UROMED-AGENTS: AN INTERACTIVE MULTI-AGENT FRAMEWORK FOR CLINICAL DECISION-MAKING TRAINING IN UROLOGY

Anonymous authors
Paper under double-blind review

ABSTRACT

Effective and standardized clinical decision-making (CDM) training is essential for preparing clinicians to diagnose and treat urological diseases. However, traditional approaches are constrained by limited access to longitudinal cases and expert supervision, making it difficult to cover traceable decision-making processes. Recent advances in large language models (LLMs) enable interactive medical environments, but most methods either evaluate autonomous artificial intelligence (AI) clinicians or focus on isolated history taking tasks, providing limited support for end-to-end, multi-turn training of human clinicians. To address this, we propose UroMed-Agents, an interactive multi-agent framework for CDM training in urology. The framework centers on human clinicians and employs four agents for patient interaction, examination result delivery, process guidance, and final evaluation. To reduce role overreach and information leakage across turns, we introduce a stateful Case Engine that maintains interaction state, enforces information boundaries for each role, and releases examination results only when requested. Furthermore, the framework separates training and testing scenarios and evaluates clinician performance based on interaction records and final reports, while diagnostic performance at the system level is assessed using metrics based on the International Classification of Diseases (ICD-10). Experimental results on 239 real longitudinal urology cases illustrate the effectiveness of our method, which achieves better performance than existing interactive approaches.

1 INTRODUCTION

Urological diseases impose a substantial global health burden, with approximately 2.52 million new cases of prostate, bladder, and kidney cancer worldwide in 2022 (Bray et al., 2024). Therefore, timely and accurate clinical decision-making (CDM) training is essential, particularly in urology, where benign and malignant conditions may present similarly and decisions often require the sequential integration of history, laboratory tests, imaging, endoscopy, and pathology. Clinicians must actively select examinations, interpret incomplete evidence, and revise diagnostic and treatment decisions as new information emerges. However, traditional training relies heavily on real cases and expert supervision, making rare or complex cases difficult to cover and the training process hard to standardize or reproduce. Although instruction based on cases is easier to implement, it often provides key findings in advance, limiting practice in active information gathering and iterative decision-making. Therefore, existing methods provide limited support for comprehensive and controllable urological CDM training (Plackett et al., 2022). In recent years, large language models (LLMs) have enabled interactive medical environments. Research such as Med-PaLM has shown that LLMs adapted to the domain perform well in medical question answering (QA) and clinical knowledge reasoning, although their factuality, safety, and reasoning reliability still require rigorous evaluation (Singhal et al., 2023; 2025; Hager et al., 2024). Concurrently, LLMs have also been applied to clinical summarization, medical record processing, and patient communication (Van Veen et al., 2024; Ayers et al., 2023). However, most studies provide complete case information upfront, whereas CDM in the real world is sequential: clinicians must actively gather information, select examinations, and revise their judgments as evidence accumulates. Recent work such as AI Hospital further reveals a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

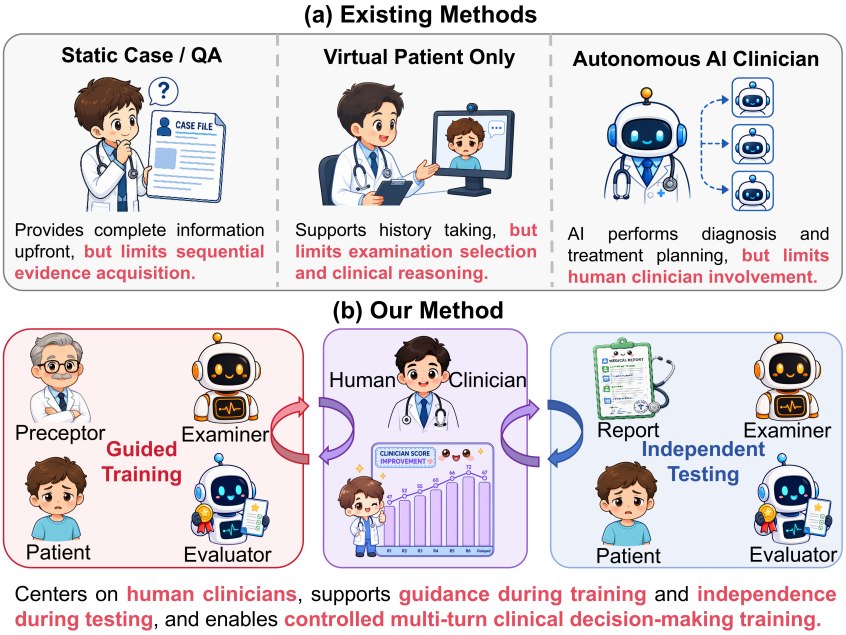


Figure 1: Comparison between UroMed-Agents and existing methods.

substantial performance gap between dynamic clinical interactions and static medical QA (Fan et al., 2025; Li et al., 2024; Tu et al., 2025; Schmidgall et al., 2026).

To address this gap, prior work has explored virtual patients, interactive diagnosis, and multi-agent collaboration (Khandwaha & Law, 2025; Wei et al., 2025; Tang et al., 2024; Kim et al., 2024b). However, most methods either focus on history taking or treat the LLM as the decision maker, providing limited support for end-to-end CDM training centered on human clinicians with controlled information disclosure, process guidance, and structured evaluation.

To this end, we propose UroMed-Agents, an interactive multi-agent framework for CDM training in urology. As shown in Figure 1, UroMed-Agents shifts the focus from interactions centered on artificial intelligence (AI) or fragmented across isolated tasks to complete CDM training centered on human clinicians. It employs four specialized agents: Patient, Examiner, Preceptor, and Evaluator, for history taking, examination result delivery, process guidance, and outcome evaluation, respectively. To prevent excessive information disclosure and uncontrolled release of examination results, we introduce a stateful Case Engine. It organizes longitudinal cases into static case memory and dynamic working memory, with information flow controlled by role boundaries, intent routing, and examination evidence gating. The framework supports both training and testing scenarios: the Preceptor Agent provides guidance during training but is disabled during testing, where clinicians' independent CDM ability is evaluated from interaction records and final structured reports.

