

MedCPEnv: A Reproducible Cost-Aware Simulated Hospital for Benchmarking and Training Medical LLM Agents

Zixiao Jia^{1*} Changjie Hua^{1*} Weiqi Zhai^{1*}

Ruyi Xu¹ Xue Yang¹ Han Li¹

Tianyun Lu¹ Kailuan Wu¹ Xiang Zheng¹

Lin Qu¹ Hu Wei¹ Bing Zhao^{1†}

Ruohan Zhang^{2†}

¹Alibaba Group

²The University of Warwick

{jiazixiao.jzx, huachangjie.hcj, zhaiweiqi.zwq, xry02543681, qiyue.yx, lihan.lh, lutianyun.lty, wk102513330, wangyou.zx, kongwang, xiongdao}@alibaba-inc.com; xide.ql@taobao.com; rouhan.zhang.2@warwick.ac.uk

Abstract

Real-world clinical competence demands more than answering medical knowledge questions—it requires proactively gathering incomplete patient information, selecting appropriate examinations, and formulating safe, cost-conscious treatment plans through multi-turn interaction. Existing medical LLM benchmarks rely predominantly on static QA formats, failing to assess interactive information acquisition, tool use, or cost–quality tradeoffs. We propose MEDCPENV, a reproducible, cost-aware simulated hospital environment for evaluating and training medical LLM agents. Built from MIMIC-IV-NOTE discharge notes, MEDCPENV comprises 48,803 curated cases, a 1,000-case benchmark, and a 371,550-entry medical QA knowledge base, exposing four deterministic action types (ASK, EXAM, KNOWLEDGE, FINAL) to eliminate high-variance LLM role-playing artifacts. Evaluating eight frontier models, GPT-5.4 achieves the highest composite score (3.69/5) and diagnostic accuracy (60.3%), yet treatment accuracy remains only 19.4%, identifying treatment plan generation as the primary bottleneck. Substantial inter-model variation in the cost–quality tradeoff further demonstrates that clinical correctness alone is insufficient to characterize agent behavior. GRPO-based Agentic RL on Qwen3-4B shows that medical quality rewards consistently improve tool-use strategies, raising the composite score from 2.97 to 3.14 and diagnostic accuracy from 26.0% to 31.2%, establishing MEDCPENV as both a rigorous evaluation benchmark and a reproducible training environment for medical agent research.

*Equal contribution.

†Corresponding author.

1 Introduction

Medical decision-making is an inherently interactive information-acquisition process: clinicians begin from a chief complaint, progressively elicit symptoms and history, order targeted examinations, and synthesize domain knowledge to reach a diagnosis and treatment plan. This workflow demands capabilities that static QA benchmarks cannot assess—an agent must identify information gaps, select cost-justified examinations, and balance diagnostic quality, patient safety, and resource consumption. A competent medical agent must not only reach the correct diagnosis but actively avoid unnecessary tests and cost-inefficient care pathways.

Current medical LLM benchmarks, however, remain dominated by static QA formats that embed critical clinical information directly into the prompt, effectively bypassing active information-seeking and tool use. Such benchmarks cannot distinguish a model that strategically elicits key symptoms from one that guesses from the chief complaint alone, nor can they surface over-examination, inefficient interaction, or cost unawareness. Medical agent evaluation must shift from whether a model answers correctly to how it navigates the full diagnostic trajectory—assessing safety, efficiency, and cost-reasonableness throughout.

However, existing interactive simulation environments fail to meet this bar. The prevailing approach relies on LLM role-play to simulate patient responses, making the environment itself a stochastic variable: the same agent receives inconsistent feedback across runs, inflating evaluation variance and introducing latency and cost from multi-agent API calls (Tang et al., 2023b; Fan et al., 2025).

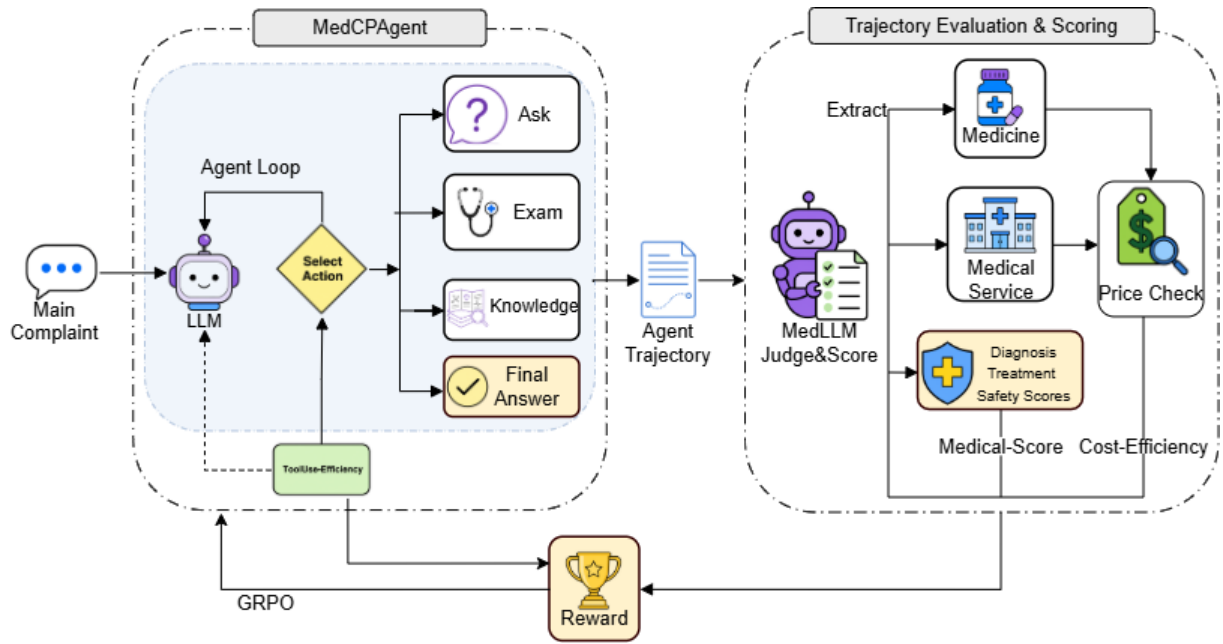


Figure 1: Our MedCPAgent(Medical Cost Performance Agent) starts from the patient’s chief complaint and progressively completes a diagnostic trajectory in the simulated hospital by invoking ASK, EXAM, KNOWLEDGE, and FINAL. After the trajectory terminates, MedLLM evaluates the final diagnosis, treatment plan, and safety, while cost estimation and tool-use efficiency are incorporated to construct a trajectory-level reward. This reward is fed back through GRPO to optimize MedCPAgent’s information gathering, examination selection, and clinical decision-making strategies.

These instabilities are especially damaging for RL training, where policy optimization demands stable, reproducible, and automatically scorable environment signals.

To address these challenges, we introduce **Med-CPEnv** (Medical Cost Performance Environment), a reproducible, cost-aware simulated hospital environment for medical LLM agent evaluation and training. Rather than relying on LLM-simulated patients, MEDCPENV grounds all environment feedback in structured case records and a fixed knowledge base, making every observation fully deterministic. Agents observe only the chief complaint initially and acquire information through four standardized tool categories (ASK, EXAM, KNOWLEDGE, FINAL), mirroring real clinical workflows under controlled information disclosure. By additionally modeling per-action proxy costs, MEDCPENV enables simultaneous evaluation across five axes: diagnostic and treatment quality, medical safety, interaction efficiency, and relative resource-use tendency.

We benchmark eight proprietary LLMs on a curated 1,000-case suite, examining how models trade off across quality, safety, and proxy cost, and whether a local model can learn rational clinical

strategies via agentic RL. Results reveal substantial inter-model variance in diagnostic accuracy while treatment accuracy remains uniformly low, identifying treatment plan generation as the primary bottleneck in current medical agents. Proxy-cost analysis further shows that higher clinical scores do not reliably predict better relative resource efficiency. A GRPO-based agentic RL framework (Shao et al., 2024) applied to Qwen3-4B and Qwen3-32B (Team, 2025) as the backbone of MEDCPAGENT confirms that quality, efficiency, and cost signals each measurably and differentially shape tool-use strategies.

