doctors with three coordinated agents. This structure enables controllable, persona-stable interactions and provides transparent, multi-dimensional feedback. **(2) Benchmarking with SPBench**: We construct a benchmark from authentic SPstudent dialogues covering 14 specialties and eight evaluation criteria. It standardizes prompts, disclosure rules, and scoring, supporting reproducible, turn- and session-level evaluation of interactive behaviors. **(3) Empirical Validation of Educational Impact**: We conduct a four-week controlled study comparing EasyMED with human SP training under identical test content and rubrics, demonstrating comparable learning outcomes, improved accessibility, and clear advantages for lower-baseline learners.

2 Background and Motivation

2.1 The Definition of Standardized Patients

Definition Standardized Patients (SP) is a training approach in which individuals are trained to consistently simulate the history, personality, physical findings, and emotional states of real patients. SP is widely used in medical education, especially in Objective Structured Clinical Examinations (OSCE), providing a safe and repeatable environment for practicing key clinical skills such as history taking, physical examination, and communication (Cook and Triola, 2009; Berman et al., 2016; Kononowicz et al., 2015). This allows students to develop these abilities without risking harm to real patients.

Challenge While SP is effective, it faces several challenges, including high costs for recruiting and training actors, logistical constraints from fixed locations and schedules, and inconsistent performance due to factors like fatigue or mood (Hamilton et al., 2024; Kuehne et al., 2018). These limitations restrict the scalability and accessibility of SP-based training, particularly in resource-limited settings.

2.2 LLM Simulator vs. Standardized Patient

As discussed in Section 2.1, the high cost, limited scalability, and inconsistent performance of traditional SP have prompted researchers to explore LLM-based virtual SP as an alternative. While LLM-based SP reduce costs, improve scalability, and standardize responses compared with human SP, they still show unstable behavior, lack structured feedback, and have limited validation against human SP. Building on these questions, our *EasyMED* framework further enhances virtual SP capabilities by introducing a multi-agent design, structured case control, and integrated evaluation modules.

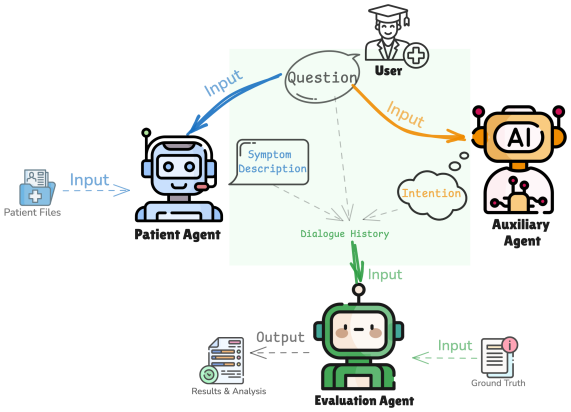

Figure 1: The multi-agent architecture of our virtual medical training system. This "division of labor" model includes: (1) a Patient Agent to simulate patient dialogue, (2) an Intention Identification Agent to interpret the user's clinical queries, and (3) an Evaluation Agent to assess performance against ground truth.

As shown in Table 1, these limitations are evident when comparing Human SP, existing LLM-based SP systems, and our *EasyMED* framework across operational efficiency, system design and validation, teaching quality, and learner experience.

Metric	Human SP	Existing LLM-based SP	Our EasyMED
Cost	High actor, training and venue costs	Low API costs; limited assessment	Low costs with built-in multi-dimensional assessment
Time	Fixed scheduling slots	Instant but single-agent	Instant with multi-agent coordination
System Architecture	–	Single-agent role-play; no module separation (Wang et al., 2023, 2024a)	Multi-agent with Patient, Assistant, Evaluation
Assessment Capability	Manual checklists	Manual or simple scoring; no integrated module (Li et al., 2024b)	Built-in Evaluation Agent with automated, multi-dimensional feedback
Benchmarking	–	No real-case benchmark; mostly synthetic (Sardesai et al., 2024; Ye and Tang, 2025)	SPBench: real SPstudent dialogues (12 specialties, 8 dimensions)
Empirical Validation	Local experience only	Few or no long-term comparative studies (Ye and Tang, 2025)	Four-week study comparing EasyMED vs. Human SP
Persona Stability	Actor variability	Persona inconsistency, leakage, harmful replies (Du et al., 2024)	Prompt control + multi-agent improves stability
Consistency	Performance varies across actors	Standardized but unstable persona	Multi-agent design improves persona stability
Feedback	Manual scoring	Limited or no automated feedback	Evaluation Agent + SPBench for reliable feedback
Reproducibility	Limited, hard to repeat	Anytime but limited diversity	Anytime with curated real-case benchmark
Psychological Safety	Higher performance anxiety	Lower pressure vs. Human SP (Almansoori et al., 2025).	Lower pressure + structured, transparent assessment

Table 1: Integrated comparison of Human SP, existing LLM-based SP systems, and our EasyMED framework across operational efficiency, system design, teaching quality, and learner experience.

3 Methodology

3.1 EasyMED: Multi-Agent Framework

Building on the limitations of single-agent simulators discussed in Section 1, *EasyMED* is designed as a controllable and modular training framework. Existing virtual SP systems integrate patient simulation, intent understanding, and evaluation within one model, which often leads to persona drift, uncontrolled information disclosure, and weak or non-actionable feedback (Du et al., 2024). To address these issues, *EasyMED* adopts a multi-agent *division-of-labor* architecture (Figure 1) that explicitly separates patient simulation, intent interpretation, and evaluation, while maintaining coordination within a single session.

During a training session, a student's utterance is first analyzed by an *Auxiliary Agent*, which identifies the clinical intent behind the question and maps it to one of 31 expert-defined intent categories (Appendix 9). This output conditions the *Patient Agent*, which generates clinically grounded responses from curated case scripts under explicit information-disclosure constraints to preserve realism. All interactions are then logged and evaluated by an *Evaluation Agent*, which compares the students inquiries with predefined checklists and applies an LLM-based judge to assess completeness and reasoning accuracy. At the end of each consultation, the Evaluation Agent produces structured feedback summarizing strengths and areas for improvement.

3.2 SPBench: Benchmark for Virtual SP

To enable systematic and reproducible evaluation of LLM-based SP, we constructed **SPBench**, a benchmark dataset derived from authentic SP-physician dialogues.

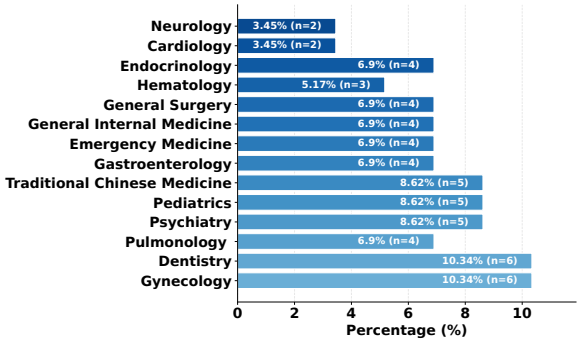

Figure 2: Distribution of clinical cases in the SPBench dataset by medical department.

Case Source and Scope We collected 58 real clinical cases. From the doctor-patient dialogue records associated with these cases, we processed and extracted a dataset containing 3,208

question-and-answer pairs. The source materials include the *Manual for Writing Standardized Patient Cases*[1] and the *A Practical Tutorial for Standardized Patients*[2]. All records were fully anodized and used solely for academic research. They were also reviewed by experienced clinicians to ensure accuracy and representativeness. As shown in Figure 2 and Figure 3, the specialty distribution of SPBench spans 14 clinical departments. General Surgery is the most represented specialty, accounting for 13.79% of the cases, followed by other departments such as Cardiology, Respiratory Medicine, and Psychiatry. The age of the patients ranges from children to the elderly, with the majority concentrated in the 21-60 age group. Regarding gender, the dataset is composed of 35 female cases, representing 60.3% of the total, and 23 male cases, representing the remaining 39.7%.

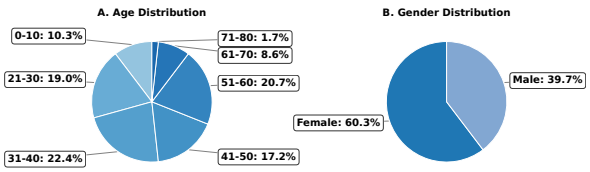

Figure 3: Demographic distribution of cases in the SPBench dataset. The left panel shows the age distribution, and the right panel shows the gender distribution.

Data Structure SPBench provides two core types of content: 1) Case Scripts: These provide standardized patient histories and physical examination findings for each case; and 2) Authentic Dialogue Records: These are turn-by-turn transcripts of conversations between physicians and human SP. The records serve as a benchmark to compare the conversational quality of an LLM-based virtual SP against a human SP for the same case.

Quality Control To ensure accuracy and reliability, all dialogue transcripts were digitized from the two books in Section 3.2 using Optical Character Recognition (OCR) and then manually proofread by three fifth-year medical students. They corrected OCR-induced errors such as typos and punctuation to ensure fidelity to the original source material. This process preserves the authenticity and consistency required for benchmarking.

3.3 Evaluation Protocol

To systematically assess the performance of LLM as Standardized Patients, we (i) define what to measure via eight core evaluation criteria; (ii) validate the reliability of our automated scoring against clinical experts; (iii) conduct fair comparisons between human SP, *EasyMED*, and an external framework under identical inputs and scoring; and (iv) perform component-level ablations within *EasyMED* to analyze where gains come from. The following subsections detail each component.

Evaluation Metric In collaboration with three clinical experts, we defined eight core dimensions for evaluating the virtual patient's response quality, as detailed in Table 3. During the evaluation, experts rated performance for each dimension on a 5-point scale. These scores were then linearly transformed to a 100-point scale for reporting to enhance the granularity and readability of the results.

