MedGUIDE: Benchmarking Clinical Decision-Making in Large Language Models

Xiaomin Li[*]
Harvard University

Mingye Gao
MIT

Yuexing Hao
Cornell University
MIT
Mayo Clinic

Taoran Li
UIUC

Guangya Wan
University of Virginia

Zihan Wang
Harvard Medical School
Abaka AI

Yijun Wang
Harvard University

Xupeng Chen
NYU

Abstract

Clinical guidelines, typically structured as decision trees, are central to evidence-based medical practice and critical for ensuring safe and accurate diagnostic decision-making. However, it remains unclear whether Large Language Models (LLMs) can reliably follow such structured protocols. In this work, we introduce **MedGUIDE**, a new benchmark for evaluating LLMs on their ability to make guideline-consistent clinical decisions. MedGUIDE is constructed from 55 curated NCCN decision trees across 17 cancer types and uses clinical scenarios generated by LLMs to create a large pool of multiple-choice diagnostic questions. We apply a two-stage quality selection process, combining expert-labeled reward models and LLM-as-a-judge ensembles across ten clinical and linguistic criteria, to select 7,747 high-quality samples. We evaluate 25 LLMs spanning general-purpose, open-source, and medically specialized models, and find that even domain-specific LLMs often underperform on tasks requiring structured guideline adherence. We also test whether performance can be improved via in-context guideline inclusion or continued pretraining. Our findings underscore the importance of MedGUIDE in assessing whether LLMs can operate safely within the procedural frameworks expected in real-world clinical settings.

1 Introduction

Clinical guidelines play a critical role in modern medicine, providing structured recommendations for diagnosis and treatment that are informed by expert consensus and current evidence. These guidelines, such as those published by the National Comprehensive Cancer Network (NCCN) or the American Heart Association (AHA), are often encoded as decision trees, offering standardized pathways for handling diverse patient scenarios [National Comprehensive Cancer Network, 2023, Arnett et al., 2019]. Adherence to these protocols is critical for ensuring consistent, safe, and high-quality clinical decision-making.

[*]Co-first authors: Xiaomin Li and Mingye Gao. Correspondence to Xiaomin Li (xiaominli@g.harvard.edu).

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

Large Language Models (LLMs) have demonstrated remarkable capabilities across general and medical natural language processing (NLP) tasks, including biomedical question-answering (QA) [Lee et al., 2020, Alsentzer et al., 2019, Shin et al., 2020], clinical dialogue modeling [Sun, 2024, He et al., 2024], and even diagnostic reasoning [Singhal et al., 2025, Yang et al., 2022, Wu et al., 2023]. However, most evaluations to date focus on factual recall or general in-context reasoning, rather than testing whether LLMs can follow domain-specific decision rules as clinicians must in practice. Emerging work has begun to explore instruction and rule-following behavior in LLMs [Zheng et al., 2023, Dong et al., 2024, Chefer et al., 2024], yet few studies directly assess LLM adherence to formal clinical pathways, particularly in complex diagnostic contexts [Fast et al., 2024, Huang et al., 2024b].

To fill this research gap, we introduce **MedGUIDE—G**uideline **U**nderstanding and **I**nference for **D**ecision **E**valuation—a benchmark designed to evaluate whether LLMs can make diagnostic decisions in accordance with established medical guidelines. MedGUIDE is constructed from 55 decision trees curated from NCCN oncology protocols, covering 17 of the most common cancer types. We transform these trees into clinical vignettes and corresponding multiple-choice questions (MCQs) that require selecting the correct next step in a patient's management plan. Rather than testing general knowledge alone, MedGUIDE probes whether models can apply structured clinical logic. We implement a rigorous and efficient two-stage filtering process using both human-annotated reward models [Glaese et al., 2022, Wang et al., 2024c] and LLM-as-a-judge ensembles [Chen et al., 2024a, Huang et al., 2024a, Polo et al., 2024] to ensure the final dataset is both clinically plausible and textually well-formed. From a raw pool of 16,000 QA samples, we retain 7,747 high-quality examples.

We evaluate 25 LLMs, including general-purpose, open-source, and medically fine-tuned models, using both standard accuracy and a weighted accuracy metric that accounts for question difficulty. To better understand the LLM's capabilities tested by MedGUIDE, we compare model performance on MedGUIDE against other established benchmarks, including IFEval [Dong et al., 2024] and MMLU-Professional Medicine [Hendrycks et al., 2020], and analyze cross-benchmark correlations. In addition, we explore two methods for further improving the model's capability of adhering to the clinical guideline: (1) *guideline-in-context prompting*, which supplies the model with the relevant decision tree during inference; and (2) *continued pretraining on guidelines*, which aims to internalize the structure and logic of clinical pathways and apply them to diagnostic or treatment planning tasks in MedGUIDE.

Our key contributions are:

- We introduce **MedGUIDE**, the first benchmark focused on evaluating LLMs' ability to follow structured clinical decision trees based on real-world medical guidelines. [2]
- We construct a high-quality multiple-choice questions (MCQs) dataset using 55 NCCN decision trees, and apply a dual-stage-filtering pipeline combining expert-labeled reward models and ensemble LLM scoring.
- We benchmark 25 diverse LLMs and reveal significant limitations in both general and medical models' ability to align with guideline-based decision logic.
- We analyze correlations with existing benchmarks and show that MedGUIDE evaluates capabilities beyond factual recall, including structured guideline comprehension and task-specific adherence.
- We evaluate whether guideline grounding, via in-context prompting or continued pertaining, improves LLM performance on MedGUIDE.

