



浙江大学
ZHEJIANG UNIVERSITY

ClinMAS: A Knowledge-Grounded Multi-Agent Simulation Framework for Evaluating Clinical Reasoning in LLMs

Yuqi Tang¹, Jing Yu¹, Zichang Su¹, Kehua Feng¹, Zhihui Zhu¹
Libin Wang¹, Lei Liang², Qiang Zhang¹, Keyan Ding¹, Huajun Chen¹

¹ Zhejiang University ² Ant Group

NeurIPS 2026

Motivation



Interactive diagnosis

Physicians ask questions, request tests, and refine diagnoses over multiple turns.

Limits of static benchmarks

Question-answering tests miss iterative reasoning and risk information leakage or contamination.

Incomplete evaluation

Existing simulators use limited cases and rarely separate physical exams from auxiliary tests.

Contributions



Multi-agent framework

Doctor, patient, and examiner agents enable scalable, multi-turn diagnostic evaluation.

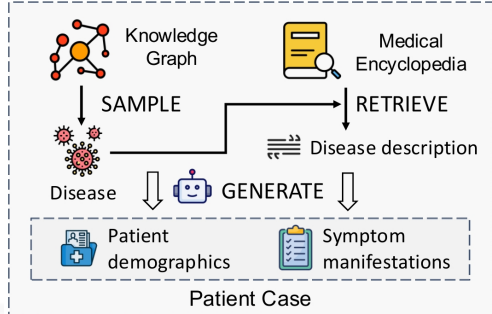
Knowledge-grounded assessment

Dynamic cases and process-aware metrics assess diagnostic efficiency and quality.

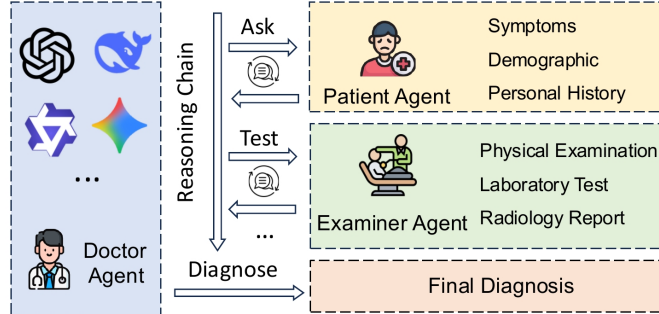
Comprehensive LLM analysis

Fine-grained evaluation of 15 LLMs reveals systematic clinical reasoning deficiencies.

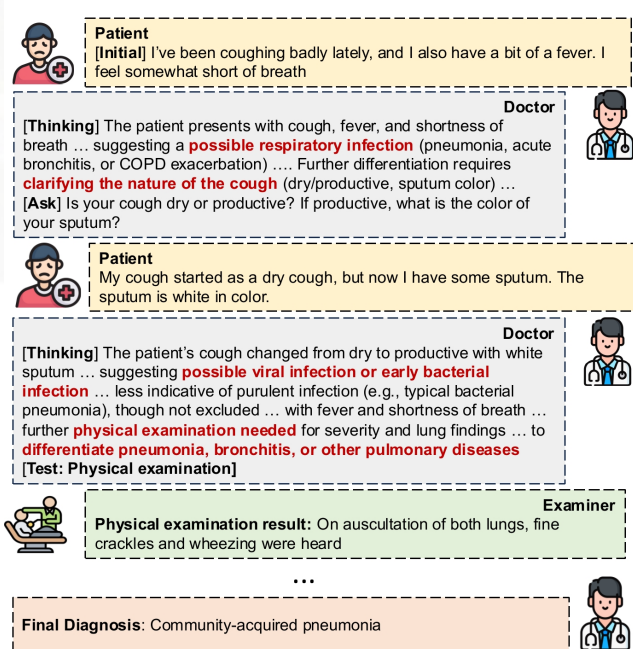
a. Knowledge-Grounded Case Generation



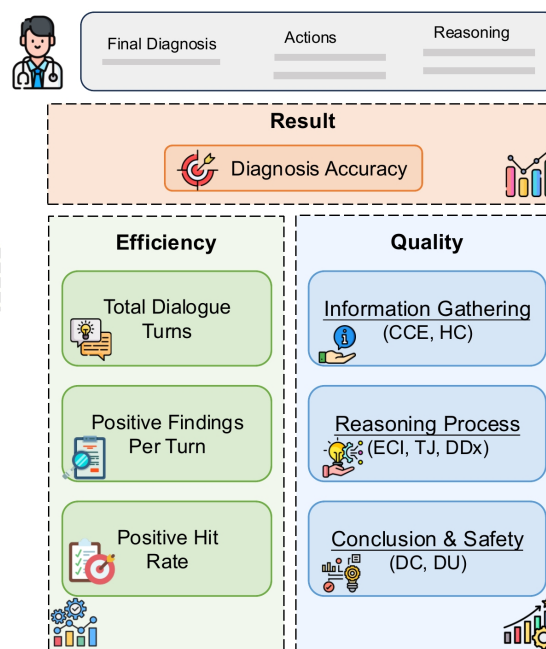
b. Multi-Agent Environment



c. Diagnostic Dialogue



d. Evaluation Metric



Knowledge-grounded case generation

A disease–symptom knowledge graph and medical encyclopedia generate coherent, previously unseen patient cases.