We evaluate UroMed-Agents in terms of agent behavioral reliability, clinician training effectiveness, and diagnostic performance at the system level. Agent reliability is assessed using a standardized clinician script on system test cases. Training effectiveness is examined in a longitudinal study of 30 medical students in their fourth year across six training and testing rounds and a delayed retention test. Students were grouped by prior knowledge, and within each prior knowledge group, students were randomly assigned to UroMed-Agents training or static case training, with five students per condition. In the static condition, students received complete case information, submitted a structured report, and then reviewed the reference answer. Clinician performance is measured using CDM scores across five dimensions, while comparative and ablation experiments assess the diagnostic performance of UroMed-Agents at the system level using precision (P), recall (R), and F1 based on the International Classification of Diseases (ICD-10) (World Health Organization, 2019). The main contributions of this paper are as follows:

- We propose UroMed-Agents, an interactive multi-agent framework for clinical decision-making training in urology, designed for human clinicians.

- We introduce a stateful Case Engine that integrates static and dynamic memory, intent routing, and examination evidence gating to support traceable information release specific to each role and available on demand across multiple turns.
- We construct a benchmark from real longitudinal urology cases and conduct extensive experiments demonstrating the reliability and effectiveness of UroMed-Agents.

2 Related Work

2.1 Interactive Multi-Agent Methods for CDM

Recent medical LLM research has expanded from static QA to interactive diagnosis and multi-agent reasoning (McDuff et al., 2025; Saab et al., 2026; Li et al., 2023; Wu et al., 2024; Wang et al., 2025). CRAFT-MD shows that LLM performance declines when static clinical vignettes are reformulated as simulated conversations, highlighting the limitations of static QA for evaluating interactive diagnostic capability (Johri et al., 2025). AMIE and AgentClinic support multi-turn information gathering and tool use, while ClinicalAgents dynamically updates diagnostic hypotheses (Tu et al., 2025; Schmidgall et al., 2026; Ge et al., 2026). AgentClinic and AI Hospital simulate sequential diagnosis using Patient Agents paired with Measurement and Examiner Agents, respectively (Schmidgall et al., 2026; Fan et al., 2025). MedAgents and MDAgents use collaboration among multiple experts, whereas ArgMed-Agents and MedGen improve interpretability through argumentation and multisource knowledge fusion (Tang et al., 2024; Kim et al., 2024b; Hong et al., 2024; Liu et al., 2024). However, these methods generally treat the LLM as the decision maker and lack a unified design combining explicit memory separation with stateful examination evidence gating. In contrast, UroMed-Agents centers on human clinicians and uses a stateful Case Engine to enforce information boundaries for each role, maintain interaction states, and release examination evidence on demand during CDM training.

2.2 LLM-Based Virtual Patients and Medical Education

LLMs also support virtual patients and medical education by providing repeatable environments for history taking and communication between clinicians and patients. Existing virtual patients interact with clinicians in natural language based on case information (Khandwaha & Law, 2025; Cross et al., 2025; Cook et al., 2025; Liu et al., 2025; Qu et al., 2026; Li & Lebai Lutfi, 2026; Zeng et al., 2025). AIPatient improves consistency between patient responses and case facts by combining real electronic health records with knowledge augmentation, while MEDCO creates a multi-agent education environment through collaboration between Patient and Expert Agents (Yu et al., 2026; Wei et al., 2025). However, they primarily focus on patient simulation and history taking, with limited support for end-to-end CDM. In contrast, UroMed-Agents combines real longitudinal cases, process guidance, and structured evaluation in a complete CDM training loop.

3 Methodology

3.1 Framework Overview

As shown in Figure 2, UroMed-Agents centers on human clinicians in a controlled multi-turn environment supported by four agents with distinct roles and a stateful Case Engine. The Case Engine maintains static case memory containing case facts and information boundaries for each role, together with dynamic working memory that tracks the dialogue state, turn control, requested examinations, released evidence, and decision records. At each turn, it identifies the clinician's intent, routes the request to the appropriate agent, and controls examination result disclosure through evidence gating. The Patient and Examiner Agents provide subjective information and requested clinical evidence, respectively. The Preceptor Agent provides guidance only during training using the current case state and external knowledge, whereas the Evaluator Agent scores the interaction record and final report. Overall, the framework comprises **agent initialization by role**, a **stateful Case Engine**, and **separate training and testing scenarios**. Additional implementation details are provided in Appendix C.

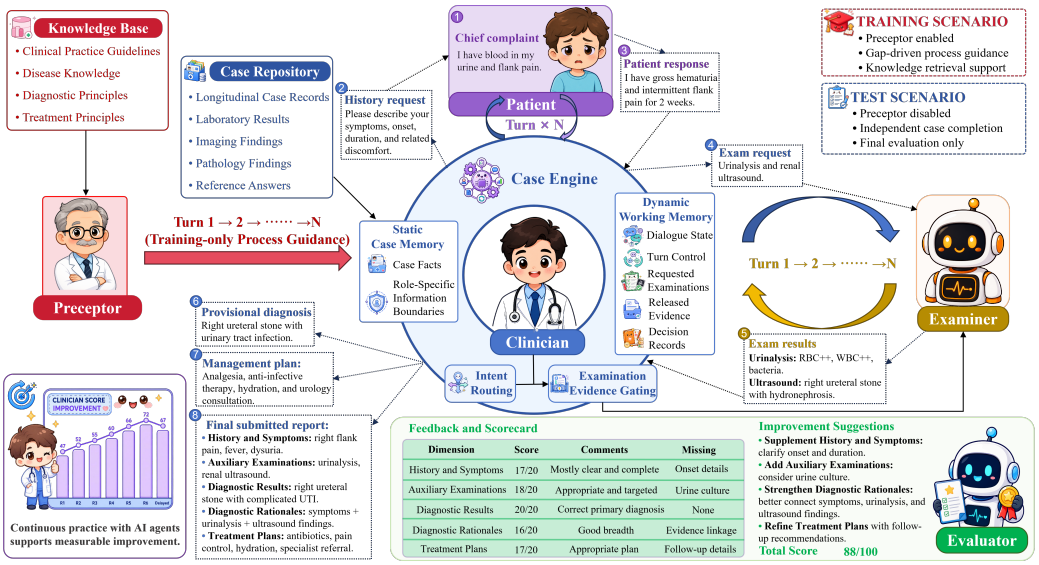


Figure 2: The overall framework of UroMed-Agents for CDM training in urology.

