# MMedAgentBench: A Multi-Stage Benchmark for Evaluating LLM-based Medical Agents

Anonymous ACL submission

## Abstract

Large language models (LLMs) have demonstrated remarkable potential in medical AI through their natural language understanding and reasoning capabilities. However, real clinical practice involves multi-stage, dynamically evolving decision-making where physicians continuously adjust pathways based on updated examination results—far more complex than single-turn question answering. Existing medical benchmarks focus on static scenarios, failing to evaluate LLMs in realistic clinical environments. We present **MMedAgentBench**, a comprehensive benchmark for evaluating LLM-based medical agents in multi-stage clinical decision-making. Unlike previous benchmarks, MMedAgentBench captures the temporal complexity of clinical workflows by simulating authentic patient journeys across multiple diagnostic and treatment stages. The benchmark comprises a high-quality dataset of over 150 de-identified clinical cases from tertiary hospitals, encompassing 500+ clinical stages, with each case meticulously structured into sequential stages and annotated by board-certified physicians. Our evaluation framework simultaneously assesses diagnostic accuracy and resource efficiency, requiring agents to plan examinations and control costs while navigating progressive information disclosure. This enables systematic evaluation of medical agents across dynamic decision-making, cost control, and multi-stage reasoning dimensions, providing a realistic testbed for advancing medical AI toward clinical deployment.

## 1 Introduction

With the rapid advancement of large language model (LLM) technology, application of AI in the medical field has gradually become a focal point. LLMs can process and understand vast amounts medical data. Particularly when faced with diverse and extensive medical information, they have potential in medical decision support through their natural language processing and reasoning capabilities (Thirunavukarasu et al., 2023; Singhal et al., 2023; Zhang et al., 2025). In medicine, doctors must analyze extensive data—including patient histories, symptom presentations, and test results—while making accurate diagnostic and treatment decisions within limited timeframes. This poses a significant challenge for manual processing. LLMs, however, leverage their robust information retrieval and reasoning capabilities to extract relevant insights from vast medical literature and case data, thereby accelerating the diagnostic process and enhancing accuracy (Tian et al., 2025b,a).

Particularly during disease diagnosis, LLM-based agent models can provide effective assistance throughout the multi-stage, dynamic decision-making process. By simulating a doctor's thought process, these agent models adapt subsequent decision paths based on current examination results, enabling more precise decisions in complex, evolving diagnostic scenarios (Li et al., 2024). Doctors must repeatedly adjust diagnostic hypotheses based on varying symptoms and test outcomes to narrow down the most probable conditions. Therefore, LLM-based agents which are capable of flexibly adapting to changes and progressively refining pathways based on each round of test results are particularly crucial in medical clinical decision-making. They can not only process medical texts but also dynamically adjust diagnostic strategies, helping doctor make the most appropriate treatment decisions at every stage.

However, most current benchmarks focus on static, single-task processing, overlooking the complexity of multi-stage decision-making inherent in medical practice. These methods fail to authentically reflect how doctors dynamically adjust examinations and treatment plans based on evolving patient conditions and test results, thereby limiting the applicability of current standards for assessing diagnostic systems.

To address this, we propose an innovative evaluation framework built upon a high-quality dataset comprising over 150 clinical cases and 500+ diagnostic stages. Unlike static approaches, our framework simulates the iterative reasoning paths of doctors, allowing diagnostic strategies to dynamically adapt based on intermediate findings. This methodology provides a comprehensive assessment of LLM agents in complex clinical scenarios, delivering more precise evaluations to advance their application in real-world decision support.

Our contributions are threefold:

- We propose **MMedAgentBench**, a benchmark comprising over 150 clinical cases and more than 500 diagnostic stages, designed to systematically evaluate LLM-based medical agents in multi-stage clinical decision-making.

- This benchmark is an LLM-based agent evaluation framework that simulates a doctor's reasoning and judgment during multi-stage decision-making. The LLM plays a central role here, handling dynamic decision-making and outcome updates at each stage.

- We established multiple evaluation dimensions within the benchmark framework, including diagnostic accuracy, cost-effectiveness, and multi-stage decision-making, to provide a more comprehensive assessment of LLM applications in medical diagnosis.

## 2 Related Work

A large body of work has focused on evaluating medical AI systems through benchmark datasets. Early efforts such as MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022) introduced large-scale multiple-choice examinations to assess factual medical knowledge, while benchmarks like PubMedQA (Jin et al., 2019) and emrQA (Pampari et al., 2018) targeted question answering from biomedical literature or clinical notes. More recently, comprehensive resources such as MultifacetEval (Zhou et al., 2024) expanded the scope to cover broader clinical reasoning tasks. However, these benchmarks are primarily designed around single-turn question answering or isolated decision-making tasks, which fail to capture the temporal dependencies and sequential nature of real-world clinical workflows.

With the rapid development of large language models (LLMs), increasing research efforts have extended medical agents from static question answering to interactive and workflow-oriented scenarios. Several representative works are particularly noteworthy in this direction.

First, MedAgentBench (Jiang et al., 2025) introduced a virtual electronic health record (EHR) environment for comprehensive evaluation of medical LLM agents. The benchmark covers 300 clinical tasks across 10 types of medical scenarios and supports interactive simulations through FHIR-compliant interfaces. Unlike traditional static QA, MedAgentBench emphasizes tool usage, information retrieval, and task planning capabilities. Second, $\tau$ bench (Yao et al., 2024) highlights the interaction ability of language agents with users and tools in real-world domains. It simulates dynamic dialogues and tool invocation processes, while also proposing new reliability metrics to assess an agent's consistency and robustness across multiple interactions. Finally, AgentClinic (Schmidgall et al., 2024) is specifically designed for medical settings, supporting doctor–patient dialogues, multimodal information acquisition and tool usage. However, it still lacks explicit control over multi-stage patient journeys, making it less aligned with the sequential nature of real-world clinical workflows.