2 Related Work

Medical LLMs for Clinical Diagnosis. The application of Large Language Models (LLMs) in healthcare has progressed from early biomedical pretraining [Lee et al., 2020, Alsentzer et al., 2019, Shin et al., 2020, Wang et al., 2023] to instruction-tuned and dialogue-optimized systems [Singhal et al., 2025, Yang et al., 2022, Wu et al., 2023, Venigalla et al., 2022] that support medical reasoning tasks. Recent efforts further expand capabilities via: (1) retrieval-augmented generation using medical corpora [Wen et al., 2023, Wu et al., 2024a, Shi et al., 2023, Ranjit et al., 2023, Ge et al., 2023],

[2] Code and data: `https://anonymous.4open.science/r/Submission-MedGUIDE-187A`.

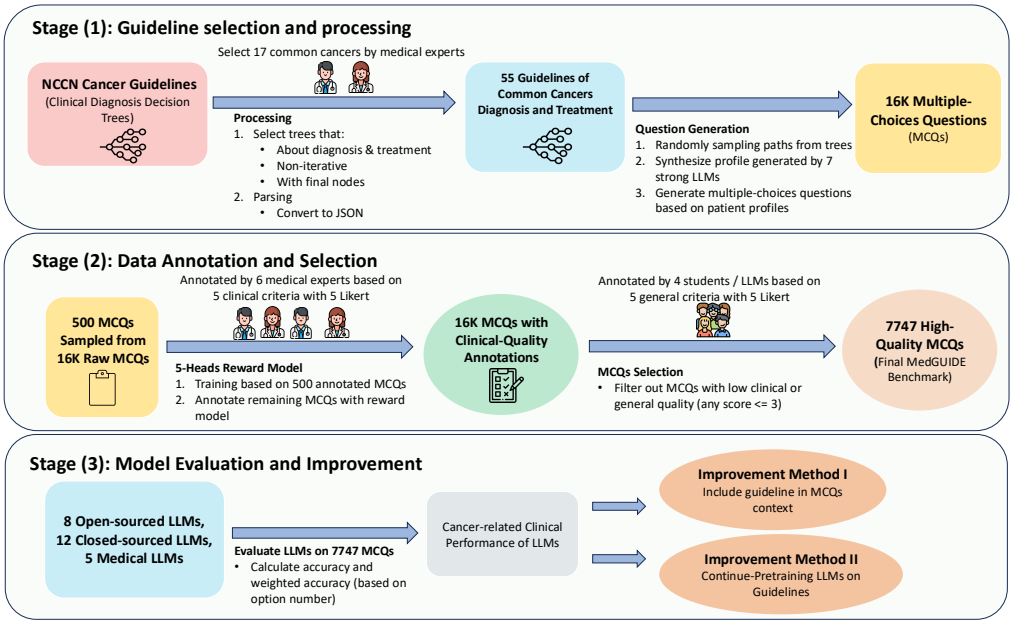


Figure 1: Overview of the MedGUIDE Benchmark Pipeline. **Stage (1)**: selecting and processing 55 NCCN clinical decision tree guidelines for 17 common cancers to generate 16K synthetic MCQs. **Stage (2)**: annotating and filtering these MCQs based on clinical and general quality criteria using expert-labeled data and a 5-head reward model, resulting in a curated set of 7,747 high-quality MCQs. **Stage (3)**: evaluating 25 LLMs (open-source, closed-source, and medical) on the benchmark and applying two improvement methods—guideline-aware prompting (Method I) and guideline-based continued pretraining (Method II).

databases [Shi et al., 2023, Rau et al., 2023, Wang et al., 2024a, Chen et al., 2025a], and knowledge graphs [Wen et al., 2023, Wu et al., 2024a, Gao et al., 2023, Zhu et al., 2024]; (2) supervised fine-tuning on clinical notes, dialogues, and multimodal inputs [Toma et al., 2023, Jiang et al., 2023b, Sun, 2024, Wu et al., 2024b]; and (3) RLHF strategies, both online [Wang et al., 2023, Zhang et al., 2023, Zhou et al., 2024] and offline [Dou et al., 2024, Yang et al., 2024b].

Medical Benchmarks for LLMs. LLM benchmarks in medicine span from knowledge-focused tasks [Jin et al., 2019, Pal et al., 2022] to more complex clinical reasoning and scenario-based evaluations [Xue et al., 2024, Hosseini et al., 2024, Liu et al., 2024c, Kratzwald et al., 2021]. A growing body of work emphasizes guideline adherence [Zhang et al., 2025, Wu et al., 2025], yet few benchmarks systematically test whether LLMs can follow structured protocols such as those from NCCN [National Comprehensive Cancer Network, 2023] or AHA [Arnett et al., 2019]. Prior work addressing guideline-following remains limited in scope [Fast et al., 2024], motivating our MedGUIDE benchmark, which uniquely evaluates both diagnostic accuracy and fidelity to established clinical pathways.

Clinical Decision-Making Benchmarks Recent work has developed benchmarks for clinical decision-making in LLMs. CliBench Chen et al. [2024c] evaluates broad clinical reasoning across diverse scenarios, while ClinicalBench Liu et al. [2024b] compares LLMs against traditional ML models in clinical prediction tasks. CLIMB Chen et al. [2024b] focuses on clinical bias evaluation. Unlike these approaches that emphasize general clinical reasoning or prediction accuracy, MedGUIDE specifically targets adherence to structured clinical guidelines—a critical capability for real-world deployment. Our benchmark uniquely derives from actual NCCN decision trees and tests step-by-step guideline following, essential for clinical safety and regulatory compliance.