3.2 Agent Initialization by Role

UroMed-Agents comprises four agents: Patient, Examiner, Preceptor, and Evaluator. To prevent role overreach, case content is divided into subjective patient information, objective examination evidence, and clinical management information. Each agent is initialized with a prompt specific to its role:

$$P_i = \{\text{Role}_i, \text{Goal}_i, \text{Context}_i, \text{Constraint}_i, \text{Format}_i\}, \tag{1}$$

where Role_i denotes the role identity, Goal_i the task objective, Context_i the accessible context, Constraint_i the prohibited behaviors, and Format_i the output format.

Patient Agent. The agent answers history taking questions in the patient's voice using only subjective information known to the patient. It responds conversationally, states when information is unavailable, refuses requests for diagnoses, examination results, or complete treatment plans, and redirects irrelevant queries to the presenting complaint.

Examiner Agent. The agent matches examination requests to the case repository and returns the corresponding laboratory, imaging, pathology, or surgical results. To control information disclosure, it reports that no matching result is available for overly broad or unmatched requests and withholds unrequested evidence, the final diagnosis, and the treatment plan.

Preceptor Agent. During training, the agent detects critical gaps in history taking, examination selection, evidence interpretation, diagnostic reasoning, or treatment planning using the current interaction state, case facts, and an external knowledge base. It addresses only the most critical weakness without revealing the complete reference answer and responds in a fixed "dimension–current problem–suggested revision" format.

Evaluator Agent. After each case, the agent evaluates the interaction record and final report across history and symptoms, auxiliary examinations, diagnostic results, diagnostic rationales, and treatment plans. It identifies missing or incorrect content and provides targeted suggestions (Liu et al., 2023; Kim et al., 2024a; Zheng et al., 2023).

3.3 Stateful Case Engine

Role-specific prompts alone cannot reliably control interactions across multiple turns: agents may forget released evidence, repeat responses, access information outside their roles, or reveal results

prematurely. Therefore, we introduce the Case Engine to maintain state, route requests, and control the information available to each agent (Park et al., 2023; Shinn et al., 2023; Ge et al., 2026). The Case Engine state comprises static case memory and dynamic working memory, as defined in Definition 1.

Definition 1 (Case Engine State). At interaction turn t, the Case Engine state is defined as

$$\begin{aligned} S_t &= \{\mathcal{M}^s, \mathcal{M}_t^w\}, \\ \mathcal{M}^s = \{\mathcal{F}, \mathcal{B}\}, \qquad \mathcal{M}_t^w &= \{D_t, T_t, O_t, \mathcal{E}_t, \mathcal{R}_t\}, \end{aligned} \tag{2}$$

where $\mathcal{M}^s$ and $\mathcal{M}_t^w$ denote the static and dynamic memory. $\mathcal{F}$ contains complete case facts and $\mathcal{B}$ denotes the role boundaries. D_t, T_t, O_t, $\mathcal{E}_t$, and $\mathcal{R}_t$ denote the dialogue state, turn control state, requested examinations, released evidence, and clinician decision records, respectively. The static memory remains fixed, whereas the dynamic working memory is updated after each clinician input.

Definition 2 (Role Routing). Given clinician input u_t, dialogue state D_t, and turn control state T_t, the Case Engine predicts the interaction intent and selects the target agent as

$$c_t = \arg\max_{c \in \mathcal{I}} p(c \mid u_t, D_t, T_t), \qquad r_t = \Pi(c_t, T_t), \tag{3}$$

where $\mathcal{I}$ denotes the set of candidate intents, c_t the predicted intent, r_t the invoked agent, and $\Pi(\cdot)$ the routing rule. Examination requests are further processed by the evidence gating mechanism.

Definition 3 (Examination Gating). When c_t indicates an examination request, the Case Engine extracts the newly requested examination items as

$$\Delta O_t = \mathrm{Exam}(u_t) \setminus O_t, \tag{4}$$

where O_t contains the examination items requested in previous turns. The newly requested items are then matched against the objective evidence set $\mathcal{F}_{\mathrm{obj}}$:

$$\Delta \mathcal{E}_t = \{e \in \mathcal{F}_{\mathrm{obj}} \mid \exists\, o \in \Delta O_t : \mathrm{sim}(e, o) \geq \tau\}. \tag{5}$$

The matching score $\mathrm{sim}(e, o)$ equals 1 for an exact or alias match and $\cos\big(h(e), h(o)\big)$ otherwise, where $h(\cdot)$ denotes the text encoder and τ the matching threshold.

Before response generation, the current request and matched evidence are provisionally incorporated as $O_t^+ = O_t \cup \Delta O_t$ and $\mathcal{E}_t^+ = \mathcal{E}_t \cup \Delta \mathcal{E}_t$. If no match is found, the Examiner reports that no corresponding result is recorded. Without an examination request, $\Delta O_t = \varnothing$ and $\Delta \mathcal{E}_t = \varnothing$.

Definition 4 (Information View by Role). Given the target role and the evidence available at the current turn, the Case Engine constructs the information view for the role as

$$x_t^{r_t} = \phi_{r_t}\big(\mathcal{F}, \mathcal{B}, \mathcal{E}_t^+\big) \oplus G_{r_t}\big(D_t, T_t, O_t^+, \mathcal{E}_t^+, \mathcal{R}_t\big), \tag{6}$$

where $\phi_{r_t}(\cdot)$ filters static case facts according to role boundaries and masks unreleased objective evidence, while $G_{r_t}(\cdot)$ extracts relevant dynamic context and $\oplus$ denotes ordered context concatenation.

When the Preceptor Agent is invoked, the retrieved knowledge K_t in Equation equation 9 is appended to $x_t^{r_t}$. The response is then generated as

$$y_t = \mathrm{LLM}_\theta\big(P_{r_t}, x_t^{r_t}, u_t\big). \tag{7}$$

Definition 5 (State Update). After generating the response at turn t, the dynamic working memory is updated as

$$\begin{cases} D_{t+1} = D_t \oplus (u_t, y_t), \\ T_{t+1} = \delta(T_t, c_t), \\ O_{t+1} = O_t^+, \\ \mathcal{E}_{t+1} = \mathcal{E}_t^+, \\ \mathcal{R}_{t+1} = \mathcal{R}_t \oplus \mathrm{Dec}(u_t). \end{cases} \tag{8}$$

Here, $\delta(\cdot)$ updates the turn control state, and $\mathrm{Dec}(\cdot)$ extracts the clinician's intermediate diagnostic or treatment decisions. If u_t contains no clinical decision, then $\mathrm{Dec}(u_t) = \varnothing$. The static case memory remains read only throughout the interaction.