## 3 MMedAgentBench: A Multi-Stage Medical Agent Benchmark

We present MMedAgentBench, a comprehensive benchmark designed to evaluate medical AI agents in realistic multi-stage clinical workflows, shown as Figure 1. Unlike existing medical benchmarks that focus on single-turn question-answering or isolated diagnostic tasks, MMedAgentBench captures the temporal complexity and resource constraints inherent in real-world clinical decision-making. Our benchmark simulates complete patient journeys through multiple clinical stages, where agents must balance diagnostic accuracy with examination costs while maintaining coherent reasoning across stages.

### 3.1 Multi-Stage Clinical Workflow Design

Real clinical practice rarely involves single-point decisions. Patients typically progress through multiple stages of care, each with distinct objectives and available resources. We model this reality by organizing evaluation scenarios into sequential stages $S = \{s_1, s_2, ..., s_n\}$ that mirror actual hospital workflows. Each stage $s_i$ represents a distinct clinical phase such as initial assessment, follow-

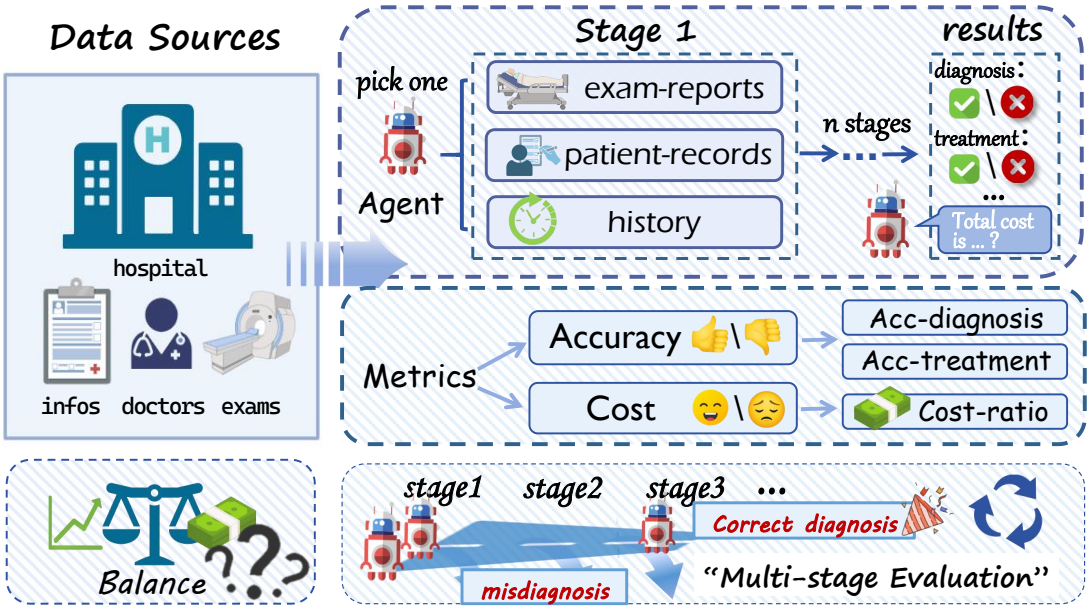


Figure 1: Overview of the MMedAgentBench. The benchmark models real-world clinical decision-making as a sequential, cost-aware reasoning process. Each case unfolds through multiple diagnostic stages, where the agent iteratively selects examinations, interprets results, and updates its diagnostic hypotheses. The evaluation metrics consider both diagnostic accuracy and resource costs, aiming to promote more balanced clinical decision-making.

up examination, or pre-discharge evaluation. The agent must navigate through these stages sequentially, making decisions based on available information at each stage while considering the accumulated evidence from previous stages.

For each stage $s_i$, we define two catalogs: $C_i^{exam}$ containing available diagnostic examinations and $C_i^{patient}$ containing accessible patient records. The agent can select items from these catalogs, with each selection revealing detailed clinical information. Formally, at stage $s_i$, the agent's action space is defined as:

$$A_i = \{\text{pick}(c) : c \in C_i^{exam} \cup C_i^{patient}\} \quad (1)$$
$$\cup \{\text{retrieve}(h) : h \in H_{i-1}\} \quad (2)$$

where $H_{i-1} = \bigcup_{j=1}^{i-1} X_j$ represents the set of historical items selected in all previous stages.

The design of MMedAgentBench addresses several critical challenges in medical AI evaluation. First, temporal coherence is enforced through our stage-wise information access control, where agents cannot access future-stage information, preventing unrealistic perfect foresight scenarios. This mirrors how clinicians must make decisions based on currently available information without knowing future developments.

Second, our historical information retrieval mechanism allows agents to access previously obtained information without additional cost, simulating how clinicians reference prior test results and observations. This design tests an agent's ability to synthesize longitudinal patient data while avoiding redundant examinations.

Finally, the transition between stages is governed by a judgment function $J : (D_i, T_i) \rightarrow \{0, 1\}$, where $D_i$ and $T_i$ represent the diagnosis and treatment outputs at stage $s_i$. The agent proceeds to stage $s_{i+1}$ only if $J(D_i, T_i) = 1$, ensuring that stage-specific clinical objectives are met before progression. This mirrors real clinical workflows where certain milestones must be achieved before proceeding to the next phase of care. The judgment function is implemented using a large language model to assess whether the agent's conclusions adequately address the clinical requirements of the current stage.

### 3.2 Dataset Construction

Our dataset is constructed from authentic clinical cases collected from real hospital settings, ensuring the benchmark reflects genuine medical complexity and decision-making scenarios. We collaborate with multiple tertiary hospitals to obtain de-identified patient records spanning various medical specialties, including internal medicine, surgery, pediatrics, and emergency care. Each case represents a complete patient journey from initial presentation through discharge or transfer, capturing the full

spectrum of clinical decision-making.

The raw clinical data undergoes a rigorous structuring process to transform unstructured medical records into our standardized format. We employ a structured representation that separates examination reports from patient records, each organized by clinical stages. For each item $x$ in the dataset, we define a tuple $(title_x, details_x, price_x, stage_x)$ where $title_x$ serves as the unique identifier, $details_x$ contains the complete clinical information, $price_x \in \mathbb{R}^+$ represents the associated cost based on actual hospital pricing, and $stage_x \in S$ indicates the stage affiliation.

