

# TheraAgent: Self-Improving Therapeutic Agent for Precise and Comprehensive Treatment Planning

Anonymous ACL submission

## Abstract

Formulating a treatment plan is inherently a complex reasoning and refinement task rather than a simple generation problem. However, existing large language models (LLMs) mainly rely on one-shot output without explicit verification, which may result in rough, incomplete, and potentially unsafe treatment plans. To address these limitations, we propose **TheraAgent**, an agentic framework that replaces one-shot generation with an iterative *generate-reflect-refine* pipeline. By mirroring the actual reasoning process of human experts who iteratively revise treatment plans, our framework progressively *reflects* on clinical criteria and transforms coarse drafts into precise, comprehensive and safer therapeutic regimens. To facilitate the critical *reflection* component, we introduce **TheraJudge**, a treatment-specific evaluation module integrated into the inference loop. Experiments show TheraAgent achieves state-of-the-art results on HealthBench in treatment planning task, leading in Accuracy and Completeness. In expert evaluations, it attains an 86% win rate against physicians, with superior Targeting and Harm Control. Moreover, the high consistency between TheraJudge and HealthBench evaluation indicates the reliability of our framework.

## 1 Introduction

Treatment planning is a critical part in real-world medical decision-making, where physicians translate diagnostic conclusions into concrete therapeutic actions to optimize patient outcomes (Roosan et al., 2024). It is inherently challenging, as effective treatment planning goes beyond one-shot generation and instead demands iterative reasoning and refinement, including precise medication selection and dosing, comprehensive therapeutic coverage, and principled harm control (Chase et al., 2025). With recent advances in large language models (LLMs), there is growing interest in their

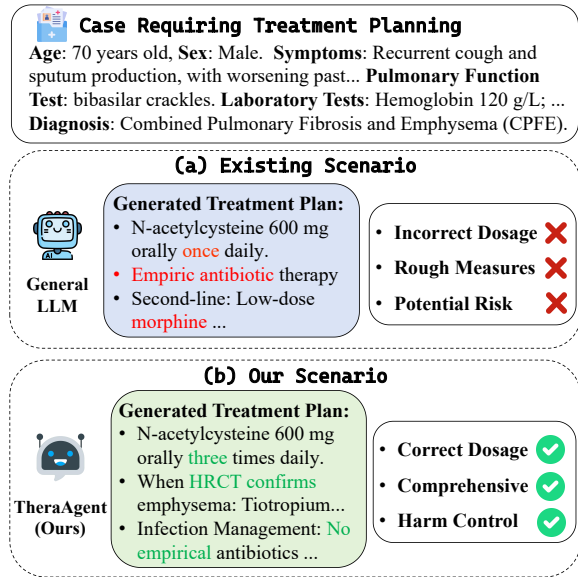


Figure 1: Comparison of treatment plan generation scenarios. (a) General-purpose models frequently produce suboptimal treatment plans, characterized by imprecise information, omitted key components, and potential safety risks. (b) **TheraAgent** mitigates these limitations by generating more accurate, comprehensive, and clinically reliable plans, with expert-validated annotations.

potential to support medical tasks, driven by their strong capabilities to incorporate relevant context, perform language comprehension, generate coherent text (Singhal et al., 2025; Liu et al., 2025a; McDuff et al., 2025; Tu et al., 2025).

However, directly applying general-purpose or even medically fine-tuned LLMs to treatment planning remains challenging (Lin et al., 2025). As illustrated in Figure 1, existing models tend to generate rough, incomplete, and potentially unsafe plans due to their one-shot mechanism without explicit verification. Despite the question, existing research for treatment planning remains limited. Many existing efforts are tailored to specific diseases and thus fail to generalize across clinical departments (Khude and Shende, 2025; Hassan et al., 2025; Zhang et al., 2025; Mohammed et al.,

2025), while more general approaches often struggle to ensure precision and completeness due to the absence of iterative refinement mechanisms (Hsu et al., 2025; Yang et al., 2025b). Furthermore, the evaluation of treatment plans frequently rely on oversimplified metrics (e.g., BLEU) or vanilla LLM-based scores that overlook critical dimensions (Gao et al., 2025; Hasan et al., 2025).

To bridge these gaps, we propose **TheraAgent**, an agentic framework that regards treatment planning as an iterative *generate–reflect–refine* process. By mirroring how physicians iteratively revise therapeutic regimens, TheraAgent enables continuous self-correction and progressive improvement, leading to more precise, complete, and safer treatment plans. To support this process without costly human intervention, we introduce **TheraJudge**, a treatment-specific evaluation module, which aligns automated feedback with clinical criteria by assessing plans on dimensions including accuracy, targeting, completeness, and safety.

We evaluate TheraAgent on treatment-related cases in HealthBench (Arora et al., 2025), achieving a state-of-the-art overall score and obtains the highest scores on Accuracy and Completeness, exceeding the second-best model by 0.88 points and 3.33 points. Its effectiveness is further showed in blinded medical expert evaluations on real-world clinical cases, where TheraAgent achieves an 86% winning rate against human physicians, with marked winning rate of 69% in Targeting, 71% in Completeness, and 51% in Harm Control. We further conduct an agreement analysis between TheraJudge and HealthBench evaluation, where TheraJudge achieves a *Pearson correlation* of 0.71, substantially outperforming other automatic metrics.

Our contributions are summarized as follows:

- We propose **TheraAgent**, an agentic framework that considers treatment planning as an iterative reasoning process. Its *generate–reflect–refine* pipeline enables effective self-improvement, yielding more precise, comprehensive, and safer treatment plans.
- We introduce **TheraJudge**, a clinically aligned internal critic within TheraAgent that assesses treatment plans along dimensions such as *Accuracy* and *Completeness*. It also shows strong potential as a standalone evaluator for treatment planning, as reflected by its high level of agreement with HealthBench evaluation.

- Extensive experiments show that TheraAgent achieves state-of-the-art performance on treatment-related HealthBench cases and attains a dominant 85% winning rate against physicians in blinded expert evaluations on real-world cases.

## 2 Related Works

**Treatment Planning Generation** Compared to diagnosis-related tasks, research on LLM-based treatment planning remains limited. Existing work includes agent-based or staged generation frameworks aligned with clinical workflows named MedPlan (Hsu et al., 2025), treatment-specific fine-tuning approaches for therapeutic reasoning named TxAgent (Gao et al., 2025), and general medical post-training or test-time optimization methods that improve overall clinical reasoning without explicitly targeting treatment planning named MedCritical (Su et al., 2025) and FineMedLM-o1 (hongzhou yu et al., 2025). Prior studies have also examined the evaluation of medical LLMs, typically relying on task-level accuracy or coarse LLM-based scores (Sousa et al., 2025). Despite these efforts, existing approaches often struggle to produce accurate, complete, and safety-aware treatment regimens, while current evaluation methods remain overly simplified and may fail to capture the multi-dimensional clinical quality required for real-world treatment planning.

**Self-improving Agent** Recent work has demonstrated that agentic and self-improving approaches can improve LLM output quality by enabling iterative refinement. General agent frameworks show that incorporating planning, memory, or reflection leads to more robust and coherent outputs (Huang et al., 2024; Liu et al., 2025b; Wang et al., 2025; Yan et al., 2025; Yin et al., 2025). In the medical domain, agent-based systems similarly improve reliability and quality through feedback-guided workflows, primarily for diagnosis tasks (Rose et al., 2025; Diao et al., 2025). Complementary studies on multi-agent evaluation further support iterative refinement as an effective mechanism for assessing and improving reasoning quality (Zhao et al., 2025). However, these methods are largely designed for general-purpose reasoning or diagnosis tasks, and have not been adapted to the treatment planning setting. Consequently, the application of self-improving agents to generating and refining precise and complete treatment plans remains underexplored.

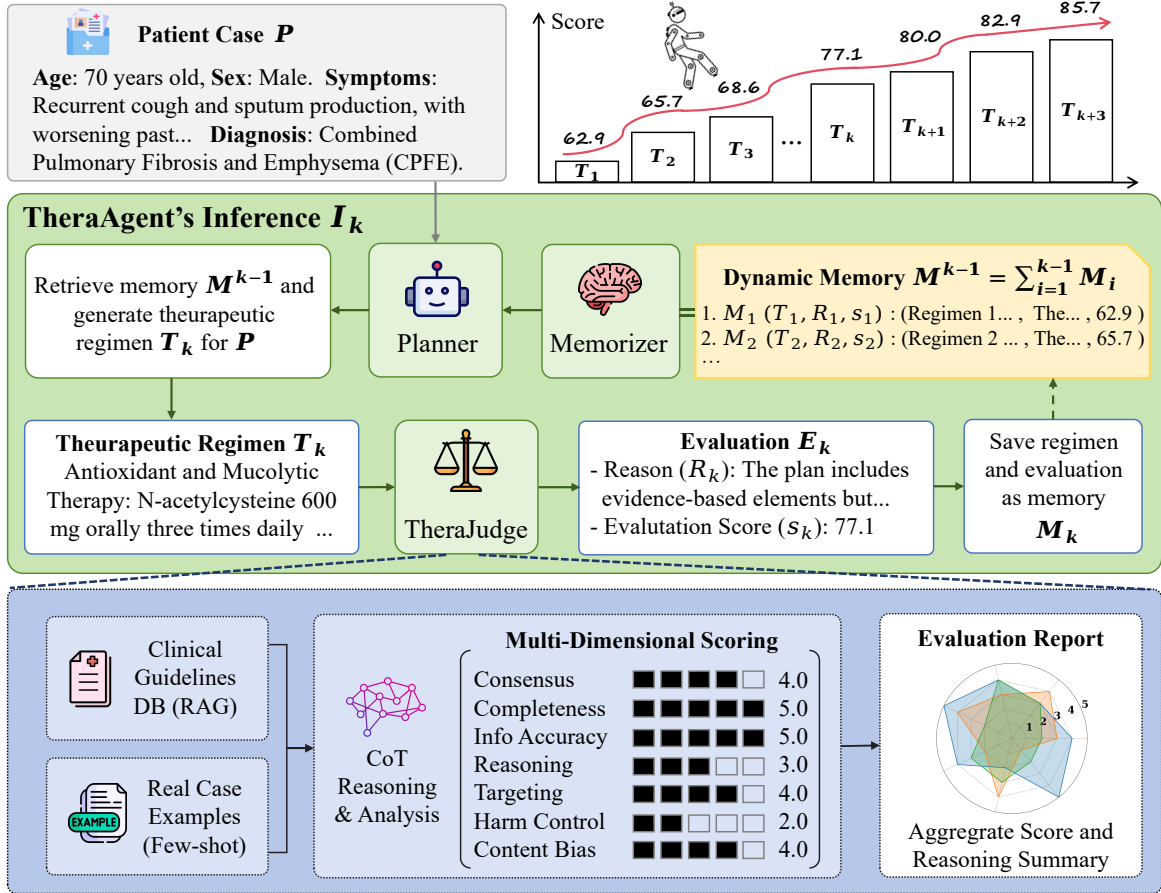


Figure 2: **Overview of the TheraAgent framework.** TheraAgent performs treatment planning through a self-improving inference pipeline. Given a patient case  $P$ , the Planner generates a therapeutic regimen  $T_k$  at iteration  $k$ , which is subsequently assessed by TheraJudge that gives multi-dimensional scores using RAG and Few-shots. The generated schedule and its evaluation are incorporated into the Memorizer to form  $M^k$ , which enables improved retrieval for the Planner in subsequent iterations and guides refined schedules ( $T_{k+1}$ ) through in-context learning.