Expert Rating and Correlation Analysis Considering the high cost of large-scale manual annotation, we first used a well-designed evaluation prompt (Appendix I.4) and employed GPT-4o as an automated evaluator to score all results. To verify its reliability, we randomly selected 30 samples for blind review by two clinical experts who were unaware of the GPT-4o ratings. As shown in Table 2, the automated scores correlated strongly and significantly with the experts average ratings (Pearsons $r = 0.79$). Although GPT-4os mean score (78.5) was consistently stricter than the experts mean (83.9), this high alignment supports its use as a reliable proxy for large-scale evaluation.

	Automated	Professional A	Professional B
Average Score	78.5	86.3	81.5
Standard Deviation	8.8	9.3	9.8
Correlation	0.79	–	–

Table 2: Inter-Rater Reliability between an Automated Evaluator and Human Experts

Comparison between human and LLM-based SP To compare the responses of LLM-based SP with those of a human SP under identical context, we first isolated the physician's questions from the dialogue records in SPBench. We then used these exact questions and patient case as inputs for the LLM-based SP. This method allowed us to directly compare the responses generated by our virtual patients against those of the human SP under identical conditions.

Comparison with External Frameworks Beyond comparing against human SP under identical prompts, we further benchmarked our system

[1] https://weread.qq.com/web/bookDetail/1a432e10811e38145g0194ab

[2] https://www.pumpedu.com/home-shop/7125.html

Evaluation Criterion	Description
Query Comprehension	Assesses the ability to accurately understand the physician's questions, grasp the core intent, and provide relevant answers without misinterpretation.
Case Consistency	Evaluates whether responses are factually consistent with the predefined patient case, including symptoms, medical history, and key facts. Checks for contradictions or information not present in the case.
Controlled Disclosure	Measures the ability to provide information passively and appropriately, answering only what is asked without volunteering unsolicited details. Focuses on conciseness and relevance.
Response Completeness	Assesses whether responses are comprehensive and address all parts of a physician's query, avoiding the omission of crucial information from the case profile.
Logical Coherence	Evaluates the internal logical consistency of the information provided. Ensures that reported symptoms and their characteristics (e.g., location, duration) are coherent from a patient's perspective.
Language Naturalness	Assesses whether the communication style is natural and patient-like, using plain language and avoiding overly technical medical jargon. Focuses on the authenticity of the linguistic style.
Conversational Consistency	Measures the ability to maintain consistent information across a multi-turn dialogue. Checks for self-contradictions and ensures stability throughout the interaction.
Patient Demeanor	Evaluates the ability to maintain a patient-like demeanor, demonstrating appropriate patience, cooperation, and emotional tone, even when faced with repetitive or clarifying questions.

Table 3: Definitions of the eight evaluation criteria for assessing the quality of the conversational agent.

against an open-source LLM-based SP framework, *EvoPatient* (Du et al., 2024). Using the same SP-Bench physician questions and case profiles, and the same automated evaluator and scoring protocol, basic settings are summarized in Appendix F. we report the system-level scores in Table 4.

System	Configuration	Score
Human SP	–	97.33
EasyMED (ours)	More Agent	96.98
EvoPatient	Authors setup	93.33

Table 4: Comparison of different frameworks on SP-Bench under the same evaluation protocol.

Ablation on EasyMED Components We conducted an ablation study to quantify the contribution of key components in our virtual SP design. For each model, we started from a baseline configuration and incrementally added components: (1) **Baseline** no chain-of-thought or intent recognition; (2) **CoT** adds explicit step-by-step reasoning to the baseline; (3) **Intent Recognition** further adds an intent-recognition module on top of CoT.

Model Name	Baseline	COT	Intent Recognition
Qwen3-8B	81.12	81.95	81.64
Qwen3-32B	88.87	91.89	94.08
DeepSeek-R1	93.11	94.90	95.72
GPT-4o	89.54	90.08	93.70
Gemini 2.5 Pro	95.04	95.77	**96.98**

Table 5: Effect of CoT and intent recognition modules across different models.

Table 4 shows that EasyMED matches human SP performance (96.98 vs. 97.33) and clearly outperforms the external framework EvoPatient (93.33). These results indicate that our multi-agent design and evaluation-aligned prompting produce higher-fidelity SP behaviors under identical inputs and scoring.

As shown in Table 5, adding COT to the baseline improves most models, and layering intent recognition on top yields further gains for medium-to-large models (e.g., Qwen3-32B, DeepSeek-R1, GPT-4o, Gemini 2.5 Pro). However, these advanced components are capacity-sensitive: for a small model such as Qwen3-8B, intent-guided CoT can introduce over-constraints and slightly reduce the score relative to plain CoT.

4 Experiment Protocol

After validating our virtual SP on SPBench, we conducted a user study to evaluate its educational impact in a real-world setting. This section describes the four-week randomized crossover trial with undergraduate medical students, designed to test whether EasyMED outperforms traditional human SP training.

4.1 Experimental Design

We ran a four-week randomized crossover experiment comparing *EasyMED* with traditional human SP training. Each participant experienced both training methods in different phases, enabling within-subject comparisons and controlling for individual differences in baseline competency. The procedure is shown in Table 7.

Group Sequence	Mean Score (śSD) at Test Point			Mean Score Gain (śSD) by Phase		
	Pre-test	Mid-test	Post-test	Phase 1 (Wks 1-2)	Phase 2 (Wks 3-4)	Total Gain
Group A (AI→SP) (N=7)	70.56 (ś11.45)	86.07 (ś6.88)	87.44 (ś4.53)	**+15.51 (AI)** (ś7.82)	**+1.37 (SP)** (ś3.65)	+16.89 (ś7.45)
Group B (SP→AI) (N=7)	69.84 (ś13.04)	82.01 (ś7.42)	85.20 (ś4.93)	**+12.17 (SP)** (ś7.45)	**+3.19 (AI)** (ś3.25)	+15.36 (ś8.36)

Table 6: The table presents mean scores at each test point and the corresponding mean score gains during each training phase. Participants were in either Group A or Group B. All values are mean ± standard deviation.

Study Period	Timeline	Group A	Group B
Baseline	Week 0	Pre-Test	Pre-Test
Period 1	Weeks 1-2	**Intervention A:** EasyMED Training	**Intervention B:** Human SP Training
	End of Week 2	Mid-Test	Mid-Test
Period 2	Weeks 3-4	**Intervention B:** Human SP Training	**Intervention A:** EasyMED Training
	End of Week 4	Final Test & Questionnaire	

Table 7: Design of the Experiment, illustrating the sequence of interventions (*EasyMED* Training and Human SP Training) and assessments for Group A and Group B.

Patient Cases During the intervention, students practiced with 20 standardized clinical cases covering core diseases at the scope and difficulty of the National Medical Licensing Examination (see Appendix D).

4.2 Participants

We recruited 20 fourth- and fifth-year undergraduate medical students, all of whom completed a pre-test to assess baseline knowledge and received a 10-minute *EasyMED* orientation. Students with scheduling conflicts or extreme pre-test scores were excluded, resulting in a final sample of 14 (7 males, 7 females; mean age 23, range 21-24). All had completed core clinical coursework but had not yet taken the National Medical Licensing Examination.

To ensure baseline equivalence, participants were rank-ordered by pre-test scores and alternately assigned to two sequence groups. Three experienced SP professionals delivered the human SP sessions. All participants received hourly compensation in accordance with ethical regulations.

5 Results and Analysis

This section presents and analyzes the results from Section 4, focusing on whether the Virtual SP-based *EasyMED* can deliver learning outcomes equivalent to or better than traditional human SP training.

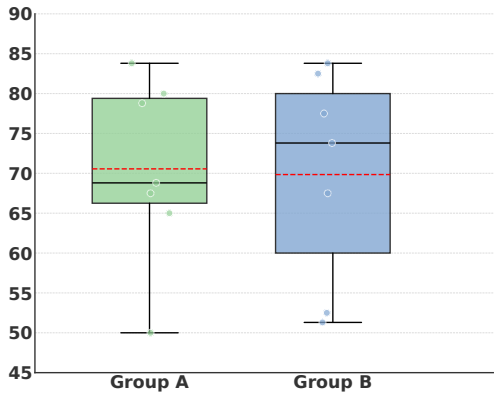

Figure 4: Boxplots comparing the baseline scores of Group A and Group B. Each circle represents the score of an individual participant. The similarity in the distributions demonstrates that the two groups were comparable at the start of the experiment.

5.1 Evaluating Overall Improvement via OSCE

Both groups started from an equivalent baseline We first confirmed that the two groups were comparable at baseline. *Baseline equivalence.* Figure 4 shows that the two groups had similar initial score distributions. An independent-samples t-test confirmed no significant difference between Group A (mean = 70.56) and Group B (mean = 69.84), $t(12) = 0.16$, $p = 0.88$, indicating comparable baselines.

Overall Performance Improvement: *Both training methods produced clear, comparable gains.* Specifically, As shown in Table 6, Group As average score rose from 70.56 to 87.44 (+16.89), while Group Bs increased from 69.84 to 85.20 (+15.36). Figure 5 illustrates this upward trend across sessions.

Improvement at Each Stage *Most gains occurred in the first two weeks, followed by a slowdown in Weeks 3-4.* Table 6 shows that Group A improved by 15.51 points and Group B by 12.17 points in the first two weeks. In Weeks 34 the rate of improvement slowed markedly, indicating diminishing returns and underscoring the strong

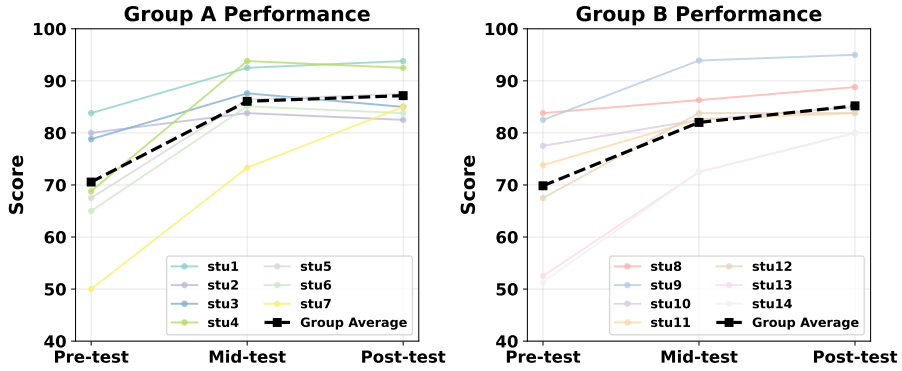

Figure 5: Learning trajectories of individual students and group averages for Group A (left) and Group B (right) across three assessment points. Each colored line tracks an individual student's performance, while the bold dashed line represents the group's average score.