Multi-agent diagnostic dialogue

The doctor uses Ask, Test, and Diagnose. Patient and examiner agents respond using evidence from the shared case profile.

Evaluation metrics

Measure diagnostic accuracy and efficiency with OSCE-inspired evaluation of reasoning quality across seven clinical dimensions.

Experiments



Settings

Models: 15 LLMs

Patient and examiner: GPT-5

Temperature = 0.0, top-p = 1.0

Cases: 5 runs × 300 per run

Dialogue limit: 15 turns

Reporting: Mean ± standard error

Model	Acc. ↑	Tot. Turns ↓	Pos. Find. ↑	Neg. Find. ↓	PHR ↑	DQS ↑
Gemini-2.5-Pro	72.87 ± 1.43	6.57 ± 0.14	5.17 ± 0.11	2.22 ± 0.17	69.93 ± 1.90	70.0 ± 0.8
GPT-4.1-mini	71.73 ± 0.98	8.27 ± 0.11	7.97 ± 0.24	4.60 ± 0.16	63.41 ± 1.31	68.6 ± 0.4
Claude-Sonnet-4	69.60 ± 2.28	6.52 ± 0.08	6.98 ± 0.09	2.84 ± 0.18	71.09 ± 1.50	65.2 ± 0.7
GPT-4.1	69.47 ± 2.36	6.35 ± 0.07	5.36 ± 0.08	3.71 ± 0.11	59.11 ± 0.91	66.7 ± 0.7
Llama-4-Maverick	69.47 ± 3.01	6.99 ± 0.10	5.36 ± 0.12	2.61 ± 0.10	67.23 ± 1.12	63.7 ± 1.0
DeepSeek-V3	68.73 ± 2.77	5.71 ± 0.08	5.66 ± 0.08	3.14 ± 0.11	64.31 ± 0.72	63.9 ± 1.2
GPT-5-mini	67.67 ± 2.36	5.59 ± 0.11	5.01 ± 0.13	3.43 ± 0.22	59.35 ± 1.22	69.5 ± 0.7
DeepSeek-R1	63.40 ± 2.55	3.86 ± 0.07	3.51 ± 0.05	1.67 ± 0.06	67.74 ± 0.93	59.8 ± 0.6
GPT-4o	63.13 ± 1.98	5.34 ± 0.08	4.57 ± 0.08	3.14 ± 0.08	59.25 ± 0.77	60.4 ± 1.1
GPT-5-nano	63.00 ± 2.51	5.74 ± 0.12	4.78 ± 0.09	3.22 ± 0.17	59.74 ± 1.44	63.5 ± 0.7
Qwen3-235B-A22B	59.87 ± 2.16	5.11 ± 0.06	5.05 ± 0.14	2.65 ± 0.09	65.58 ± 1.30	58.4 ± 0.7
Qwen3-Next-80B-A3B	57.87 ± 1.73	4.84 ± 0.08	3.76 ± 0.06	2.92 ± 0.19	56.25 ± 1.77	57.0 ± 0.7
Llama-4-scout	55.73 ± 2.66	6.07 ± 0.13	5.31 ± 0.22	3.26 ± 0.11	62.00 ± 1.49	56.4 ± 0.8
Qwen3-8B	50.93 ± 2.47	4.32 ± 0.07	4.05 ± 0.08	3.44 ± 0.12	54.06 ± 1.02	54.7 ± 0.6
Qwen2.5-7B	48.27 ± 1.55	5.66 ± 0.08	5.07 ± 0.10	3.62 ± 0.15	58.35 ± 1.46	52.7 ± 0.4