Instruction and Guidance Following. Instruction following is a fundamental competency of LLMs [IBM, 2023], while guidance following requires models to adhere to structured, domain-specific rules [Organization, 2024], such as clinical practice guidelines [Hager et al., 2024]. Evaluating guideline adherence remains challenging, often relying on proxy tasks or human review. Recent

benchmarks study rule-following under varying task structures [Chen et al., 2024a, Huang et al., 2024a, Weyssow et al., 2024, Polo et al., 2024, Zhang et al., 2024, Chefer et al., 2024, Ram et al., 2023]. Our work extends this line by examining how well LLMs follow multistep clinical decision trees under real-world constraints.

Reward Models and LLM-as-a-Judge. Prior work has employed multi-head reward models to score generated outputs along several attributes during post-training stages [Glaese et al., 2022, Wang et al., 2024c, Li et al., 2024b, 2025b, Wang et al., 2024b, Li et al., 2025a], an approach we adopt for our dataset's quality-based filtering. Complementing this, *LLM-as-a-judge* methods directly prompt a large model to rate sample quality, which has proven effective for large-scale evaluation and bias reduction [Zheng et al., 2023, Li et al., 2025b, Chen et al., 2024a, Huang et al., 2024a, Polo et al., 2024, Dong et al., 2024, Thakur et al., 2024].

3 MedGUIDE: A Guideline-Based Clinical Decision-Making Benchmark

3.1 NCCN Guidelines for Cancer

The National Comprehensive Cancer Network (NCCN) guidelines [National Comprehensive Cancer Network, 2023] are comprehensive, regularly updated protocols for cancer care—spanning prevention, diagnosis, treatment, and supportive care—and are organized as decision trees, with root nodes capturing key clinical variables (e.g., symptoms, labs, history) and branches splitting according to defined criteria or thresholds.

For example, Figure 2 illustrates a decision tree for first relapse in acute promyelocytic leukemia (APL), distinguishing among early relapse with or without arsenic trioxide exposure and late relapse. Because these pathways are designed for general clinical applicability rather than individualized care, they provide broad management strategies rather than patient-specific recommendations. The guidelines are published in PDF format behind a login portal, which limits direct ingestion by LLMs. To construct MedGUIDE, two medical experts selected 17 common cancers (listed in Table 2 in the Appendix) and their associated diagnostic decision trees. We excluded trees with loops or ambiguous leaf nodes, resulting in 55 well-structured guideline trees used for generating synthetic multiple-choice questions (MCQs).

3.2 Synthetic Prompt Generation

We convert each decision tree into JSON format and use GPT-4o to enumerate all valid paths from the root to each leaf. These paths are manually reviewed to ensure correctness. Each path corresponds to a plausible clinical scenario that matches the criteria described along its trajectory. For example, in the case illustrated in Figure 2, the patient follows the highlighted path because he experienced a **first relapse** five months after completing treatment with **ATRA and arsenic trioxide**—qualifying as an **early relapse (<6 mo)** after ATRA without anthracycline exposure. As a result, the recommended regimen is **arsenic trioxide** $\pm$ **ATRA** $\pm$ **gemtuzumab ozogamicin**, and if a **second morphologic remission** is achieved and the patient is **not a transplant candidate**, the guideline suggests **arsenic trioxide consolidation (total of 6 cycles)** as the next step.

Given a path (excluding the leaf), we prompt multiple LLMs to generate a clinical profile consistent with the scenario. We use the following models: `GPT-4o-mini` [OpenAI, 2024b], `GPT-4o` [OpenAI, 2024a], `GPT-4.1` [OpenAI, 2025a], `Claude-3.5-Haiku` [Anthropic, 2024], `Claude-3.7-Sonnet` [Anthropic, 2025], `DeepSeek-V3` [Liu et al., 2024a], `Gemini-2.5-Flash` [Google DeepMind, 2025], `Llama-3.2-1B-Instruct` [Meta AI, 2024b], `Llama-3.1-Instruct` (8B, 70B) [Meta AI, 2024a], `Qwen2.5-Instruct` (7B, 32B) [Yang et al., 2024a], `Mistral-7B-Instruct` [Jiang et al., 2023a], and `Mixtral-8x7B-Instruct` [Jiang et al., 2024]. This model diversity helps ensure variation in language and scenario framing.

Each profile is converted into a multiple-choice QA format by appending a question about the appropriate next clinical step. The answer options include all possible leaf nodes from the same decision tree, with the correct answer corresponding to the actual leaf node on the source path. This process yields a total of 16,000 multiple-choice QA pairs.

Example of QA Data

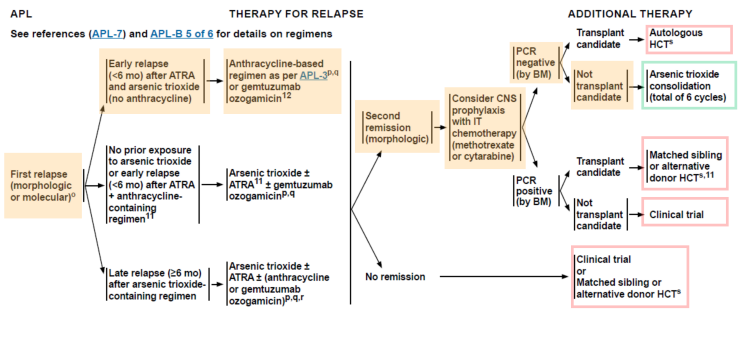


Figure 2: **NCCN Acute Myeloid Leukemia (AML) Guideline.** The orange boxes illustrate the workflow through which the sample QA dataset is generated. Red and green annotations represent the correct and incorrect options.