To ensure clinical accuracy and relevance, our annotation process involves multiple rounds of expert review. Board-certified physicians from relevant specialties annotate each case with clinical key points rather than exact text matches. Specifically, for stage $s_i$, we define the ground truth as $G_i = (P_i^{diag}, P_i^{treat})$, where $P_i^{diag} = \{p_{i,1}^{diag}, p_{i,2}^{diag}, ..., p_{i,m}^{diag}\}$ represents diagnostic key points and $P_i^{treat} = \{p_{i,1}^{treat}, p_{i,2}^{treat}, ..., p_{i,k}^{treat}\}$ represents treatment key points. This point-based representation allows for semantic flexibility while maintaining evaluation rigor.

The annotation guidelines ensure consistency across different annotators. Cases with low agreement undergo additional review and discussion until consensus is reached. Furthermore, we implement quality control measures including random audits and cross-validation to maintain high annotation standards throughout the dataset construction process.

## 3.3 Evaluation Framework

Our evaluation framework integrates both clinical accuracy and resource efficiency, reflecting the dual objectives of real-world medical practice. We assess agent performance through comprehensive metrics that balance diagnostic precision with cost-conscious decision-making.

### 3.3.1 Cost Tracking and Resource Utilization

Healthcare resources are inherently limited and expensive. To evaluate cost efficiency, we establish a baseline cost for each stage through expert annotation. For each stage $s_i$, experienced physicians annotate a reasonable reference cost $\text{Cost}_i^{ref}$ that represents the expected expense for appropriate clinical decision-making at that stage. This reference cost serves as a normalization factor for evaluating the agent's resource utilization.

For stage $s_i$, let $X_i$ be the set of unique items selected by the agent. The actual stage cost is calculated as:

$$\text{Cost}_i = \sum_{x \in X_i \cap C_i^{exam}} price_x \tag{3}$$

Note that patient record items typically have zero cost, reflecting their nature as readily available clinical information. To prevent artificial inflation of costs, we implement a deduplication mechanism where multiple selections of the same item within a stage incur cost only once.

The cost efficiency for each stage is then evaluated through a cost ratio:

$$\text{CostRatio}_i = \frac{\text{Cost}_i}{\text{Cost}_i^{ref}} \tag{4}$$

For the overall evaluation, we compute the average cost ratio across all stages that the agent has experienced. Let $m$ denote the number of stages the agent successfully completes, the average cost ratio is:

$$\text{CostRatio}_{avg} = \frac{1}{m} \sum_{i=1}^{m} \text{CostRatio}_i \tag{5}$$

### 3.3.2 Clinical Accuracy Metrics

We evaluate clinical accuracy using LLM-based semantic matching to assess coverage of key points. Let $\hat{D}_i$ and $\hat{T}_i$ denote the agent's diagnosis and treatment outputs at stage $s_i$. We define a semantic matching function $M : (\text{prediction}, \text{key point}) \rightarrow \{0, 1\}$ that determines whether a key point is covered by the prediction.

For stages that the agent successfully completes, the stage-wise diagnostic and treatment accuracies are calculated as:

$$\text{Acc}_i^{diag} = \frac{1}{|P_i^{diag}|} \sum_{p \in P_i^{diag}} M(\hat{D}_i, p) \tag{6}$$

$$\text{Acc}_i^{treat} = \frac{1}{|P_i^{treat}|} \sum_{p \in P_i^{treat}} M(\hat{T}_i, p) \tag{7}$$

For stages that the agent fails to reach or complete, we assign an accuracy of 0, reflecting the failure to address those clinical scenarios. Let $n$ be the total number of stages in the complete patient journey. The aggregate accuracies are computed

as:

$$\text{Acc}_{avg}^{diag} = \frac{1}{n}\sum_{i=1}^{n}\text{Acc}_i^{diag}, \quad (8)$$

$$\text{Acc}_{avg}^{treat} = \frac{1}{n}\sum_{i=1}^{n}\text{Acc}_i^{treat} \quad (9)$$

where $\text{Acc}_i = 0$ for all $i > m$ (stages not reached by the agent). This formulation ensures that agents are penalized for incomplete patient journeys, encouraging comprehensive clinical coverage while maintaining accuracy at each completed stage.

# 4 Experiments

## 4.1 Experimental Setup

### 4.1.1 Benchmark Details

MMedAgentBench represents a high-quality, comprehensive medical agent benchmark dataset specifically designed for evaluating AI agents in clinical decision-making scenarios. The dataset comprises 167 carefully curated patient cases, as shown in Table 1.

| Description | Value |
|---|---|
| Number of Patients | 167 |
| Number of Clinical Stages | 527 |
| Number of Examination Records | 4,019 |
| Number of Patient Records | 701 |
| Average Stages per Patient | 3.16 |
| Average Exam Records per Patient | 24.07 |
| Average Exam Records per Patient | 4.19 |

Table 1: Summary of MMedAgentBench:

### 4.1.2 Models

We conduct an extensive evaluation covering a diverse spectrum of large language models that vary in scale to ensure a comprehensive capability assessment. Among open-weight models, our benchmark includes the Qwen2.5 family (14B/72B) (Team, 2024) alongside the next-generation Qwen3 series (14B/235B-A22B) (Yang et al., 2025), the Llama series (Touvron et al., 2023; Meta AI, 2024; Dubey et al., 2024) encompassing Llama-3.1-70B and Llama-3.1-405B, as well as GLM-Z1-32B (GLM et al., 2024), DeepSeek-v3.1 (DeepSeek-AI, 2024), Hunyuan-A13B (Tencent Hunyuan Team, 2025), and Kimi-K2 (Team et al., 2025). For closed-weight models, we evaluate the GPT family—including GPT-3.5, GPT-4o, GPT-4.1, and GPT-5 (OpenAI, 2023a,b)—alongside the Claude-Sonnet series (3.5, 3.7, and 4) (Anthropic, 2025), Gemini-2.5-Flash (Comanici et al., 2025), Grok-4 (xAI, 2025), and Doubao-1.5-Pro (ByteDance Seed, 2025). All models are tested under unified prompting strategies and standardized evaluation protocols to guarantee fairness and comparability across results.