3.4 Separate Training and Testing Scenarios

UroMed-Agents supports both training and testing scenarios using the same case representation, Case Engine, and evaluation criteria. The Preceptor Agent provides guidance on critical omissions during training but is disabled during testing to evaluate independent CDM. A representative training scenario and its evaluation workflow across five dimensions are provided in Appendix E.

Training Scenario. The training scenario provides process guidance driven by gaps. When a critical omission is detected in history taking, examination selection, evidence interpretation, diagnostic reasoning, or treatment planning, the Preceptor Agent identifies the most urgent weakness and provides a targeted suggestion without revealing the reference diagnosis or a complete treatment plan.

To provide medically grounded guidance, UroMed-Agents incorporates a retrieval-augmented generation (RAG) mechanism that combines a urology knowledge base with reference answers and scoring rubrics specific to each case. The knowledge base integrates authoritative Chinese and international clinical guidelines, including those from the European Association of Urology (EAU), and leading urology monographs, particularly *Campbell-Walsh-Wein Urology*. It also summarizes dozens of core medical references covering diagnostic imaging, pathology, operative techniques, and surgical atlases. All materials were cleaned, deduplicated, organized by disease and clinical task, and segmented into retrievable units covering clinical manifestations, examination selection, diagnostic reasoning, and treatment principles. Two experienced urologists reviewed the selected sources and segmented content to ensure clinical accuracy, relevance, and consistency.

At interaction turn t, the Case Engine retrieves relevant knowledge according to the current clinician input, dialogue state, released evidence, and clinician decision records:

$$K_t = \mathrm{Retrieve}\big(\mathcal{K}, u_t, D_t, \mathcal{E}_t^+, \mathcal{R}_t\big), \tag{9}$$

where $\mathcal{K}$ denotes the urology knowledge base and K_t denotes the knowledge segments relevant to the current case state (Lewis et al., 2020). When the Preceptor Agent is invoked, the retrieved segments are appended to the information view for its role before response generation.

Testing Scenario. The testing scenario evaluates the clinician's independent CDM ability. The Preceptor Agent remains disabled and the clinician needs to gather information from the Patient Agent and request results from the Examiner Agent. The clinician must submit a structured decision report within a limited number of interaction turns.

Evaluation Stage. After the interaction and final report submission, the Evaluator assesses the interaction record and final report. The output of the Evaluator Agent is defined as

$$o^{\mathrm{eval}} = \{s_i, c_i, m_i, a_i\}_{i=1}^{5}, \tag{10}$$

where s_i denotes the score for dimension i, c_i the evaluation comment, m_i the missing or incorrect content, and a_i the improvement suggestion. Thus, even if the final diagnosis is correct, the corresponding dimensions still reflect deficiencies when essential examinations are not requested or the diagnostic rationale is insufficient.

4 Experiments

4.1 Experimental Details

Implementation Details. Unless otherwise specified, all agents in UroMed-Agents use GPT-5.4 (OpenAI, 2026) with prompts specific to each role. All LLM-based agents share the same decoding configuration, with temperature set to 0.4 and top-p set to 1.0, to isolate backbone effects from decoding differences. The examination evidence matching threshold is set to 0.8. Each case is limited to 10 turns, and the clinician submits a structured report covering five dimensions before or upon reaching the turn limit.

Dataset. We collected 255 longitudinal urology cases from three medical centers in China, covering 24 common urological diseases. All cases were de-identified and converted into a unified structured

format. Urologists then conducted a comprehensive quality review of the diagnostic results, diagnostic rationales, and treatment plans. Each component was rated as either "basically correct" or "clearly incorrect." Cases were retained only when all three components were basically correct and the records contained no major information gaps. Finally, 239 longitudinal cases that satisfied the quality requirements were retained. Additional details on case quality control, disease distribution, and structured case construction are provided in Appendix A.

The 239 cases were stratified by disease distribution, case difficulty, and source center, and then split into 120 training, 60 immediate test, 10 delayed test, and 49 system test cases, with no overlap.

Evaluation Metrics. We used two complementary evaluation schemes. First, urological CDM was assessed across five dimensions: history and symptoms, auxiliary examinations, diagnostic results, diagnostic rationales, and treatment plans (Fan et al., 2025). Each dimension was scored from 0 to 20 for completeness and correctness, yielding a total score of 0–100. All participant records were scored using the same frozen Evaluator and rubrics specific to each case. The scoring reliability of the Evaluator was validated against expert consensus on the disjoint system test subset. Second, diagnostic performance was evaluated using an ICD-10 codebook specific to the benchmark and validated by two urologists, with one canonical code assigned to each of the 24 target diseases. This scheme included micro-averaged P, R, and F1 across all system test cases, with all metrics scaled to 0–100. Under this scheme, predicted and reference diagnoses were represented as deduplicated code sets. Unsupported or unmapped predictions were counted as false positives, whereas omitted reference diagnoses were counted as false negatives. The complete diagnostic evaluation protocol and metric definitions are provided in Appendix B.

4.2 Agent Behavioral Reliability

Before the longitudinal clinician training experiment, we evaluated the reliability of each agent on 49 system test cases using a fixed clinician script with 10 turns. The first five turns covered history taking, whereas the final five focused on clinical evidence acquisition. For controlled evaluation of the Preceptor Agent, predefined decision gaps were introduced at turns 3, 5, and 8, corresponding to omissions in history taking, examination selection, and diagnostic reasoning, respectively.

For the Patient Agent, role fidelity and memory consistency assessed whether responses maintained the patient role and consistency with case facts and dialogue history. For the Examiner Agent, result specificity and request matching evaluated the accuracy and relevance of examination evidence. For the Preceptor Agent, intervention accuracy and non-overreach rate evaluated whether guidance addressed the decision gap without revealing the diagnosis or treatment plan. For the Evaluator Agent, rubric alignment evaluated adherence to the case-specific rubric. Agreement with expert consensus was assessed using the agreement rate within two points, mean absolute error (MAE), and a two-way random-effects, absolute-agreement intraclass correlation coefficient (ICC). The annotation procedure and scoring rubric are described in Appendix D.