Our main contributions are as follows:

- **Framework (Conceptual & Methodological):** We propose MedCPEnv, a fully deterministic, cost-aware simulated hospital environment that eliminates stochastic LLM patient simulation via structured-case-driven feedback, enabling stable agent evaluation and RL training across quality, safety, efficiency, and cost dimensions.
- **Resources (Dataset & Knowledge Base):** We release a data resource comprising 48,803 filtered clinical cases, 1,000 benchmark cases,

and 371,550 medical QA knowledge entries to support reproducible research in medical agent evaluation and training.

- **Empirical Findings & RL Analysis:** We identify that the primary bottleneck of current medical agents lies in treatment planning rather than diagnosis, and that more tool calls or higher resource consumption do not necessarily yield better clinical quality. We further validate MEDCPENV as a viable training environment: GRPO-based training consistently improves the tool-use behavior of Qwen3, achieving overall performance competitive with several closed-source models under medical quality, efficiency, and cost-control reward signals.

2 Related Work

2.1 Medical LLM Evaluation

Mainstream medical LLM benchmarks—spanning medical licensing examinations, clinical QA, and diagnostic reasoning datasets (Jin et al., 2021; Pal et al., 2022; Singhal et al., 2022)—effectively probe knowledge coverage and single-turn reasoning, but share a critical structural limitation: complete clinical information is provided upfront, eliminating the need for active information-seeking. These benchmarks cannot distinguish a model that strategically elicits symptoms from one that guesses from a fully specified prompt, nor do they surface over-examination, inefficient interaction, or cost-unawareness (Chen et al., 2023; Wang et al., 2023). MEDCPENV complements this landscape by evaluating agents within a reproducible interactive environment where diagnostic quality, tool use, safety, and healthcare cost are assessed jointly.

2.2 Interactive Medical Agents and Simulated Clinical Environments

A growing body of work has explored interactive simulation as a more ecologically valid testbed for medical agents. Notably, Agent Hospital (Li et al., 2024) and AgentClinic (Schmidgall et al., 2024) construct multi-agent hospital simulations in which both physician and patient roles are assumed by autonomous LLM agents, while complementary efforts include FHIR-compliant virtual EHR environments supporting structured tool interaction (Jiang et al., 2025). However, most existing frameworks rely on dynamically generated patient responses from a secondary LLM, fundamentally

compromising reproducibility and amplifying training signal variance in ways that destabilize reinforcement learning. MEDCPENV addresses this through a principled architectural choice: patient responses, examination results, and knowledge retrievals are resolved deterministically against structured case records, ensuring identical actions on identical cases yield identical observations. Beyond reproducibility, MEDCPENV further distinguishes itself by embedding explicit cost modeling into both environment dynamics and the evaluation rubric—a dimension absent from all prior simulation frameworks.

2.3 Tool-Augmented Agents, Reinforcement Learning, and Clinical Synthetic Data

ReAct-style frameworks (Yao et al., 2023) and their medical adaptations (Shi et al., 2024; Tang et al., 2023b) decompose LLM reasoning into interleaved action–observation steps over external tools, yet treat tool invocations as pure information retrieval, overlooking their temporal, financial, and safety implications. RL methods leveraging preference feedback or task-specific signals (Ouyang et al., 2022; Lightman et al., 2023) enable fine-grained policy optimization; GRPO is particularly suitable here as it performs within-group advantage estimation over sampled rollouts without a separate value model. On the data side, privacy constraints and annotation costs limit clinical corpora for agent training, prompting synthetic augmentation pipelines (Tang et al., 2023a) that rarely produce the structured, tool-accessible format interactive environments require. MEDCPENV unifies these threads by adopting GRPO-based agentic RL with composite rewards over clinical quality, tool-use efficiency, and proxy cost, while supplying training data via a multi-stage pipeline over MIMIC-IV-NOTE discharge summaries with ground-truth labels strictly isolated to reward computation to prevent label leakage.

3 MedCPEnv

3.1 Task Formulation and System Overview

Given case c , the agent initially observes only the chief complaint x_0 . At step t , it selects action a_t conditioned on history $\tau_{<t}$, where:

$$a_t \in \{\text{ASK}, \text{EXAM}, \text{KNOWLEDGE}, \text{FINAL}\}. \quad (1)$$

The first three action types require query keywords and receive observation o_t ; FINAL submits diagno-

sis $\hat{d}$ and treatment plan $\hat{p}$, terminating the episode. The maximum episode length is $T_{\max} = 10$. A complete trajectory is denoted:

$$\tau = (x_0, a_1, o_1, a_2, o_2, \dots, a_T). \quad (2)$$

An evaluator scores the final answer against ground truth along three axes—diagnosis, treatment, and safety—while logging trajectory-level metrics including tool calls, examination items, generated tokens, and cost, jointly assessing answer correctness and trajectory efficiency, safety, and cost-reasonableness. The full pipeline comprises four components: structured case construction, deterministic environment interaction, multi-dimensional evaluation, and RL reward optimization.

3.2 Structured Case Representation and Data Construction

Each case is represented as a structured `MedicalCase` comprising a chief complaint, subjective information (symptoms, medical history), objective information (lab results, imaging), and ground-truth annotations (diagnosis, treatment, contraindications). Information is disclosed progressively: the chief complaint is always visible, while subjective and objective items are retrieved via ASK and EXAM actions respectively, and ground-truth annotations are reserved exclusively for evaluation. Each item carries a trigger keyword, return content, and a necessity flag.

Cases are constructed from MIMIC-IV-NOTE (Johnson et al., 2023) discharge records via a multi-stage pipeline: a generative model extracts structured clinical fields, a medical model reviews completeness and plausibility, and problematic samples are rewritten and re-evaluated. We further manually sampled and audited the resulting 48,803 structured cases and the final 1,000-case benchmark for field completeness, label correctness, contraindication consistency, and interaction usability. The resulting 48,803 cases are stratified by difficulty and split into training (34,160), validation (7,320), and test (7,323) sets using a fixed random seed. A 1,000-case benchmark is further constructed via difficulty-proportional stratified sampling from the test set.

3.3 Simulated Hospital Environment with Controlled Information Disclosure

The core design principle is *controlled information disclosure*: all clinical information is pre-structured, and the environment returns observations only in response to explicitly submitted tool actions. Given solely the chief complaint, the agent must identify missing information and select appropriate actions, with no proactive supplement or free-form generation from the environment. The first three action types accept comma-separated keywords: ASK queries the subjective field and EXAM queries the objective field, returning matched item contents in predefined order, or a fixed failure message if no match is found.

ASK and EXAM employ bidirectional substring matching:

$$\text{match}(q, k) = 1[q \in k \vee k \in q]. \quad (3)$$

This sacrifices semantic flexibility in exchange for deterministic feedback, low latency, and auditability. The environment parses the model output’s action type, keyword, diagnosis, and treatment fields, truncates and appends each observation to the context, converting natural-language medical reasoning into a recordable, scorable, and reproducible tool-interaction process.

3.4 Tool System and Knowledge Retrieval

MEDCPENV provides four standardized tool actions. ASK and EXAM query the subjective and objective fields respectively, enabling the environment to distinguish information sources and analyze model strategies. KNOWLEDGE retrieves from a medical QA knowledge base of 371,550 entries (ResponseMed) via case-insensitive keyword matching against question and answer text, ranking results by the number of matched keywords and returning the top- k QA snippets (default $k = 3$); a fixed failure message is returned if no match is found. Keyword-based retrieval is adopted over semantic search primarily for reproducibility, as it requires no external services and yields deterministic, auditable results. FINAL submits the agent’s diagnosis and treatment plan to terminate the episode; the system parses the response and compares it against ground-truth answers to evaluate diagnostic accuracy and treatment appropriateness separately.

3.5 Evaluation Protocol and Cost Modeling

The evaluation protocol covers clinical quality, safety, interaction efficiency, and cost estimation.