For the stage judgment function $J : (D_i, T_i) \rightarrow 0, 1$ and the semantic matching function $M$ : $(\text{prediction}, \text{key point}) \rightarrow 0, 1$, we implemented them using Gemini-2.5-Flash and GPT-4o, respectively. The former requires high-level integration and reasoning over long clinical texts, while the latter focuses solely on evaluating semantic similarity and matching accuracy.

## 4.2 Results

The benchmark consists of a vast array of complex medical data, ranging from simple outpatient records to intricate multi-disease co-occurrence patterns. Based on MMedAgentBench, we evaluated multiple large models' diagnostic capability ($\text{Acc}^{diag}$), treatment capability ($\text{Acc}^{treat}$), and cost ratio (CostRatio), results are shown as Table 2. We run the benchmark for 5 times to get the average value. Even the best-performing models, Qwen3-235B-A22B and GPT-5, achieve only 36.67% and 35.62% diagnostic accuracy respectively, and 31.69% and 38.33% treatment accuracy, highlighting the substantial challenges posed by our multi-stage clinical diagnosis dataset. This relatively modest performance ceiling underscores the complexity of real-world clinical decision-making, where agents must navigate through sequential stages, synthesize evolving patient information, and make interconnected diagnostic and treatment decisions. Furthermore, the wide variation in CostRatio across models (ranging from 0.44 to 1.53) provides valuable insights into different models' examination ordering strategies, enabling researchers to better understand the trade-offs between diagnostic thoroughness and resource efficiency, and identify models that can achieve clinical objectives while maintaining cost-conscious medical practice.

## 4.3 Model Scale Effects

Evaluation results reveal a clear correlation between model scale and clinical performance within the same model families. Examining the Llama-3.1 series, we observe substantial improvements as parameters increase from 70B to 405B: $\text{Acc}^{diag}$ im-

| Open-Weight Models | | | | Closed-Weight Models | | | |
|---|---|---|---|---|---|---|---|
| Model | $Acc^{diag}$ | $Acc^{treat}$ | Cost↓ | Model | $Acc^{diag}$ | $Acc^{treat}$ | Cost↓ |
| Llama-3.1-70B | 7.75 | 4.47 | **0.46** | GPT-3.5 | 15.30 | 10.50 | **0.44** |
| Llama-3.1-405B | 16.07 | 8.60 | 0.75 | GPT-4o | 22.13 | 18.68 | 0.74 |
| Qwen2.5-14B | 17.47 | 14.04 | 0.66 | GPT-4.1 | 34.90 | 37.00 | 0.68 |
| Qwen2.5-72B | 21.76 | 17.77 | 0.85 | GPT-5 | **35.62** | **38.33** | 0.93 |
| Qwen3-14B | 17.66 | 16.28 | 0.63 | Claude-Sonnet-3.5 | 28.10 | 28.10 | 1.45 |
| Qwen3-235B-A22B | **36.67** | **31.69** | 1.53 | Claude-Sonnet-3.7 | 30.10 | 28.70 | 1.32 |
| Hunyuan-A13B | 13.55 | 18.68 | 0.75 | Claude-Sonnet-4 | 31.89 | 30.13 | 1.30 |
| GLM-Z1-32B | 22.59 | 20.69 | 1.28 | Gemini-2.5-Flash | 32.88 | 28.67 | 1.41 |
| DeepSeek-V3.1 | 25.12 | 23.99 | 0.69 | Grok-4 | 31.35 | 29.11 | 0.98 |
| Kimi-K2 | 28.30 | 31.10 | 1.14 | Doubao-1.5-Pro | 33.25 | 30.78 | 1.24 |

Table 2: Evaluation results on MMedAgentBench. $Acc^{diag}$ and $Acc^{treat}$ denote diagnostic and treatment accuracy respectively (in %). CostRatio measures resource utilization efficiency, where smaller values indicate better cost control relative to expert baseline. We use **bold** to highlight the best-performing model in each category, and underline for the second-best.

proves from 7.75% to 16.07% , while $Acc^{treat}$ rises from 4.47% to 8.60%. Similarly, the Qwen2.5 series demonstrates consistent scaling benefits, with $Acc^{diag}$ increasing from 17.47% (14B) to 21.76% (72B) and $Acc^{treat}$ from 14.04% to 17.77%. These within-family improvements underscore that parameter scaling directly enhances models' ability to process complex medical information and generate accurate clinical decisions.

However, the results also highlight that base model capabilities and training methodologies are equally crucial for medical tasks. Notably, Qwen2.5-14B achieves superior performance ($Acc^{diag}$: 17.47%, $Acc^{treat}$: 14.04%) compared to Llama-3.1-70B despite having significantly fewer parameters, suggesting that architectural design and pre-training quality on medical-relevant data can compensate for raw parameter count.

The scaling effects observed across different model families suggest that while increasing parameters generally improves medical reasoning capabilities, the efficiency of this scaling varies significantly based on the underlying model architecture and training approach. This indicates that future development of medical AI systems should balance parameter scaling with targeted improvements in base model design and domain-specific pre-training strategies.

### 4.4 Comparison of Open-Weight and Closed-Weight Models

The performance gap between open and closed-weight models has narrowed significantly. Notably, open-weight Qwen3-235B-A22B achieves the highest diagnostic accuracy ($Acc^{diag}$ 36.67%), surpassing the leading closed-weight GPT-5 (35.62%). However, GPT-5 retains the top spot in treatment recommendations ($Acc^{treat}$ 38.33%), indicating closed models still excel in therapeutic strategies.

Among closed-weight models, GPT-4.1 ranks second ($Acc^{diag}$: 34.90%, $Acc^{treat}$: 37.00%) with remarkably low cost (CostRatio: 0.68). Doubao-1.5-Pro ($Acc^{diag}$: 33.25%, $Acc^{treat}$: 30.78%) also outperforms prominent models like Gemini-2.5-Flash and Claude-Sonnet-4, proving that top-tier performance does not necessitate high costs.