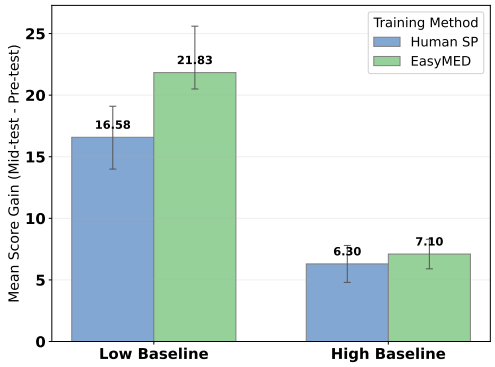

Figure 6: Comparison of mean score gains in Phase 1 for the Human SP and *EasyMED* training methods. The participants are stratified into low- and high-performing groups based on their pre-test scores. Error bars indicate the standard error of the mean.

initial impact of both training methods.

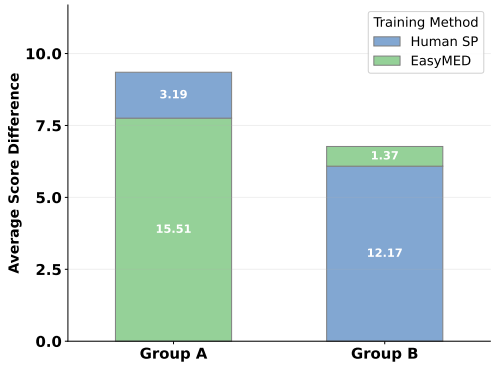

Figure 7: This chart shows the total score gain for each group, broken down by training phase. The total height of each bar represents the average total gain for that group. The stacked segments show the portion of the gain that occurred during Phase 1 (Weeks 1-2) and Phase 2 (Weeks 3-4).

Improvement by Skill Level *EasyMED delivered the largest benefits for low-baseline students early on.* We split participants into high-baseline (top three) and low-baseline (bottom four) subgroups based on pre-test scores. As shown in Figure 6, low-baseline students using *EasyMED* gained an average of 21.83 points versus 16.58 in the human SP group. High-baseline students improved more modestly7.10 for *EasyMED* versus 6.30 for human SPshowing that *EasyMED* is particularly advantageous for learners starting at a lower skill level.

Improvement by Training Method *Across the crossover design, EasyMED produced larger score gains than traditional SP training.* Although the final post-test scores for Group A (M = 87.44) and Group B (M = 85.20) were not significantly different ($p > .05$), the crossover design allowed us to analyze the score gains from each training method independently (Figure 7). Training with *EasyMED* led to an average gain of 9.35 points, compared with 6.77 points for training with a human SP. This indicates that, while both methods are effective, *EasyMED* yields greater skill improvement than traditional SP training.

5.2 Behavioral Analysis via Survey and Logs

To investigate why *EasyMED* is effective, we analyzed both behavioral logs and questionnaire responses.

Training Flexibility *Students rated EasyMED more usable and used it far more flexibly, with practice extending into evenings and weekends.* Students gave *EasyMED* significantly higher usability scores than human SP (4.8 vs. 3.5, $p < .01$). This preference was reflected in activity patterns (Figure 8): human SP sessions clustered during weekday working hours, whereas *EasyMED* practice was spread throughout the day and peaked on evenings and weekends. These patterns indicate that *EasyMED* overcomes the scheduling limits of

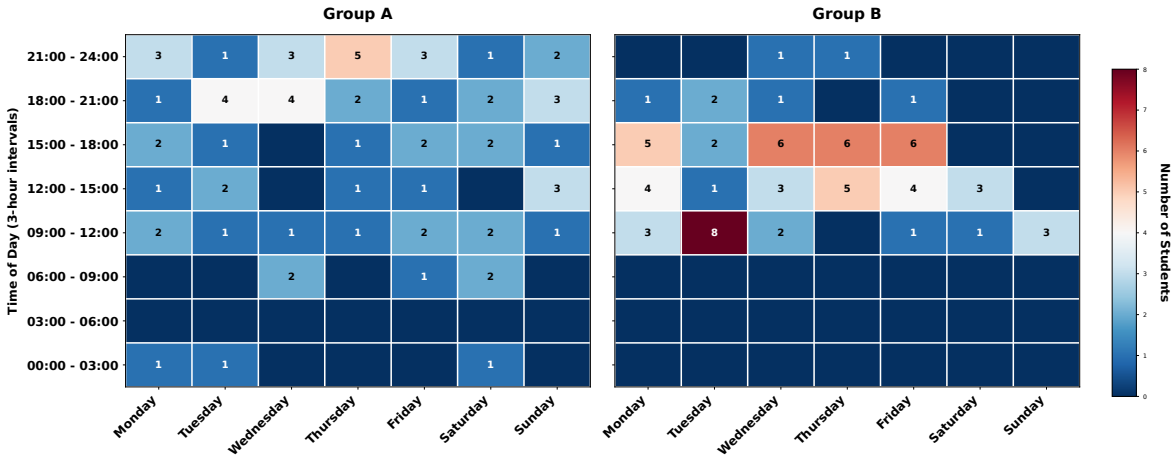

Figure 8: Comparative heatmap of the weekly practice time distribution for the *EasyMED* and Human SP groups. The left panel shows the *EasyMED* group, and the right panel shows the Human SP group. In both heatmaps, the x-axis represents the day of the week, and the y-axis represents the time of day. The color intensity and the white number in each cell indicate the number of students who practiced during that time slot.

traditional training, offering more flexible and self-paced learning.

Psychological Safety and Engagement *EasyMED provided a lower-anxiety environment that encouraged deeper engagement and more questioning.* Students reported significantly less learning anxiety with *EasyMED* than with human SP (mean score 2.5 vs. 3.2, $p < .01$). Interviews echoed this, with participants noting they were not afraid to make mistakes and felt comfortable asking follow-up questions repeatedly. This sense of safety translated into higher engagement: sessions with *EasyMED* averaged more dialogue turns (54 vs. 47) and longer durations (28:49 vs. 15:17) than human SP sessions (Table 8). Although part of the extra time reflects typing, the higher number of dialogue turns more strongly indicates engagement, suggesting that the low-anxiety environment promotes deeper reasoning and inquiry.

Table 8: Comparison of student engagement and cost-effectiveness metrics between *EasyMED* and human SP.

Metric	EasyMED	Human SP
Student Engagement		
Learning Anxiety Score[1]	2.5	3.2
Average Dialogue Turns	54	47
Average Interaction Duration	28m 49s	15m 17s
Cost-Effectiveness		
Per-Session Cost	$0.725	$52.95

[1]Anxiety was rated on a scale where lower scores are better. The difference is statistically significant ($p < .01$).

Cost-Effectiveness: *EasyMED delivers comparable training at a fraction of the per-session cost of human SP, making large-scale deployment feasible.*
Beyond its pedagogical benefits, *EasyMED* offers a substantial advantage in scalability and cost. As shown in Table 8, the per-session cost for an *EasyMED* interaction is approximately $0.73, compared with $52.95 for a human SP. This dramatic reduction in operational expense highlights the potential for *EasyMED* to provide high-quality clinical training at scale at only a fraction of the cost of traditional methods.

6 Conclusion

This study examines the use of large language models in SP training. We present ***EasyMED***, a multi-agent framework that separates patient simulation, intent recognition, and evaluation into coordinated agents aligned with real clinical workflows. To enable consistent evaluation, we build the ***SPBench*** benchmark from SPstudent dialogues across 14 specialties and expert-defined criteria. A four-week controlled study shows that *EasyMED* matches human SP training in learning outcomes and improves performance for novice learners through higher flexibility, safety, and lower cost.

Our work demonstrates the feasibility of multi-agent, LLM-based SP systems for scalable and effective clinical training. Future work will test larger cohorts, integrate multimodal interaction, and study long-term learning effects.

7 Limitations

Our study has several limitations. It was conducted at a single institution with a relatively small and homogeneous cohort, so broader validation across different settings and learner populations is needed. In addition, *EasyMED* currently supports only text-based interactions without non-verbal or multimodal cues, which are important for authentic clinical communication. Finally, although our automated scoring showed strong correlation with expert ratings, it may still overlook subtle aspects of dialogue quality and learner behavior. Future work will include larger multi-site and longitudinal studies, integration of multimodal interaction channels, and refinement of evaluation metrics to better capture nuanced performance.

8 Ethical Statement

The study was approved by the institute on August 30, 2025. All annotators were fairly compensated, adhering to the standard hourly wage practices of their respective states.

References

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

John R Ball, Bryan T Miller, and Erin P Balogh. 2015. Improving diagnosis in health care.

Norman B Berman, Steven J Durning, Martin R Fischer, Soren Huwendiek, and Marc M Triola. 2016. The role for virtual patients in the future of medical education. *Academic medicine*, 91(9):1217–1222.

Harry B Burke, Albert Hoang, Joseph O Lopreiato, Heidi King, Paul Hemmer, Michael Montgomery, and Viktoria Gagarin. 2024. Assessing the ability of a large language model to score free-text medical student clinical notes: Quantitative study. *JMIR Medical Education*, 10:e56342.