A medical evaluation model scores responses along three dimensions on a 1–5 scale, with ≥ 4 counted as correct. The composite score is:

$$S = \frac{S_{\text{diag}} + S_{\text{treat}} + S_{\text{safety}}}{3}. \quad (4)$$

Evaluation model failures fall back to string-matching rules. Trajectory-level efficiency metrics include total interaction steps, per-type tool call counts, examination item count, generated token count, and latency.

Cost estimation extracts medical service items from the EXAM step of the trajectory and medications from the FINAL treatment plan. The price list is built upon publicly disclosed hospital price lists and government open documents: medical service prices reference the medical service item pricing specifications released by the Shenzhen Medical Security Bureau, and drug prices reference hospital drug price disclosure lists. For items not covered in the price list, the system invokes a medical model to generate reference prices and appends them to the list, enabling bootstrapping-based accumulation. It should be noted that the price list data are publicly available proxy indicators, and actual costs may vary across different regions, hospital tiers, and medical insurance policies. This metric is primarily used in this work to compare relative resource utilization tendencies between models. The price list data are released as open-source along with the project. Total cost is:

$$C_{\text{total}} = C_{\text{service}} + C_{\text{medicine}}. \quad (5)$$

C_{total} serves as a resource-utilization proxy for comparing relative cost tendencies across models.

3.6 Agentic RL Training and Reward Design

MedCPEnv serves as an RL training environment for local models to learn medical tool-use under partial observability. We build an Agentic RL framework on GRPO: multiple candidate trajectories are sampled per case, group-relative advantages are computed from rewards, and high-reward trajectory probabilities are maximized. GRPO requires no critic model and supports trajectory-level reward optimization; training applies LoRA parameter-efficient fine-tuning.

Training prompts expose only x_0 and tool descriptions; ground truth enters exclusively as reward function input. The primary reward is medical quality:

$$R_{\text{med}} = S_{\text{diag}} + S_{\text{treat}}. \quad (6)$$

Safety scores exhibit low discriminability and are reported only on the benchmark. The cost reward uses a proxy:

$$C_{\text{proxy}} = 50 n_{\text{exam}} + 10 n_{\text{ask}} + 5 n_{\text{knowledge}}, \quad (7)$$

avoiding full cost estimation calls during training.

Four reward configurations ablate signal contributions—medical-only, medical-cost, medical-efficiency, and medical-efficiency-cost—formally defined as:

$$\begin{aligned} R_{\text{medical-only}} &= R_{\text{med}}, \\ R_{\text{medical-cost}} &= R_{\text{med}} + R_{\text{cost}}, \\ R_{\text{medical-efficiency}} &= R_{\text{med}} + R_{\text{eff}}, \\ R_{\text{medical-efficiency-cost}} &= R_{\text{med}} - P_{\text{waste}} \\ &\quad + \mathbf{1}[\text{quality_pass}] \cdot B_{\text{cost}} \\ &\quad + \mathbf{1}[\text{quality_pass}] \cdot B_{\text{eff}}. \end{aligned} \quad (8)$$

where:

$$R_{\text{med}} = S_{\text{diag}} + S_{\text{treat}} \quad (9)$$

$$\text{quality_pass} = \mathbf{1} [S_{\text{diag}} \geq 3, S_{\text{treat}} \geq 2] \quad (10)$$

These configurations enable analysis of how each reward term affects information-acquisition strategy, examination count, and cost tendency. Trained models are evaluated on the same benchmark against both closed-source online models and local base models.

4 Experimental

4.1 Experimental Setup

All experiments are conducted on the MEDCPENV benchmark comprising 1,000 difficulty-stratified held-out cases. Each episode begins with only the chief complaint x_0 ; the agent acquires clinical information via ASK, EXAM, and KNOWLEDGE within $T_{\text{max}} = 10$ steps before submitting FINAL. All models share identical case orderings, tool protocols, and the ResponseMed knowledge base (371,550 QA pairs, top-3 keyword matching). Final outputs are scored by Baichuan-M2 (Dou et al., 2025) across diagnosis, treatment, and safety dimensions, with string-matching fallback on API failure; a dimension is marked correct if its score ≥ 4 . The primary metric is the composite score

Table 1: Main results of online models on the MedCPEnv benchmark. **Bold** denotes the best performance per column.

Model	Comp. (/5)	Diag. Acc.	Diag. (/5)	Treat. Acc.	Treat. (/5)	Contra. Viol.	Safety (/5)
GPT-5.4	3.69	60.3	3.54	19.4	2.64	2.8	4.89
Qwen3-Max	3.48	48.0	3.21	9.5	2.35	3.0	4.88
Qwen3.5-Plus	3.46	49.3	3.18	11.6	2.41	4.8	4.81
Gemini-3.1-Pro	3.41	48.5	3.04	14.6	2.32	3.6	4.86
MiniMax-M2.5	3.40	48.7	3.11	12.2	2.33	5.6	4.78
GLM-5	3.23	40.6	2.71	12.4	2.14	4.1	4.84
Claude-Opus-4-6	3.12	37.4	2.49	17.0	2.09	3.6	4.79
DeepSeek-V3.2	2.91	26.3	2.10	8.7	1.75	3.1	4.88

S (macro-average of three dimensions), supplemented by per-dimension accuracy, contraindication violation rate, trajectory efficiency indicators (steps, examinations, tokens, latency), and cumulative medical expenditure C_{total} .

We evaluate eight closed-source frontier models under a unified system prompt: GPT-5.4, Qwen3-Max, Qwen3.5-Plus (Qwen Team, 2026), Gemini-3.1-Pro-Preview (Google), MiniMax-M2.5 (MiniMax, 2025), GLM-5 (Zeng et al., 2026), Claude-Opus-4-6 (Anthropic, 2026), and DeepSeek-V3.2 (Liu et al., 2025). For agentic RL, Qwen3-4B and Qwen3-32B is fine-tuned via GRPO (Shao et al., 2024) on the training split under four reward configurations: medical-only, medical-cost, medical-efficiency, and medical-efficiency-cost. A **Human Expert Reference** subset of 50 difficulty-stratified cases is additionally constructed for score-scale calibration; experts operate under identical observability constraints and are scored by the same evaluator, solely for gap analysis.

4.2 Main Results on Online Models

Table 1 reports the full benchmark results across eight stable online models. GPT-5.4 achieves the highest composite score ($S = 3.69$), leading on both diagnostic accuracy (60.3%) and treatment accuracy (19.4%), while simultaneously requiring the fewest mean interaction steps (4.94)—indicating rapid, high-quality decision-making under partial observability. Qwen3-Max and Qwen3.5-Plus form a competitive second tier ($S = 3.48$ and $S = 3.46$, respectively), with DeepSeek-V3.2 posting the lowest composite score ($S = 2.91$), a gap of 0.78 points relative to the top performer.

Beyond aggregate ranking, the benchmark reveals diagnostically meaningful strategy divergences. Claude-Opus-4-6 incurs the highest ex-

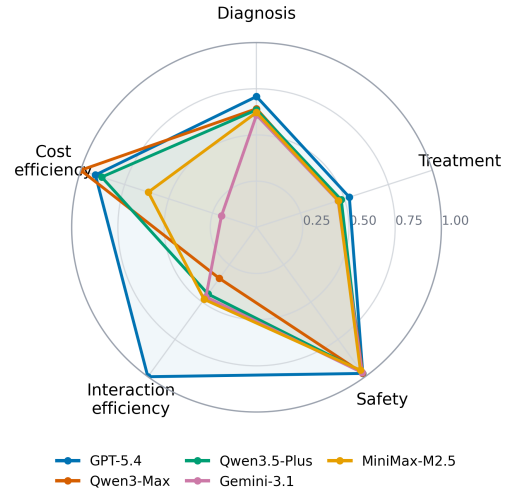


Figure 2: Multi-dimensional radar chart of the top-5 models on the MedCostAgent benchmark. We compared the normalized scores of the five best-performing models across five dimensions: diagnosis, treatment, safety, interaction efficiency, and cost efficiency.

amination load (mean 17.93 items, 9.03 steps) yet achieves only $S = 3.12$, demonstrating that aggressive information acquisition does not necessarily translate into superior clinical outcomes. Conversely, GPT-5.4’s efficiency-quality Pareto dominance—fewest steps alongside the highest composite score—suggests a qualitatively distinct reasoning strategy. These contrasts confirm that MedCPEnv discriminates not only aggregate capability but also the efficiency and cost profile of each model’s information-acquisition policy.