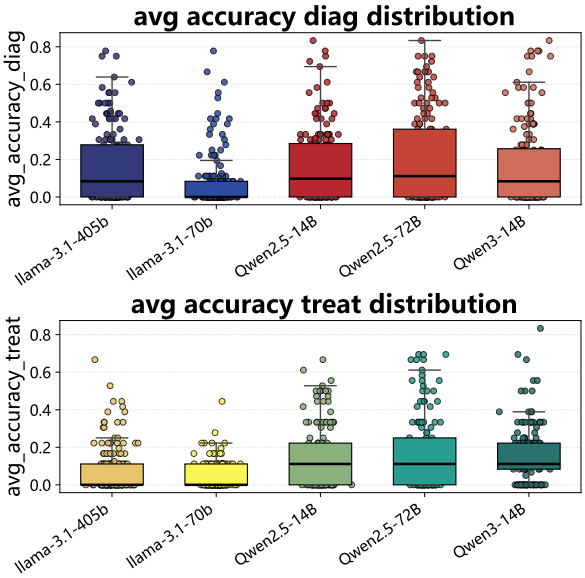


Figure 2: $Acc^{diag}$ and $Acc^{treat}$ distribution of Qwen and Llama family.

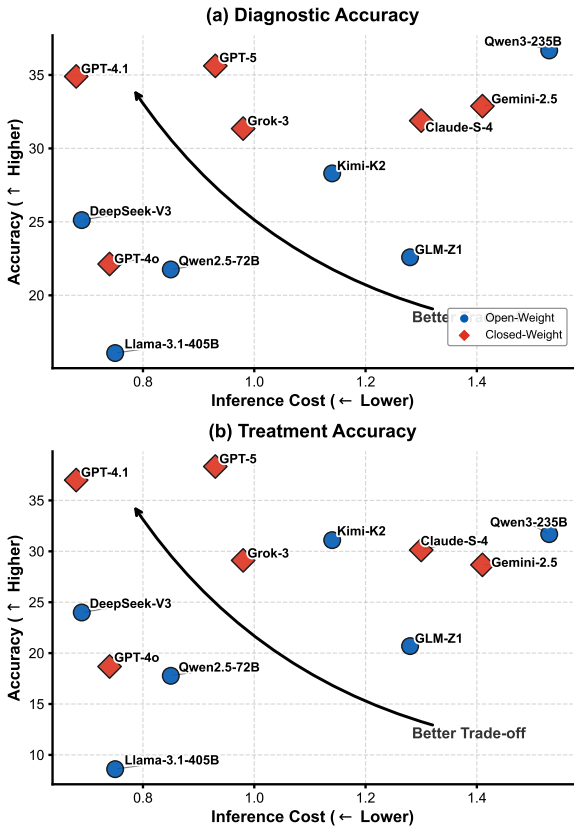


Figure 3: Cost-Efficiency Analysis. Comparison of (a) Diagnostic and (b) Treatment accuracy vs. cost. The arrow points to the optimal balance. Legend: Blue circles = Open-weight; Red diamonds = Closed-weight.

In the open-weight domain, Kimi-K2 achieves an $\text{Acc}^{treat}$ of 31.10%, surpassing closed counterparts like Gemini-2.5-Flash and Claude-Sonnet-4. Meanwhile, DeepSeek-V3.1 offers a strong efficiency-accuracy balance (25.12%, 23.99%), outperforming GPT-4o. These results demonstrate open-weight models effectively rivaling top-tier proprietary systems.

#### 4.4.1 Cost Sensitivity and Diagnostic-Therapeutic Balance

In this study, CostRatio represents the relative cost of models during medical diagnosis and treatment recommendation generation. Results reveal distinct trade-offs between performance and cost across different models. We visualize acc vs cost in Figure 3.

Overall, closed-weight models like Claude-Sonnet-4 (1.30), Gemini-2.5-Flash (1.41), and GPT-5 (0.93) exhibit cost ratios close to or slightly exceeding real-world medical scenarios. In contrast, certain open-weight models (e.g., DeepSeek-V3.1 at 0.69, Qwen2.5-14B at 0.66) demonstrate lower costs while maintaining competitive performance.

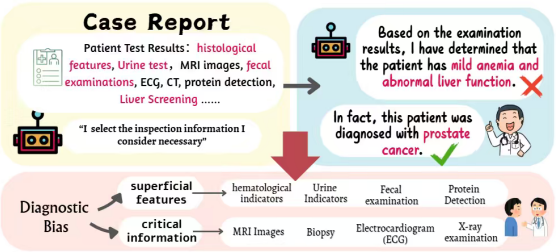


Figure 4: Case illustration of diagnostic bias. By focusing on superficial features like routine blood work instead of MRI/biopsy results, the model incorrectly diagnosed prostate cancer as anemia and liver issues.

However, high cost does not directly equate to high efficiency: at similar cost levels, significant gaps exist in diagnostic accuracy among different models. For instance, GPT-5 achieves a significantly higher diagnostic accuracy (35.62%) than Claude-Sonnet-4 (31.89%) at a lower cost (0.93 vs. 1.30), demonstrating superior performance-cost ratio.

It is noteworthy that DeepSeek-V3.1 (CostRatio = 0.69) and GPT-4o (CostRatio = 0.74) present a compelling contrast: both models incur similar medical examination costs, yet DeepSeek-V3.1 outperforms GPT-4o in both diagnostic accuracy (25.12% vs. 22.13%) and treatment recommendation accuracy (23.99% vs. 18.68%). Similarly, comparing Qwen2.5-72B (CostRatio = 0.85) with GPT-5 (CostRatio = 0.93), we observe that despite similar cost levels, GPT-5 significantly outperforms Qwen2.5-72B in diagnostic accuracy (35.62% vs. 21.76%) and treatment recommendation accuracy (38.33% vs. 17.77%). These contrasts demonstrate that effective utilization of medical examinations is more critical than merely minimizing costs. The key challenge in medical AI is not just reducing examination costs, but effectively synthesizing the resulting information to reach accurate clinical conclusions—achieving the optimal balance between diagnostic accuracy and resource efficiency.