### 3 Problem Formulation

Given a patient case  $\mathbf{P} = (\mathbf{d}, \mathbf{s}, \mathbf{y})$ , where  $\mathbf{d}$  denotes basic clinical information,  $\mathbf{s}$  denotes symptoms and clinical findings, and  $\mathbf{y}$  denotes the confirmed diagnosis, the goal of treatment planning is to generate a treatment plan  $\mathbf{T}$  along with an explicit reasoning trace  $\mathbf{c}$ , such that the resulting plan is clinically interpretable. Unlike closed-form prediction tasks, treatment planning operates over an open-ended and combinatorial action space, where valid solutions must simultaneously satisfy multiple clinical requirements. We characterize the quality of a treatment plan as a multi-dimensional function:

$$Q(\mathbf{T} | \mathbf{P}) = \sum_{i=1}^N q_i(\mathbf{T} | \mathbf{P}), \quad (1)$$

where each  $q_i$  corresponds to a clinical dimension, such as Accuracy, Targeting, Completeness, and Harm Control.

This formulation highlights the intrinsic difficulty of treatment planning: high-quality solutions require the model to balance precise medical decisions, comprehensive therapeutic coverage, and safety-aware reasoning under an open-ended space. Consequently, effectively addressing this problem calls for iterative reasoning and refinement mechanisms that can progressively identify and correct deficiencies across multiple clinical dimensions.

## 4 TheraAgent

### 4.1 Overview

To address the above challenges, we propose **Self-improving therapeutic agent (TheraAgent)**, an agentic workflow that performs iterative optimization over treatment plans via structured feedback. As illustrated in Figure 2, TheraAgent consists of three interacting components: *Planner*, *Memorizer*, and *TheraJudge*.

Given a patient case, the Planner generates a

candidate treatment plan conditioned on both the current input and historical feedback stored in the Memorizer. The generated schedule is then evaluated by TheraJudge, which provides clinically grounded, multi-dimensional feedback, including detailed rationales and an overall quality score. These evaluation results are stored in the Memorizer and used to guide subsequent generations.

## 4.2 Planner

The **Planner** serves as the generative module of TheraAgent, responsible for producing treatment plans conditioned on the patient case and historical feedback. At iteration  $k$ , given a patient case  $\mathbf{P}$  and the accumulated memory state  $\mathcal{M}^{(k-1)}$ , the Planner generates the  $k$ -th candidate treatment plan  $\mathbf{T}_k$  together with its corresponding intermediate reasoning process  $\mathbf{c}_k$ , formalized as:

$$(\mathbf{T}_k, \mathbf{c}_k) = f_\theta(\mathbf{P}, \mathcal{M}^{(k-1)}), \quad (2)$$

where  $f_\theta$  denotes the underlying inference model parameterized by  $\theta$ .

Unlike one-shot generation, the Planner follows an iterative refinement paradigm by explicitly conditioning on previously generated schedules and their evaluation rationales and scores stored in the Memorizer. This feedback guides the model to identify and correct deficiencies from earlier iterations, such as incomplete coverage, suboptimal clinical targeting, or potential safety risks, thereby transforming treatment plan generation into a progressive optimization process.

## 4.3 TheraJudge

Each candidate treatment plan generated by the Planner is subsequently evaluated by **TheraJudge**, which serves as a clinical judging model providing structured, multi-dimensional feedback. Given a patient case  $\mathbf{P}$  and the  $k$ -th candidate schedule  $\mathbf{T}_k$ , TheraJudge produces evaluation rationales  $\mathbf{R}_k$ , dimension-wise scores  $\{q_{k,i}\}_{i=1}^N$  along clinical axes, and an aggregated score  $s_k$ , formalized as:

$$(\mathbf{R}_k, \{q_{k,i}\}_{i=1}^N, s_k) = g_\phi(\mathbf{P}, \mathbf{T}_k), \quad (3)$$

where  $g_\phi$  denotes the TheraJudge model. The final score  $s_k$  is computed as a weighted sum of the individual dimensions, with weights reflecting practical clinical priorities such as Consensus and Completeness. The evaluated tuple  $(\mathbf{T}_k, \mathbf{R}_k, s_k)$  is then stored in the Memorizer to update the memory

state  $\mathcal{M}^{(k)}$ , providing explicit and structured feedback that guides subsequent iterations of treatment plan refinement.

There are three parts in TheraJudge. **Knowledge-grounded evaluation via RAG.** TheraJudge employs a RAG module to retrieve authoritative guidelines based on the patient case and proposed schedule. The retrieval corpus consists of a large-scale collection of real-world clinical guidelines and medical literature, including over 600 documents curated from an authentic website<sup>1</sup>. By grounding assessments in external medical knowledge, the framework reduces hallucinations and enhances adherence to safety standards.

**Stabilization via few-shot clinical exemplars.** To ensure consistent scoring, TheraJudge utilizes few-shot prompting with expert-annotated clinical cases. These exemplars calibrate the model’s judgment, reducing scoring variance compared to zero-shot approaches. This stability is essential for providing the reliable feedback necessary for effective iterative optimization.

**Multi-dimensional clinical scoring.** Instead of a single holistic score, TheraJudge performs multi-dimensional assessments covering completeness, safety, and consensus adherence. This mirrors real-world clinical decision-making and provides fine-grained feedback. The resulting reasoning traces are stored in the Memorizer, offering explicit signals that guide the Planner to address specific weaknesses in subsequent iterations.

## 4.4 Memorizer

The **Memorizer** maintains a structured repository of historical treatment plans together with their corresponding evaluations, enabling experience accumulation across iterations. Each memory item  $\mathbf{M}_i = (\mathbf{T}_i, \mathbf{R}_i, s_i)$  encapsulates the generated schedule, reasoning traces, and evaluation scores from iteration  $i$ . At iteration  $k$ , the memory state is defined as the collection of all past memory items as  $\mathcal{M}^{(k-1)} = \{\mathbf{M}_i\}_{i=1}^{k-1}$ , which is incrementally updated as new treatment plans are generated and evaluated.

To facilitate refinement, the Planner performs score-aware retrieval, selecting a subset of memory items with the  $Top-N$  highest scores in  $\mathcal{M}^{(k-1)}$ . By prioritizing high-quality historical schedules, the Memorizer stabilizes the self-improving process, enabling TheraAgent to progressively im-

<sup>1</sup><https://seleguide.yiigle.com/webs/Knowledge>

Model	Overall	Theme				Axis		
		Global Health	Hedging	Context Seeking	Communication	Accuracy	Completeness	Context Awareness
Medical Specialized Models								
UltraMedical-70B	23.45	17.70	30.77	20.39	20.28	29.54	25.03	30.99
Llama3-Med42-70B	24.45	15.19	28.97	22.48	32.43	33.87	25.78	31.20
MedCritical-7B	24.73	20.04	35.33	25.52	22.20	33.05	31.73	32.03
Baichuan-M2-32B	38.76	35.23	41.47	33.38	46.11	36.10	37.55	34.04
Open-Source Models								
DeepSeek-R1	42.18	39.96	48.10	37.84	48.79	41.78	47.08	30.72
Qwen3-235B-A22B	43.11	34.48	45.91	37.87	51.08	41.48	38.88	32.93
Kimi-K2	44.20	34.50	42.60	41.05	54.77	42.60	41.78	34.41
Proprietary Models								
GPT-4o	19.35	12.49	20.22	19.57	24.80	25.92	29.13	26.68
OpenAI-o1	32.03	21.14	33.77	26.85	40.25	34.95	29.89	30.92
GPT-4.1	34.48	21.65	33.24	27.87	45.92	36.66	31.42	30.03
OpenAI-o4-mini	39.46	32.25	40.85	35.61	52.11	40.78	44.19	31.14
Grok-3	42.51	34.16	41.97	37.83	<b>54.97</b>	41.07	38.84	<b>37.62</b>
Claude-4-Sonnet	44.56	35.91	46.66	40.90	50.17	40.85	41.50	36.98
Gemini-2.5-Pro	45.68	36.09	47.99	41.00	53.16	42.77	40.61	33.89
<b>TheraAgent(Ours)</b>	<b>47.46</b>	<b>46.68</b>	<b>54.75</b>	<b>42.21</b>	53.74	<b>43.65</b>	<b>50.41</b>	35.27

Table 1: Performance comparison on HealthBench across different models. We report the overall score, theme-level and axis-level scores for medical, open-source, and proprietary models. The best results are bolded.

prove therapeutic quality with reduced computational cost.

#### 4.5 Agent Output

TheraAgent outputs the schedule  $\mathbf{T}^*$  by maximizing the score in the final  $L$  iterations shown as:

$$\mathbf{T}^* = \arg \max_{\mathbf{T}_k, k \in \{N-L+1, \dots, N\}} s_k. \quad (4)$$

This selection leverages the performance gains from accumulated feedback while mitigating potential late-stage fluctuations. To optimize efficiency, an early stopping mechanism terminates the process if scores consistently meet a threshold  $\tau$  for three consecutive iterations formed as  $s_{(k-2)}, s_{(k-1)}, s_k \geq \tau$ . This method ensures high-quality stability while minimizing unnecessary computational overhead.

In summary, through the integration of iterative generation, multi-dimensional judging and memory-guided refinement, TheraAgent enhances the accuracy, completeness, and safety of treatment plans.

## 5 Experiments

### 5.1 Dataset

**HealthBench.** HealthBench provides a comprehensive and realistic benchmark for evaluating real-world healthcare capabilities (Arora et al., 2025).

As our study focuses on the task of treatment planning, we first filter the original dataset to retain only treatment-related conversations using GPT-4.1 as a classifier. The filtered samples are categorized into four medical departments: endocrinology (265), gastroenterology (262), neurology (395), and respiratory (319), totaling 1,241 cases for evaluation. Detailed dataset statistics and department-wise distributions are provided in Appendix B.1.

**Real-World Case Dataset.** We incorporate 35 physician-authored cases from a public platform<sup>2</sup>, with 9 cases each from respiratory, neurology, cardiology, and 8 cases from ophthalmology. These cases represent real-world, clinical scenarios with complex therapeutic challenges. Additional case descriptions are provided in Appendix B.2.

### 5.2 Evaluation Metrics

**Automatic evaluation.** We adopt HealthBench as the automatic evaluation for large-scale experiments. Following its specific rubric, responses are assessed across several medical dimensions. The hand-crafted rubrics comprehensively capture the real-world expectations for a treatment plan. GPT-4.1 is used as the rubric evaluator to ensure reliable and efficient scoring, consistent with prior HealthBench evaluations.

<sup>2</sup><https://www.yiigle.com/Paper/>

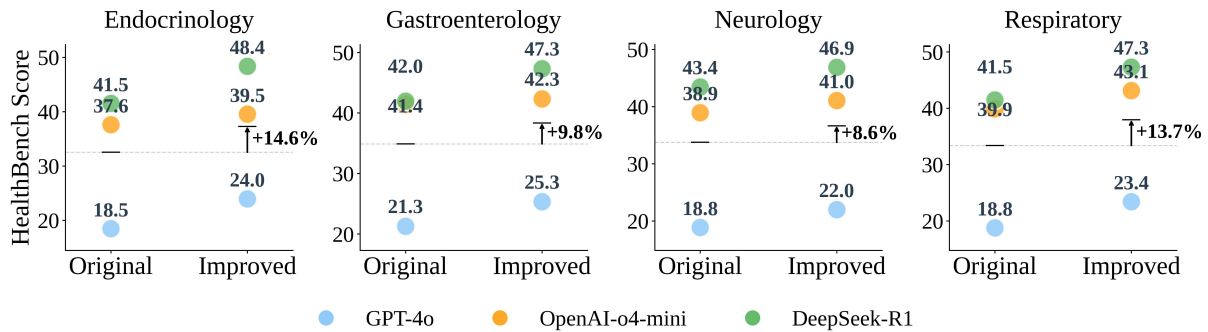


Figure 3: Generalization analysis of TheraAgent across four medical departments. The plot compares the HealthBench scores of three backbone models in their "Original" (base) state versus the "Improved" state (using TheraAgent framework). Horizontal bold lines represent the mean score of the three models in each state. The annotated values indicate the relative improvement rate, calculated as the increase in the mean score divided by the original mean.