Table 1: Behavioral reliability across LLM backbones (%). RF: role fidelity, MC: memory consistency, RS: result specificity, RM: request matching, IA: intervention accuracy, NO: non-overreach, RA: rubric alignment, HA: human agreement.

LLM	Patient		Examiner		Preceptor		Evaluator	
	RF	MC	RS	RM	IA	NO	RA	HA
DeepSeek-V4-Pro	93.5	93.1	93.9	94.3	94.6	95.9	93.5	93.1
Qwen3-Max	91.8	91.4	92.2	92.7	92.5	93.2	91.8	91.4
Claude Sonnet 4.6	94.3	93.9	94.7	95.1	95.2	95.9	94.7	93.9
ERNIE 4.5 Turbo	90.6	90.2	91.0	91.4	90.5	91.2	90.6	90.2
GPT-5.4	**95.5**	**95.1**	**96.3**	**96.7**	**97.3**	**98.0**	**95.9**	**95.5**

To evaluate robustness across backbones, we selected DeepSeek-V4-Pro, Qwen3-Max, Claude Sonnet 4.6, ERNIE 4.5 Turbo (128K), and GPT-5.4 as LLM backbones spanning a wide capability range (DeepSeek-AI, 2026; Alibaba Cloud, 2026; Anthropic, 2026; Baidu, 2026; OpenAI, 2026). All models were evaluated using the same cases, scripts, and prompts. Two experienced urologists independently annotated the relevant behavior units and assigned scores at the dimension level, with disagreements resolved by consensus.

As shown in Table 1, all models achieved high reliability across most metrics. GPT-5.4 performed best, achieving 95.5% agreement within two points with expert consensus, an MAE of 0.83, and an ICC of 0.91. It was therefore used in the longitudinal clinician training experiment.

Table 2: CDM performance across five dimensions, training rounds, and prior knowledge groups. Values are mean (SD), reported as UroMed-Agents/Static Case Training. HS: history and symptoms, AE: auxiliary examinations, DR: diagnostic results, DRA: diagnostic rationales, TP: treatment plans. $^{*}p_{\text{adj}} < 0.05$ for differences between conditions in change from R1.

Prior Knowledge	Round	HS ↑	AE ↑	DR ↑	DRA ↑	TP ↑	Overall ↑
Low	R1	11.0 (1.8)/11.2 (1.7)	7.4 (1.9)/7.2 (1.8)	10.2 (1.7)/10.1 (1.9)	9.6 (2.0)/9.4 (1.9)	8.9 (1.8)/8.8 (2.0)	47.1 (4.2)/46.7 (4.4)
	R2	10.8 (1.5)/11.0 (1.9)	8.6 (1.8)/7.8 (1.6)	11.3 (1.6)/10.7 (1.8)	10.7 (1.9)/10.0 (2.0)	9.6 (1.7)/9.3 (1.8)	51.0 (4.0)/48.8 (4.1)
	R3	11.9 (1.7)/11.5 (1.6)	8.1 (1.5)/8.0 (1.9)	11.0 (1.8)/10.8 (1.7)	12.1 (1.6)/10.7 (1.8)	11.5 (1.9)/9.9 (2.0)	54.6 (3.8)/50.9 (4.2)
	R4	11.5 (1.4)/11.4 (1.3)	10.0 (1.7)/8.8 (1.6)	13.0 (1.5)/11.4 (1.9)	12.6 (1.7)/11.2 (1.7)	11.8 (1.6)/10.6 (1.8)	58.9 (3.6)/53.4 (3.9)
	R5	13.2 (1.6)/12.0 (1.8)	10.6 (1.4)/9.0 (1.5)	12.8 (1.6)/11.6 (1.6)	14.2 (1.3)/**11.8 (2.0)**	13.3 (1.5)/10.7 (1.7)	64.1 (3.4)/55.1 (4.0)
	R6	**13.8 (1.2)/12.5 (1.5)**	**13.0 (1.6)/9.8 (1.7)**	**14.9 (1.3)/12.3 (1.5)**	**14.6 (1.5)**/11.7 (1.8)	**14.5 (1.4)/11.0 (1.6)**	**70.8 (3.1)/57.3 (3.8)***
	Delayed	12.9 (1.5)/11.9 (1.4)	12.0 (1.7)/8.9 (1.9)	13.9 (1.4)/11.7 (1.8)	13.8 (1.6)/11.1 (1.9)	13.2 (1.8)/10.2 (1.7)	65.8 (3.5)/53.8 (4.1)*
Medium	R1	13.4 (1.6)/13.6 (1.5)	8.8 (1.7)/8.6 (1.8)	12.0 (1.5)/11.5 (1.6)	12.4 (1.8)/11.8 (1.7)	9.8 (1.7)/10.0 (1.8)	56.4 (3.6)/55.5 (3.8)
	R2	14.5 (1.3)/14.0 (1.7)	9.3 (1.6)/9.0 (1.5)	13.2 (1.4)/12.7 (1.4)	12.9 (1.7)/12.8 (1.6)	10.8 (1.5)/10.3 (1.9)	60.7 (3.4)/58.8 (3.6)
	R3	14.1 (1.5)/14.2 (1.4)	11.0 (1.3)/9.6 (1.7)	12.9 (1.6)/12.9 (1.5)	14.7 (1.4)/13.2 (1.8)	12.5 (1.6)/10.8 (1.5)	65.2 (3.2)/60.7 (3.5)
	R4	15.8 (1.2)/14.8 (1.6)	10.6 (1.5)/10.1 (1.4)	15.0 (1.3)/13.5 (1.6)	15.4 (1.5)/13.8 (1.5)	13.0 (1.4)/11.8 (1.7)	69.8 (3.0)/64.0 (3.3)
	R5	15.5 (1.4)/15.1 (1.5)	13.2 (1.2)/11.6 (1.6)	**15.9 (1.1)**/14.0 (1.8)	16.1 (1.3)/14.5 (1.3)	13.8 (1.5)/12.4 (1.5)	74.5 (2.8)/67.6 (3.6)
	R6	**16.4 (1.1)/15.6 (1.3)**	**14.0 (1.4)/12.3 (1.5)**	15.7 (1.5)/**14.3 (1.4)**	**16.8 (1.2)/15.0 (1.6)**	**15.2 (1.3)/12.9 (1.4)**	**78.1 (2.7)/70.1 (3.1)***
	Delayed	15.4 (1.4)/14.2 (1.5)	13.1 (1.6)/10.5 (1.7)	14.9 (1.3)/13.6 (1.6)	15.2 (1.5)/14.3 (1.4)	14.4 (1.6)/11.5 (1.8)	73.0 (3.0)/64.1 (3.7)*
High	R1	15.4 (1.4)/15.1 (1.3)	10.0 (1.5)/9.9 (1.6)	13.2 (1.3)/12.8 (1.4)	14.6 (1.5)/13.8 (1.6)	10.1 (1.4)/10.3 (1.5)	63.3 (3.1)/61.9 (3.3)
	R2	15.0 (1.2)/15.7 (1.5)	11.1 (1.4)/10.3 (1.3)	14.3 (1.1)/13.2 (1.5)	15.4 (1.3)/14.4 (1.4)	11.3 (1.4)/10.5 (1.2)	67.1 (2.9)/64.1 (3.0)
	R3	16.6 (1.0)/16.0 (1.2)	11.6 (1.3)/10.7 (1.5)	13.8 (1.4)/13.9 (1.3)	16.3 (1.2)/15.3 (1.5)	13.4 (1.1)/11.0 (1.4)	71.7 (2.8)/66.9 (3.1)
	R4	16.2 (1.2)/16.3 (1.4)	13.4 (1.1)/11.3 (1.2)	15.7 (1.0)/14.4 (1.4)	16.0 (1.4)/15.4 (1.2)	13.9 (1.3)/12.0 (1.6)	75.2 (2.7)/69.4 (3.0)
	R5	**17.4 (0.9)**/16.6 (1.1)	13.1 (1.3)/11.6 (1.4)	**16.5 (1.2)**/14.7 (1.2)	**17.2 (1.0)**/15.9 (1.3)	14.8 (1.2)/12.6 (1.4)	79.0 (2.5)/71.4 (2.8)
	R6	16.9 (1.1)/**16.8 (1.3)**	**15.4 (1.2)/12.0 (1.1)**	16.2 (1.3)/**15.1 (1.4)**	17.0 (1.1)/**16.3 (1.2)**	**14.9 (1.4)/13.1 (1.3)**	**80.4 (2.6)/73.3 (2.9)***
	Delayed	16.0 (1.3)/15.2 (1.4)	14.2 (1.5)/11.4 (1.6)	15.5 (1.2)/14.4 (1.5)	16.2 (1.4)/15.7 (1.3)	14.1 (1.5)/11.5 (1.6)	76.0 (2.9)/68.2 (3.4)*