#### 4.4.2 Case Study: Misleading Information Distorting Diagnostic Direction

When confronted with complex medical records, models are susceptible to misleading clues that can cause significant deviations in diagnostic direction.

In this case shown as Figure 4, the data input to the model included comprehensive clinical information: the patient's complete blood count results, MRI imaging descriptions, and biopsy reports. The blood test results indicated mild ane-

| Model | Gemini-2.5-Flash (Judge) + GPT-4o (Match) | | | GPT-5 (Judge) + Claude-Sonnet-4 (Match) | | |
|---|---|---|---|---|---|---|
| | $Acc^{diag}$ (%) | $Acc^{treat}$ (%) | CostRatio | $Acc^{diag}$ (%) | $Acc^{treat}$ (%) | CostRatio |
| Qwen2.5-14B | 17.47 | 14.04 | 0.66 | 18.16 | 15.00 | 0.67 |
| Qwen2.5-72B | 21.76 | 17.77 | 0.85 | 20.93 | 17.61 | 0.85 |
| Qwen3-14B | 17.66 | 16.28 | 0.63 | 17.59 | 16.05 | 0.63 |
| GPT-3.5 | 15.30 | 10.50 | 0.44 | 15.15 | 8.95 | 0.45 |
| GPT-4o | 22.13 | 18.68 | 0.74 | 23.02 | 18.59 | 0.74 |
| GPT-5 | 35.62 | 38.33 | 0.93 | 37.04 | 38.15 | 0.95 |

Table 3: Sensitivity analysis of different LLM judges. Relative rankings remain consistent across configurations.

mia and mild liver function abnormalities, while the MRI and biopsy information revealed abnormal prostate signals and related histological features, with the actual diagnosis being prostate cancer. However, when generating its diagnosis, the model primarily focused on hematological indicators, outputting a diagnosis and treatment recommendation of "mild anemia and liver function abnormalities," completely ignoring the malignant tumor risk suggested by the imaging and record.

This outcome demonstrates that even with comprehensive clinical information, the model may exhibit systemic bias due to overemphasis on superficial features. Its reasoning process showed excessive sensitivity to prominent data (e.g., blood indicators) while undervaluing critical information (e.g., imaging and pathology reports). The diagnostic logic lacked hierarchical judgment capabilities, failing to perform causal reasoning and diagnostic hypothesis verification based on the overall clinical picture as clinical physicians do.

#### 4.4.3 Evaluation Validity and Robustness

We validate our metrics through a hybrid strategy. First, human-defined ground truth ensures reliability; reference points are explicitly extracted by experts, strictly limiting the judge to semantic matching rather than generation. Second, judge sensitivity analysis confirms that relative model rankings remain robust across different judge models, despite minor absolute score variations (Table 3; more results in Appendix A.2). Finally, human alignment checks reveal high Fleiss' kappa agreement between medical experts and our automated judge, confirming clinical consistency (details in Appendix A.2).

## 5 Conclusion

We presented MMedAgentBench, a comprehensive benchmark designed to evaluate LLM-based medical agents in realistic multi-stage clinical workflows. Unlike existing medical benchmarks that focus on static, single-turn question answering, our framework captures the temporal complexity and resource constraints inherent in real clinical practice. Through a high-quality dataset comprising over 150 clinical cases and 500+ diagnostic stages, we enable systematic evaluation of medical agents across dynamic decision-making, cost control, and multi-stage reasoning dimensions.

MMedAgentBench establishes a crucial testbed for advancing medical AI toward clinical deployment. By requiring agents to navigate sequential clinical stages while balancing accuracy and resource efficiency, our benchmark pushes the development of more sophisticated medical reasoning capabilities. Future work can leverage this framework to develop specialized medical agents that better handle temporal dependencies, cost constraints, and the inherent uncertainty in clinical workflows, ultimately contributing to more practical and deployable medical AI systems.

## 6 Limitations

**Simplification of Clinical Complexity.** Real clinical practice involves nuances that benchmark cannot fully capture, including non-verbal patient cues, urgency considerations, and real-time collaborative decision-making among healthcare teams. We assumes perfect information retrieval once an examination is ordered, whereas real-world scenarios often involve incomplete, ambiguous, or delayed results.

**Static Ground Truth Assumption.** Our benchmark assumes fixed ground truth diagnoses and treatments annotated by experts. However, medical knowledge evolves, and legitimate clinical disagreements exist even among specialists. The benchmark may not adequately capture this inherent uncertainty and variability in medical practice.

## References


Anthropic. 2025. Claude 4. https://www.anthropic.com/claude. Available at https://www.anthropic.com/claude.

ByteDance Seed. 2025. Doubao-1.5-pro. https://seed.bytedance.com/en/special/doubao_1_5_pro. Accessed: 2026-01-06.

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

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and 37 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Yixing Jiang, Kameron C Black, Gloria Geng, Danny Park, James Zou, Andrew Y Ng, and Jonathan H Chen. 2025. Medagentbench: a virtual ehr environment to benchmark medical llm agents. *NEJM AI*, 2(9):AIdbp2500144.

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.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Junkai Li and 1 others. 2024. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

Meta AI. 2024. Introducing meta llama 3: The most capable openly available llm to date. Accessed: 2025-07-28.

OpenAI. 2023a. Chatgpt. https://chat.openai.com/chat.

OpenAI. 2023b. Gpt-4 technical report. Technical report, OpenAI. Available at https://arxiv.org/abs/2303.08774.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. *arXiv preprint arXiv:1809.00732*.

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

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, and 150 others. 2025. Kimi k2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Tencent Hunyuan Team. 2025. Hunyuan-a13b technical report. Technical report, Tencent. Accessed: 2025-01-06.

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

Yuan Tian, Kaiyuan Ji, Rongzhao Zhang, Yankai Jiang, Chunyi Li, Xiaosong Wang, and Guangtao Zhai. 2025a. Towards all-in-one medical image re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30774–30786.