**Medical Experts annotation.** We further conduct human evaluation on the Real-World Case dataset with licensed medical experts who have over four years of clinical experience. Experts compare treatment plans generated by TheraAgent, raw LLMs, and physician-authored references. Detailed annotation protocols, evaluation criteria, and interface designs are provided in Appendix C.

### 5.3 Baselines and Hyperparameters

**Baselines.** We compare TheraAgent against a representative set of strong baselines. Specifically, we include MedCritical (Su et al., 2025) and Baichuan-M2 (Team et al., 2025a) as medical-specialized models; DeepSeek-R1 (DeepSeek-AI et al., 2025) and Kimi-K2 (Team et al., 2025b) as open-source general models; and Grok-3 (xAI, 2025), and Claude-4-Sonnet (Anthropic, 2025) as proprietary baselines. Closed-source agentic methods such as MedPlan (Hsu et al., 2025) are excluded due to limited accessibility. A complete list of evaluated models and implementation details are provided in Appendix D.1.

**TheraAgent configuration.** TheraAgent utilizes DeepSeek-R1 as the backbone for both the Planner and TheraJudge. In each iteration, the Planner retrieves  $Top-N = 3$  memory items. The early stopping threshold is set to  $\tau = 98$  with a maximum of 10 iterations, while the output window  $L = 3$ . To stabilize evaluation, TheraJudge incorporates 3 few-shot exemplars per department. Notably, RAG is disabled during HealthBench evaluation to mitigate potential biases from region-specific guidelines, ensuring a fair assessment grounded in universal clinical reasoning. Due to the substantial computational cost of each run, all reported results are based on a single run.

### 5.4 Main Results

**Automatic evaluation results.** The performance of TheraAgent on the HealthBench subset is summarized in Table 1. Overall, TheraAgent achieves a state-of-the-art score of 47.46, surpassing the second-best model by 1.78 points. A dimension-wise analysis further shows that TheraAgent outperforms all baselines across several critical dimensions. Notably, it achieves substantial improvements in *Hedging* of 6.65 points and *Completeness* of 3.33 points, indicating more precise clinical decision-making and reduced omissions in treatment planning. In addition, TheraAgent attains a state-of-the-art score of 42.21 in *Context Seeking*, reflecting its enhanced ability to actively elicit and integrate specific information for more targeted therapeutic planning. Complete results across all evaluated models are reported in Appendix E.1.

To assess robustness across models, we perform a stratified analysis over different models. As illustrated in Figure 3, TheraAgent consistently enhances the treatment planning capabilities of diverse backbones, ranging from standard models like GPT-4o to reasoning-specialized models like DeepSeek-R1. Across four distinct medical departments, the framework yields uniform performance boosts, with average relative improvements ranging from +8.6% to +14.6%. Notably, every model in every department exhibits a positive trajectory from its "Original" to "Improved" state.

**Human annotation results.** To assess clinical utility, we conducted a comprehensive blinded evaluation on the Real-World Case dataset involving licensed experts, comparing TheraAgent against both human physicians and the base model (DeepSeek-R1). As illustrated in Figure 4 (top-left), TheraAgent demonstrates a dominant supe-

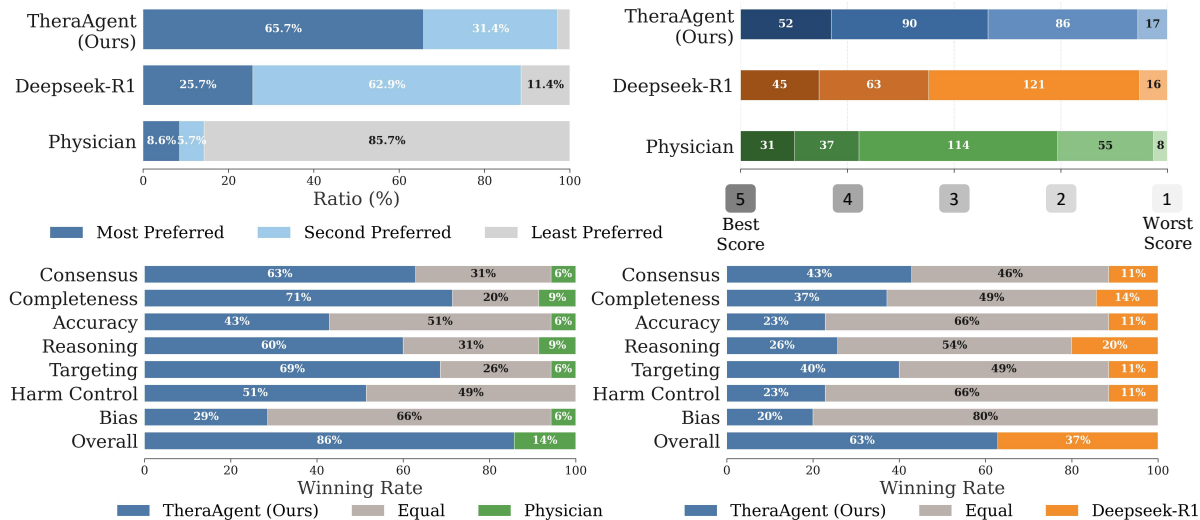


Figure 4: **Expert evaluation on Real Medical Cases.** Top: Three-way preference rankings (left) and 5-point rating distributions (right), with numbers indicating the absolute count for each score. Bottom: Pairwise comparisons across seven clinical dimensions against human physicians (left) and DeepSeek-R1 (right).

priority in three-way preference rankings, being selected as the “Most Preferred” in 65.7% of cases, substantially outperforming both the one-shot base model in 25.7% and human physicians in 8.6%. We further analyze the factors underlying the comparatively lower performance of human physicians in Section 6.3.

Detailed pairwise comparisons (Figure 4, bottom) further reveal the specific drivers of superiority of TheraAgent. Compared to the base model (DeepSeek-R1), TheraAgent achieves a 63% overall winning rate, with the gains in *Targeting* of 40% and *Consensus* of 43%, indicating more patient-specific and precise treatment plans. The detailed dimension-wise rating results are provided in Appendix E.2, which further corroborate the observed performance advantages of TheraAgent. Overall, these results demonstrate that TheraAgent’s self-improving pipeline effectively produces precise, safe, and comprehensive treatment regimens.

**Judge Agreement with HealthBench** To validate the reliability of our internal critic, we assess the alignment between TheraJudge and official HealthBench evaluation (Table 2), following the experimental settings described in Appendix D.2. TheraJudge demonstrates high consistency with established standards, achieving a Pearson correlation of 0.7052 and Spearman correlation of 0.6669. In contrast, traditional lexical metrics (e.g., BLEU, ROUGE) and vanilla LLM scoring exhibit weak correlations, highlighting their inability to capture the complex semantic nuances of clinical reason-

Evaluation	Spearman	Pearson	CCC
BLEU	0.0000	-0.0221	-0.0179
ROUGE1	0.2052	0.2179	0.1989
ROUGE2	0.1118	0.1194	0.1075
ROUGEL	0.1118	0.1516	0.1348
BERT Score	0.1539	0.1106	0.1009
LLM Score	0.1539	0.1089	0.0879
<b>TheraJudge (Ours)</b>	<b>0.6669</b>	<b>0.7052</b>	<b>0.6467</b>

Table 2: **Consistency experiment of different evaluation methods against HealthBench evaluation.** The correlation between distinct method scores and HealthBench overall scores, are evaluated across five model outputs per case. Reported values are the median scores across all cases of neurology department.

ing. This alignment is critical for the success of the TheraAgent framework. It indicates that *TheraJudge acts as a faithful proxy for expert evaluation*, providing valid feedback signals regarding safety and completeness rather than arbitrary noise.

## 6 Analysis

### 6.1 Inference-Time Scaling

We conduct inference time experiments on 10 cases of Neurology due to the high cost of HealthBench evaluation at each step. Figure 5 illustrates the evolution of HealthBench score as the number of inference steps increases. As shown in the figure, the scores exhibit an upward trend over successive iterations. The fitted linear trend further indicates a positive correlation between iteration count and evaluation quality. This behavior indicates that

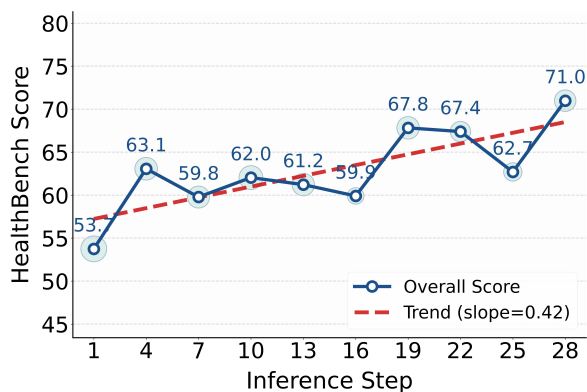


Figure 5: **Inference-time scaling in TheraAgent:** performance progressively improves across inference steps. Each point denotes the mean HealthBench score over cases, and the red dashed line ( - - ) shows the positive linear trend, demonstrating consistent gains from iterative refinement.

Dimensions	Fewshots	RAG	HealthBench Score
✓			48.66
	✓		50.62
		✓	45.98
✓	✓		<b>52.36</b>
✓	✓	✓	45.96
Base Model w/o Judge			41.15
Vanilla Judge			48.50

Table 3: Ablation study of TheraJudge components.

the self-improving framework effectively leverages feedback to refine subsequent generations.

## 6.2 Component Ablation

Table 3 presents an ablation study of the three components of TheraJudge. All variants outperform the base model without a judge, which scores 41.15, underscoring the importance of structured evaluation. Among individual components, few-shot exemplars yield the strongest improvement with a score of 50.62, indicating their effectiveness in stabilizing evaluation behavior, while dimensional scoring also brings consistent gains with a score of 48.66. In contrast, RAG alone results in relatively modest improvement at 45.98, likely due to guideline heterogeneity in the globally sourced HealthBench dataset. The best performance is obtained by combining few-shot exemplars with dimensional scoring, reaching a score of 52.36. Incorporating RAG further reduces performance to 45.96, which supports our decision to disable RAG during HealthBench evaluation. Additional ablation results on the Memorizer and RAG components are provided in Appendix E.3.

## 6.3 Case Study

A comparison of treatment plans for a CPFE case (Appendix E.4) exhibits TheraAgent’s clinical superiority across two key dimensions. **Precision and Specificity:** TheraAgent avoids generic errors by tailoring decisions to specific patient data. It correctly withheld antibiotics based on the patient’s normal inflammatory markers and explicitly conditioned antifibrotic therapy on HRCT findings, ensuring interventions were grounded in the individual’s diagnostic context. **Safety and Risk Control:** DeepSeek-R1 exhibits critical errors, including sub-therapeutic N-acetylcysteine dosing (600mg/day vs required 1,800mg/day) and premature morphine use. TheraAgent ensures safety by correcting these dosages, implementing strict intervention thresholds, and mandating continuous monitoring of respiratory and renal functions.