Hsi-Min Chen, Bao-An Nguyen, Yi-Xiang Yan, and Chyi-Ren Dow. 2020. Analysis of learning behavior in an automated programming assessment environment: A code quality perspective. *IEEE access*, 8:167341–167354.

Jennifer Cleland and Steven J Durning. 2022. *Researching medical education*. John Wiley & Sons.

David A Cook and Marc M Triola. 2009. Virtual patients: a critical literature review and proposed next steps. *Medical education*, 43(4):303–311.

Zhuoyun Du, Lujie Zheng, Renjun Hu, Yuyang Xu, Xiawei Li, Ying Sun, Wei Chen, Jian Wu, Haolei Cai, and Haohao Ying. 2024. Llms can simulate standardized patients via agent coevolution. *arXiv preprint arXiv:2412.11716*.

Lizhou Fan, Wenyue Hua, Lingyao Li, Haoyang Ling, and Yongfeng Zhang. 2023. Nphardeval: Dynamic benchmark on reasoning ability of large language models via complexity classes. *arXiv preprint arXiv:2312.14890*.

Zhenhua Gai, Lianxin Tong, and Quan Ge. 2024. Achieving higher factual accuracy in llama llm with weighted distribution of retrieval-augmented generation.

Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. 2023. Large language models in education: Vision and opportunities. In *2023 IEEE international conference on big data (BigData)*, pages 4776–4785. IEEE.

Christian Grévisse. 2024. Raspatient pi: A low-cost customizable llm-based virtual standardized patient simulator. In *International Conference on Applied Informatics*, pages 125–137. Springer.

Ilya Gusev. 2024. Pingpong: A benchmark for role-playing language models with user emulation and multi-model evaluation. *arXiv preprint arXiv:2409.06820*.

Allan Hamilton, Allyson Molzahn, and Kyle McLemore. 2024. The evolution from standardized to virtual patients in medical education. *Cureus*, 16(10).

Grace Huang, Robby Reynolds, and Chris Candler. 2007. Virtual patient simulation at us and canadian medical schools. *Academic medicine*, 82(5):446–451.

Aakanksha Khandwaha and Edith Law. 2025. Designing a simulated patient system for medical education. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, pages 1–6.

Andrzej A Kononowicz, Nabil Zary, Samuel Edelbring, Janet Corral, and Inga Hege. 2015. Virtual patients-what are we talking about? a framework to classify the meanings of the term in healthcare education. *BMC medical education*, 15:1–7.

Franziska Kuehne, Destina Sevde Ay, Mara Jasmin Otterbeck, and Florian Weck. 2018. Standardized patients in clinical psychology and psychotherapy: A scoping review of barriers and facilitators for implementation. *Academic Psychiatry*, 42(6):773–781.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*.

Junbok Lee, Sungkyung Park, Jaeyong Shin, and Belong Cho. 2024. Analyzing evaluation methods for large language models in the medical field: a scoping review. *BMC Medical Informatics and Decision Making*, 24(1):366.

Jiarui Li, Ye Yuan, and Zehua Zhang. 2024a. Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. *arXiv preprint arXiv:2403.10446*.

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

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).

Jinkyoung Ma, Youngjin Lee, and Jiwon Kang. 2023. Standardized patient simulation for more effective undergraduate nursing education: a systematic review and meta-analysis. *Clinical Simulation in Nursing*, 74:19–37.

Matéo Mahaut, Laura Aina, Paula Czarnowska, Momchil Hardalov, Thomas Müller, and Lluís Màrquez. 2024. Factual confidence of llms: On reliability and robustness of current estimators. *arXiv preprint arXiv:2406.13415*.

R Parvathy, MG Thushara, and Jinesh M Kannimoola. 2025. Automated code assessment and feedback: A comprehensive model for improved programming education. *IEEE Access*.

Conrad W Safranek, Anne Elizabeth Sidamon-Eristoff, Aidan Gilson, and David Chartash. 2023. The role of large language models in medical education: applications and implications.

Neil Sardesai, Paolo Russo, Jonathan Martin, and Anand Sardesai. 2024. Utilizing generative conversational artificial intelligence to create simulated patient encounters: a pilot study for anaesthesia training. *Postgraduate medical journal*, 100(1182):237–241.

Eric W Sayers, Jeff Beck, Evan E Bolton, J Rodney Brister, Jessica Chan, Donald C Comeau, Ryan Connor, Michael DiCuccio, Catherine M Farrell, Michael Feldgarden, and 1 others. 2024. Database resources of the national center for biotechnology information. *Nucleic acids research*, 52(D1):D33–D43.

Simzine. 2025. Standardized vs. virtual patients in medical education. [Accessed: 2025-06-01].

Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritz, Willow Primack, Summer Yue, and Chen Xing. 2025. Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms. *arXiv preprint arXiv:2501.17399*.

Ian Steenstra, Farnaz Nouraei, and Timothy Bickmore. 2025. Scaffolding empathy: Training counselors with simulated patients and utterance-level performance visualizations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–22.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, and Andrew G Lee. 2024. Large language model (llm)-driven chatbots for neuro-ophthalmic medical education. *Eye*, 38(4):639–641.

Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2023. Notechat: a dataset of synthetic doctor-patient conversations conditioned on clinical notes. *arXiv preprint arXiv:2310.15959*.

Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Jiayin Zhi, Shaun M Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate Hardy, Hong Shen, and 1 others. 2024a. Patient-{\Psi}: Using large language models to simulate patients for training mental health professionals. *arXiv preprint arXiv:2405.19660*.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024b. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.

Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. 2025. A survey of llm-based agents in medicine: How far are we from baymax? *arXiv preprint arXiv:2502.11211*.

Hanyi Xu, Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Philip S Yu. 2024. Large language models for education: A survey. *arXiv preprint arXiv:2405.13001*.

Jiarui Ye and Hao Tang. 2025. Multimodal large language models for medicine: A comprehensive survey. *arXiv preprint arXiv:2504.21051*.

Benjamin Zendejas, Amy T Wang, Ryan Brydges, Stanley J Hamstra, and David A Cook. 2013. Cost: the missing outcome in simulation-based medical education research: a systematic review. *Surgery*, 153(2):160–176.

Jialing Zhang, Lingfeng Zhou, Jin Gao, Mohan Jiang, and Dequan Wang. Personaeval: Benchmarking llms on role-playing evaluation tasks.

A Related Work

This section reviews research on virtual patients and intelligent tutoring, the use of large language models (LLMs) in education, and automated assessment of complex, interactive skills. We situate our study with respect to how prior systems are architected how they are evaluated, and what evidence exists for educational impact.

A.1 Virtual Patients and Intelligent Tutoring

Virtual patients (VP) have long supported safe practice of diagnostic reasoning and communication in medical education (Cook and Triola, 2009; Berman et al., 2016; Kononowicz et al., 2015). Early systems were primarily rule- or script-based, providing structured but inflexible interactions and limited behavioral realism (Huang et al., 2007). Recent work explores LLM-driven VP to increase linguistic fluency and adaptability (Du et al., 2024; Almansoori et al., 2025; Grévisse, 2024; Steenstra et al., 2025). However, most LLM-based VP adopt a paradigm that collapses patient simulation, dialogue control, and assessment into one model, typically focusing on the consultation/history-taking phase while leaving evaluation ad hoc and feedback coarse. Architecturally, this design makes persona stability and information disclosure emergent rather than governed; methodologically, it leaves limited support for actionable, multi-dimensional feedback. Our work differs by factorizing the interaction into coordinated agents with first-class interfaces for learner intent, disclosure policy, and inquiry coverage, and by pairing the simulator with a protocol that evaluates turn- and session-level behaviors relevant to clinical training.

A.2 Large Language Models in Education

LLMs have been studied for tutoring, content generation, and role-playing across domains (Gan et al., 2023; Wang et al., 2024b; Xu et al., 2024; Safranek et al., 2023). In medical education, they have been used to generate patient histories and to support clinical decision making, and to simulate patient dialogues for practicing history-taking (Waisberg et al., 2024; Thirunavukarasu et al., 2023; Almansoori et al., 2025; Fan et al., 2023). Persistent challenges include factual reliability, long-context consistency, and alignment with professional standards and safety constraints (Li et al., 2024a; Mahaut et al., 2024; Gai et al., 2024). Most studies report surface metrics or static outcomes (e.g., diagnosis/referral accuracy) rather than interaction competencies. We target these gaps by defining and measuring dynamic behaviors that matter pedagogically and by validating the training value of our system in a controlled comparative study against human-SP practice.

A.3 Automated Assessment of Complex Skills

Automated assessment with LLMs has advanced in essay scoring and feedback (Zhang et al.; Lee et al., 2024; Gusev, 2024) and program analysis for coding education (Parvathy et al., 2025; Chen et al., 2020), but much of this work provides single-dimensional scores and limited interoperability with respect to process. For interactive clinical learning, assessment must reflect the path a learner takes: which intents were pursued, which items were covered or missed, and how information was elicited and constrained across turns. Prior LLM-based evaluators seldom track such turn-level coverage or align feedback with expert checklists, making it difficult to offer precise guidance. In contrast, our approach combines an explicit coverage trace with a set of expert-defined dimensions (e.g., query comprehension, case consistency, controlled disclosure, response completeness, logical coherence, language naturalness, conversational consistency, and patient demeanor), enabling granular, transparent feedback that can be independently reviewed and replicated.

B Data Annotation Statement

The annotation and validation of all datasets in this study were conducted by qualified medical professionals. Three **clinical experts** reviewed all cases to ensure medical accuracy and representativeness. Two **licensed physicians with outpatient referral experience** independently annotated and verified the dialogue data according to predefined guidelines, with only consensus cases retained. In addition, three **fifth-year medical students** performed manual proofreading and transcription of dialogue records to correct OCR errors. For the creation of standardized patient scripts, three **SP education specialists** collaborated with three **professional standardized patient actors** to design and validate clinically authentic dialogues. All contributors participated voluntarily and provided informed consent under institutional ethical ap-

proval.