Prompt: A 32-year-old male patient with acute promyelocytic leukemia (APL) was initially diagnosed one year ago and achieved complete remission after treatment with all-trans retinoic acid (ATRA) and arsenic trioxide. He experienced his first relapse five months after completing these treatments, with both morphologic and molecular evidence of disease recurrence. He has no prior exposure to anthracycline therapy, making him eligible for a more aggressive treatment approach. Upon reevaluation, he received an anthracycline-based regimen according to the APL-3 protocol, resulting in a second morphologic remission confirmed by bone marrow biopsy. Given his potential central nervous system involvement and current PCR negativity in bone marrow, the team is contemplating prophylactic measures. He is not a transplant candidate due to comorbidities and overall health status. As consolidation therapy, the plan is to initiate arsenic trioxide for a total of six cycles. Given this clinical scenario, what would be the appropriate next step in management for consolidation therapy?

Options:

(A) Clinical trial

(B) Arsenic trioxide consolidation

(C) Clinical trial or matched sibling/alternative donor HCT

(D) Matched sibling or alternative donor HCT

(E) Autologous HCT

Correct Answer: (B) Arsenic trioxide consolidation.

3.3 Quality-Based Selection via Reward Models

3.3.1 Clinical and General Criteria

To ensure the high quality of our MCQs data, we adopt a quality-based data selection framework. In collaboration with six medical experts, we define five clinically grounded evaluation criteria; meanwhile, we introduce five general quality criteria to evaluate language and structure of MCQs (Table 1). Each criterion is rated on a 5-point Likert scale. Detailed rubrics are provided in Appendix D.

Table 1: Clinical and General Evaluation Criteria

Clinical Evaluation Criteria	General Quality Criteria
Clinical Plausibility	Clarity and Detail Level
Clinical Utility	Consistency and Internal Logic
Quality of Decision Path	Safety and Toxicity
Alignment to Decision Path	Textual Quality and Professionalism
Clinical Accuracy of Correct Answer	Option Distinctiveness

3.3.2 Training Reward Models

Due to time and resource constraints, we propose an efficient data-annotation pipeline. Six medical experts from Harvard Medical School, MIT Biology, and Harvard Stem Cell Institute (detailed demographic information provided in Appendix B) first labeled a randomly-selected subset of 500 MCQs according to the 5 clinical evaluation criteria shown in Table 1; each MCQ is assigned a single score for each criterion, resulting in a vector $[R_1(x), R_2(x), \ldots, R_5(x)] \in \{1, 2, \ldots, 5\}^5$. We then trained a 5-head reward model using `Qwen2.5-7B-Instruct` as the backbone, with 5 epochs and a learning rate of 2×10^{-5} (hyperparameter tuning details in Appendix E). The trained model was then used to rate the remaining 15,500 MCQs.

For general quality, we used an ensemble of LLM-as-a-judge models. Specifically, we queried `GPT-4o-mini`, `Claude 3.5-Haiku`, `Gemini 2.5-Flash`, and `DeepSeek-V3`, and averaged their scores across each criterion. The ensemble methods help reduce rating bias [Schoenegger et al., 2024, Li et al., 2024a, Chen et al., 2025b]. To validate these automatic ratings, four human annotators independently reviewed a random subset of 500 samples, achieving over 96% agreement across all annotators.

3.3.3 Data Selection

After obtaining quality scores from the reward models, we apply a filtering step to retain only the highest-quality samples for the final dataset. For each sample $x \in \mathcal{D}$, we collect ten rating scores covering both clinical and general evaluation criteria, denoted by $\{R_1(x), R_2(x), \ldots, R_{10}(x)\}$. To ensure robustness and consistency across all dimensions, we apply two selection conditions: (1) the minimum of the ten scores must be strictly greater than 2, and (2) the average score must exceed 3. This dual criterion ensures that selected samples exhibit both overall strength and no major weaknesses in any specific dimension. Formally, the selection rule is:

$$\mathcal{D}^* = \left\{ x \in \mathcal{D} : \min_{1 \le i \le 10} R_i(x) > 2 \text{ and } \sum_{i=1}^{10} R_i(x) > 3 \right\} \tag{1}$$

This filtering step yields 7,747 high-quality samples, which constitute our final MedGUIDE benchmark dataset. The distribution of these samples across various dimensions (such as cancer types, number of answer options, and LLMs used for generation) is illustrated in Figure 3 below.

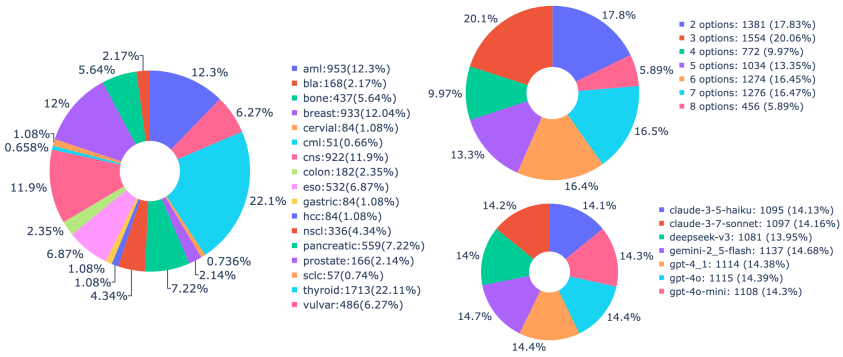


Figure 3: Distributions of cancer types (left), MCQ option counts (top right), and LLMs used for question generation (bottom right).