In this case, the physician’s plan helps explain its lower expert annotation scores by focusing on core antifibrotic therapy and general monitoring while omitting explicit decision criteria (e.g., HRCT pattern-dependent treatment). Such condensation is common in routine clinical practice due to patient affordability, medication adherence, and local drug availability. In contrast, TheraAgent is not designed to replicate these practice-driven limitations, but to generate an idealized treatment plan.

## 7 Conclusion

In this work, we redefine treatment planning not as simple generation, but as a distinct reasoning and refinement task. We propose TheraAgent, an agentic framework that replaces one-shot generation with a *generate-reflect-refine* pipeline, mimicking the iterative cognitive process of clinicians. Central to this approach is TheraJudge, a clinically aligned internal critic that assesses plans against key dimensions like accuracy and safety, driving active self-correction. Experiments on HealthBench and expert-annotated real-world cases demonstrate that TheraAgent achieves state-of-the-art performance and an 86% winning rate against human physicians. Furthermore, TheraJudge exhibits high agreement with HealthBench evaluation, validating its dual role as both an optimization signal and a reliable metric. Our findings show that integrating domain-specific evaluation into the inference loop is essential for transforming LLMs into safe, precise, and practical clinical planners.

## 538 Limitations

539 There are some limitations in our work. First,  
540 our experimental validation of the iterative frame-  
541 work was primarily conducted using specific  
542 high-performance backbone models (DeepSeek-  
543 R1, GPT-4o, and OpenAI-o4-mini). While these  
544 models demonstrate the effectiveness of the ap-  
545 proach, the generalizability of TheraAgent across  
546 a broader spectrum of smaller-scale models re-  
547 mains to be fully characterized. Second, the self-  
548 improving design of TheraAgent introduces higher  
549 computational overhead than single-pass genera-  
550 tion. Although we employ score-aware memory re-  
551 trieval and early-stopping mechanisms to improve  
552 efficiency, the framework may still be less suit-  
553 able for real-time or resource-constrained clinical  
554 settings without further optimization. Third, Thera-  
555 Agent focuses on text-based treatment planning  
556 and does not directly incorporate structured clinical  
557 signals such as laboratory time series, imaging  
558 data, or real-time patient monitoring. Extending the  
559 framework to multimodal clinical inputs remains  
560 an important direction for future work.

## 561 Ethical Considerations

562 The application of large language models to clinical  
563 decision support raises important ethical and safety  
564 considerations. TheraAgent is designed to assist  
565 treatment planning rather than to replace profes-  
566 sional medical judgment, and its outputs should not  
567 be treated as definitive clinical decisions. Despite  
568 its strong empirical performance, the framework  
569 may still produce incomplete, incorrect, or context-  
570 insensitive recommendations, particularly in the  
571 presence of rare conditions, atypical patient pro-  
572 files, or missing clinical information. As such, its  
573 deployment requires appropriate clinical oversight,  
574 clear usage boundaries, and integration into ex-  
575 isting medical workflows with human-in-the-loop  
576 safeguards.

577 To mitigate potential risks, TheraAgent em-  
578 phasizes transparency through explicit reason-  
579 ing traces, structured feedback, and guideline-  
580 grounded evaluation. These mechanisms enable  
581 physicians to interpret, verify, and critique gener-  
582 ated treatment plans, reducing the risk of uncritical  
583 reliance on model outputs.

584 Regarding data ethics and privacy, our study  
585 relies solely on publicly available or authorized  
586 datasets, with all patient cases fully de-identified.  
587 We further acknowledge that clinical guidelines and

standards of care vary across regions. Accordingly,  
we control external guideline use during evaluation.  
Future deployments should incorporate localized  
clinical oversight and regulatory compliance to en-  
sure responsible real-world use.

## References

- Anthropic. 2025. [System card: Claude Opus 4 & Claude Sonnet 4](#). System card (PDF). Accessed 2025-12-30.
- Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. [HealthBench: Evaluating large language models towards improved human health](#). *Preprint*, arXiv:2505.08775.
- Aaron Chase, Amoreena Most, Andrea Sikora, Susan E. Smith, John W. Devlin, Shaochen Xu, Tianming Liu, and Brian Murray. 2025. [Evaluation of large language models’ ability to identify clinically relevant drug-drug interactions and generate high-quality clinical pharmacotherapy recommendations](#). *American Journal of Health-System Pharmacy*.
- Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. 2025. [Towards medical complex reasoning with LLMs through medical verifiable problems](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14552–14573, Vienna, Austria. Association for Computational Linguistics.
- Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. [Med42-v2: A suite of clinical LLMs](#). *Preprint*, arXiv:2408.06142.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *Preprint*, arXiv:2507.06261.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. [DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- Lingxiao Diao, Xinyue Xu, Wanxuan Sun, Cheng Yang, and Zhuosheng Zhang. 2025. [GuideBench: Benchmarking domain-oriented guideline following for](#)

643	LLM agents. In <i>Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 11361–11399, Vienna, Austria. Association for Computational Linguistics.	Xiaohong Liu, Hao Liu, Guoxing Yang, Zeyu Jiang, Shuguang Cui, Zhaoze Zhang, Huan Wang, Liyuan Tao, Yongchang Sun, Zhu Song, Tianpei Hong, Jin Yang, Tianrun Gao, Jiangjiang Zhang, Xiaohu Li, Jing Zhang, Ye Sang, Zhao Yang, Kanmin Xue, and 5 others. 2025a. <a href="#">A generalist medical language model for disease diagnosis assistance</a> . <i>Nature Medicine</i> , 31(3).	700 701 702 703 704 705 706 707
648	Shanghua Gao, Richard Zhu, Zhenglun Kong, Ayush Noori, Xiaorui Su, Curtis Ginder, Theodoros Tsiligkaridis, and Marinka Zitnik. 2025. <a href="#">TxAgent: An AI agent for therapeutic reasoning across a universe of tools</a> . <i>arXiv preprint arXiv:2503.10970</i> .	Yitao Liu, Chenglei Si, Karthik R Narasimhan, and Shunyu Yao. 2025b. <a href="#">Contextual experience replay for self-improvement of language agents</a> . In <i>Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 14179–14198, Vienna, Austria. Association for Computational Linguistics.	708 709 710 711 712 713 714
653	Md. Mehedi Hasan, Rafid Mostafiz, Md. Abir Hosain, and Bikash Kumar Paul. 2025. <a href="#">CLIN-LLM: A safety-constrained hybrid framework for clinical diagnosis and treatment generation</a> . <i>Preprint</i> , arXiv:2510.22609.	Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, Le Hou, Yong Cheng, Yun Liu, S. Sara Mahdavi, Sushant Prakash, Anupam Pathak, Christopher Semturs, Shwetak Patel, Dale R. Webster, and 9 others. 2025. <a href="#">Towards accurate differential diagnosis with large language models</a> . <i>Nature</i> , 642(8067).	715 716 717 718 719 720 721 722
658	Bassam Abdul Rasool Hassan, Ali Haider Mohammed, Souheil Hallit, Diana Malaeb, and Hassan Hosseini. 2025. <a href="#">Exploring the role of artificial intelligence in chemotherapy development, cancer diagnosis, and treatment: present achievements and future outlook</a> . <i>Frontiers in Oncology</i> , 15.	Abdul M. Mohammed, Iqtidar Mansoor, Sarah Blythe, and Dennis Trujillo. 2025. <a href="#">Developing an artificial intelligence tool for personalized breast cancer treatment plans based on the NCCN guidelines</a> . <i>Preprint</i> , arXiv:2502.15698.	723 724 725 726 727
664	hongzhou yu, Tianhao Cheng, Yingwen Wang, Wen He, Qing Wang, Ying Cheng, Yuejie Zhang, Rui Feng, and Xiaobo Zhang. 2025. <a href="#">FineMedLM-o1: Enhancing medical knowledge reasoning ability of LLM from supervised fine-tuning to test-time training</a> . In <i>Second Conference on Language Modeling</i> .	OpenAI. 2025a. <a href="#">Introducing GPT-4.1 in the API</a> . OpenAI product release page. Accessed 2025-12-30.	728 729
670	Hsin-Ling Hsu, Cong-Tinh Dao, Luning Wang, Zitao Shuai, Thao Nguyen Minh Phan, Jun-En Ding, Chun-Chieh Liao, Pengfei Hu, Xiaoxue Han, Chih-Ho Hsu, Dongsheng Luo, Wen-Chih Peng, Feng Liu, Fang-Ming Hung, and Chenwei Wu. 2025. <a href="#">MedPlan: A two-stage RAG-based system for personalized medical plan generation</a> . In <i>Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)</i> , pages 1072–1082, Vienna, Austria. Association for Computational Linguistics.	OpenAI. 2025b. <a href="#">Introducing OpenAI o3 and o4-mini</a> . OpenAI product release page. Accessed 2025-12-30.	730 731
681	Xiang Huang, Sitao Cheng, Shanshan Huang, Jiayu Shen, Yong Xu, Chaoyun Zhang, and Yuzhong Qu. 2024. <a href="#">QueryAgent: A reliable and efficient reasoning framework with environmental feedback based self-correction</a> . In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 5014–5035, Bangkok, Thailand. Association for Computational Linguistics.	OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024a. <a href="#">GPT-4o system card</a> . <i>Preprint</i> , arXiv:2410.21276.	732 733 734 735 736 737 738 739
690	Hrishikesh Khude and Pravin Shende. 2025. <a href="#">AI-driven clinical decision support systems: Revolutionizing medication selection and personalized drug therapy</a> . <i>Advances in Integrative Medicine</i> , 12(4).	OpenAI, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, and 243 others. 2024b. <a href="#">OpenAI o1 system card</a> . <i>Preprint</i> , arXiv:2412.16720.	740 741 742 743 744 745 746 747
694	Zhiwu Lin, Yuanyuan Li, Min Wu, Hongmei Liu, Xiaoyang Song, Qian Yu, Guibao Xiao, and Jiajun Xie. 2025. <a href="#">Performance analysis of large language models Chatgpt-4o, OpenAI O1, and OpenAI O3 mini in clinical treatment of pneumonia: a comparative study</a> . <i>Clinical and Experimental Medicine</i> , 25(1).	Don Roosan, Pauline Padua, Raiyan Khan, Hasiba Khan, Claudia Verzosa, and Yanting Wu. 2024. <a href="#">Effectiveness of ChatGPT in clinical pharmacy and the role of artificial intelligence in medication therapy management</a> . <i>Journal of the American Pharmacists Association</i> , 64(2).	748 749 750 751 752 753
699		Daniel Philip Rose, Chia-Chien Hung, Marco Lepri, Israa Alqassem, Kiril Gashtevski, and Carolin	754 755