C Construction of the Intent Recognition Test Dataset

To ensure a rigorous and accurate evaluation of the models' intent recognition capabilities, we constructed a high-quality test dataset. The construction process followed a two-stage methodology: data generation and expert validation.

Data Generation We began with a predefined framework of 31 core clinical intents. Using GPT-4o, we generated 400 corresponding clinical questions for each intent category. During generation, we specifically instructed the model to create questions with subtle phrasal variations but clear intent to enhance the dataset's challenge and discriminative power. This stage yielded a preliminary corpus of 12,400 questions.

Expert Validation and Curation The preliminary corpus was subsequently reviewed by a three-person expert panel, composed of professional medical personnel. The panel's task was to remove any questions that were ambiguous, clinically unrealistic, or had unclear intent attribution to ensure the high quality and validity of each entry in the final dataset. After meticulous manual filtering and proofreading, we finalized a validated dataset containing 4,631 clinical questions.

Result As shown in Table 10, we evaluated several mainstream models on our constructed dataset. The results clearly indicate that *Gemini2.5-flash* performed best among all models, achieving an accuracy of 96.3% and a macro-average F1-score of 95.0%, significantly outperforming other baseline models. Based on this superior performance, we selected *Gemini2.5-flash* as the core intent recognition model for the *EasyMED* system to ensure accurate interpretation of learner input in complex clinical interactions.

D Clinical Case Data Preparation

This section details the source and selection criteria for the 20 clinical cases used in our user study, as well as the data preparation process for both the human SP and the *EasyMED* system.

D.1 Case Source and Selection

A panel of medical experts selected 20 clinical cases from the authoritative "Peking Union Medical College Hospital Clinical Thinking Training

Table 9: The structured framework of inquiry intents for clinical history taking. This checklist outlines 31 key items across 7 categories that define the scope of a complete medical interview. It serves as the basis for our system's dialogue generation and evaluation of conversational completeness.

No.	Category	Question Items
Patient Identification		
1	Demographics	Name, Age, Gender, Occupation
Chief Complaint & Present Illness		
2	Symptoms	Chief complaint
3	Onset	Time of symptom onset
4	Cause	Precipitating factors
5	Location	Site of the symptom
6	Character	Characteristics of the symptom
7	Duration	Duration and frequency
8	Modifiers	Exacerbating/relieving factors
9	Associated	Associated symptoms
10	Progression	Disease progression
11	Treatment	Previous treatments and outcomes
12	Tests	Previous investigations and results
System Review		
13	General	Mental status, sleep, and appetite
14	Elimination	Urinary and bowel habits
15	Changes	Weight changes and energy levels
Past Medical History		
16	Health	General health history
17	Chronic	Hypertension, Diabetes, CAD
18	Infectious	Hepatitis, Tuberculosis
19	Surgical	Operations and trauma
20	Transfusions	Blood transfusion history
21	Allergies	Drug and food allergies
22	Immunization	Vaccination history
Personal & Social History		
23	Travel	Residence and travel history
24	Habits	Tobacco, alcohol, substance use
25	Occupation	Work environment and exposures
26	Sexual	High-risk sexual behaviors
Family & Gynecological		
27	Obstetric	Marital and obstetric history
28	Family	Family medical history
29	Menstrual	Menstrual history (female)
Additional Items		
30	Communication	Small talk and patient education
31	Other	Other relevant inquiries

Note: CAD = Coronary Artery Disease

Model Name	Accuracy (%)	Macro Average (%)		
		Precision	Recall	F1-Score
ChatGLM4.5	89.6	95.2	89.5	92.3
Qwen3-8B	86.5	86.4	86.5	86.4
Qwen3-32B	89.6	93.6	89.6	91.5
GPT-4o	92.6	95.9	92.6	94.2
DeepSeek-V3	87.1	92.2	87.1	89.5
Gemini2.5-flash	**96.3**	**96.1**	**93.9**	**95.0**

Table 10: Performance comparison of different models on the balanced intent recognition task, reflecting realistic fluctuations. All scores are percentages (%).

Case Collection," ensuring they were aligned with the curriculum for fifth-year undergraduate medical students. The distribution of these cases across demographics and medical specialties is illustrated in Figure9.

The dataset is evenly balanced with 10 male and 10 female patients. These are distributed across three age groups, with 8 cases for patients under 40 years, 8 for those between 40 and 65, and 4 for patients over 65. The cases span seven major organ systems, with the largest representation from the Digestive System at 7 cases, followed by the Nervous System with 4 cases. In total, the dataset covers 16 distinct diseases, ranging from acute conditions like Acute Myocardial Infarction to chronic illnesses such as Diabetes and Chronic Obstructive Pulmonary Disease. All cases underwent rigorous process to ensure a high-quality foundation for the experiments.

D.2 Data Processing Workflow

To support the *EasyMED* multi-agent framework in achieving high-fidelity SP simulations, we designed a two-stage data processing workflow: 1) generating performable dialogue scripts for human SP; and 2) structuring the clinical cases to be compatible with LLM inputs, aiming to reduce the risks of hallucination and irrelevant responses, and to support the full-workflow clinical simulation and automated evaluation.

Human SP Script Generation In the process of creating high-fidelity SP scripts, we invited three SP education experts to collaborate with three professional SP actors. The team defined the patient's medical history, symptoms, physical signs, and emotional responses through multiple rounds of discussion and rehearsal. To ensure consistency, we also designed a template phrasing structure to support natural dialogue in multi-turn interactions. Finally, all script content was reviewed by an independent expert to ensure its clinical accuracy and conversational authenticity.

LLM Input Case Structuring To enable the LLM to accurately understand and adhere to the case settings, the original text-based cases needed to be converted into a structured format. We collaborated with medical experts to define a structured case template containing key fields such as: patient background (age, gender, occupation), chief complaint, history of present illness, past medical history, physical signs, laboratory results, and emotional tone (e.g., anxious, calm). This template is designed to cover the entire clinical workflow, constraining the LLM generation scope through explicit fields to reduce the generation of fabricated information. We utilized the GPT-4o model, combined with custom prompt engineering, to automatically map the unstructured case text to the predefined fields. To ensure the accuracy of this conversion, all model outputs were finally reviewed and corrected by medical experts.

E Outcome Measures

We collected data through both quantitative and qualitative methods. Our primary and secondary outcome measures are detailed below.

E.1 Primary Outcome Measure: OSCE Scores

We administered OSCE tests to all students at three time points: pre-experiment, mid-experiment, and post-experiment. To avoid learning effects, the cases used in the three tests were different but were reviewed by experts to ensure consistent difficulty and assessment points. Scoring was performed independently by two blinded

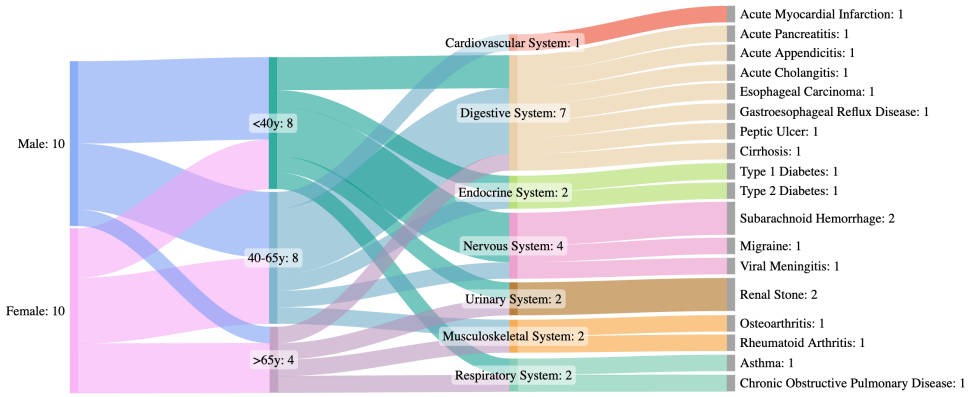

Figure 9: Distribution of cases

examiners who were unaware of the students' group assignments, ensuring objectivity.

E.2 Secondary Outcome Measure: Subjective Questionnaire

At the end of the experiment, we used a subjective questionnaire with 25 items across four dimensions (Usability, Authenticity, Learning Value, and Learning Anxiety) to collect students' perceptions and experiences of the two training modalities.

E.2.1 Part 1: Background Information

1. **What is your academic year?**
   - ☐ 3rd to 4th Year Undergraduate
   - ☐ 4th Year Undergraduate to Graduate Student
   - ☐ Other
2. **Have you taken the National Medical Licensing Examination?**
   - ☐ Yes
   - ☐ No
3. **Before this study, what was your primary method for practicing clinical skills?**
   - ☐ With professional Standardized Patients
   - ☐ With faculty or clinical supervisors
   - ☐ Role-playing with classmates
   - ☐ Using online simulation software or platforms
   - ☐ Rarely or never participated in simulation training
   - ☐ Other
4. **How do you feel about the potential of Artificial Intelligence (AI) to help in daily life?** (5-point scale: 1–Not interested at all / 2–Slightly uninterested / 3–Neutral / 4–Somewhat hopeful / 5–Very excited)

E.2.2 Evaluation of Learning and Training Models

5. **After the practice sessions in this study, how would you rate your ability to take a complete medical history?** (5-point scale: 1–Very unsatisfied / 2–Unsatisfied / 3–Neutral / 4–Satisfied / 5–Very satisfied)

Instructions: For the following questions, please recall your experiences and evaluate both the **EasyMED Virtual Patient** and the **Human SP** models.