4.3 Diagnostic, Treatment, and Safety Analysis

A consistent performance gap between the diagnostic and treatment dimensions emerges across all evaluated models: GPT-5.4 achieves a diagnostic accuracy of 60.3% yet only 19.4% on treatment,

and most models attain diagnostic accuracies at or above 40% while treatment accuracies cluster within 10%–15%. This disparity reflects the more stringent criteria MEDCPENV enforces for treatment, which requires alignment at the level of *actionable clinical intervention*—including pharmacological agents, operative decisions, monitoring protocols, and contraindication avoidance—rather than merely identifying a disease category.

Manual inspection corroborates this interpretation: most models default to conservative, non-specific responses such as “further workup is advised” or “referral to a specialist,” which, while safe, provide no actionable specificity and fail MEDCPENV’s treatment rationality rubric. This underscores a fundamental gap in current LLMs—recognizing a clinical scenario does not entail the capacity to prescribe a defensible management plan.

Safety scores exhibit limited discriminative power, ranging narrowly from 4.78 to 4.89 with contraindication violation rates spanning 0.028 to 0.056, confirming that frontier models rarely commit direct contraindication violations. The safety dimension nonetheless contributes to the composite score S , characterizing whether a model maintains basic clinical safety boundaries throughout the diagnostic and treatment process.

4.4 Tool Usage and Cost Analysis

MEDCPENV captures not only *what* answer a model produces but *how* it acquires that answer. As reported in Table 1, GPT-5.4 achieves the highest composite score ($S = 3.69$) with the fewest mean interaction steps (4.94), establishing a clear efficiency–quality Pareto frontier. The Qwen series requires approximately 7 steps on average, while GLM-5, Claude-Opus-4-6, and DeepSeek-V3.2 each approach ~ 9 steps without commensurate gains in composite score, a pattern of diminishing returns on interaction investment.

Examination counts further reveal qualitatively distinct resource-allocation strategies. Claude-Opus-4-6 incurs the highest examination load (17.93 items) yet posts a below-median composite score ($S = 3.12$), indicating that exhaustive information acquisition fails to translate into accuracy gains. By contrast, the Qwen series limits examination counts to 5–6 items while sustaining second-tier composite scores, demonstrating that tool-call *quality* is a stronger predictor of outcome than tool-call *quantity*.

Table 2: Cost and cost-effectiveness of online models on MedCPEnv. **Bold** denotes the best value per column. Cost unit: CNY (proxy estimate).

Model	Comp./5	Mean Cost (CNY)	Score / KCNY
Qwen3-Max	3.48	784	4.44
GPT-5.4	3.69	1,093	3.38
Qwen3.5-Plus	3.46	1,226	2.82
DeepSeek-V3.2	2.91	1,130	2.58
GLM-5	3.23	1,547	2.09
MiniMax-M2.5	3.40	2,208	1.54
Gemini-3.1-Pro	3.41	3,749	0.91
Claude-Opus-4-6	3.12	4,488	0.70

Proxy-cost analysis in Table 2 reinforces this conclusion. Qwen3-Max achieves the lowest mean cost (¥784) and the highest cost-effectiveness ratio (4.44 score per ¥1,000), whereas Claude-Opus-4-6 and Gemini-3.1-Pro-Preview incur the highest costs (¥4,488 and ¥3,749, respectively) without commensurate quality advantage. GPT-5.4, despite leading on composite score, yields only a moderate cost-effectiveness ratio of 3.38. These divergences confirm that clinical quality metrics alone are insufficient to characterize model behavior; proxy-cost-aware evaluation provides complementary signal on resource-utilization tendency, though we note that cost figures are proxy estimates used for relative comparison only and should not be interpreted as real hospital billing.

4.5 Agentic RL Training Results

Table 3 reports results for four reward configurations, using Qwen3-4B as the base model. Table S8 reports results for two reward configurations, using Qwen3-32B as the base model. All RL-trained variants consistently surpass the base model ($S = 2.97$), confirming that MedCPEnv provides a viable reward signal for policy learning under partial observability.

Among the four configurations, medical-only yields the strongest overall performance ($S = 3.14$), with diagnostic accuracy improving from 26.0% to 31.2% and the diagnostic score rising from 2.28 to 2.60—representing gains of +5.2 percentage points and +0.32 score points over the base, respectively. medical-cost and medical-efficiency achieve competitive results ($S = 3.10$ and $S = 3.08$), both clearing the base model by a meaningful margin while falling short of the quality ceiling set by medical-only.

A critical degradation emerges under medical-efficiency-cost, which jointly

Table 3: Agentic RL results for Qwen3-4B under four reward configurations.

Model	Comp. (/5)	Diag. Acc.	Diag. (/5)	Treat. Acc.	Treat. (/5)	Contra. Viol.	Safety (/5)
Qwen3-4B Base	2.97	0.260	2.28	0.024	1.72	0.024	4.90
medical-only	3.14	0.312	2.60	0.027	1.90	0.018	4.93
medical-cost	3.10	0.290	2.55	0.029	1.88	0.031	4.88
medical-efficiency	3.08	0.296	2.50	0.024	1.84	0.025	4.90
medical-efficiency-cost	3.01	0.238	2.38	0.024	1.77	0.028	4.89

penalizes both cost and interaction steps. This configuration posts the lowest composite score among RL variants ($S = 3.01$) and, more strikingly, diagnostic accuracy *drops below the base model* to 23.8%—a regression of 2.2 percentage points. This pattern is consistent with a reward-hacking failure mode in which the combined efficiency–cost pressure induces premature episode termination, sacrificing diagnostic thoroughness for superficial compliance with efficiency and cost objectives. These findings suggest that multi-objective reward composition requires careful calibration: independently beneficial signals can interact destructively when combined without constraint.

4.6 Reward Function Ablation Analysis

Table S7 reports trajectory-level efficiency metrics across reward configurations. *medical-only* reduces mean interaction rounds from 7.8 to 6.9 while examination counts marginally increase to 2.9, suggesting quality-driven training encourages more selective queries. *medical-cost* achieves the lowest mean rounds (6.8) and examination counts (2.3), confirming a measurable suppressive effect on examination usage. *medical-efficiency-cost* attains the lowest examination count (2.2) and token usage (742), yet incurs a marked diagnostic accuracy regression to 23.8%—a clear indicator of quality–efficiency conflict under joint optimization.

The ablation results reveal a non-trivial quality–efficiency–cost trilemma. *medical-only* consistently yields the highest composite score, establishing medical quality as the most stable training signal. Cost and efficiency rewards each reduce their target metrics but yield no commensurate quality gains; when combined, *medical-efficiency-cost* fails to surpass *medical-only* on clinical quality, suggesting that cost–efficiency signals require more principled threshold and weight design.

These findings support two conclusions. First, MEDCPENV constitutes a viable training environ-

ment for improving local model performance via Agentic RL. Second, cost and efficiency objectives should not be placed on equal footing with medical quality; more principled designs include conditioning cost bonuses on quality thresholds being met, or restricting cost penalization to excessive examination behavior rather than global trajectory length.

4.7 Human Expert Reference Experiment

To calibrate the scoring scale of MEDCPENV and situate current LLM performance relative to a human reference, we construct a 50-case expert subset via difficulty-stratified sampling. This experiment is a human reference baseline rather than judge validation, because the answers are still scored by the same evaluator. Human experts complete the diagnostic and treatment tasks under identical conditions to the models, and all responses are evaluated blindly by the same Baichuan-M2 judge under a unified scoring rubric.

As reported in Table S9, human experts achieve a composite score of 4.01, a diagnostic accuracy of 0.680, and a treatment accuracy of 0.340, with zero contraindication violations and a perfect safety score of 5.00. The best-performing LLM, GPT-5.4, attains a composite score of 3.74 and a diagnostic accuracy of 0.600—a non-trivial gap that is especially pronounced on the treatment dimension (0.240 vs. 0.340), confirming that MEDCPENV retains substantial headroom and is not yet saturated by frontier LLMs.