4 Experiments

We conduct experiments to evaluate the ability of various LLMs to reason about next steps in clinical diagnosis using the MedGUIDE benchmark.

4.1 Models:

We evaluated a total of 25 models (at inference temperature 0) spanning a range of sizes and architectures, aiming to ensure broad coverage and diversity for comprehensive assessment. For analysis, we loosely categorize these models into the following groups:

- **General LLMs (closed-source)**: `GPT-4o-mini`, `GPT-4.1`, `O1` and `O4-mini` [OpenAI, 2025b], `Claude-3.5-Haiku`, `Claude-3.7-Sonnet`, and `Deepseek` (V3, R1) [Liu et al., 2024a, Guo et al., 2025].
- **General LLMs (open-source)**: `Llama-3.2-Instruct` (1B, 3B), `Llama-3.1-Instruct` (8B, 70B), `Mistral-7B-Instruct`, `Mixtral-8x7B-Instruct`, `Qwen2.5-Instruct` (1.5B, 7B), and `Qwen3` (4B, 8B, 14B, 32B) [Qwen Team, 2025].
- **Medical LLMs:** `ClinicalCamel-70B` [Toma et al., 2023], `Medalpaca` (7B, 13B) [Han et al., 2023], and `Meditron` (7B, 70B) [Chen et al., 2023].

4.2 Evaluation Metrics

Each MedGUIDE sample is a multiple-choice question (MCQ), making *accuracy* the natural baseline metric. However, since the number of answer options varies across questions, we introduce a *weighted accuracy* metric that accounts for the difficulty of each sample. Let $c(x)$ denote the number of options in sample x. We define the difficulty function as

$$f(x) \stackrel{\text{def}}{=} 1 - \frac{1}{c(x)}, \tag{2}$$

which reflects the margin above random guess performance. Then we assign weight to each sample according to:

$$w_i \stackrel{\text{def}}{=} \frac{f(x_i)}{\sum_{i=1}^{|\mathcal{D}^*|} f(x_i)} = \frac{1 - \frac{1}{c(x_i)}}{|\mathcal{D}^*| - \sum_{i=1}^{|\mathcal{D}^*|} \frac{1}{c(x_i)}}. \tag{3}$$

Denote $1(\cdot)$ as the indicator function. Then the final weighted accuracy is defined as:

$$\text{WeightedAccuracy}(\mathcal{D}^*) \stackrel{\text{def}}{=} \frac{\sum_{i=1}^{|\mathcal{D}^*|} w_i \cdot 1(\hat{y}_i = y_i)}{|\mathcal{D}^*|} \tag{4}$$

4.3 Results

The results of all 25 models under both accuracy metrics are plotted in Figure 4. We observe that closed-source models, which are generally state-of-the-art across many tasks [Achiam et al., 2023, OpenAI, 2024c, Guo et al., 2025, Liu et al., 2024a, Anthropic, 2025], outperform others. In particular, `GPT-4.1` achieves the best performance, followed by reasoning-augmented models like `O1` and `O4-mini`. While these reasoning models perform well, they incur higher inference costs due to the extra thinking tokens. Among open-source models, we find that increasing model size does not always lead to better performance. For instance, `Qwen3-14B` and `Qwen3-32B` do not significantly outperform `Qwen3-4B`, and `Mixtral-8x7B` shows no large advantage over `Mistral-7B`. Notably, the `Qwen3` series shows strong performance despite smaller sizes.

Surprisingly, all medical LLMs underperform, regardless of size. We hypothesize two reasons: 1. MedGUIDE emphasizes reasoning over clinical decision paths, which requires models to **apply medical knowledge contextually**. Medical LLMs may possess domain knowledge but lack the reasoning capabilities required for next-step prediction. 2. Most medical LLMs are based on older backbones (e.g., Llama2), which may **struggle with instruction-following and logical deduction**. These reasoning skills are essential for serving as reliable medical assistants. Thus, MedGUIDE not only evaluates medical knowledge but also tests fundamental LLM capabilities crucial for clinical support tasks.

4.4 Correlation with Other Benchmarks

To better understand what capabilities MedGUIDE evaluates, we examine its correlation with other established benchmarks. In particular, we ask: Does performance on MedGUIDE reflect general

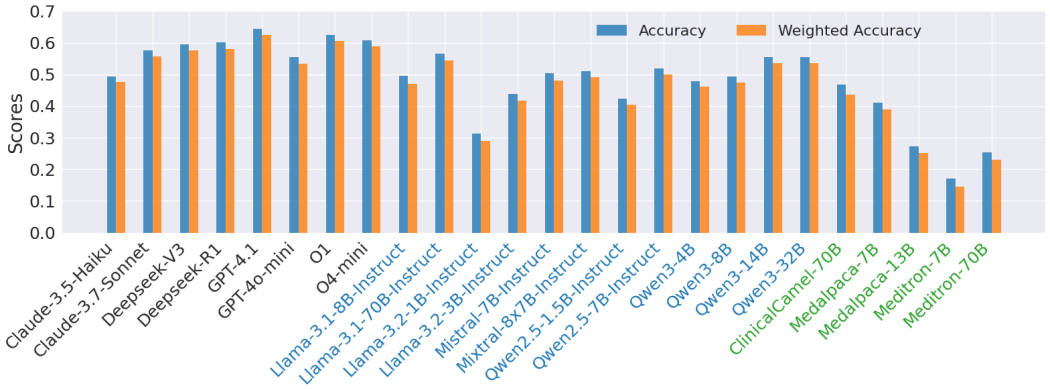


Figure 4: Benchmark evaluation results. For x-axis label, models in black, blue and green are closed-sources LLMs, open-sources LLMs and medical LLMs, respectively

medical knowledge alone, or does it additionally require broader reasoning or instruction-following abilities? To explore this, we evaluate the models on two complementary benchmarks: **IFEval**, which targets instruction-following under factual constraints, and **MMLU-Professional Medicine**, a knowledge-heavy subtask from MMLU that tests factual recall across a wide range of clinical topics. Detailed descriptions of both benchmarks are provided in Appendix G.1.