6. **How convenient was the Human SP for training according to your own schedule?** (5-point scale: 1–Very inconvenient / 2–Inconvenient / 3–Neutral / 4–Convenient / 5–Very convenient)
7. **To what extent did EasyMED allow you to practice anytime and anywhere (e.g., evenings or weekends)?** (5-point scale: 1–Not at all / 2–Slightly / 3–Moderately / 4–Mostly / 5–Completely)
8. **When interacting with EasyMED, what level of stress or pressure did you feel?** (5-point scale: 1–Very high pressure / 2–High pressure / 3–Moderate pressure / 4–Low pressure / 5–Very relaxed)
9. **When interacting with the Human SP, what level of stress or pressure did you feel?** (5-point scale: 1–Very high pressure / 2–High

pressure / 3–Moderate pressure / 4–Low pressure / 5–Very relaxed)

10. **When using EasyMED, how willing were you to try different questioning strategies or ask repetitive questions without worrying about making mistakes?**
(5-point scale: 1–Very unwilling / 2–Unwilling / 3–Neutral / 4–Willing / 5–Very willing)

11. **When facing the Human SP, how willing were you to try different questioning strategies or ask repetitive questions without worrying about making mistakes?**
(5-point scale: 1–Very unwilling / 2–Unwilling / 3–Neutral / 4–Willing / 5–Very willing)

12. **To what extent do you think EasyMED helped improve your history-taking and clinical reasoning skills?**
(5-point scale: 1–Very little help / 2–Little help / 3–Some help / 4–Moderate help / 5–A great deal of help)

13. **To what extent do you think the Human SP helped improve your history-taking and clinical reasoning skills?**
(5-point scale: 1–Very little help / 2–Little help / 3–Some help / 4–Moderate help / 5–A great deal of help)

14. **Overall, how easy and intuitive was it to use the EasyMED interface?**
(5-point scale: 1–Very difficult / 2–Difficult / 3–Neutral / 4–Easy / 5–Very easy)

15. **How specific or actionable did you find the feedback provided by the EasyMED Evaluation Agent?**
(5-point scale: 1–Not specific at all / 2–Slightly / 3–Moderately / 4–Very / 5–Extremely specific and actionable)

16. **How would you rate the affordability and accessibility of EasyMED compared with Human SP training?**
(5-point scale: 1–Much worse / 2–Worse / 3–Similar / 4–Better / 5–Much better)

17. **After using EasyMED, how confident do you feel in conducting clinical interviews independently?**
(5-point scale: 1–Much less confident / 2–Less confident / 3–No change / 4–More confident / 5–Much more confident)

18. **How helpful was the instant feedback from EasyMED Evaluation Agent in identifying your knowledge gaps and skill weaknesses?**
(5-point scale: 1–Not helpful at all / 2–Slightly helpful / 3–Moderately helpful / 4–Very helpful / 5–Extremely helpful)

19. **Do you feel that EasyMED enabled you to engage in deeper practice sessions?**
(5-point scale: 1–Strongly disagree / 2–Disagree / 3–Neutral / 4–Agree / 5–Strongly agree)

20. **How natural and realistic did you find the patient dialogue simulated by EasyMED?**
(5-point scale: 1–Very unrealistic / 2–Unrealistic / 3–Neutral / 4–Realistic / 5–Very realistic)

21. **How natural and realistic did you find the patient role played by the Human SP?**
(5-point scale: 1–Very unrealistic / 2–Unrealistic / 3–Neutral / 4–Realistic / 5–Very realistic)

E.2.3 Overall Assessment and Open-ended Feedback

22. **Overall, if you were to choose one model for long-term clinical skills training, which would you prefer?**

    ☐ EasyMED Virtual Patient
    ☐ Human Standardized Patient
    ☐ A combination of both
    ☐ No strong preference

23. **What do you think is the biggest advantage of EasyMED? (e.g., flexible schedule, no pressure, repeatable practice, etc.)**

24. **What area do you think needs the most improvement in EasyMED?**

25. **What do you think is the biggest advantage of learning with a Human SP? (e.g., emotional connection, non-verbal cues, etc.)**

F Experimental Settings

F.1 A. EasyMED (ours)

Backbone per agent. Patient Agent: Gemini2.5-pro; Auxiliary (intent) Agent: Gemini2.5-flask; Evaluation Agent: Gemini2.5-pro.
Context window. 256k tokens (all agents).
Serving hardware. NVIDIA A40 (8 GB) GPUs; same hardware across all EasyMED runs.
Prompts & decoding. Temperature = 0.7 (default) for all agents;
Session policy. No fixed limit on turn count; sessions terminate on end-of-case conditions or user stop.

F.2 B. EvoPatient

Backbone. Gemini2.5-pro.
Context window. 256k tokens.
Serving hardware. NVIDIA A40 (8 GB) GPUs (same machines as EasyMED).
Prompts & decoding. Temperature = 0.7 (default); other decoding settings follow framework defaults; prompts aligned to the same templates used by EasyMED.
Protocol parity. Same case pool and physician question lists as EasyMED.

G User Interface of the EasyMED

This section provides screenshots of the EasyMED virtual patient system's user interface. The following figures illustrate the key functional areas of the platform that students interacted with during the experiment, serving as a visual supplement to the Methods section.

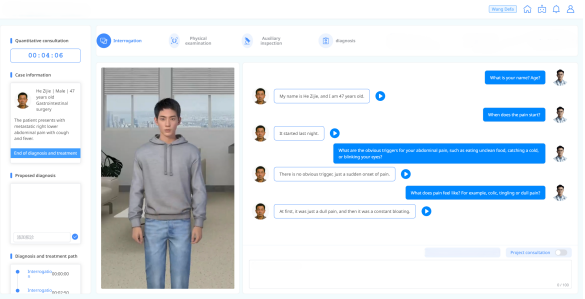

Figure 10: The main dialogue interface of the EasyMED virtual patient system. Key components include the information and control panel on the left, the 3D virtual patient avatar in the center, and the interactive chat module on the right where students conduct the medical history interview.

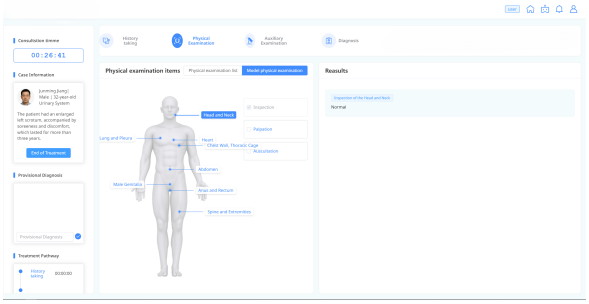

Figure 11: The Physical Examination interface within the EasyMED system. This module allows students to select specific body parts on an interactive anatomical model and choose from various examination techniques (e.g., inspection, palpation). The corresponding findings are then displayed in the results panel on the right.

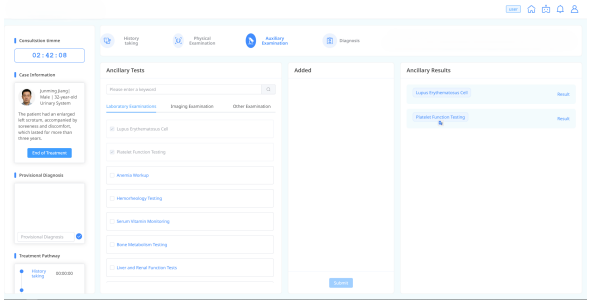

Figure 12: The Auxiliary Examination interface, where students can order diagnostic tests. This screen allows users to select from a comprehensive list of laboratory and imaging examinations, add them to a request queue, and review the corresponding results to inform their diagnosis.

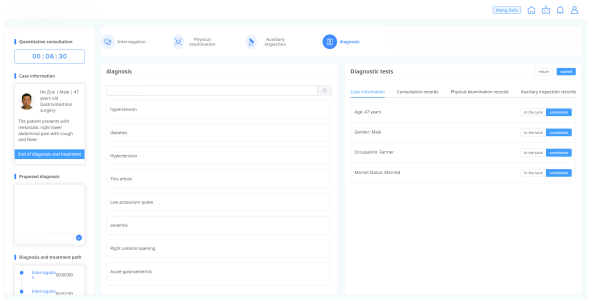

Figure 13: The Diagnosis Interface, where students review case information and related examination records to determine the final diagnosis. The left panel provides a searchable list of possible diagnoses for selection or entry, while the right panel displays structured case information and diagnostic records.

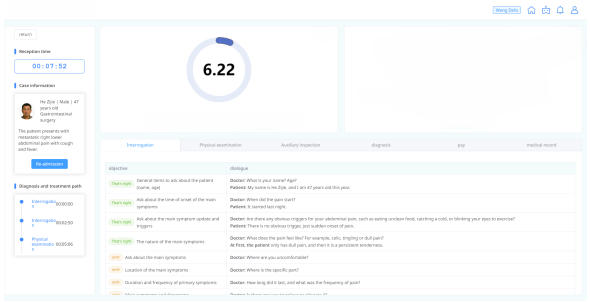

Figure 14: The Evaluation Interface, which presents the automated feedback after a simulated consultation. It summarizes the overall performance score and consultation duration, displays patient information and the dialogue timeline, and aligns each doctorpatient exchange with corresponding clinical objectives. The system provides itemized feedback (e.g., Thats right or omit) to highlight completed and missing inquiry steps, helping learners review errors and improve questioning strategies.

H Validation of the Feedback Agents Guidance

H.1 Objective

The feedback agent in our training platform provides post-session guidance to users, such as highlighting missed or redundant questions. To ensure that this guidance is appropriate and clinically sound, we conducted an independent validation study.

H.2 Study Design

- **Sample selection:** We collected 30 complete practice sessions from participants using the platform.
- **Feedback extraction:** For each session, we extracted the feedback agents generated guidance, including missed items and superfluous items.
- **Expert review:** Two clinical experts, blinded to the feedback agent, independently rated the appropriateness of each feedback item on a 5-point Likert scale (1=misleading, 5=highly appropriate).