Yuan Tian, Shuo Wang, and Guangtao Zhai. 2025b. Medical manifestation-aware de-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26363–26372.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

xAI. 2025. Grok 4. Online. Official announcement – Grok 4m̈odel news.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. $\tau$-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*.

Zicheng Zhang, Junying Wang, Yijin Guo, Farong Wen, Zijian Chen, Hanqing Wang, Wenzhe Li, Lu Sun, Yingjie Zhou, Jianbo Zhang, Bowen Yan, Ziheng Jia, Jiahao Xiao, Yuan Tian, Xiangyang Zhu, Kaiwei Zhang, Chunyi Li, Xiaohong Liu, Xiongkuo Min, and 2 others. 2025. Aibench: Towards trustworthy evaluation under the 45° law. https://aiben.ch/.

Yuxuan Zhou, Xien Liu, Chen Ning, and Ji Wu. 2024. Multifaceteval: Multifaceted evaluation to probe llms in mastering medical knowledge. *arXiv preprint arXiv:2406.02919*.

# A Appendix

## A.1 Annotation Guidelines and Protocol

The creation of MMedAgentBench relied on a rigorous annotation protocol executed by board-certified physicians. These guidelines were provided to each annotator to ensure consistency, clinical accuracy, and relevance across all cases in the dataset. The objective was to establish a gold-standard ground truth for evaluating an AI agent's multi-stage clinical reasoning.

### A.1.1 General Principles

Annotators were instructed to approach each case as a real-world clinical scenario. The primary goal was to define what a competent, cost-conscious physician would conclude and do at each stage of the patient's journey, based only on the information available at that point.

### A.1.2 Per-Stage Annotation Form

For each clinical stage ($s_i$) of a patient case, annotators were required to complete a structured form with three main components:

**Part 1: Clinical Key Points.**

This section captures the essential diagnostic and therapeutic conclusions for the stage. Annotators were asked to provide a list of short, distinct points.

**Diagnostic Key Points ($P_i^{diag}$):**

- **Objective:** List the most critical diagnoses, differential diagnoses, or significant findings that should be identified at this stage.
- **Instructions:**
  1. Focus on the most probable diagnoses and critical "can't-miss" diagnoses (e.g., myocardial infarction, pulmonary embolism).
  2. Points should be justifiable based on the evidence presented up to stage $s_i$.
  3. Avoid vague statements. Instead of "abnormal lab results," specify "leukocytosis with a left shift suggesting infection."

**Treatment Key Points ($P_i^{treat}$):**

- **Objective:** List the essential treatment actions, management plans, or further consultations that should be initiated.
- **Instructions:**
  1. List specific actions (e.g., "Administer intravenous antibiotics," "Consult cardiology").
  2. Include key monitoring plans (e.g., "Monitor vital signs every 4 hours").
  3. The treatment plan should directly address the diagnostic findings.

**Part 2: Reference Cost ($\text{Cost}_i^{ref}$).**

This section establishes a cost baseline for efficient and effective clinical investigation.

- **Objective:** Identify a set of examinations that represents a reasonable and cost-effective diagnostic strategy for the stage.
- **Instructions:**
  1. From the list of all available examinations at stage $s_i$, select only those that are **essential** to confirm or rule out the primary differential diagnoses.
  2. Balance diagnostic yield against cost. Avoid ordering overly broad or low-yield tests unless clinically crucial.
  3. The final $\text{Cost}_i^{ref}$ is the sum of the prices of the selected essential examinations. Patient record items have a cost of zero.

**Part 3: Stage Progression Criteri.**

This defines the minimum requirements for an agent's performance to be considered successful for the current stage, allowing it to proceed to the next.

- **Objective:** Define the primary clinical goal of the current stage.

| Model | GPT-4o (Judge) + GPT-3.5 (Match) | | |
|---|---|---|---|
| | $\text{Acc}^{diag}$ (%) | $\text{Acc}^{treat}$ (%) | CostRatio |
| Qwen2.5-14B | 16.27 | 14.26 | 0.66 |
| Qwen2.5-72B | 20.15 | 17.22 | 0.84 |
| Qwen3-14B | 18.05 | 18.21 | 0.59 |
| GPT-3.5 | 15.90 | 9.87 | 0.45 |
| GPT-4o | 22.75 | 18.94 | 0.75 |
| GPT-5 | 35.12 | 38.57 | 0.93 |

Table 4: Sensitivity analysis with an alternative judge/-match configuration (appendix).

| Model | $\kappa_J$ | $\kappa_S$ |
|---|---|---|
| Qwen2.5-14B | 0.88 | 0.81 |
| Qwen2.5-72B | 0.91 | 0.85 |
| GPT-4o | 0.90 | 0.88 |
| GPT-5o | 0.94 | 0.91 |

Table 5: Human-LLM agreement (Fleiss' kappa) for selected models. ($\kappa_J$: Stage Judgment; $\kappa_S$: Semantic Matching).

- **Instructions:**
  1. Answer the question: "What is the minimum level of understanding or action required at this stage to ensure patient safety and effective progression of care?"
  2. This is a holistic judgment. For example, the criterion might be "The agent must recognize the signs of sepsis and initiate fluid resuscitation and antibiotics" or "The agent must correctly interpret the MRI as highly suspicious for malignancy and recommend a biopsy."
  3. An agent failing to meet this criterion would not proceed, simulating a critical clinical error.

### A.2 Evaluation Validity and Robustness

This section provides supplementary details on the validation experiments summarized in Section 4.4.3 of the main text.

To verify that our evaluation framework produces robust rankings independent of specific model choices, we conducted additional sensitivity experiments using an alternative configuration: GPT-4o as the stage judge and GPT-3.5 as the semantic matcher. As shown in Table 4, while absolute scores differ from our primary configuration reported in the main text, the relative ranking of models remains consistent. GPT-5 maintains a substantial lead across all metrics, followed by GPT-4o and Qwen2.5-72B, while GPT-3.5 continues to show the weakest performance. This consistency across different judge configurations demonstrates that our evaluation methodology captures genuine capability differences rather than artifacts of specific judge-model pairings.