4.3 Training Effects across Levels of Prior Knowledge

This longitudinal repeated-measures experiment examined CDM performance across levels of prior knowledge and training conditions. We recruited 30 medical students in their fourth year from one cohort at a medical school in China (15 male and 15 female students, aged 21–24 years with a mean age of 23 years). All students were compensated on an hourly basis, and the study was conducted in accordance with institutional ethical guidelines.

Prior knowledge was approximated by cumulative performance in core medical courses completed before the study. Students were stratified by class rank into high, medium, and low prior knowledge groups, with 10 students in each group. Within each stratum, students were randomly assigned to UroMed-Agents or static case training, with five students in each condition. Before the study, all students completed standardized interface familiarization and two system test cases to reduce bias from unfamiliarity with the interactive environment. The study comprised six training and testing rounds over two weeks, each with 20 training cases and 10 immediate test cases, totaling 120 training cases and 60 immediate test cases per student. Two weeks after Round 6, students completed 10 delayed test cases without further training.

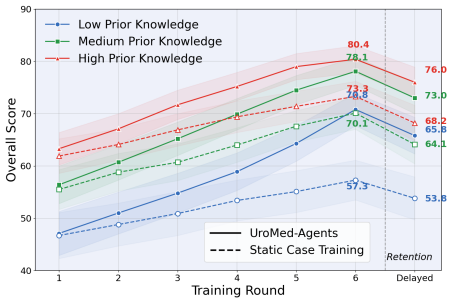


Figure 3: Overall CDM scores across six rounds and delayed testing by training condition and prior knowledge.

As shown in Table 2 and Figure 3, both training conditions produced progressive score gains, but UroMed-Agents consistently outperformed static case training across all three prior knowledge groups. Exact Wilcoxon rank-sum tests on changes at the participant level with Holm correction showed significantly greater gains for UroMed-Agents from R1 to R6 and from R1 to the delayed test ($p_{\text{adj}} < 0.05$). The consistent gains across successive rounds indicate that the observed improvement reflects cumulative training effects rather than an isolated advantage in a single round. Within the UroMed-Agents condition, the group with low prior knowledge showed the largest gains, whereas the group with high prior knowledge showed smaller gains, possibly reflecting its higher initial performance. Delayed test scores remained above R1 in both training conditions, indicating partial retention after two weeks.

4.4 Comparison Results

To evaluate adaptability and effectiveness across models, we used GPT-5.4 and DeepSeek-V4-Pro as backbones and compared UroMed-Agents with representative prompting and multi-agent baselines under the same data split and evaluation protocol (OpenAI, 2026; DeepSeek-AI, 2026). The prompting baselines included Zero-Shot, Few-Shot, Chain-of-Thought, and Chain-of-Logic (Brown et al., 2020; Kojima et al., 2022; Wei et al., 2022; Servantez et al., 2024). The multi-agent baselines included AI Hospital, AgentClinic, MedAgentSim, and MEDCO (Fan et al., 2025; Schmidgall et al., 2026; Almansoori et al., 2025; Wei et al., 2025). Full Information served as a reference condition in which all observable case evidence was provided at once.

Table 3: Performance comparison of UroMed-Agents with prompting and multi-agent baselines.