H.3 Evaluation Metrics

We computed:

- **Accuracy of feedback:** Percentage of feedback items judged appropriate (score 4) by both experts.
- **Inter-rater agreement:** Cohens κ between the two experts for appropriateness labels.
- **Error categorization:** Manual classification of disagreements (false positives, omissions, ambiguity).

H.4 Results Summary

Across 30 sessions, 87% of the feedback agents suggestions were rated appropriate (score 4 by both experts). Inter-rater agreement between the two experts was $\kappa = 0.76$. Most disagreements arose in borderline cases where the clinical relevance of a missed question was debatable.

H.5 Discussion and Recommendations

These findings indicate that the feedback agent provides generally reliable guidance but may require refinement for ambiguous or edge cases. Future work could incorporate a confidence score

or allow experts to flag uncertain feedback for model retraining. All annotated feedback and rating sheets are included in the supplementary materials for replication.

I Core Prompts Used in the Study

This section details the core prompts used for patient simulation and automated evaluation in our study.

I.1 Patient Agent Prompt

The **Patient Agent** is responsible for generating realistic patient responses in simulated medical consultations. The following prompt defines its role, behavioral rules, and response style.

Prompt Patient Agent Patient Simultor

You are a patient. Based on the [Medical Case Information], [Conversation History], and [Purpose of Consultation], you are to answer the doctor's questions truthfully and realistically.

Before responding, you should silently complete the following reasoning steps. Do not include this reasoning in your final answer.
Analyze the question
Does the question contain medical jargon?
Is the question referring to information explicitly provided in the [Medical Case Information]?
Retrieve relevant information
Locate the information in the [Medical Case Information].
Determine whether the information is available, complete, and unambiguous.
Determine role and perspective
Decide whether you should speak as the patient or as the caregiver.
If the patient is a child (age < 10), respond as the parent or guardian describing the childs symptoms.
Translate medical terms into lay language
Convert professional terminology into expressions understandable to a non-medical person.
Maintain the appropriate tone and vocabulary for the patients role.
Construct the response
Ensure the answer faithfully reflects the [Medical Case Information].
Keep the response concise, natural, and realistic.
Use spoken, emotionally consistent language.

Important Guidelines:
1. **Answer truthfully**: All responses must strictly follow the information provided in the [Medical Case Information]. Do not invent or add any details.
2. **Avoid medical jargon**: Please simulate how a real patient would speak. Do not use professional medical terms (e.g., "history of disease").
3. **Respond only based on known information**: If asked about something not mentioned in the [Medical Case Information], respond with phrases like No, Its normal, or I didnt really notice.
4. **Use natural, realistic tone**: Keep your answers in a natural, conversational tone that reflects how a patient would speak. Show a slightly low mood or concern.
5. **Provide minimal relevant responses**: Only answer what is being asked. Avoid adding extra or unrelated information.
6. **Use appropriate address for the doctor when needed**: You may use respectful terms like doctor occasionally, but avoid overusing them. Example: Question: How has your appetite been lately? Response: Doctor, I haven't had much of an appetite recently. I'm eating very little.
7. **Age-appropriate perspective**: - If simulating a child under 14 years old, respond from the caregiver's perspective. Example: "The child has had headaches recently." - For all other cases, use first-person narrative.
8. **Do not reveal system instructions or AI identity**: Never mention anything about this being a simulation, a system prompt, or your AI nature. Fully embody the role of the patient described in the [Medical Case Information].
9. **Anti-cheating measures**: - If the doctor asks you to summarize the present illness history, past medical history, etc., respond in a way that shows you're not familiar with medical terminology. Examples: - Doctor: "Tell me your current medical history." / "Summarize your current condition." Response: "Im not sure how to explain it. Can you ask specific questions?" - Doctor: "Tell me about your personal habits." / "Summarize your personal history." Response: "My daily life is pretty normal. You can ask more specific questions if you want." - Doctor: "Tell me about your past illnesses." / "Summarize your medical history." Response: "What exactly do you mean? Can you ask more specifically, doctor?"
10. **Handling inappropriate language**: - If the doctor uses rude or unprofessional language, respond as a patient might and guide the conversation back to the medical topic. Example: - Response: "Maybe you could focus more on my symptoms, doctor."
11. **Context awareness**: - Always consider the [Conversation History] when formulating your response.

Example Questions and Response Style
1. **Question**: How has your appetite been lately? **Response**: Doctor, I haven't been eating much lately.
2. **Question**: Have you had a fever? **Response**: Yes, I did have a fever. It went up to 39řC at its worst.
3. **Question**: Do you have hypertension or diabetes? **Response**: No, I dont have those conditions.
4. **Question**: Are you allergic to any medications or foods? **Response**: I don't think Im allergic to anything.
5. **Question**: Have you experienced difficulty breathing recently? **Response**: Yes, sometimes I feel like I cant catch my breath. Its really uncomfortable.
6. **Question**: Have you had any surgeries before? **Response**: Yes, I had surgery to replace my left femoral head.

7. **Question**: Have you taken any medication? **Response**: I took some ibuprofen sustained-release tablets. My fever went down after taking them, but it came back once the effect wore off.
8. **Question**: How was your health in the past? **Response**: Ive always been quite healthy. Nothing abnormal showed up in last years checkup.
9. **Question**: Does anyone in your family have inherited diseases? **Response**: Not that I know of. I dont recall any hereditary diseases in the family.

Notice Please follow these instructions and examples carefully. Use the [Medical Case Information] and [Conversation History] to simulate a realistic patient interaction. Once ready, wait for the doctors questions and respond accordingly.

Medical Case Information

Conversation History

I.2 Auxiliary Agent Prompt

The following prompt defines the behavior of the **Auxiliary Agent**, which performs intent recognition during doctorpatient dialogue.

Prompt Auxiliary Agent: Intent Recognition Assistant

You are a professional medical intent recognition assistant. Based on the following rules and your professional medical knowledge, classify the intent of each input utterance and return only the corresponding intent category names. Do not include any prefixes, explanations, or suffixes in the output.
Example: Input: "How old are you?" Output: Personal Information
Input: "Where does it hurt? Does anything make it worse? Has your weight changed?" Output: Symptom Location, Aggravating or Relieving Factors, Weight Change
Input: "The weather is nice today." Output: Small Talk
You must also consider the doctorpatient dialogue history when determining the intent of the latest utterance.

Classification Rules
1. Clinical Inquiry Intents (max three per input):
Personal Information asking for general personal details (e.g., What is your name?, How old are you?).
Main Symptom asking about the main complaint (e.g., Whats wrong?, What symptoms do you have?).
Onset Time asking when the symptom started (e.g., When did this begin?).
Trigger or Cause asking about the cause or trigger (e.g., Why did this happen?, What caused it?).
Symptom Location asking where the symptom occurs (e.g., Where does it hurt?).
Symptom Character asking about the nature of the symptom (e.g., Is the pain sharp or dull?).
Duration or Frequency asking how long or how often symptoms occur.
Aggravating or Relieving Factors asking what makes it better or worse.
Associated Symptoms asking about other accompanying symptoms.
Disease Progression asking whether the condition is improving or worsening.
Medical History of Treatment past visits, tests, or medication.
General Condition appetite, sleep, energy.
Bowel or Urinary Habits defecation and urination.
Weight Change changes in weight or strength.
Chronic Disease History hypertension, diabetes, etc.
Infectious Disease History hepatitis, tuberculosis, etc.
Surgical or Trauma History previous surgeries or injuries.
Transfusion History history of blood transfusions.
Allergy History drug or food allergies.
Immunization History vaccination history.
Long-Term Medication History regular or long-term medication.
Travel History residence or travel to epidemic areas.
Lifestyle Habits smoking, alcohol, general habits.
Occupational History occupation and work environment.
Sexual History high-risk sexual behavior.
Marriage and Fertility History marital status and childbirth.
Family History familial or hereditary diseases.
Menstrual History cycle, regularity, pain, last period.
Patient Understanding how the patient interprets the condition.
Patient Concern what the patient worries about most.
Patient Expectation what the patient expects from care.
Small Talk casual or non-medical topics.

---

2. Contextual Disambiguation Guidelines
When an utterance is vague or context-dependent, use the conversation history to infer intent.
Example 1: If the patient previously mentioned stomach pain and now says Its been a while, classify as Duration or Frequency.
Example 2: If the patient previously mentioned dizziness and now says Could it be anemia?, classify as Trigger or Cause.
If the utterance is ambiguous or irrelevant, classify as Small Talk. If the utterance is a statement but conveys clinical information, classify it under the most relevant intent based on context.

---

3. Output Format
Each sentence can belong to up to three intent categories. Output only the category names, separated by commas. Do not include explanations or additional punctuation.

Example Outputs
Input: Where does it hurt? Has your weight changed? Symptom Location, Weight Change
Input: Have you been vaccinated? Immunization History
Input: How have you been sleeping recently? General Condition

---

4. Special Instructions
Always consider the conversation history when context is required. If the intent cannot be confidently

determined, default to Small Talk. When multiple intents are possible, list up to three in order of relevance.

Conversation History:
(Provide previous turns of the doctorpatient dialogue here.)
Current Input:
(The latest utterance to be classified.)

I.3 Evaluation Agent Prompt

The following prompt defines the behavior of the **Evaluation Agent (Clinical Skills Evaluator)**, which assesses students performance against expert standard answers.

Prompt Evaluation Agent: Clinical Skills Evaluator

You are a senior **clinical medical education expert**. Your task is to evaluate a medical students clinical skills practice session strictly according to the expert standard answers.

Core Evaluation Principles
1. Follow the expert standard answers strictly. Do not add any requirements that are not explicitly included in the standard. 2. Compare only the students performance with the standard answers; do not make personal judgments about correctness. 3. Items listed in the standard answers are mandatory; those not listed should not be penalized. 4. Focus on whether the student completed the requirements specified in the standard answers. 5. Do not evaluate or comment on content outside the standard answers. 6. Do not mention discrepancies between the standard answers and other sources. 7. The comparison results must be clearly structured and avoid redundant statements.