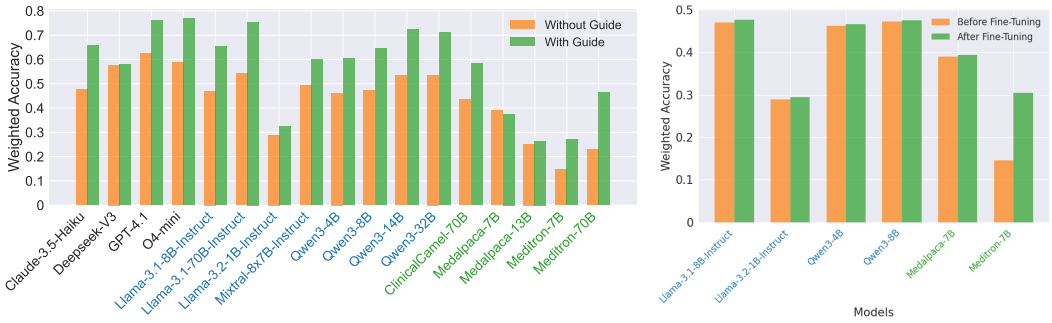
To quantify alignment, we compute three correlation metrics: **Spearman's** ρ [Spearman, 1904] , **Kendall's** τ [Kendall, 1938], and **Pearson's** r [Pearson, 1895], and report full results in Appendix G.3. Here, we focus on **Spearman's** ρ, which reflects rank-order agreement between benchmarks. We find that MedGUIDE exhibits strong Spearman correlation with MMLU-Professional Medicine ($\rho = 0.85$), suggesting that medical knowledge remains an important component. However, its nontrivial correlation with IFEval ($\rho = 0.71$) indicates that instruction-following and decision-step reasoning also contribute meaningfully to performance. Together, these results imply that MedGUIDE captures a unique intersection of knowledge recall and structured clinical decision-making not fully represented by either benchmark alone.

5 Improve Performance

5.1 Method I: Include the Guideline in Context

Our first strategy for boosting performance is to provide the relevant guideline decision tree (in JSON form) directly in the model's prompt. In this case, the model likewise does not need to recall the protocol from scratch but can parse and follow the tree structure in the context. In clinical practice, a physician would readily know which guideline applies. To mimic real-world usage, we also trained a lightweight Qwen-4B classifier to select the appropriate guideline given the patient vignette, achieving 98% validation accuracy, demonstrating that identifying the correct guideline is straightforward for both clinicians and capable LLMs.

Results: Improvements are clear across most models (see Figure 5a). For example, Meditron-70B's weighted accuracy rises from 0.230 to 0.462 (a 102% relative increase) while Meditron-7B improves by 86%. Detailed percentage gains for both metrics are shown in Figure 8 in the Appendix. We also include a detailed case study in Appendix H.1, comparing model outputs before and after applying Method I. In that case study, we ask the model to generate both the answer and an explanation. With the guideline in context, the model selects the correct answer and provides a step-by-step explanation that mirrors the decision-tree logic—from metastatic diagnosis through local control and adjuvant therapy to lung-only complete response and the final recommendation. In contrast, without the guideline, the model chooses an incorrect option and offers a plausible but guideline-inconsistent rationale, overlooking the patient's complete pulmonary response. This demonstrates the value of including structured guideline information to enhance both answer accuracy and explanation quality.



(a) Method I

(b) Method II

Figure 5: Weighted accuracy before and after using Method I (top) and Method II (bottom)

5.2 Method II: Fine-tuning with Guideline Data

In the second approach, we investigate whether a model can internalize clinical guideline knowledge through continued pretraining and subsequently apply it to MedGUIDE without receiving explicit guideline context at inference time. Rather than injecting the guideline into the prompt (as in Method I), we perform lightweight fine-tuning using the 55 NCCN decision trees (in structured JSON format). This serves as a form of domain-adaptive pretraining, aimed at infusing the model with the logical structure and semantics of the guidelines. We fine-tune a subset of models due to resource constraints, selecting representative general and medical LLMs. Each model is trained for 8 epochs with a learning rate of 1×10^{-5} (hyperparameters were selected via a grid search).

Results: Figure 5b and Appendix Figure 10 show the accuracy gains post-finetuning. While most models show only marginal improvements, suggesting their limited ability to transfer guideline knowledge to the downstream QA task, one notable exception is `Meditron-7B`, whose performance nearly doubles. We hypothesize two contributing factors: (1) `Meditron-7B` starts from a relatively low baseline, leaving greater room for improvement; and (2) despite being trained on medical corpora, the model may lack structured clinical reasoning skills required by MedGUIDE, which our continued pretraining helps reinforce. These results further underscore the importance of MedGUIDE—not only as a test of general medical knowledge, but as a benchmark for evaluating whether LLMs can follow structured, guideline-based clinical logic.