Leading Accuracy

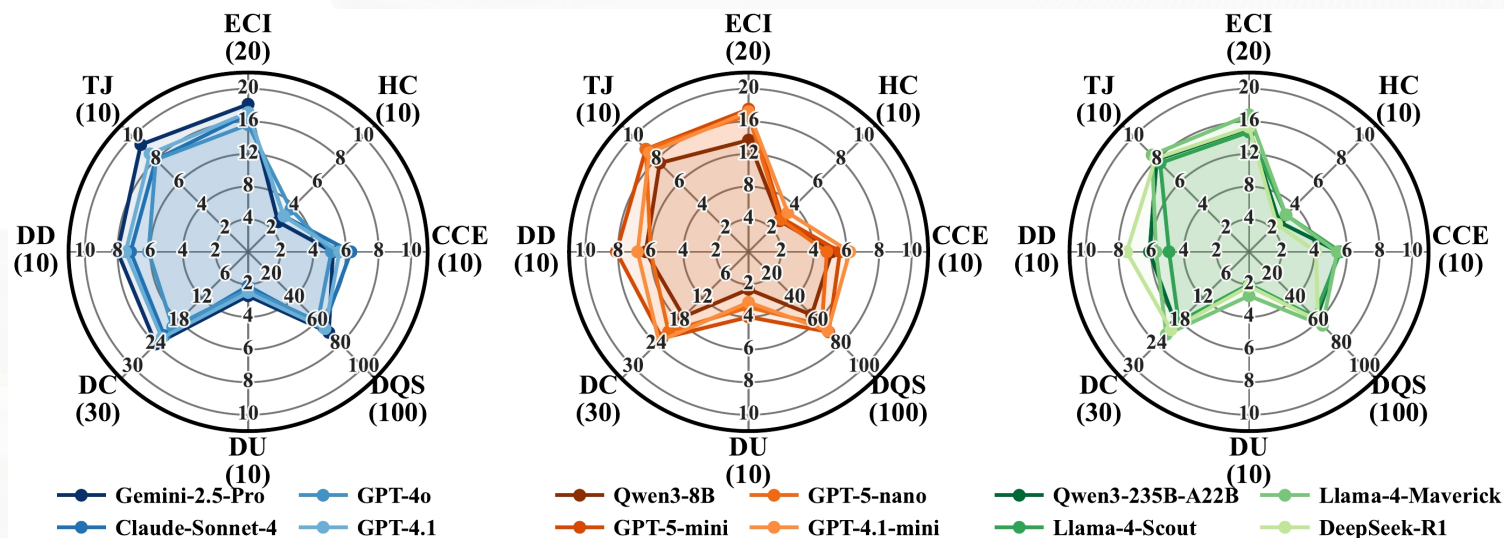
Gemini-2.5-Pro: **72.87%**.
Strong open-source models
narrow the gap.

Efficiency–accuracy trade-off

DeepSeek-R1: **3.86 turns**
and **63.40%** accuracy, with
less evidence gathered.

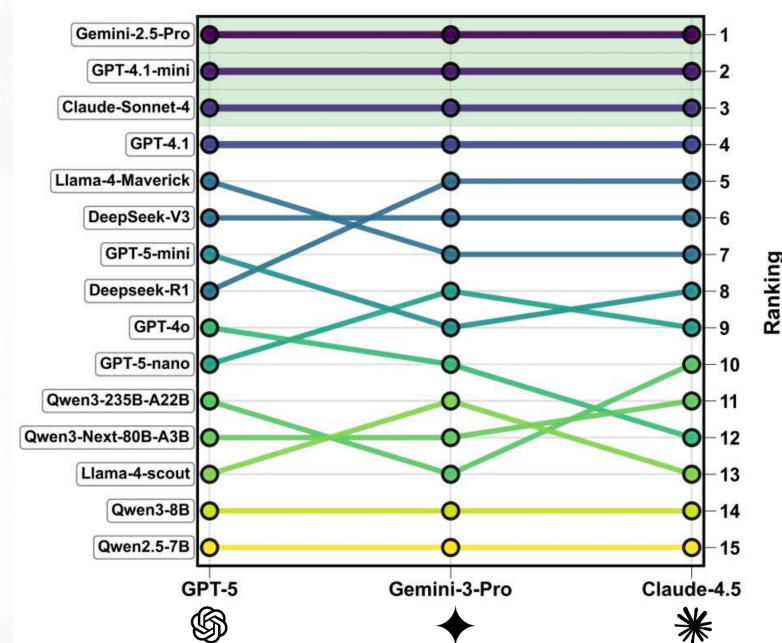
Table 1. Diagnostic accuracy, efficiency, and quality on ClinMAS

Clinical Reasoning Quality



- **Information gathering:** Active inquiry remains weak; the best chief complaint exploration score is only **6.3/10**.
- **Core reasoning:** Top models integrate evidence better; Gemini-2.5-Pro reaches **18.1/20** in Evidence Chain Integrity.
- **Conclusion and safety:** Models overstate certainty; even GPT-5-mini scores only **4.0/10** in Diagnostic Uncertainty.

Simulator Robustness



Model comparisons remain stable
across patient-examiner backbones.

Conclusion



Interactive evaluation

Knowledge-grounded cases and multi-agent simulation evaluate clinical reasoning.

Process-aware assessment

Metrics capture accuracy, efficiency, and seven dimensions of diagnostic quality.

Persistent reasoning gaps

LLMs struggle with active inquiry, evidence use, and calibrated diagnostic uncertainty.

Future Work



Across the clinical workflow

Extend evaluation to treatment decisions and prognostic follow-up.

Multimodal clinical reasoning

Integrate medical imaging and other inputs for a holistic assessment of reasoning.