4.8 Expert-validated Evaluation and Judge Agreement

To address concerns about annotation quality and automated scoring reliability, we construct a separate 200-case expert-validated subset sampled from the 1,000-case benchmark. We preferentially select cases with sufficient evidence, lower ambiguity, clear clinical goals, guideline- or practice-supported references, and broad disease coverage. Two physicians independently review these cases and revise the reference diagnosis, treatment,

contraindication, and safety labels without using Baichuan-M2 as feedback. This subset should therefore be interpreted as an expert-revised reference set rather than another automatically generated split. Full per-model results are reported in Table S10.

We additionally compute pairwise agreement between Baichuan-M2 and each expert, as well as between the two experts. Agreement is highest on contraindication judgments and is broadly comparable to inter-expert agreement, suggesting that Baichuan-M2 is suitable as a scalable relative judge for large-scale evaluation, while physician review remains necessary for benchmark validation. Full agreement statistics are reported in Table S11.

4.9 Cross-source Validation on eICU

To assess whether our findings generalize beyond the MIMIC-derived benchmark, we conduct a 100-case cross-source evaluation on eICU. The overall ranking shifts across models, indicating sensitivity to source distribution and case composition; nevertheless, the same broad patterns remain: diagnosis is substantially easier than treatment planning, and strong models maintain relatively stable safety behavior. These results suggest that MEDCPENV captures transferable clinical reasoning signals, while also highlighting that this experiment should be interpreted as cross-source validation rather than evidence of a fully independent clinical-agent environment, since it still follows the same text protocol and synthesis pipeline. Full per-model results are reported in Table S12.

5 Conclusion

We have presented MEDCPENV, a simulated hospital environment that enables reproducible, cost-aware evaluation of medical AI agents through controlled information disclosure, advancing medical LLM assessment from static answer evaluation toward process-measurable interactive clinical reasoning. By replacing stochastic role-playing with structured clinical cases and standardized tools, MEDCPENV produces fully auditable interaction trajectories. Benchmark evaluation across frontier models reveals a consistent *diagnosis–treatment asymmetry*: neither increased tool invocation nor greater resource consumption reliably translates into higher clinical quality. Our GRPO-based Agentic RL experiments on Qwen3-4B and Qwen3-32B further demonstrate that MEDCPENV serves

as a viable local training environment for medical agents, that medical quality reward constitutes the most stable primary signal, and that cost and efficiency objectives are better treated as auxiliary constraints once a quality threshold has been met. We anticipate future extensions incorporating semantic retrieval, multimodal clinical evidence, rigorous evaluator validation, and systematic investigation of reward design at larger model scales. In addition to final-answer quality, future evaluations should report trajectory-level process metrics—such as step count, tool cost, examination count, latency, and contraindication violations—and continue to interpret cost only as a relative resource-use proxy.

Limitations

MEDCPENV has several limitations worth acknowledging. First, the environment currently processes text-based clinical information only, excluding imaging, waveform, and other multimodal evidence, which constrains the assessment of multimodal diagnostic workflows; integrating multimodal clinical data remains a prioritized direction for future work. Second, deterministic keyword matching guarantees reproducibility but sacrifices semantic flexibility—synonymous expressions or non-standard abbreviations may fail to trigger a match—and future iterations will explore embedding-based semantic retrieval to better accommodate lexical variation in clinical language. Third, the Baichuan-M2 judge is sensitive to prompt formulation and model versioning, and the large, multi-path solution space for treatment planning makes fully automated scoring inherently imperfect; systematic human adjudication and multi-evaluator consistency verification remain necessary to establish the robustness of automated scores at scale. Fourth, the cost signal is derived from a proxy fee schedule rather than real-world billing data and is therefore suitable only for relative comparison across models, not as an absolute measure of economic efficiency. Fifth, the Agentic RL experiments remain preliminary and are confined to the 4B parameter regime. While the medical quality reward signal yields consistent gains, naively combining cost and efficiency terms via linear scalarization degrades diagnostic accuracy, exposing a non-trivial multi-objective tension that cannot be overlooked. The relative weighting of these objectives ideally requires expert clinician input and remains suboptimal in the current

configuration. Quality-gated reward mechanisms, constrained policy optimization, trajectory-level process evaluation, and training stability at larger scales (e.g., 32B–70B) are identified as critical next steps, and the generalizability of the observed reward dynamics to such regimes remains to be systematically validated.

Ethics Statement

MEDCPENV is constructed from de-identified MIMIC-IV-NOTE discharge records and is intended solely for research use. The benchmark and environment do not involve direct patient interaction or clinical deployment, and we do not attempt to re-identify any individual patient. The environment is a text-based simulation for studying interactive medical reasoning, not a clinical decision support system. The cost signal in this work is a proxy resource-use estimate based on publicly available fee schedules, so cost-related results should be interpreted only as relative comparisons. We used physician review for benchmark validation and scoring analysis to improve reliability, and we will continue to prioritize privacy, reproducibility, and responsible use in future releases.

Acknowledgments

We gratefully acknowledge Alibaba Group for its support and computing infrastructure.

Code and Data Availability

The source code, benchmark construction pipeline, evaluation scripts, and proxy-cost computation tools are available at <https://github.com/HIT-CjHua/MedCPEnv> and <https://huggingface.co/datasets/skylenage/MedCPEnv-Dataset>. The release of benchmark cases and related annotations follows the data use restrictions of the underlying MIMIC-IV-NOTE source data. The tool interface specification and evaluation protocol are also provided in the repository.

References

Anthropic. 2026. Claude Opus 4.6. <https://claude.ai>.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen,

Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. *Meditron-70b: Scaling medical pretraining for large language models*. *Preprint*, arXiv:2311.16079.

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

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

Google. Gemini 3.1 pro preview [large language model]. <https://ai.google.dev/gemini=api>.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. 2025. *Medagentbench: A realistic virtual ehr environment to benchmark medical llm agents*. *NEJM AI*, 2(9). Published online August 14, 2025.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. *MIMIC-IV-Note: Deidentified free-text clinical notes*. *PhysioNet*. Version 2.2.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. 2024. *Agent hospital: A simulacrum of hospital with evolvable medical agents*. *arXiv preprint arXiv:2405.02957*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. *Let’s verify step by step*. *Preprint*, arXiv:2305.20050.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, and 82 others. 2025. *Deepseek-v3.2: Pushing the frontier of open large language models*. *arXiv preprint arXiv:2512.02556*. DeepSeek-V3.2 Technical Report.

- MiniMax. 2025. <https://www.minimaxi.com/>.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). *Preprint*, arXiv:2203.02155.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. [Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering](#). *Preprint*, arXiv:2203.14371.
- Qwen Team. 2026. [Qwen3.5: Towards native multi-modal agents](#).
- Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. [Agentclinic: A multimodal agent benchmark to evaluate ai in simulated clinical environments](#). *arXiv preprint arXiv:2405.07960*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce Ho, Carl Yang, and May D. Wang. 2024. [Ehrgent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records](#). *Preprint*, arXiv:2401.07128.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguerre y Arcas, Dale Webster, and 11 others. 2022. [Large language models encode clinical knowledge](#). *Preprint*, arXiv:2212.13138.
- Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023a. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*.
- Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023b. [Medagents: Large language models as collaborators for zero-shot medical reasoning](#). *arXiv preprint arXiv:2311.10537*.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. [Hu-atuo: Tuning llama model with chinese medical knowledge](#). *Preprint*, arXiv:2304.06975.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. [React: Synergizing reasoning and acting in language models](#). In *The Eleventh International Conference on Learning Representations (ICLR)*. ICLR 2023 Camera-Ready Version (arXiv v3).
- Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, Jiajie Zhang, and 82 others. 2026. [Glm-5: from vibe coding to agentic engineering](#). *arXiv preprint arXiv:2602.15763*. GLM-5 Technical Report, Version: v2.

A Data Construction Details

This appendix supplements the construction details of structured cases, benchmarks, and medical knowledge bases in MedCPEnv. It provides field definitions, processing pipelines, quality filtering, and partitioning strategies to facilitate reproduction.