6 Conclusion

In this work, we introduce **MedGUIDE**, a benchmark for evaluating the clinical reasoning abilities of LLMs grounded in standardized medical guidelines. Unlike prior benchmarks that primarily test factual recall or domain-specific knowledge, MedGUIDE emphasizes stepwise diagnostic reasoning and adherence to expert-defined clinical pathways. Through a high-quality QA dataset generated from 55 NCCN decision trees and a dual-stage quality filtering pipeline, we evaluate 25 general and medical LLMs, revealing significant gaps in existing models' ability to reason through structured decision logic. Notably, medical LLMs often underperform despite their specialized training, and guideline grounding via contextual input or continued pretraining yields limited improvements, with few exceptions. These findings suggest that, beyond domain knowledge, robust guideline-following and sequential reasoning remain open challenges. We hope MedGUIDE serves as a valuable resource for driving progress toward clinically useful, safety-critical LLM applications.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint*

arXiv:1904.03323, 2019.

Anthropic. Introducing computer use, a new claude 3.5 sonnet, and claude 3.5 haiku, 10 2024. URL https://www.anthropic.com/news/3-5-models-and-computer-use. Upgraded Claude 3.5 Sonnet with improved coding capabilities and new computer use feature allowing Claude to interact with computers like humans.

Anthropic. Claude 3.7 sonnet, 2 2025. URL https://www.anthropic.com/claude/sonnet. Hybrid reasoning model with state-of-the-art coding skills, computer use, and 200K context window.

Donna K Arnett, Roger S Blumenthal, Michelle A Albert, Andrew B Buroker, Zachary D Goldberger, Ellen J Hahn, Cheryl D Himmelfarb, Amit Khera, Donald Lloyd-Jones, J William McEvoy, et al. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation*, 140(11):e596–e646, 2019.

Hila Chefer, Nir Ratner, Ori Ram, and Yoav Goldberg. Rulebench: A benchmark for inferential rule-following. *arXiv*, abs/2407.08440, 2024. URL https://arxiv.org/abs/2407.08440.

Chen Chen, Lei Li, Marcel Beetz, Abhirup Banerjee, Ramneek Gupta, and Vicente Grau. Large language model-informed ecg dual attention network for heart failure risk prediction. *IEEE Transactions on Big Data*, 2025a.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024a.

Yuwen Chen, Yiran Wang, Kai Zhang, et al. Climb: A benchmark of clinical bias in large language models. *arXiv preprint arXiv:2407.05250*, 2024b.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023.

Zeyu Chen, Zhenyu Huang, Sibo Zhang, et al. Clibench: A multifaceted and multigranular evaluation of large language models for clinical decision making. *arXiv preprint arXiv:2406.09923*, 2024c.

Zhijun Chen, Jingzheng Li, Pengpeng Chen, Zhuoran Li, Kai Sun, Yuankai Luo, Qianren Mao, Dingqi Yang, Hailong Sun, and Philip S Yu. Harnessing multiple large language models: A survey on llm ensemble. *arXiv preprint arXiv:2502.18036*, 2025b.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. Can llm be a personalized judge? *arXiv preprint arXiv:2406.11657*, 2024.

Chengfeng Dou, Zhi Jin, Wenpin Jiao, Haiyan Zhao, Yongqiang Zhao, and Zhenwei Tao. Integrating physician diagnostic logic into large language models: Preference learning from process feedback. *arXiv preprint arXiv:2401.05695*, 2024.

Dennis Fast, Lisa C Adams, Felix Busch, Conor Fallon, Marc Huppertz, Robert Siepmann, Philipp Prucker, Nadine Bayerl, Daniel Truhn, Marcus Makowski, et al. Autonomous medical evaluation for guideline adherence of large language models. *NPJ Digital Medicine*, 7(1):1–14, 2024.

Yanjun Gao, Ruizhe Li, Emma Croxford, Samuel Tesch, Daniel To, John Caskey, Brian W Patterson, Matthew M Churpek, Timothy Miller, Dmitriy Dligach, et al. Large language models and medical knowledge grounding for diagnosis prediction. *medRxiv*, pages 2023–11, 2023.

Jin Ge, Steve Sun, Joseph Owens, Victor Galvez, Oksana Gologorskaya, Jennifer C. Lai, Mark J Pletcher, and Ki Lai. Development of a liver disease-specific large language model chat interface using retrieval augmented generation. *medRxiv*, 2023.

Amelia Glaese, Nat McAleese, Melanie Trätner, Jonathan Uesato, Shane Legg, and Geoffrey Irving. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

Google DeepMind. Gemini 2.5 flash: Speed and value at scale, 4 2025. URL https://deepmind.google/technologies/gemini/flash/. Low latency, cost-efficient thinking model with 1-million token context window and multimodal capabilities.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30:2613–2622, 2024. doi: 10.1038/s41591-024-03097-1.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.

Yuhong He, Yongqi Zhang, Shizhu He, and Jun Wan. Bp4er: Bootstrap prompting for explicit reasoning in medical dialogue generation. *arXiv preprint arXiv:2403.19414*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Pedram Hosseini, Jessica M. Sin, Bing Ren, Bryceton G. Thomas, Elnaz Nouri, Ali Farahanchi, and Saeed Hassanpour. A benchmark for long-form medical question answering. *arXiv*, abs/2411.09834, 2024. URL https://arxiv.org/html/2411.09834v2.

Hui Huang, Yingqi Qu, Xingyuan Bu, Hongli Zhou, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge model is not a general substitute for gpt-4. *arXiv preprint arXiv:2403.02839*, 2024a.