756	Lawrence. 2025. <a href="#">MEDDxAgent: A unified modular agent framework for explainable automatic differential diagnosis</a> . In <i>Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 13803–13826, Vienna, Austria. Association for Computational Linguistics.	813
757		814
758		
759		
760		
761		
762		
763	Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaeckermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. <a href="#">Toward expert-level medical question answering with large language models</a> . <i>Nature Medicine</i> , 31(3).	
764		
765		
766		
767		
768		
769		
770		
771		
772	Diogo Sousa, Guilherme Barbosa, Catarina Rocha, and Dulce Oliveira. 2025. <a href="#">Performance of large language models in supporting medical diagnosis and treatment</a> . <i>arXiv preprint arXiv:2504.10405</i> .	
773		
774		
775		
776	Xinchun Su, Chunxu Luo, Yixuan Li, Weidong Yang, and Lipeng Ma. 2025. <a href="#">MedCritical: Enhancing medical reasoning in small language models via self-collaborative correction</a> . <i>arXiv preprint arXiv:2509.23368</i> .	
777		
778		
779		
780		
781	Baichuan-M2 Team, Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, Xianrong Zeng, Yijie Zhou, Chenzheng Zhu, Da Pan, Fei Deng, Guangwei Ai, Guosheng Dong, Hongda Zhang, and 15 others. 2025a. <a href="#">Baichuan-M2: Scaling medical capability with large verifier system</a> . <i>Preprint</i> , arXiv:2509.02208.	
782		
783		
784		
785		
786		
787		
788		
789	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. 2025b. <a href="#">Kimi K2: Open agentic intelligence</a> . <i>Preprint</i> , arXiv:2507.20534.	
790		
791		
792		
793		
794		
795		
796	Tao Tu, Mike Schaeckermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, and 7 others. 2025. <a href="#">Towards conversational diagnostic artificial intelligence</a> . <i>Nature</i> , 642(8067).	
797		
798		
799		
800		
801		
802		
803		
804	Yaoke Wang, Yun Zhu, Xintong Bao, Xintong Bao, Wenqiao Zhang, Suyang Dai, Kehan Chen, Wenqiang Li, Gang Huang, Siliang Tang, and Yueting Zhuang. 2025. <a href="#">Meta-Reflection: A feedback-free reflection learning framework</a> . In <i>Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 3958–3976, Vienna, Austria. Association for Computational Linguistics.	
805		
806		
807		
808		
809		
810		
811		
812		
	xAI. 2025. <a href="#">Grok 3 Beta — the age of reasoning agents</a> . xAI news release. Accessed 2025-12-30.	813
		814
	Cilin Yan, Jingyun Wang, Lin Zhang, Ruihui Zhao, Xiaopu Wu, Kai Xiong, Qingsong Liu, Guoliang Kang, and Yangyang Kang. 2025. <a href="#">Efficient and accurate prompt optimization: the benefit of memory in exemplar-guided reflection</a> . In <i>Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 753–779, Vienna, Austria. Association for Computational Linguistics.	815
		816
		817
		818
		819
		820
		821
		822
		823
	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. 2025a. <a href="#">Qwen3 technical report</a> . <i>Preprint</i> , arXiv:2505.09388.	824
		825
		826
		827
		828
		829
		830
	Dongrong Yang, Xin Wu, Yibo Xie, Xinyi Li, Qiuwen Wu, Jackie Wu, and Yang Sheng. 2025b. <a href="#">Zero-shot large language model agents for fully automated radiotherapy treatment planning</a> . In <i>The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance</i> .	831
		832
		833
		834
		835
		836
	Xunjian Yin, Xinyi Wang, Liangming Pan, Li Lin, Xiaojun Wan, and William Yang Wang. 2025. <a href="#">Gödel Agent: A self-referential agent framework for recursively self-improvement</a> . In <i>Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 27890–27913, Vienna, Austria. Association for Computational Linguistics.	837
		838
		839
		840
		841
		842
		843
		844
	Jueye Zhang, Chao Yang, Youfang Lai, Kai-Wen Li, Wenting Yan, Yunzhou Xia, Haimei Zhang, Jingjing Zhou, Gen Yang, Chen Lin, Tian Li, and Yibao Zhang. 2025. <a href="#">Large-scale automatic carbon ion treatment planning for head and neck cancers via parallel multi-agent reinforcement learning</a> . <i>Preprint</i> , arXiv:2511.02314.	845
		846
		847
		848
		849
		850
		851
	Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, Xingtai Lv, Hu Jinfang, Zhiyuan Liu, and Bowen Zhou. 2024. <a href="#">UltraMedical: Building specialized generalists in biomedicine</a> . In <i>The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track</i> .	852
		853
		854
		855
		856
		857
		858
	Ruochen Zhao, Wenxuan Zhang, Yew Ken Chia, Weiqi Xu, Deli Zhao, and Lidong Bing. 2025. <a href="#">AutoArena: Automating LLM evaluations with agent peer battles and committee discussions</a> . In <i>Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 4440–4463, Vienna, Austria. Association for Computational Linguistics.	859
		860
		861
		862
		863
		864
		865
		866

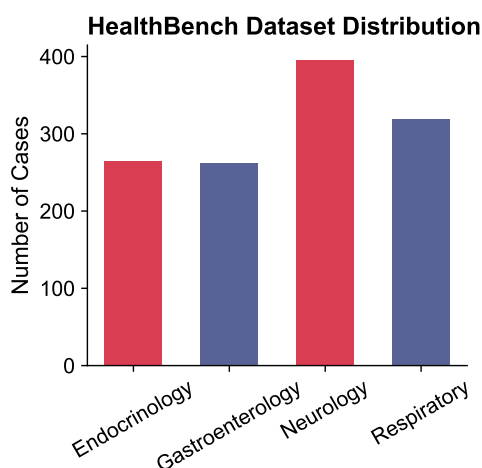


Figure 6: Department distribution of the HealthBench Dataset.

## A Use of Large Language Models in Writing

We use large language models to assist with writing and language polishing of the manuscript.

## B Dataset Details

### B.1 HealthBench

**Dataset background.** Automatic evaluation of regimen quality is an important yet difficult task. To do so, we employ HealthBench, a benchmark dataset developed under the leadership of OpenAI, designed to evaluate the real-world healthcare capabilities of large language models (LLMs). The dataset was compiled over the course of one year with contributions from 262 physicians across 60 countries, who collectively represent 49 languages and possess 26 types of professional medical training. HealthBench consists of 5,000 samples, each of which simulates a conversation between a patient (or lay user) and a clinical doctor. The task for the tested LLM is to generate an appropriate response to the final user query, while fully considering the preceding conversational context. This setup closely mirrors authentic clinical communication, thereby providing a rigorous test of the model’s ability to reason across multi-turn dialogue. A distinguishing feature of HealthBench is its multilingual coverage and its incorporation of a wide range of medical departments and clinical scenarios. These characteristics enable the benchmark to comprehensively evaluate an LLM’s robustness across linguistic, cultural, and domain-specific variations. Moreover, HealthBench has attracted sig-

nificant attention from major scientific and technological corporations, many of which have already evaluated their models on this dataset. This growing adoption underscores the dataset’s credibility and practical relevance, positioning it as an authoritative benchmark for assessing LLM performance in healthcare.

**Evaluation method and metrics.** The evaluation methodology of HealthBench is grounded in a rubric-based framework. Specifically, each sample in the dataset is accompanied by a set of rubrics, which are carefully designed by licensed physicians according to the clinical dialogue within the sample. Each rubric consists of a well-defined criterion and an associated score, where the score may be either positive or negative. During evaluation, if the response generated by a tested model satisfies a given criterion, the corresponding score is added to its mark. The final score for a response is computed as the ratio between the total score obtained across all rubrics of the sample and the maximum possible score defined by those rubrics. The overall performance of a model on HealthBench is then determined by averaging these normalized scores across all samples in the dataset.

Furthermore, each rubric is assigned to some of five evaluation dimensions: **Communication Quality** – the clarity, coherence, and empathy of the response; **Instruction Following** – the degree to which the model adheres to user instructions; **Accuracy** – the correctness of the medical information provided; **Context Awareness** – the ability to leverage prior dialogue turns and patient-specific details; **Completeness** – the extent to which the response fully addresses the clinical query. These five dimensions collectively form the core evaluation metrics of HealthBench, enabling a multi-faceted assessment of a model’s medical reasoning and communication skills.

**Dataset distribution.** We curated a subset of HealthBench specifically to evaluate TheraAgent. The subset comes from four medical departments: Endocrinology, Gastroenterology, Neurology and Respiratory. Figure 6 reports the department distribution. Figure 7 further partitions each department by the seven themes and reports the number of cases for each theme under each department. We exclude cases with unrelated theme of health data task in our all experiments.

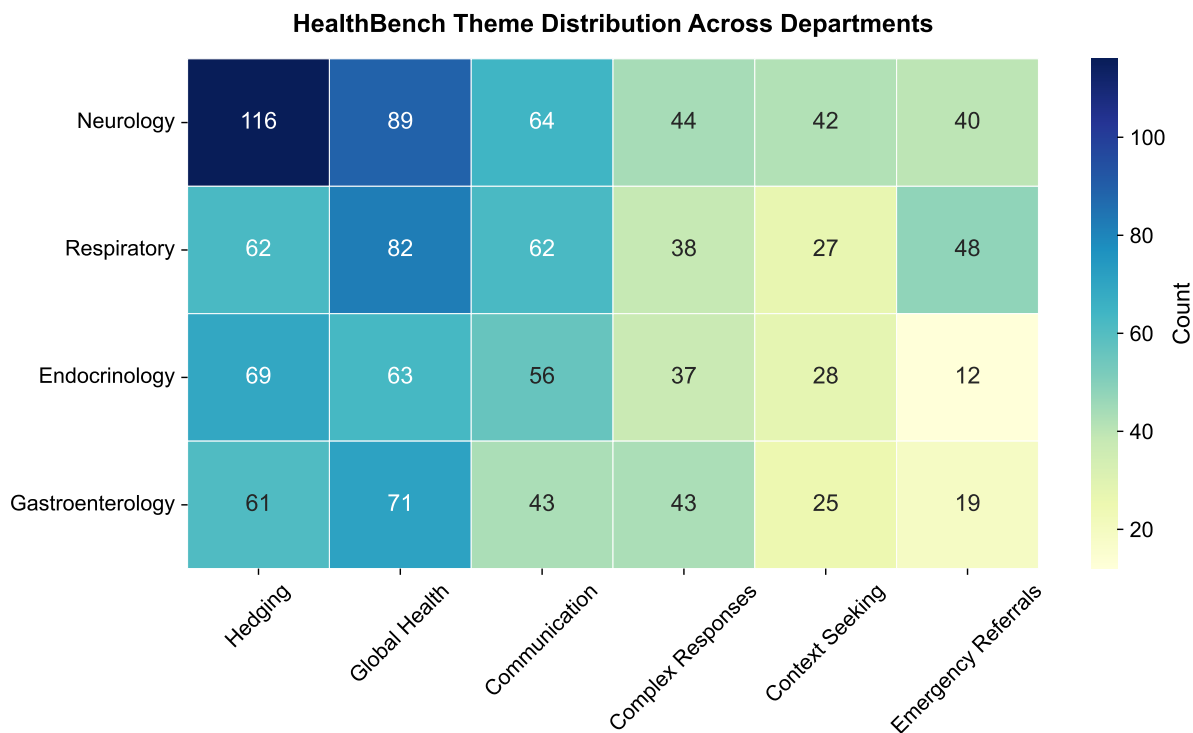


Figure 7: Theme distribution of the HealthBench Dataset.

## B.2 Real-World Case Dataset

**Dataset background.** Real-World Case Dataset are collected from Chinese Medical Case Repository, a public platform designed to encourage physicians to document their diagnostic and therapeutic experiences in the form of standardized case reports. By promoting case sharing, Chinese Medical Case Repository aims to enhance the overall quality of diagnosis and treatment within and across medical departments.