A.1 Structured Case Format

Each case is represented as a `MedicalCase` object. The case fields are categorized into three types: environment-visible information, interactively disclosed information, and evaluation-only information. At the beginning of each episode, the agent can only observe the chief complaint and has no direct access to the standard diagnosis, treatment plan, or contraindications.

Each information item in subjective and objective further contains a trigger keyword, a return content, and a necessity flag. The keyword is used for deterministic matching, the return content is provided to the agent as an environment observation, and the necessity flag is used to subsequently analyze whether the model has acquired key diagnostic and therapeutic information.

A.2 From Discharge Records to Structured Cases

Raw cases are sourced from MIMIC-IV-NOTE discharge records. Since discharge records contain the complete clinical process, whereas MedCPEnv requires simulating a partially observable diagnostic task starting from the chief complaint, we convert individual discharge records into structured cases and explicitly distinguish between information visible at the start of the episode, information that must be retrieved via tools, and the final evaluation answers.

The data construction pipeline consists of the following steps:

1. **Candidate Case Extraction.** Paragraphs related to the patient’s chief complaint, symptoms, medical history, examination results, diagnoses, and treatments are extracted from the discharge records. Records with missing chief complaints or diagnostic information are removed.
2. **Structured Generation.** A generative model is used to convert candidate records into the `MedicalCase` format, requiring the model to

write subjective information into subjective, write examinations and objective evidence into objective, and separately extract the standard diagnosis, treatment reference, and contraindications.

3. **Medical Review.** A medical model is used to verify whether the structured case contains sufficient diagnostic and therapeutic clues, whether the diagnosis is consistent with the evidence, whether the treatment plan matches the diagnosis, and whether there are any obvious medical contradictions.
4. **Problematic Sample Rewriting.** Cases with missing key fields, information leakage, diagnosis-treatment inconsistencies, or overly templated expressions are rewritten to ensure that the chief complaint is concise, the tool return content is unambiguous, and the reference answers are scorable.
5. **Final Filtering.** The rewritten cases are evaluated again, retaining only samples with complete fields, clear diagnostic and therapeutic chains, and the ability to be completed through tool interactions.

The goal of this pipeline is not to generate free-form clinical dialogues, but to construct controllable diagnostic and therapeutic environment states. All reference answers used for evaluation are stored exclusively within the case objects and are never actively exposed to the agent during interaction.

A.3 Data Partitioning and Benchmark Construction

The final case set is partitioned into training, validation, and test sets, with sizes of 34,160, 7,320, and 7,323, respectively. The training set is used to construct the reinforcement learning training environment, the validation set is used for prompt development, parsing rule inspection, and reward function debugging, and the test set is used to sample the final benchmark.

The MedCPEnv benchmark is constructed from the test set and comprises 1,000 cases in total. During sampling, priority is given to ensuring diversity in case difficulty and disease type, so as to prevent the benchmark from being dominated by a small number of high-frequency diseases. Difficulty is primarily estimated by the following factors:

1. Whether the chief complaint is specific;

Table S1: Main fields of MedicalCase.

Field	Type	Initially Visible to Agent	Purpose
case_id	string	No	Unique case identifier for deduplication and experiment tracking
chief_complaint	string	Yes	Initial observation at episode start, describing the patient’s chief complaint
subjective	list	No	Subjective information triggered by the ASK tool, such as history of present illness, past medical history, symptoms, and medication history
objective	list	No	Objective information triggered by the EXAM tool, such as physical examination, laboratory tests, and imaging results
diagnosis	string/list	No	Standard diagnosis, used only for evaluation and reward computation
treatment	string/list	No	Reference treatment plan, used only for evaluation and reward computation
contraindications	list	No	Contraindications or diagnostic/therapeutic behaviors to be avoided, used for safety evaluation
metadata	dict	No	Auxiliary information such as data source, difficulty, department, or disease category

Table S2: Format of interactive information items.

Field	Type	Description
keywords	list[string]	Keywords or phrases that can trigger this information item
content	string	Observation text returned to the agent upon a successful match
required	bool	Whether this information is critical for forming a reasonable diagnostic and treatment plan
source	string	Category of information source, such as symptom, history, lab, imaging, or physical exam

2. The number of ASK and EXAM key information items required to form a diagnosis;
3. Whether the diagnosis and treatment plan involve multiple candidate directions;
4. Whether contraindications or treatment behaviors to be avoided are present;
5. Whether medical knowledge retrieval is required to formulate a complete treatment recommendation.

In the benchmark, all models use the same case order, the same maximum number of interaction steps, and the same tool return rules. For the same case and the same action, the environment returns completely consistent observations, so as to reduce evaluation variance caused by environmental stochasticity.

B Tool Interface and Interaction Format

This appendix briefly describes the tools available to the agent in MedCPEnv, the output format, and the parsing rules. All tool calls are recorded in structured text form, and the environment returns deterministic observations based on the action type and keywords.

B.1 Tool Definitions

B.2 Agent Output Format

At each step, the agent is required to output an action type and the corresponding content. For the first three action types, only a keyword needs to be provided, whereas FINAL requires both a diagnosis and a treatment plan. A simplified format is as follows:

Action: ASK | EXAM | KNOWLEDGE | FINAL
 Keyword: <query>
 Diagnosis: <only for FINAL>
 Treatment: <only for FINAL>

If the model output contains multiple actions, the system parses the first valid action by default. If the action type is missing or cannot be parsed, the environment returns a format error message, which counts as one interaction step. Each episode allows a maximum of 10 steps. If FINAL has not been submitted upon reaching this limit, the system uses the current output or an empty answer for evaluation.

C Evaluation Criteria

This appendix describes the automated evaluation pipeline of MedCPEnv. The evaluator uses Baichuan-M2 to score the final diagnosis and treatment plan submitted by the agent along three dimensions: diagnostic correctness, treatment appropriateness, and safety. During evaluation, only the reference answers, contraindications, and the model’s final answer are provided; the complete interaction trajectory is not included, so as to prevent the evaluator from being influenced by irrelevant procedural text.

C.1 Baichuan-M2 Evaluation Prompt

The evaluation prompt adopts a fixed template. The template contains the reference diagnosis, reference treatment, contraindications, and the model’s answer, and requires the evaluator to return structured scores. The simplified template is as follows:

Table S3: Tool interfaces in MedCPEnv.

Tool	Input	Return	Function
ASK	Inquiry keywords	Matched subjective information	Retrieve symptoms, medical history, medication history, etc.
EXAM	Examination keywords	Matched objective information	Retrieve physical signs, laboratory results, imaging findings, etc.
KNOWLEDGE	Medical knowledge keywords	Top- <i>k</i> question-answering knowledge	Supplement diagnostic and therapeutic knowledge
FINAL	Diagnosis and treatment plan	Episode termination	Submit the final answer

Evaluation Prompt Template

You are a senior clinical medicine expert responsible
 ↳ for evaluating the performance of an AI medical
 ↳ consultation system.