Method	GPT-5.4			DeepSeek-V4-Pro		
	P	R	F1	P	R	F1
Prompting Methods						
Zero-Shot	68.6	75.8	72.0	63.8	71.2	67.3
Few-Shot	70.4	77.0	73.6	65.7	72.9	69.1
CoT	71.7	79.1	75.2	66.4	74.8	70.4
CoL	73.6	78.4	75.9	68.2	74.1	71.0
Multi-Agent Methods						
AI Hospital	75.0	81.3	78.0	70.5	77.6	73.9
AgentClinic	76.4	82.5	79.3	72.1	79.2	75.5
MedAgentSim	77.8	84.0	80.8	73.8	81.1	77.3
MEDCO	78.6	84.5	81.4	74.6	81.5	77.9
UroMed-Agents (ours)	**82.2**	**87.4**	**84.7**	**79.0**	**84.9**	**81.8**
Full Information Reference						
Full Information	86.2	91.0	88.5	83.2	88.5	85.8

To control for questioning strategy and case exposure, all evaluated conditions used identical system test cases. For UroMed-Agents and the interactive multi-agent baselines, evidence was elicited using the same standardized clinician script with 10 turns without predefined decision gaps. Prompting baselines received the resulting evidence in a standardized format, whereas Full Information received all observable evidence at once. The Preceptor Agent was disabled in the comparison experiments. After each case, the resulting record was converted into a structured report by the same frozen report generation module using GPT-5.4, without access to reference diagnoses or scoring rubrics. Diagnostic performance was then evaluated using the ICD-10 metrics described above. As shown in Table 3, UroMed-Agents outperformed all interactive baselines under both backbones. The standardized interaction script and additional analyses at the system level are provided in Appendix F.

4.5 Ablation Study

To assess the contributions of initialization by role and the stateful Case Engine, we compared the full framework with three variants: (I) without both modules, (II) without the Case Engine, and (III) without initialization by role. Without the Case Engine, the fixed clinician script directly invoked the target agent using the current request, dialogue history, and case content appropriate to the role. All other settings matched the comparative experiment.

Table 4: Results of the ablation study.

Module		GPT-5.4			DeepSeek-V4-Pro		
Role Init	Case Engine	P	R	F1	P	R	F1
×	×	62.1	72.8	67.0	61.7	70.8	65.9
✓	×	72.2	80.0	75.9	68.4	77.5	72.7
×	✓	76.8	83.1	79.8	73.3	80.9	76.9
✓	✓	**82.2**	**87.4**	**84.7**	**79.0**	**84.9**	**81.8**

As shown in Table 4, removing either module reduced P, R, and F1 under both backbones. Removing the Case Engine caused a larger decline, underscoring the importance of state tracking and controlled evidence release. The full framework performed best, showing that initialization by role and the Case Engine are complementary. Extended ablation settings for additional LLM backbones are provided in Appendix G.

5 Conclusion

We present UroMed-Agents, a multi-agent framework centered on human clinicians for CDM training in urology. By coordinating Patient, Examiner, Preceptor, and Evaluator Agents through a stateful Case Engine, the framework enables traceable disclosure of clinical information under role constraints and on demand across multi-turn interactions. It further separates guided training from independent testing, thereby supporting both feedback at the process level and structured outcome assessment. Experiments on 239 real longitudinal urology cases demonstrated high agent reliability and improved CDM performance across prior knowledge groups, while UroMed-Agents outperformed most existing prompting and multi-agent baselines.

THE STATEMENTS

LLM USAGE STATEMENT

In this work, LLMs were used solely as writing assistants to polish the language of the manuscript. They were not involved in research ideation, experiment design, implementation, analysis, or any other scientific contribution.

ETHICS STATEMENT

This study involved a longitudinal training experiment with medical students and the use of de-identified longitudinal urology cases. All student participants were clearly informed of the purpose, procedures, and requirements of the study, and their participation was voluntary with informed consent obtained beforehand. Participants were compensated on an hourly basis. All clinical cases were de-identified before use, and no personally identifiable patient information was included in the study. The study was conducted in accordance with institutional ethical guidelines and the ICLR Code of Ethics, with particular attention to participant autonomy, data privacy, fairness, and transparency.

REFERENCES

Alibaba Cloud. Qwen3-Max model and api documentation. Alibaba Cloud Model Studio, 2026. URL https://help.aliyun.com/zh/model-studio/model-pricing.

Mohammad Almansoori, Komal Kumar, and Hisham Cholakkal. MedAgentSim: Self-evolving multi-agent simulations for realistic clinical interactions. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15968 of *Lecture Notes in Computer Science*, pp. 362–372. Springer, 2025. doi: 10.1007/978-3-032-05114-1_35.

Anthropic. Claude sonnet 4.6 system card. Technical report, Anthropic, February 2026. URL https://www-cdn.anthropic.com/78073f739564e986ff3e28522761a7a0b4484f84.pdf.

John W. Ayers, Adam Poliak, Mark Dredze, Eric C. Leas, Zechariah Zhu, Jessica B. Kelley, Dennis J. Faix, Aaron M. Goodman, Christopher A. Longhurst, Michael Hogarth, and Davey M. Smith. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, 183(6):589–596, 2023.

Baidu. ERNIE 4.5 Turbo model documentation. Baidu AI Cloud Qianfan, 2026. URL https://cloud.baidu.com/doc/qianfan/s/rmh4stp0j.

Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74 (3):229–263, 2024.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

David A. Cook, Jacob Overgaard, V. Shane Pankratz, Guilherme Del Fiol, and Christopher A. Aakre. Virtual patients using large language models: Scalable, contextualized simulation of clinician-patient dialogue with feedback. *Journal of Medical Internet Research*, 27:e68486, 2025.

James Cross, Thomas Kayalackakom, R. E. Robinson, A. Vaughans, R. Sebastian, R. Hood, C. Lewis, S. Devaraju, P. Honnavar, S. Naik, et al. Assessing ChatGPT's capability as a new age standardized patient: Qualitative study. *JMIR Medical Education*, 11:e63353, 2025.

DeepSeek-AI. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence, April 2026. URL https://arxiv.org/abs/2606.19348.

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, and Fei Huang. AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 10183–10213, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.680/.

Z. Ge, H. Li, Y. Wang, N. Hu, C. J. Zhang, and Q. Li. ClinicalAgents: Multi-agent orchestration for clinical decision making with dual-memory. *arXiv preprint arXiv:2603.26182*, 2026.