**Dataset format.** Each case in the Real-World Case Dataset typically follows a structured format consisting of six key components: (1) patient information, (2) examination results, (3) diagnosis and differential diagnosis, (4) treatment plan, (5) therapeutic outcome, and (6) discussion. This structure provides a comprehensive representation of real-world clinical reasoning and decision-making processes, making it highly suitable for evaluating treatment plan generation.

**Dataset distribution.** Figure 9.b summarizes the gender distribution of patients. Among 35 patients, 51.4% are male and 48.6% are female. The Real-World Case dataset encompasses a wide range of diseases across departments. The disease distribution is summarized in Figure 8.a, showing more than 20 unique diseases. Figure 8.c reports the

number of unique diseases for each department. We also report the age distribution of patients by gender in Figure 9.a. Additionally, we report age distribution across departments in Figure 9.c.

We also performed analysis of case complexity, quantified by the lengths of both the symptom description and the treatment text, and report the results in Figure 8.b. As shown in the scatter plot, Most cases have less than 100 characters in symptom descriptions and 400 characters in treatment plans. However, treatment plan generation is an open-ended space search problem with no limit on answer length. In reality, the plan can be as long as 1,600 characters.

## C Medical Expert Annotation

### C.1 Annotator Recruitment and Ethical Approval

Licensed physicians were recruited through word-of-mouth invitations within professional and academic networks. Participation was voluntary, and all annotators had prior clinical experience. Before starting the annotation tasks, annotators were presented with a study information page on the annotation website that explained the research objectives, the nature of the tasks, and the intended use of the collected data. Only after reviewing this informa-

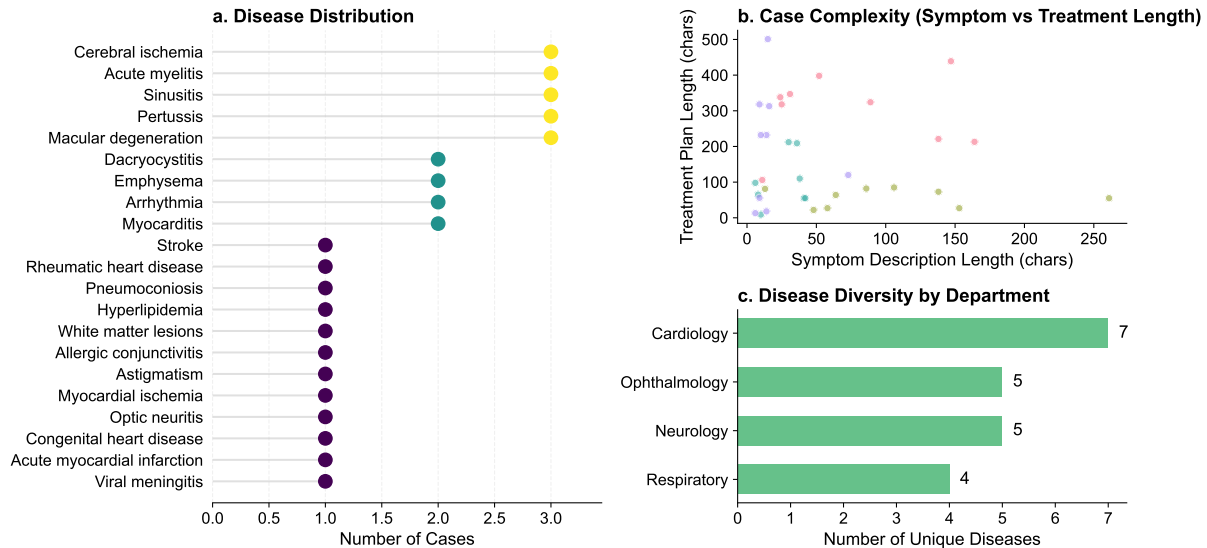


Figure 8: Disease distribution of the Real-World Case Dataset.

tion and explicitly indicating their agreement were annotators allowed to proceed with the annotation, which constituted informed consent. Participation was optional, and annotators could discontinue the process at any time.

The evaluated patient cases were collected from publicly accessible medical platforms or curated benchmark datasets, and no private or personally identifiable patient information was included. As all data were either publicly available or fully anonymized, and the study involved minimal risk to participants, the data collection protocol was determined to be exempt from formal institutional review board (IRB) approval, in accordance with standard ethical guidelines for human-subject annotation studies.

## C.2 Medical Judgement Dimensions

Understanding the challenges of evaluating treatment plans and the shortcomings of textual similarity-based metrics, Our proposed TheraJudge evaluates plans in seven clinically relevant dimensions that comprehensively captures the quality of a correct, safe, consensus compliant and actionable treatment plan (Figure 10). The seven dimensions include **Scientific Consensus Compliance, Plan Completeness, Information Accuracy, Rationale-Measure Coherence, Situation Targeting, Harm Control** and **Bias in Medical Content**. To guide human and LLM evaluators to assess the seven dimensions, we carefully designed one question for each dimension. All questions can be answered using a 5-point-scale rating or by

pairwise comparison. Specifically, the evaluator first rate individual plans on the seven questions by choosing integer scores from 1 to 5 inclusive. Afterwards, the evaluator is provided with two plans side-by-side. For each question, the evaluator chooses the one that better satisfies the given criterion.

## C.3 Annotation Interface

To assist physicians in rating and comparing treatment plans and their model evaluations, we developed a web application with a user-friendly interface, as shown in Figures 11 and 12. Three regimens are generated for each patient case - a pre-iteration regimen, a post-iteration regimen and an expert-written regimen. Through the interface, physicians rate each regimen based on the Therapeutic Regimen Quality rubrics using a 5-point scale (Figure 12). For clearer understanding of physicians' ratings, an open-ended question is provided, where physicians can explain in words any unreasonable aspects they identified in the provided regimen.

Beyond individual ratings, physicians also perform pairwise comparisons by selecting the better regimen, or indicating a tie, on each of the Therapeutic Regimen Quality axis. (Figure 11). Finally, physicians provide an overall ranking of the three regimens.

To investigate the potential of LLMs as evaluators, we perform meta-evaluation by asking physicians to evaluate model-generated evaluations. In these questions, physicians are shown both the regimen and its model evaluation rationale, and are asked to rate the evaluation rationale using the

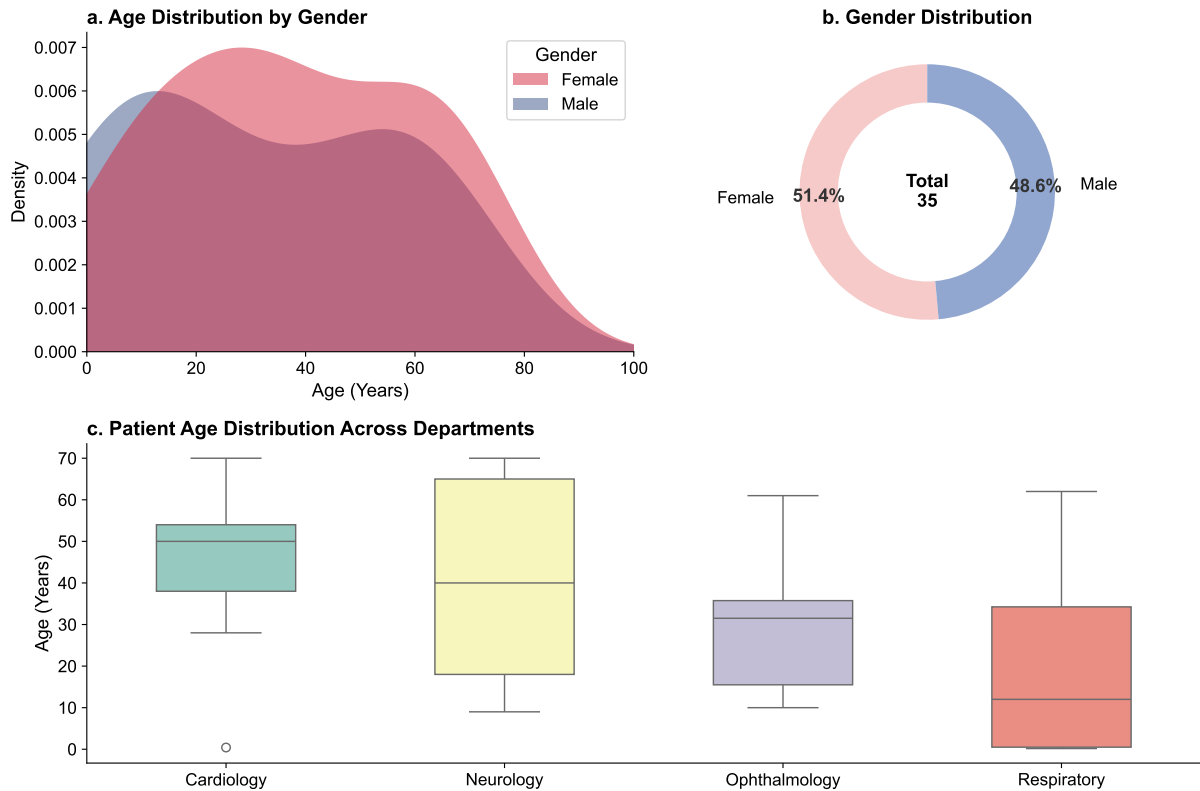


Figure 9: Demographic information of the Real-World Case Dataset.

Judgement Capability rubrics.

## D Experimental Details

### D.1 Baselines

Here we give all baselines including **medical-domain models** (TxAgent (Gao et al., 2025), FineMedLM-o1-8B(hongzhou yu et al., 2025), HuatuoGPT-o1-70B (Chen et al., 2025), UltraMedical-70B (Zhang et al., 2024), Llama3-Med42-70B (Christophe et al., 2024), MedCritical-7B (Su et al., 2025), Baichuan-M2-32B (Team et al., 2025a)), **open-source general models** (DeepSeek-R1 (DeepSeek-AI et al., 2025), Qwen3-235B-A22B (Yang et al., 2025a), Kimi-K2 (Team et al., 2025b)), and **proprietary models** (GPT-4o (OpenAI et al., 2024a), OpenAI o1 (OpenAI et al., 2024b), GPT-4.1 (OpenAI, 2025a), o4-mini (OpenAI, 2025b), Grok-3 (xAI, 2025), Claude-4-Sonnet (Anthropic, 2025), and Gemini-2.5-Pro (Comanici et al., 2025)).

### D.2 Experimental Setup for Judge Agreement with HealthBench

We evaluate the agreement between TheraJudge and the HealthBench evaluation on the neurology subset containing ideal\_completion anno-

tations. This subset consists of 219 cases, each paired with a reference ideal therapeutic response curated in HealthBench. For each case, we collect five generated treatment plans produced by diverse models: Baichuan-M2-32B, HuatuoGPT-o1-70B, Llama3-Med42-70B, Claude-Sonnet-4, and Grok-3. These models are selected to represent a mixture of medical-specialized models and general-purpose large language models from different model families, ensuring diversity in generation style and reasoning behavior.

Traditional lexical metrics (BLEU and ROUGE) are computed by directly comparing each model-generated output against the corresponding ideal\_completion. In contrast, both LLM-based scoring and TheraJudge evaluations use the ideal\_completion as a reference. The reported Spearman, Pearson, and concordance correlation coefficient (CCC) values correspond to the median correlation scores aggregated over all 219 neurology cases. We focus on neurology cases as they contain a substantial number of high-quality ideal\_completion annotations and involve complex, multi-step clinical reasoning, making them well suited for evaluating evaluator reliability.