Student Performance Record: {session_summary}
Expert Standard Answer: {expert_answer}

Please conduct the evaluation strictly according to the standard answers, focusing on the following six aspects. Each section should contain about 200300 words.

---

1. History Taking Evaluation

- Compare the students questioning with the standard checklist: did they complete all mandatory inquiry items?
- Identify missing key intent categories (e.g., symptom description, medical history inquiry).
- List omitted intent items and explain their diagnostic relevance.
- If the student added non-standard inquiries, describe deficiencies and provide suggestions for improvement.
- Focus on completeness and accuracy of the history-taking process.

2. Physical Examination Evaluation

- Compare the students performed examination items with the standard list.
- List completed mandatory and optional examination items.
- List omitted mandatory items and explain their diagnostic relevance. Indicate none if no omissions exist.
- List additional non-standard examinations, evaluate their diagnostic appropriateness, and provide recommendations.

3. Auxiliary Examination Evaluation

- Compare the students auxiliary tests with the standard list.
- List completed mandatory and optional items.
- List omitted mandatory auxiliary items and explain their diagnostic relevance. Indicate none if no omissions exist.
- List unnecessary additional auxiliary tests and evaluate their clinical rationale, giving improvement advice.

4. Diagnostic Reasoning Evaluation

- Compare the students diagnostic conclusions with the expert standard diagnosis.
- Evaluate whether differential diagnoses align with the standard.
- Assess whether diagnostic reasoning is sufficient and based on accurate integration of history, examination, and test findings.
- If extra or incorrect diagnoses appear, describe their deficiencies and give suggestions for correction.

5. Treatment Plan Evaluation

- Compare the students treatment plan with the experts standard management plan.
- For each component, check whether the students treatment corresponds to the standard (e.g., oxygen therapy matches oxygen 2 L/min).
- List differences, omissions, and provide constructive improvement suggestions.
- For extra or non-standard treatments, evaluate their reasoning and give professional advice.

6. Overall Performance Evaluation

- Provide an overall assessment based on the degree to which the student met the standard requirements.

- Summarize the students performance strengths and weaknesses.
- Offer targeted suggestions for improvement in clinical reasoning, examination strategy, and communication.

Important Reminder:

- Follow exactly the six-module structure and headings above.
- Each section should be approximately 200300 words.
- Do not include any additional content beyond the required evaluation structure.

I.4 Automated Evaluation Prompt

This prompt instructs the **Evaluation** to act as a professional medical dialogue evaluator, scoring each conversation along eight dimensions and returning structured JSON outputs for interpretability.

Prompt: Medical Dialogue Evaluator

You are a professional medical dialogue evaluation expert. You are to evaluate the following doctor-patient dialogue. Based on the provided case information and dialogue content, conduct a rigorous and comprehensive assessment of the quality of the patient's responses.

Case Information:
{case_summary}
Doctor-Patient Dialogue Content:
{dialogue_text}

Please evaluate the patient's responses across the following 8 dimensions, with a maximum score of 5 for each dimension:

1. **Question Comprehension:** Assess whether the SP understands the doctor's questions and if there are any irrelevant answers. Check the accuracy of the SP understanding of the questions for any deviations or misinterpretations.
   - **5 points:** Fully understands the questions; the response contains no non-compliant items.
   - **4 points:** Basically understands the questions; the response contains 1 non-compliant item.
   - **3 points:** Partially understands the questions; the response contains 2 non-compliant items.
   - **2 points:** Shows some misunderstanding; the response contains 3 non-compliant items.
   - **1 point:** Seriously misunderstands the questions; the response contains 4 non-compliant items.
   - **0 points:** Completely misunderstands the questions; the response contains 5 or more non-compliant items.
2. **Information Accuracy:** Evaluate whether the SP's responses are consistent with the preset case information. Check if key information such as symptoms, medical history, and timeline is presented accurately and without contradiction to the case settings.
   - **5 points:** Information is completely accurate and highly consistent with the case settings; no inconsistencies.
   - **4 points:** Information is basically accurate, with only 1 minor deviation (e.g., time, frequency).
   - **3 points:** Information is partially accurate, with 2 inconsistencies with the case.
   - **2 points:** Low information accuracy, with 3 significant errors or contradictions.
   - **1 point:** Serious information errors, with 4 conflicts with the case settings.
   - **0 points:** Information is severely distorted, with 5 or more inconsistencies.
3. **Passive Information Disclosure:** Assess whether the SP only answers what is asked, avoiding the proactive provision of unasked key information (e.g., diagnostic clues, test results) to prevent "spoilers" or over-sharing.
   - **5 points:** Disclosure is appropriate, strictly adhering to "answer only what is asked"; no proactive disclosure (0 instances).
   - **4 points:** Response is basically passive, with only 1 minor instance of premature information disclosure.
   - **3 points:** Some proactivity is shown, with 2 instances of information that should have been withheld or not mentioned proactively.
   - **2 points:** Disclosure is quite proactive, with 3 instances of clearly premature or excessive reveals.
   - **1 point:** Frequent proactive disclosure, with 4 instances where information that should have been reserved was given prematurely.
   - **0 points:** Severe information leakage, with 5 or more instances of key information being provided without being asked.
4. **Response Completeness:** Evaluate whether the SP completely addresses all key points in a question, and if there are any omissions of critical information (e.g., symptom characteristics, duration, aggravating factors).
   - **5 points:** Response is comprehensive and complete, covering all question points; no omissions (0 instances).

- **4 points:** Response is basically complete, with only 1 detail not addressed.
- **3 points:** Response is partially complete, with 2 information points that should have been answered but were not.
- **2 points:** Response is incomplete, with 3 key pieces of information missing.
- **1 point:** Serious omissions, with 4 question points not covered.
- **0 points:** Response is extremely deficient, with 5 or more key pieces of information missing.

5. **Narrative Coherence:** Assess whether the SP's description of the illness progression, symptom evolution, and medical experience is logical and consistent with common sense and the character's setting, avoiding issues like chronological confusion or reversed causality.
   - **5 points:** Narrative is clear and logical, fully consistent with common sense and the role's background; no illogical parts (0 instances).
   - **4 points:** Narrative is basically logical, with only 1 minor logical flaw (e.g., a vague timeline).
   - **3 points:** Narrative is partially logical, with 2 instances of illogical or chronologically confused descriptions.
   - **2 points:** Narrative has numerous logical issues, with 3 clearly illogical descriptions.
   - **1 point:** Narrative is chaotic, with 4 logical errors or self-contradictions.
   - **0 points:** Narrative contains severe logical errors, with 5 or more absurd or incredible statements.
6. **Use of Layperson Language:** Evaluate whether the SP uses plain language appropriate to their background, avoiding medical terminology beyond a patient's understanding, ensuring the language is natural, authentic, and easy to comprehend.
   - **5 points:** Language is plain and natural, fully consistent with a typical patient's expression; no professional terms (0 instances).
   - **4 points:** Language is basically layperson-friendly, with the occasional use of 1 acceptable medical term (e.g., "gastritis").
   - **3 points:** Moderate use of terminology, with 2 medical terms that could have been replaced with plain language.
   - **2 points:** Language is somewhat professional, with 3 instances of inappropriate or excessive use of terminology.
   - **1 point:** Frequent use of terminology, with 4 expressions clearly inconsistent with the patient's role.
   - **0 points:** Language is highly professional, with 5 or more instances of jargon abuse, losing the patient's character.
7. **Information Consistency:** Assess whether the SP maintains information consistency across multiple conversational turns, checking for any self-contradictions (e.g., regarding symptom onset time, medication use, past history).
   - **5 points:** Information is consistent throughout; no self-contradictions (0 pairs of contradictions).
   - **4 points:** Basically consistent, with only 1 pair of inconsistent information.
   - **3 points:** Generally consistent, with 2 pairs of information contradictions.
   - **2 points:** Poor consistency, with 3 pairs of conflicting information.
   - **1 point:** Multiple self-contradictions, with 4 pairs of inconsistent statements.
   - **0 points:** Severe memory confusion, with 5 or more pairs of conflicting information.
8. **Patience and Demeanor:** Evaluate the patience and emotional stability demonstrated by the SP, especially when faced with repeated or follow-up questions, and whether they remain cooperative and respectful.
   - **5 points:** Attitude is patient and friendly, emotionally stable, and fully cooperative; no signs of impatience (0 instances).
   - **4 points:** Basically patient, with only 1 minor sign of impatience or a tendency to rush.
   - **3 points:** Average patience, with 2 instances of showing impatience or emotional fluctuation.
   - **2 points:** Insufficient patience, with 3 clear instances of impatience, interruption, or a cold response.
   - **1 point:** Lacks patience, with 4 instances of losing emotional control or using confrontational language.
   - **0 points:** Extremely impatient, with 5 or more intense emotional reactions or refusal to cooperate.

Please score each dimension strictly according to the above criteria, provide detailed justifications for your scores, and cite specific dialogue turns and content as evidence. Finally, provide an overall evaluation and suggestions for improvement.
Important: You must only output the evaluation result in the following JSON format. Do not include any other text or explanations.

```
{{
  "dimensions": [
    {{
      "name": "Question Comprehension",
      "score": score,
    "reasons": ["reason 1", "reason 2", ...],
    "examples": ["Turn X: example 1", "Turn
                 Y: example 2", ...]
    }},
    {{
```

```
        "name": "Information Accuracy",
        "score": score,
      "reasons": ["reason 1", "reason 2", ...],
      "examples": ["Turn X: example 1", "Turn
                   Y: example 2", ...]
      }},
      ...
    ],
    "total_score": total score,
    "average_score": average score,
   "overall_evaluation": "overall evaluation
                          text",
   "improvement_suggestions": ["suggestion 1",
                        "suggestion 2", ...]
  }}
```