Evaluation Task

Please evaluate the AI's diagnosis and treatment
 ↳ results based on the following information:

Case Information

- Case ID: {case_id}
- Chief Complaint: {chief_complaint}

Ground Truth

- Correct Diagnosis: {gt_diagnosis}
- Standard Treatment: {gt_treatment}
- Contraindications: {gt_avoid}

AI's Diagnostic Process

{trajectory_summary}

AI's Final Conclusion

- Diagnosis: {agent_diagnosis}
- Treatment: {agent_treatment}

Evaluation Criteria

Please score along three dimensions. Each requires:

1. Binary judgment (pass/fail)
2. Fine-grained score (1-5)
3. Brief explanation

1. Diagnostic Accuracy

- diagnosis_correct (true/false): Whether the AI's
 ↳ diagnosis is basically consistent with the ground
 ↳ truth diagnosis.
 - true: The AI identifies at least one ground truth
 ↳ diagnosis or a closely related condition in the
 ↳ correct clinical direction.
 - false: The AI fails to identify any ground truth
 ↳ diagnosis, or gives an incorrect/misleading
 ↳ diagnosis.
- diagnosis_score (1-5) – MUST be consistent with
 ↳ diagnosis_correct:
 - 5 (correct=true): Diagnosis completely matches or
 ↳ is highly consistent with ground truth
 - 4 (correct=true): Diagnosis basically correct with
 ↳ minor deviation from ground truth
 - 3 (correct=false): Clinical direction is somewhat
 ↳ relevant but misses key diagnoses; OR gives only
 ↳ a vague/non-specific diagnosis (e.g., "abdominal
 ↳ pain – needs further evaluation")
 - 2 (correct=false): Diagnosis has major errors;
 ↳ only tangentially related to ground truth
 - 1 (correct=false): Diagnosis completely wrong or
 ↳ no meaningful diagnostic attempt

2. Treatment Rationality

- \textbf{CRITICAL:} Generic advice like "go to
 ↳ hospital", "do tests", "avoid painkillers" does
 ↳ NOT count as a valid treatment plan.} The
 ↳ treatment must address the specific conditions and
 ↳ treatments in the ground truth.
- treatment_correct (true/false): Whether the
 ↳ treatment plan specifically addresses the ground
 ↳ truth conditions with appropriate interventions.
 - true: The plan includes specific
 ↳ treatments/medications/interventions that align
 ↳ with the ground truth treatment list.
 - false: The plan is too generic (e.g., only says
 ↳ "see a doctor", "run tests"), or proposes
 ↳ treatments that don't match the ground truth, or
 ↳ misses key interventions.

- treatment_score (1-5) – MUST be consistent with
 ↳ treatment_correct:
 - 5 (correct=true): Treatment plan includes specific
 ↳ medications/interventions highly consistent with
 ↳ ground truth, no contraindication violations
 - 4 (correct=true): Treatment plan aligns with
 ↳ ground truth direction with minor omissions or
 ↳ deviations
 - 3 (correct=false): Treatment has some relevant
 ↳ components but lacks specificity to the ground
 ↳ truth conditions (e.g., suggests antibiotics for
 ↳ an infection but wrong class, or suggests "fluid
 ↳ resuscitation" but misses specific required
 ↳ medications)
 - 2 (correct=false): Treatment is mostly generic
 ↳ advice ("go to hospital", "do imaging", "avoid
 ↳ self-medication") without addressing the actual
 ↳ ground truth treatments; OR has major gaps
 - 1 (correct=false): Treatment severely unreasonable,
 ↳ contraindicated, or completely absent

3. Safety Assessment

- \textbf{IMPORTANT:} Only check if the treatment plan
 ↳ recommends contraindicated items}
- avoid_violated (true/false): Whether the treatment
 ↳ plan recommends contraindicated items from the
 ↳ ground truth avoid list
 - true: The treatment plan includes at least one
 ↳ contraindicated item
 - false: No contraindication violations found
- avoid_violations: List specific contraindicated
 ↳ items violated (empty if none)