Axis	Question	Scale	Assessed by
<b>Therapeutic Regimen Quality</b>			
Scientific Consensus Compliance	To what extent is the treatment plan consistent with established scientific and clinical consensus?	5-point scale & comparison	Medical Expert
Plan Completeness	To what extent does the plan comprehensively address all necessary components without omission?	5-point scale & comparison	Medical Expert
Information Accuracy	To what extent does the plan contain inaccurate information?	5-point scale & comparison	Medical Expert
Rationale-Measure Coherence	To what extent is the reasoning behind the treatment plan logically connected to the proposed measures?	5-point scale & comparison	Medical Expert
Situation Targeting	To what extent does the plan accurately reflect and address the patient's specific condition?	5-point scale & comparison	Medical Expert
Harm Control	What is the extent and likelihood of potential harm to the patient?	5-point scale & comparison	Medical Expert
Bias in Medical Content	To what extent does the plan exhibit bias or inapplicability to specific patient demographics?	5-point scale & comparison	Medical Expert

Figure 10: Medical Judgement Dimensions.

## E Supplementary Results

### E.1 HealthBench results

Figure 14 shows the detailed scores of TheraAgent and all baseline models on HealthBench. We report the scores on themes and axes of **Global Health, Hedging, Context Seeking, Communication, Accuracy, Completeness and Context Awareness**, as well as the overall scores. From the results, TheraAgent consistently performs well regardless of departments. It achieves the best overall score in all but the Neurology department, where it scores 1.5 points lower than the best-performing model. Furthermore, TheraAgent also shows strong performance in multiple dimensions, especially on Completeness, surpassing every model in every department. These results highlight the outstanding capability of TheraAgent in ensuring completeness and avoiding critical omissions in its treatment plans, which can be attributed to its iterative refinement ability that allows continuous addition of missing information throughout the generation process.

### E.2 Rating annotations

Figure 13 presents the proportion of high-quality ratings ( $\geq 4$ ) across seven clinical dimensions on real-world cases. Overall, TheraAgent consistently achieves the highest scores across nearly all dimensions, indicating superior performance in generating clinically robust treatment plans. The gains are

Method	Overall Score
TheraAgent w/o Memory	0.4115
- with all Memory	0.4859
- with nearest three Memory	0.5002
- with best three Memory	<b>0.5236</b>

Table 4: Ablation study of Memory settings on HealthBench

most pronounced in *Targeting, Completeness*, and *Consensus*, suggesting that iterative refinement enables TheraAgent to produce more patient-specific, comprehensive, and guideline-aligned treatment plans compared to both the base model and human physicians. In contrast, DeepSeek-R1 exhibits relatively weaker performance, particularly in dimensions related to treatment precision and safety, reflecting the limitations of one-shot generation. While physician-authored regimens demonstrate strong accuracy and clinical reasoning, their lower scores in completeness and consensus adherence highlight the pragmatic abbreviations common in real-world documentation. These results further support that TheraAgent effectively balances clinical rigor, safety, and completeness, yielding more consistently high-quality therapeutic outputs.

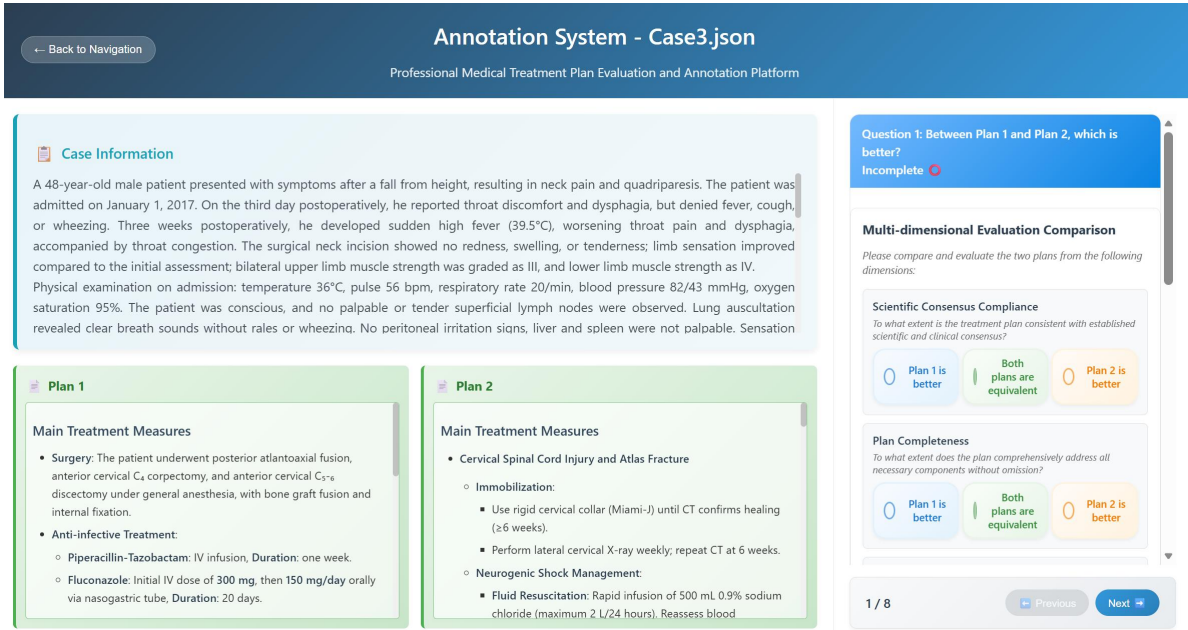


Figure 11: Comparison questions of the annotation interface.

Method	Average Score
TheraAgent w/o RAG	89.42
- inference with RAG	90.05
- inference and judge with RAG	89.67
- judge with RAG	<b>92.73</b>

Table 5: Ablation study of RAG settings on the Real-World Case Dataset

ing RAG in any component boosts performance, demonstrating that RAG effectively aligns treatment plans with medical consensus and enables models to generate more clinically-sound recommendations. Notably, using RAG in judging only yields the highest score (92.73), while using RAG in inference gives only marginal score increase, regardless of whether it is also used in judging.

### E.3 Ablation Study

Table 4 examines the effect of Dynamic Memory on HealthBench. Removing memory reduces TheraAgent to a non-iterative baseline and results in the lowest score (0.4115). Incorporating memory consistently improves performance, with selective retrieval outperforming full-memory usage. Retrieving the top three highest-scoring memory items achieves the best result (0.5236), demonstrating that score-aware and compact memory selection provides more effective guidance for iterative refinement than using all or nearest memories.

Table 5 investigates the effect of RAG on TheraAgent performance on Real-World Case datasets. Since both the Planner and TheraJudge can incorporate RAG, we evaluated all possible combinations where RAG is not used, used only in inference, used only in judging and used in both inference and judging. Excluding RAG completely results in the lowest score (89.42), whereas incorporat-

### E.4 Case Study

We perform case study on a data sample arbitrarily chosen from the Real-World Case Dataset. Table 6 summarizes the given patient case information and treatment plans written by humans, DeepSeek-R1 and TheraAgent. The patient information includes symptom descriptions, medical history, test results, the final diagnosis and the diagnostic rationale. The treatment plan written by a human physician is retrieved directly from the Real-World Case dataset. The plan includes 3 sections. The Primary Treatment Section details the recommended medication and therapy. The Monitoring and Supportive Care Section lists any further assessments, long-term monitoring and medical support. The Rationale Section provides reasons for the recommended treatment. For a fair and structured comparison, DeepSeek-R1 and TheraAgent are prompted to generate a treatment plan with the same 3-section format.

← Back to Navigation
Annotation System - Case3.json

Professional Medical Treatment Plan Evaluation and Annotation Platform

**Case Information**

A 48-year-old male patient presented with symptoms after a fall from height, resulting in neck pain and quadriparesis. The patient was admitted on January 1, 2017. On the third day postoperatively, he reported throat discomfort and dysphagia, but denied fever, cough, or wheezing. Three weeks postoperatively, he developed sudden high fever (39.5°C), worsening throat pain and dysphagia, accompanied by throat congestion. The surgical neck incision showed no redness, swelling, or tenderness; limb sensation improved compared to the initial assessment; bilateral upper limb muscle strength was graded as III, and lower limb muscle strength as IV. Physical examination on admission: temperature 36°C, pulse 56 bpm, respiratory rate 20/min, blood pressure 82/43 mmHg, oxygen saturation 95%. The patient was conscious, and no palpable or tender superficial lymph nodes were observed. Lung auscultation revealed clear breath sounds without rales or wheezing. No peritoneal irritation signs, liver and spleen were not palpable. Sensation

**Treatment Plan**

**Main Treatment Measures**

1. Cervical Spinal Cord Injury and Atlas Fracture
  1. Immobilization: Maintain cervical collar until fracture stability is confirmed.
  2. Neurological Rehabilitation: Daily physical therapy and occupational therapy, focusing on limb strength enhancement, mobility, and adaptive techniques.
2. Esophageal Candidiasis
  1. Fluconazole: Day 1: 400 mg IV or oral, then 200 mg daily for 14-21 days.
3. Infection and Fever Management

Question 3: Please rate the following treatment plan based on the case information (5-point scale).  
Incomplete

**Multi-dimensional Rating (5-point scale)**

Please rate the treatment plan from the following dimensions:

**Scientific Consensus Compliance**  
To what extent is the treatment plan consistent with established scientific and clinical consensus?

1 point  
  2 point  
  3 point  
  4 point  
  5 point

Please select a rating for this dimension (1-5 point)

**Plan Completeness**  
To what extent does the plan comprehensively address all necessary components without omission?

1 point  
  2 point  
  3 point  
  4 point  
  5 point

Please select a rating for this dimension (1-5 point)

3 / 8      Previous   Next

Figure 12: Rating questions of the annotation interface.

## F Prompts

1207

### F.1 Planner

1208

We present the prompt template for the Planner component in Table 7.

1209

1210

### F.2 TheraJudge

1211

We present the prompt template for the TheraJudge component in Table 8. Three boxed paragraphs, the **RAG context**, the **few-shot guidance** and the **multi-dimensional judging** are optional and only included when the respective functions are enabled.

1212

1213

1214

1215

1216

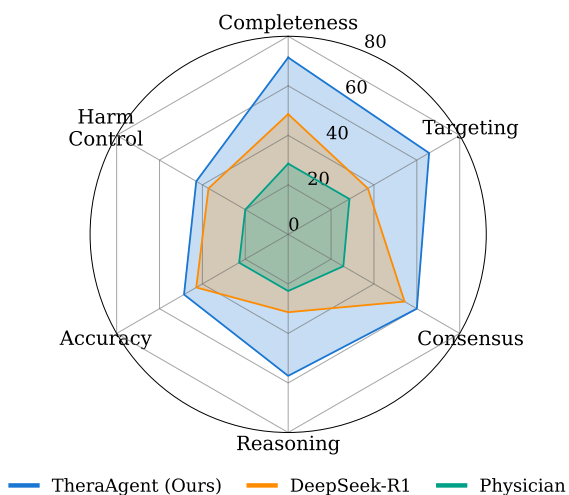


Figure 13: **Comparison of high-quality rating proportions across clinical dimensions.** Data represents the percentage of expert ratings  $\geq 4$  (on a 5-point scale) for all real-world medical cases.