To ensure clinical validity, we further conducted human alignment studies with three board-certified physicians who independently evaluated a random subset of 50 diagnostic conversations. Table 5 reports Fleiss' kappa coefficients measuring agreement between human experts and our automated evaluation pipeline. For stage judgment ($\kappa_J$), which assesses whether models appropriately transition through diagnostic phases, we observe strong agreement ranging from 0.88 to 0.94. Similarly, semantic matching ($\kappa_S$), which evaluates whether model outputs align with ground-truth clinical findings, achieves substantial agreement between 0.81 and 0.91. These kappa values exceed the conventional threshold of 0.80 for "almost perfect" agreement, confirming that our LLM-based evaluation closely mirrors expert clinical judgment and supporting the validity of our automated metrics for large-scale assessment.

### A.3 Recruitment and Payment

All data annotation for MMedAgentBench was conducted manually by a team of trained medical annotators with relevant clinical backgrounds. Each annotation underwent a multi-stage verification pipeline, including internal cross-checking and random sampling review, to minimize potential bias and human error. A subset of the annotated data was further proofread and refined by the authors to ensure accuracy and adherence to medical logic. Reasonable compensation was provided to all annotators according to their workload and contribution, ensuring fairness and quality throughout the entire annotation process.

### A.4 Dataset Statistics

Table 6 summarizes the distribution of patients across clinical departments in our dataset. The 167 cases span six major departments, covering a diverse range of conditions with varying diagnostic complexity and multi-stage reasoning requirements.

Table 6: Distribution of patients across clinical departments.

| Department | Patients | Percentage |
|---|---|---|
| Internal Medicine | 42 | 25.1% |
| Surgery | 35 | 21.0% |
| Cardiology | 28 | 16.8% |
| Neurology | 24 | 14.4% |
| Oncology | 21 | 12.6% |
| Emergency Medicine | 17 | 10.2% |

### A.5 Prompt Design

Figure 5 presents the system prompts for our diagnostic reasoning framework. The Controller prompt guides the agent in iteratively selecting relevant medical information from a structured catalog, enforcing strict title matching and cost-effective retrieval. The Finalize prompt instructs the model to synthesize collected evidence into structured diagnostic conclusions.

Figure 6 shows the evaluation prompts. The Stage-Level Judge assesses whether candidate answers align with reference answers derived from clinical guidelines, allowing semantic equivalence rather than exact matching. The emantic Point Matching prompt enables fine-grained evaluation of whether specific clinical points are adequately covered.

## B LLM Usage

LLM was used for drafting and language polishing. Except these uses, no AI tools were involved in study design, core experiments, result analysis, or interpretation. The authors are fully responsible for the accuracy of the work.

### Controller System Prompt

You are a medical reasoning assistant. The system provides:

- Current stage available catalog (titles only, without prices or parenthetical information);
- Historical known titles list (items selected in previous stages, used for retrieving historical information).

**Task:** Based on the patient's condition and cost-effectiveness, select one precise item title from the ``Current Stage Catalog'' (without any prices or parentheses), or finalize the current stage.

**Available Actions** (output strictly in JSON format):

```
{"action": "pick_report"|"pick_patient"|"get_history"|"finalize",
 "title": "exact title from catalog",
 "reason": "brief rationale"}
```

**Rules:**

1. You may ONLY select from titles in the ``Current Stage Catalog'' (must be exact title match); do NOT invent items outside the catalog; do NOT include prices or parentheses.
2. Historical known titles are ONLY for retrieving historical information.
3. Avoid repeatedly selecting the same item; when additional information provides limited value, select **finalize**.

### Finalize System Prompt

You are a clinical physician. Based on the provided evidence, generate the conclusion for the current stage in strict JSON format:

```
{"diagnosis": "...",
 "treatment": "...",
 "used_evidence_titles": ["..."],
 "notes": "..."}
```

**Requirements:**

1. **used_evidence_titles** must ALL come from the provided evidence titles.
2. Diagnosis and treatment should be concise and definitive.
3. If evidence is insufficient or multiple retrievals return empty results, state in **notes**: ``No additional relevant records/reports found within the current scope.''

Figure 5: System prompts for the Controller and Finalize modules in our diagnostic reasoning framework.

**Stage-Level Judge Prompt**

```
You are a medical evaluation judge. Based on the reference answer (diagnosis + treatment
plan) derived from authoritative clinical guidelines and the candidate's submitted answer,
determine whether the current diagnostic stage passes the evaluation.

Evaluation Criteria:

1. Assess whether the diagnostic reasoning is clinically sound and logically coherent.

2. Evaluate whether the proposed treatment plan is appropriate, safe, and aligns with
   evidence-based medicine principles.

3. Consider whether key differential diagnoses have been appropriately addressed.

4. Minor variations in terminology or phrasing are acceptable if the clinical meaning is
   preserved.

Output Format (strict JSON only):

{"pass": true|false,
 "reasoning": "detailed explanation of the judgment"}

Input:
Reference Answer: {gold_answer}
Candidate Answer: {pred_answer}
Output JSON only.
```

**Semantic Point Matching Prompt**

```
You are a medical evaluation assistant. Your task is to determine whether the candidate's
answer text covers or equivalently expresses the given evaluation point, according to the
reference clinical guidelines.

Evaluation Guidelines:

1. Semantic equivalence is sufficient; exact wording match is NOT required.

2. Consider synonyms, paraphrases, and clinically equivalent expressions.

3. Partial coverage does NOT count as a hit; the core meaning must be fully addressed.

4. Be strict but fair: the candidate should demonstrate understanding of the key clinical
   concept.

Output Format (strict JSON only):

{"hit": true|false,
 "reasoning": "explanation of why the point is/is not covered"}

Input:
Candidate Answer Text: {pred_text}
Evaluation Point: {point}
Output JSON only.
```

Figure 6: Evaluation prompts for stage-level judgment and semantic point matching used in our automated assessment pipeline. Both prompts are grounded in reference clinical guidelines to ensure evaluation validity.