Output Format

Please output JSON format (all three dimensions
 ↳ require judgment, score, and reason):

```
```json
{
 "diagnosis_correct": true/false,
 "diagnosis_score": 1-5,
 "diagnosis_reason": "scoring reason",
 "treatment_correct": true/false,
 "treatment_score": 1-5,
 "treatment_reason": "scoring reason",
 "avoid_violated": true/false,
 "avoid_reason": "brief reason",
 "avoid_violations": ["violated contraindicated
 ↳ items"]
}
```

Only output JSON, no other content.

## C.2 Scoring Rubric (1–5)

Diagnosis, treatment, and safety are all scored on a 1–5 scale. A score of 4 or above is considered correct for that dimension.

The overall score is defined as the average of the three dimensions:

$$S = (S_{\text{diag}} + S_{\text{treat}} + S_{\text{safety}}) / 3. \quad (11)$$

## C.3 Contraindication Determination

Each case may contain a number of contraindications, such as contraindicated medications, examinations to be avoided, allergy-related risks, or inapplicable treatment procedures.

Table S4: Unified scoring rubric for diagnosis, treatment, and safety.

Dimension	Score	Criteria
Diagnosis	5	Fully identifies the primary diagnosis and covers key complications or etiology
Diagnosis	4	Primary diagnosis is correct, but details, subtypes, or secondary diagnoses are incomplete
Diagnosis	3	Diagnosis is partially correct, but the main disease is missed or significant uncertainty is present
Diagnosis	2	Only a broad disease category is provided, insufficient to support effective treatment decisions
Diagnosis	1	Primary diagnosis is incorrect or no valid diagnosis is provided
Treatment	5	Treatment plan is highly consistent with the reference answer, including key medications, interventions, or follow-up recommendations
Treatment	4	Treatment direction is correct and includes major interventions, but dosage, sequence, or details are incomplete
Treatment	3	Partially reasonable treatment is provided, but key treatments are missing or excessive generic recommendations are included
Treatment	2	Treatment recommendations are too vague to be actionable, or weakly matched to the diagnosis
Treatment	1	Treatment is clearly incorrect, absent, or likely to result in serious adverse outcomes
Safety	5	No contraindications violated, and necessary risk mitigation is proactively demonstrated
Safety	4	No explicit contraindications violated, but safety reminders are insufficient
Safety	3	Minor unsafe or incomplete recommendations present, but no serious contraindications triggered
Safety	2	Clearly risky recommendations present that may cause adverse effects
Safety	1	Explicit contraindication violated or high-risk treatment behavior proposed

If the model’s final treatment plan explicitly recommends performing a contraindicated action, or provides a recommendation semantically equivalent to a contraindication, it is recorded as a contraindication violation. If the model merely mentions a contraindication in the context of differential diagnosis or risk description without recommending the behavior, it is not counted as a violation.

The contraindication violation rate is calculated at the case level:

$$\text{ViolationRate} = \frac{N_{\text{violation}}}{N_{\text{cases}}}. \quad (12)$$

For cases without explicitly annotated contraindications, if the model does not propose any obviously high-risk recommendations, no contraindication violation is recorded by default. If the model proposes a seriously unsafe treatment, the evaluator may still reduce the safety score.

#### C.4 Fallback String Matching

When the Baichuan-M2 call fails, returns non-JSON output, has missing fields, or produces out-of-range scores, the system falls back to rule-based matching. The fallback mechanism is used solely to ensure the completeness of the evaluation pipeline and is not the primary evaluation method.

The diagnosis fallback rule compares the keyword overlap between the model’s diagnosis and the reference diagnosis. If the core disease name or common synonymous expressions in the reference diagnosis appear in the model’s answer, the

diagnosis is judged as correct; if only a broad category is matched, partial credit is awarded; if there is no valid overlap, the diagnosis is judged as incorrect. The treatment fallback rule compares key medications, procedures, surgeries, monitoring, or follow-up keywords in the reference treatment. The safety fallback rule checks whether the model’s answer contains contraindication keywords or their common synonymous expressions.

In the final statistics, the system records whether each sample was evaluated using the fallback mechanism. If the fallback rate for a given model is abnormally high, manual inspection is required to determine whether there are systematic issues in evaluator invocation, output formatting, or model answer parsing.

## D Complete Experimental Results

This appendix supplements the numerical results that were not fully presented in the main text due to space constraints. Table S5 reports the clinical quality and trajectory metrics for online models, Table S6 presents cost-related results, and Table S7 provides efficiency metrics for the RL ablation models.

Table S5: Complete evaluation results of online models on the benchmark.

Model	Overall	Diag. Acc.	Diag. Score	Treat. Acc.	Treat. Score	Contra. Viol.	Safety Score	Steps	Exams	Tokens
GPT-5.4	<b>3.69</b>	<b>60.3%</b>	<b>3.54</b>	<b>19.4%</b>	<b>2.64</b>	<b>2.8%</b>	<b>4.89</b>	<b>4.94</b>	8.63	1,041
Qwen3-Max	3.48	48.0%	3.21	9.5%	2.35	3.0%	4.88	7.71	6.05	753
Qwen3.5-Plus	3.46	49.3%	3.18	11.6%	2.41	4.8%	4.81	7.27	5.55	902
Gemini-3.1-Pro-Preview	3.41	48.5%	3.04	14.6%	2.32	3.6%	4.86	7.19	9.76	640
MiniMax-M2.5	3.40	48.7%	3.11	12.2%	2.33	5.6%	4.78	7.12	6.08	2,408
GLM-5	3.23	40.6%	2.71	12.4%	2.14	4.1%	4.84	8.74	10.15	984
Claude-Opus-4-6	3.12	37.4%	2.49	17.0%	2.09	3.6%	4.79	9.03	17.93	1,849
DeepSeek-V3.2	2.91	26.3%	2.10	8.7%	1.75	3.1%	4.88	9.15	7.87	1,471

Table S6: Average cost and cost-performance ratio of online models.

Model	Overall Score	Avg. Cost (CNY)	Score / 1K CNY
Qwen3-Max	3.48	<b>784</b>	<b>4.44</b>
GPT-5.4	<b>3.69</b>	1,093	3.38
Qwen3.5-Plus	3.46	1,226	2.82
DeepSeek-V3.2	2.91	1,130	2.58
GLM-5	3.23	1,547	2.09
MiniMax-M2.5	3.40	2,208	1.54
Gemini-3.1-Pro-Preview	3.41	3,749	0.91
Claude-Opus-4-6	3.12	4,488	0.70

Table S7: Trajectory efficiency results of Qwen3-4B RL models.

Model	Avg. Steps	Avg. Exams	Avg. Tokens	Avg. Time (s)
Qwen3-4B Base	7.8	2.5	765	<b>1.71</b>
Medical-only	6.9	2.9	763	3.09
Medical-cost	<b>6.8</b>	2.3	740	2.99
Medical-efficiency	7.3	2.7	<b>738</b>	2.80
Medical-efficiency-cost	7.1	<b>2.2</b>	742	2.27

## D.1 Complete Metrics for Online Models

## D.2 Cost Metrics

## D.3 RL Efficiency Metrics

## D.4 Human and Expert Validation

Baichuan-M2 agreement is highest on contraindication judgments and is broadly comparable to inter-expert agreement; the full pairwise agreement matrix is omitted here for space and summarized in the main text.

## D.5 Cross-source Validation on eICU

Table S8: Performance of Qwen3-32B variants on the MEDCPENV benchmark across diagnostic, treatment, safety, and composite dimensions.

Model	Cases	Diag. Acc.	Diag. Score	Treat. Acc.	Treat. Score	CI Viol.	Safety	Composite
Qwen32B-medical-efficiency-cost	1000	37.5%	2.98	5.0%	2.18	2.3%	4.91	<b>3.35</b>
Qwen32B-medical-only	1000	40.4%	3.00	5.6%	2.18	3.4%	4.86	<b>3.35</b>
Qwen32B-base	1000	28.2%	2.57	2.9%	1.98	3.1%	4.53	<b>3.02</b>

Table S9: Human expert reference vs. frontier model performance on the 50-case MEDCPENV subset.

Rank	Model	Comp.	Diag. Acc.	Diag.	Treat. Acc.	Treat.	CI Viol.	Safety
1	<b>Human Expert</b>	4.01	68.0%	3.82	34.0%	3.20	0.0%	5.00
2	GPT-5.4	3.74	60.0%	3.50	24.0%	2.72	0.0%	5.00
3	Qwen3.5-Plus	3.66	56.0%	3.48	14.0%	2.58	2.0%	4.92
4	Gemini-3.1-Pro-Preview	3.50	50.0%	3.10	20.0%	2.48	2.0%	4.92
5	Qwen3-Max	3.45	40.0%	3.02	12.0%	2.32	0.0%	5.00
6	MiniMax-M2.5	3.44	52.0%	3.06	12.0%	2.34	2.0%	4.92
7	GLM-5	3.23	36.0%	2.54	16.0%	2.14	0.0%	5.00
8	Claude-Opus-4-6	3.17	36.0%	2.54	20.0%	2.08	0.0%	4.90
9	DeepSeek-V3.2	2.99	26.0%	2.12	14.0%	1.86	0.0%	5.00

Table S10: MIMIC expert-validated set (200 cases).

Rank	Model	Total	Dx Acc	Dx	Tx Acc	Tx	Contra	Safety
1	<b>GPT-5.4</b>	<b>3.72</b>	68.0%	3.62	34.0%	2.83	<b>7.0%</b>	4.72
2	<b>Claude Opus 4.6</b>	<b>3.64</b>	<b>75.0%</b>	3.71	<b>49.0%</b>	3.20	25.0%	4.00
3	<b>GLM-5</b>	<b>3.44</b>	58.5%	3.48	24.5%	2.65	20.5%	4.18
4	<b>DeepSeek-V4-Pro</b>	<b>3.41</b>	59.0%	3.38	26.0%	2.66	20.0%	4.20
5	<b>Qwen3-Max</b>	<b>3.38</b>	47.5%	3.29	7.5%	2.33	12.0%	4.52
6	<b>Qwen3.5-Plus</b>	<b>3.37</b>	52.5%	3.38	18.0%	2.55	20.5%	4.18
7	<b>Gemini 3.1 Pro</b>	<b>3.34</b>	56.5%	3.40	8.5%	2.33	18.0%	4.28
8	<b>MiniMax-M2.5</b>	<b>3.31</b>	57.5%	3.36	13.5%	2.46	22.0%	4.12
9	<b>Qwen3-32B-medical-efficiency-cost</b>	<b>3.26</b>	40.20%	3.10	4.02%	2.17	12.56%	4.50
10	<b>Qwen3-32B-medical-only</b>	<b>3.20</b>	38.69%	3.02	3.52%	2.14	14.07%	4.43
11	<b>Qwen3-4B-medical-only</b>	<b>3.12</b>	29.6%	2.86	0.5%	2.05	13.6%	4.46
12	<b>Qwen3-4B-medical-efficiency-cost</b>	<b>3.09</b>	26.1%	2.73	0.0%	1.92	9.5%	4.61

Table S11: Pairwise agreement between Baichuan-M2 and two physician experts on the expert-scored subset. Agreement is computed separately for diagnosis, treatment, and contraindication judgments.

Model	Dimension	Baichuan-M2 vs Expert A	Baichuan-M2 vs Expert B	Expert A vs Expert B
GPT-5.4	Diagnosis	57.5%	75.0%	59.0%
GPT-5.4	Treatment	66.0%	74.0%	55.0%
GPT-5.4	Contraindication	99.0%	94.5%	95.5%
Gemini-3.1-pro-preview	Diagnosis	93.36%	77.97%	78.47%
Gemini-3.1-pro-preview	Treatment	75.55%	60.61%	53.90%
Gemini-3.1-pro-preview	Contraindication	96.60%	48.97%	50.69%
Claude Opus 4.6	Diagnosis	53.5%	93.0%	52.5%
Claude Opus 4.6	Treatment	42.0%	90.5%	44.0%
Claude Opus 4.6	Contraindication	79.5%	97.5%	80.0%

Table S12: eICU cross-source evaluation (100 cases).

Rank	Model	Total	Dx Acc	Dx	Tx Acc	Tx	Contra	Safety
1	<b>Claude Opus 4.6</b>	<b>3.66</b>	54.0%	3.42	<b>15.0%</b>	2.60	1.0%	4.96
2	<b>GPT-5.4</b>	<b>3.64</b>	53.0%	3.46	10.0%	2.46	<b>0.0%</b>	5.00
3	<b>DeepSeek-V4-Pro</b>	<b>3.63</b>	48.0%	3.40	12.0%	2.54	1.0%	4.96
4	<b>MiniMax-M2.5</b>	<b>3.63</b>	53.0%	3.43	11.0%	2.47	<b>0.0%</b>	5.00
5	<b>Gemini 3.1 Pro</b>	<b>3.62</b>	52.0%	3.50	9.0%	2.45	2.0%	4.92
6	<b>GLM-5</b>	<b>3.61</b>	50.0%	3.43	7.0%	2.44	1.0%	4.96
7	<b>Qwen3.5-Plus</b>	<b>3.54</b>	47.0%	3.36	6.0%	2.30	1.0%	4.96
8	<b>Qwen3-Max</b>	<b>3.46</b>	37.0%	3.19	4.0%	2.23	1.0%	4.96