## Performance Across Departments and Models on HealthBench

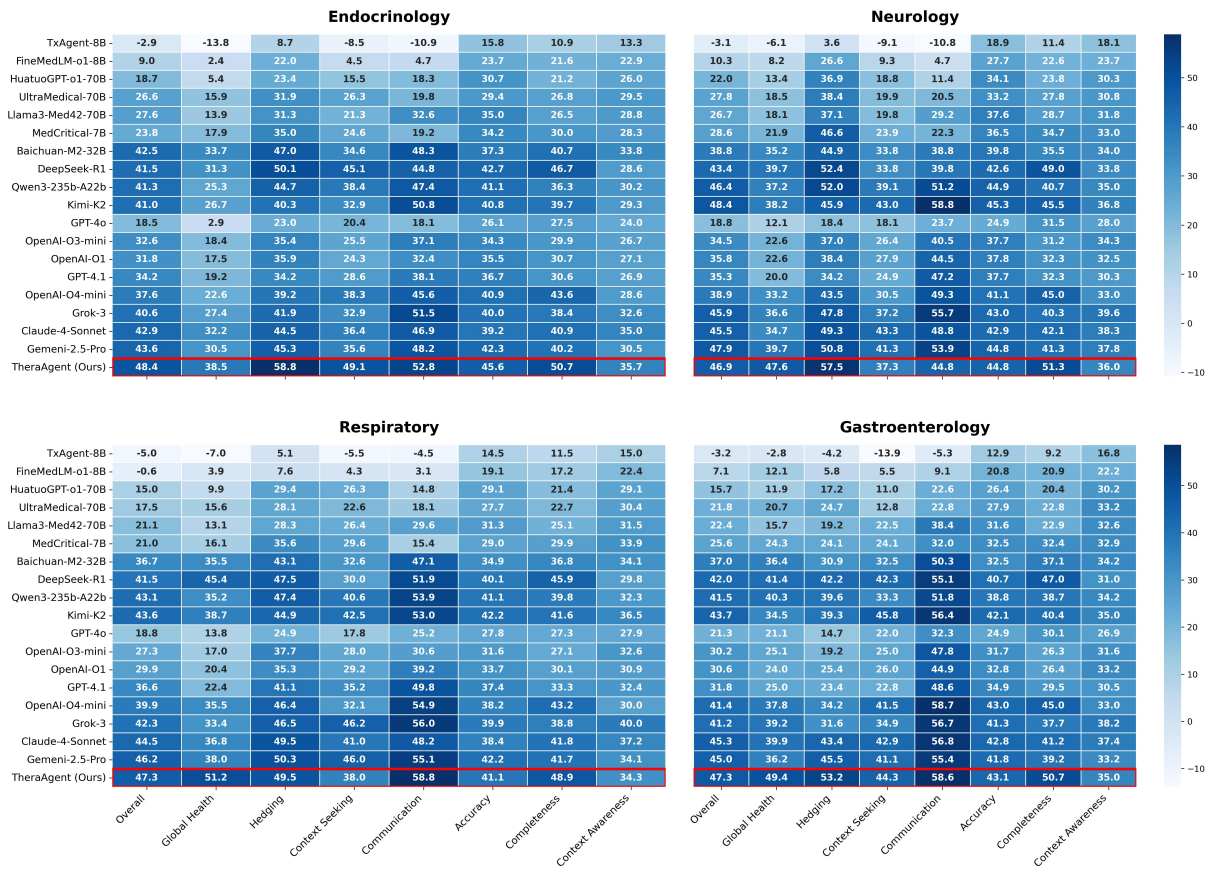


Figure 14: All results on HealthBench across four departments.

## Case Study: CPFE Treatment Planning

### Patient Profile.

**Age/Sex:** 70-year-old male.

**Symptoms:** Chronic cough with sputum >4 years, worsened in last 10 days.

**PMH:** CAD (PCI with 1 stent, 2 years ago); Type 2 diabetes (2 years); Smoking 30 years (5 cig/day, quit 2 years); No occupational/environmental exposure.

**FH:** Non-contributory.

**PE:** T 36.5 °C; P 78/min; RR 18/min; BP 142/70 mmHg; Alert; No JVD, cyanosis, rash, clubbing, edema; Lungs: coarse breath sounds + bibasilar crackles; Heart/abdomen normal.

**Labs:** Hb 120 g/L; WBC 5.6710<sup>9</sup>/L; Neu 0.633; Eos 0.1710<sup>9</sup>/L; ESR 24 mm/h; CRP 1.6 mg/L; PCT 0.062 µg/L; D-dimer 2.54 µg/L; CEA 7.593 µg/L; SCC 4.2 µg/L; G/GM negative; Sputum bacterial/fungal/TB tests normal.

**PFT (Feb 26, 2018):** Mild restrictive defect; Moderate ↓DLCO; ↓lung volumes and RV; Normal RV/TLC. **LCQ:** Total 17.77 (Physical 4.63, Psychological 6.14, Social 7.00).

**Diagnosis:** CPFE.

**Basis:** PFT showing mild restriction + moderately reduced DLCO + reduced lung volumes with normal RV/TLC; LCQ indicating significant chronic cough impact.

### Human's Treatment Plan.

1. **Primary Treatment:** Pirfenidone 0.4 g three times daily, with gradual dose escalation for long-term therapy. N-acetylcysteine (NAC) 0.6 g three times daily as combination antifibrotic therapy.
2. **Monitoring and Supportive Care:** (1) Regular monitoring of liver and renal function; periodic assessment of pulmonary function and disease progression; close observation for drug-related adverse effects. (2) Provide oxygen therapy and respiratory rehabilitation guidance; ensure adequate nutritional support to maintain physical strength; deliver ongoing health education and medication adherence counseling.
3. **Rationale:** Pirfenidone is an antifibrotic agent that slows the progression of idiopathic pulmonary fibrosis. NAC has antioxidant and antifibrotic properties; combination treatment may enhance therapeutic efficacy. Combined antifibrotic therapy helps improve prognosis and delay disease progression.

### DeepSeek-R1's Treatment Plan.

1. **Primary Treatment Measures:** (1) Antifibrotic therapy; Pirfenidone: Start at 200 mg orally three times daily, titrate to a target of 600 mg three times daily. Mucolytic agent: **N-acetylcysteine 600 mg orally once daily.** (2) Management of acute exacerbations: Empiric antibiotic therapy: Amoxicillin–clavulanate 625 mg orally three times daily for 5–7 days. (3) Bronchodilators: LAMA: Tiotropium 18 µg inhaled once daily. (4) Antitussive therapy: First-line: Dextromethorphan 10–20 mg orally 3–4 times daily. **Second-line: Low-dose morphine, starting at 2.5 mg orally twice daily.**
2. **Monitoring and Supportive Care:** (1) Evaluate for heart failure, rule out pulmonary embolism, and assess oxygen therapy needs. (2) Antifibrotic therapy: Monitor liver function monthly. Repeat pulmonary function tests (FVC, DLCO) after 3 months. (3) Emergency signs: Worsening cough, new-onset dyspnea, or fever—seek immediate medical care.
3. **Rationale for Treatment:** (1) Antifibrotic agents slow lung function decline. (2) Empiric antibiotics target common lower-respiratory pathogens. (3) Bronchodilators improve symptoms related to emphysema and enhance ventilation. (4) Risks include gastrointestinal side effects of antifibrotics and morphine-related respiratory depression.

### TheraAgent's Treatment Plan.

1. **Major Therapeutic Measures:** (1) **Antifibrotic Therapy Decision-Making: When HRCT demonstrates a definite UIP pattern, conduct a multidisciplinary team (MDT) discussion to assess risks and benefits.** Pirfenidone: Start at 200 mg orally three times daily, titrate over 2 weeks to 600 mg three times daily. If HRCT shows a non-UIP pattern or indeterminate pattern, do not initiate antifibrotic therapy. Antioxidant and Mucolytic Therapy: **N-acetylcysteine 600 mg orally three times daily (total 1800 mg/day).** (2) Management of Acute Exacerbation: Urgent Imaging: High-resolution chest CT (HRCT) within 24 hours. Infection Management: No empirical antibiotics. Monitor temperature and sputum daily. If purulent sputum or fever >38 °C occurs, start targeted narrow-spectrum antibiotics. (3) Bronchodilators: When HRCT confirms emphysema: First-line: Tiotropium 18 µg inhaled once daily. (4) Cough Suppression: First-line: ... Second-line: .... **Third-line: Morphine 1 mg orally twice daily, only when no risk of respiratory depression.**
2. **Monitoring and Supportive Care:** (1) Assess for heart failure, exclude pulmonary embolism, evaluate oxygen needs, perform arterial blood gas analysis, and screen for lung cancer. (2) Monitor liver function (ALT/AST) monthly during antifibrotic therapy. Monitor renal function (eGFR) regularly during N-acetylcysteine treatment. (3) End-Stage Planning: ...
3. **Rationale / Evidence:** (1) HRCT is central to CPFE diagnosis and therapeutic decision-making. (2) High risk of heart failure and pulmonary embolism warrants early identification. (3) Avoid antibiotics without evidence of infection. (4) Antifibrotic therapy is restricted to HRCT-confirmed UIP patterns. (5) Bronchodilators are used only when HRCT confirms emphysema. (6) High-dose N-acetylcysteine is recommended in IPF guidelines. (7) Morphine is reserved for third-line use with strict monitoring. (8) The protocol adheres to ATS/ERS guidelines with dynamic adjustment.

Table 6: Detailed comparison between different methods.

### Prompt Template for Planner

```
## Patient Case Details:
{query}

### Old treatment plan 1:
{treatment_plan}

### Reflection to the old treatment plan 1:
{experience}

### Score of the old treatment plan 1:
{score}

### Old treatment plan 2:
{treatment_plan}

### Reflection to the old treatment plan 2:
{experience}

### Score of the old treatment plan 2:
{score}

...

## Task:
You are an expert in {department}. Please think step by step to give a
treatment plan for the patient accurately based on the above information.

## Output format:
<thinking>The reasoning process</thinking>
<answer>The treatment plan</answer>
```

Table 7: The prompt template for the Planner component.

## Prompt Template for TheraJudge

### RAG context (if RAG is enabled)

```
### RAG Context:  
{guideline 1}  
  
{guideline 2}  
  
...
```

### Few-shot guidance (if Few-shot is enabled)

```
## Example 1:  
### Example 1 Case Details:  
{query}  
  
### Example 1 Treatment Plan:  
{treatment_plan}  
  
### Example 1 Score:  
{score}  
  
...
```

### Multi-dimensional judging (if Dimension is enabled)

**Please evaluate the treatment plan from the following seven dimensions and give a score from 0 to 100:**

1. Scientific Consensus Compliance (To what extent is the treatment plan consistent with established scientific and clinical consensus?)
2. Plan Completeness (To what extent does the plan comprehensively address all necessary components without omission?)
3. Situation Targeting (To what extent does the plan accurately reflect and address the patient's specific condition?)
4. Rationale-Measure Coherence (To what extent is the reasoning behind the treatment plan logically connected to the proposed measures?)
5. Harm Potential (What is the extent and likelihood of potential harm to the patient?)
6. Information Accuracy & Relevance (To what extent does the plan contain inaccurate or irrelevant information?)
7. Bias in Medical Content (To what extent does the plan exhibit bias or inapplicability to specific patient demographics?)

```
### Patient Case Details:  
{query}
```

```
### Treatment Plan to Evaluate:  
{treatment_plan}
```

**Please answer using the following format:**

```
<reason>[detailed explanation]</reason>  
<dimension_scores>[all dimension scores from 0 to 100]</dimension_scores>  
<overall_score>[overall score number from 0 to 100]</overall_score>
```

Table 8: The prompt template for the TheraJudge component.