Paul Hager, Florian Jungmann, Robbie Holland, K. Bhagat, I. Hubrecht, M. Knauer, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30:2613–2622, 2024.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

S. Hong, L. Xiao, X. Zhang, and J. Chen. ArgMed-Agents: Explainable clinical decision reasoning with LLM disscusion via argumentation schemes. In *2024 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 5486–5493, 2024.

Qiao Jin, Won Kim, Qingyu Chen, Donald C. Comeau, Lana Yeganova, W. John Wilbur, and Zhiyong Lu. MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023. doi: 10.1093/bioinformatics/btad651.

S. Johri, J. Jeong, B. A. Tran, D. I. Schlessinger, S. Wongvibulsin, L. A. Barnes, H.-Y. Zhou, Z. R. Cai, E. M. Van Allen, D. Kim, R. Daneshjou, and P. Rajpurkar. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature Medicine*, 31(1):77–86, 2025.

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

Seungone Kim, Jamin Shin, Y. Cho, J. Jang, S. Longpre, H. Lee, S. Yun, S. Shin, S. Kim, J. Thorne, and M. Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations*, 2024a.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Y. S. Chan, X. Xu, D. McDuff, H. Lee, M. Ghassemi, C. Breazeal, and H. W. Park. MDAgents: An adaptive collaboration of LLMs for medical decision-making. In *Advances in Neural Information Processing Systems*, volume 37, pp. 79410–79452, 2024b.

T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213, 2022.

Terry K. Koo and Mae Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163, 2016. doi: 10.1016/j.jcm.2016.02.012.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

D. Li and S. Lebai Lutfi. Large language model-based virtual patient systems for history-taking in medical education: Comprehensive systematic review. *JMIR Medical Informatics*, 14:e79039, 2026. doi: 10.2196/79039.

G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem. CAMEL: Communicative agents for "mind" exploration of large scale language model society. In *Advances in Neural Information Processing Systems*, volume 36, pp. 51991–52008, 2023.

S. S. Li, V. Balachandran, S. Feng, J. S. Ilgen, E. Pierson, P. W. Koh, and Y. Tsvetkov. MediQ: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 28858–28888, 2024.

Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023.

Y. Liu, C. Shi, L. Wu, X. Lin, X. Chen, Y. Zhu, H. Tan, and W. Zhang. Development and validation of a large language model-based system for medical history-taking training: Prospective multicase study on evaluation stability, human-ai consistency, and transparency. *JMIR Medical Education*, 11:e73419, 2025. doi: 10.2196/73419.

Z. Liu, L. Xiao, R. Zhu, H. Yang, and M. He. MedGen: An explainable multi-agent architecture for clinical decision support through multisource knowledge fusion. In *2024 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 6474–6481, 2024.

D. McDuff, M. Schaekermann, T. Tu, A. Palepu, A. Wang, J. Garrison, K. Singhal, et al. Towards accurate differential diagnosis with large language models. *Nature*, 642:451–457, 2025.

OpenAI. GPT-5.4 model documentation. OpenAI API Documentation, March 2026. URL `https://developers.openai.com/api/docs/models/gpt-5.4`.

Joon Sung Park, Joseph O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023.

R. Plackett, A. P. Kassianos, S. Mylan, M. Kambouri, R. Raine, and J. Sheringham. The effectiveness of using virtual patient educational tools to improve medical students' clinical reasoning skills: A systematic review. *BMC Medical Education*, 22:365, 2022.

Y. Qu, X. Xu, Y. Long, Y. Wang, J. Li, and X. Lu. A large language model-powered multiagent framework emulating standardized patients in clinical communication skills training: Development and evaluation study. *Journal of Medical Internet Research*, 28:e84747, 2026. doi: 10.2196/84747.

Khaled Saab, Chunjong Park, Tim Strother, Jan Freyberg, David G. T. Barrett, Yong Cheng, Wei-Hung Weng, et al. Advancing conversational diagnostic AI with multimodal reasoning. *Nature Medicine*, 32:1726–1736, 2026. doi: 10.1038/s41591-026-04371-0.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. AgentClinic: A multimodal benchmark for tool-using clinical AI agents. *npj Digital Medicine*, 9:499, 2026. doi: 10.1038/s41746-026-02674-7.

S. Servantez, J. Barrow, K. Hammond, and R. Jain. Chain of logic: Rule-based reasoning with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2721–2733, 2024.

N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652, 2023.

K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180, 2023.

K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31:943–950, 2025.

X. Tang, A. Zou, Z. Zhang, Z. Li, Y. Zhao, X. Zhang, A. Cohan, and M. Gerstein. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 599–621, 2024.

T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P.-C. Chang, A. Carroll, et al. Towards conversational diagnostic artificial intelligence. *Nature*, 642:442–450, 2025.

D. Van Veen, C. Van Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E. P. Reis, A. Seehofnerová, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4):1134–1142, 2024.

W. Wang, Z. Ma, Z. Wang, C. Wu, J. Ji, W. Chen, X. Li, and Y. Yuan. A survey of LLM-based agents in medicine: How far are we from Baymax? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10345–10359, 2025.

H. Wei, J. Qiu, H. Yu, and W. Yuan. MEDCO: Medical education copilots based on a multi-agent framework. In *Computer Vision – ECCV 2024 Workshops*, volume 15630 of *Lecture Notes in Computer Science*, pp. 119–135. Springer, 2025.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

World Health Organization. *International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10), Version 2019*. World Health Organization, Geneva, 2019.

Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, A. H. Awadallah, R. W. White, D. Burger, and C. Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *Proceedings of the First Conference on Language Modeling*, 2024.

Huizi Yu, Jiayan Zhou, Lingyao Li, Shan Chen, Jack Gallifant, Anye Shi, Jie Sun, Xiang Li, et al. Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education. *Communications Medicine*, 6:27, 2026. doi: 10.1038/s43856-025-01283-x.

J. Zeng, W. Qi, S. Shen, X. Liu, S. Li, B. Wang, C. Dong, X. Zhu, Y. Shi, X. Lou, et al. Embracing the future of medical education with large language model-based virtual patients: Scoping review. *Journal of Medical Internet Research*, 27:e79091, 2025. doi: 10.2196/79091.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023.