Yining Huang, Keke Tang, Meilian Chen, and Boyuan Wang. A comprehensive survey on evaluating large language model applications in the medical industry, 2024b. URL https://arxiv.org/abs/2404.15777.

IBM. What is instruction tuning? *IBM*, 2023. URL https://www.ibm.com/think/topics/instruction-tuning.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 10 2023a.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Lavender Yao Jiang, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Antony Riina, Ilya Laufer, Paawan Punjabi, et al. Health system-scale language models are all-purpose prediction engines. *Nature*, 619(7969):357–362, 2023b.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. pages 2567–2577, 2019.

Maurice G. Kendall. *A new measure of rank correlation*, volume 30. Oxford University Press, 1938.

Bernhard Kratzwald, Jens Göhner, and Stefan Feuerriegel. Towards a multilingual benchmark for medical knowledge assessment. *Research Square*, 2021. URL https://doi.org/10.20944/PREPRINTS202105.0498.V1.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024a.

Xiaomin Li, Mingye Gao, Zhiwei Zhang, Chang Yue, and Hong Hu. Rule-based data selection for large language models. *arXiv preprint arXiv:2410.04715*, 2024b.

Xiaomin Li, Xupeng Chen, Jingxuan Fan, Eric Hanchen Jiang, and Mingye Gao. Multi-head reward aggregation guided by entropy. *arXiv preprint arXiv:2503.20995*, 2025a.

Xiaomin Li, Mingye Gao, Zhiwei Zhang, Jingxuan Fan, and Weiyu Li. Data-adaptive safety rules for training reward models. *arXiv preprint arXiv:2501.15453*, 2025b.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Canyu Liu, Ziyuan Wang, Jian Tan, et al. Clinicalbench: Can llms beat traditional ml models in clinical prediction? *arXiv preprint arXiv:2411.06469*, 2024b.

Wei Liu, S. Sara Mahdavi, Nishant Subramani, Laichee Man, Karan Singhal, Alan Karthikesalingam, Vivek Natarajan, and Ryutaro Tanno. Clinical calculation qa: A benchmark for clinical calculation tasks. *arXiv*, abs/2406.12036, 2024c. URL `https://doi.org/10.48550/arXiv.2406.12036`.

Meta AI. Introducing llama 3.1: Our most capable models to date, 7 2024a. URL `https://ai.meta.com/blog/meta-llama-3-1/`. New models including flagship 405B parameter model, along with upgraded 8B and 70B models featuring 128K context length and multilingual capabilities.

Meta AI. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models, 9 2024b. URL `https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/`. First vision-capable Llama models with 11B and 90B parameters, alongside lightweight text-only models (1B and 3B) for edge and mobile devices.

National Comprehensive Cancer Network. NCCN Clinical Practice Guidelines in Oncology. `https://www.nccn.org/guidelines/guidelines-detail`, 2023. Accessed: 2023-05-10.

OpenAI. Hello gpt-4o, 5 2024a. URL `https://openai.com/index/hello-gpt-4o/`. New flagship multimodal model capable of reasoning across audio, vision, and text in real time.

OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence, 7 2024b. URL `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`. Smart, affordable model outperforming GPT-3.5 Turbo at significantly lower cost.

OpenAI. Openai o1, 9 2024c. URL `https://openai.com/o1/`. Full version released on December 5, 2024.

OpenAI. Introducing gpt-4.1 in the api, 4 2025a. URL `https://openai.com/index/gpt-4-1/`. Models featuring improved coding, instruction following, and 1 million token context window.

OpenAI. Introducing o3 and o4-mini, 4 2025b. URL `https://openai.com/index/introducing-o3-and-o4-mini/`. Advanced reasoning models with tool use capabilities including web browsing, Python, image understanding, and image generation.

World Health Organization. Who releases ai ethics and governance guidance for large multi-modal models. *World Health Organization*, 2024. URL `https://www.who.int/news/item/18-01-2024-who-releases-ai-ethics-and-governance-guidance-for-large-multi-modal-models`.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical question answering. 174:248–260, 07–08 Apr 2022. URL `https://proceedings.mlr.press/v174/pal22a.html`.

Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun, and Mikhail Yurochkin. Efficient multi-prompt evaluation of llms. *arXiv preprint arXiv:2405.17202*, 2024.

Qwen Team. Qwen3: Think deeper, act faster, 4 2025. URL `https://qwenlm.github.io/blog/qwen3/`. Latest generation of large language models with hybrid thinking modes, including dense models (0.6B to 32B parameters) and Mixture-of-Experts models (30B-A3B and 235B-A22B).

Or Ram, Hila Chefer, Nir Ratner, Gal Chechik, and Yoav Goldberg. Llmbar: A meta-evaluation benchmark for llm-based evaluation of instruction adherence. *arXiv*, abs/2310.07641, 2023. URL `https://arxiv.org/abs/2310.07641`.

Mercy Prasanna Ranjit, Gopinath Ganapathy, Ranjit Frederick Manuel, and Tanuja Ganu. Retrieval augmented chest x-ray report generation using openai gpt models. *arXiv*, abs/2305.03660, 2023.

Alexander Rau, Stephan Rau, Daniela Zoeller, Anna Fink, Hien Tran, Caroline Wilpert, Johanna Nattenmueller, Jakob Neubauer, Fabian Bamberg, Marco Reisert, and Maximilian Frederik Russe. A context-based chatbot surpasses trained radiologists and generic chatgpt in following the acr appropriateness guidelines. *Radiology*, 308(1):e230970, 2023.

Philipp Schoenegger, Indre Tuminauskaite, Peter S Park, Rafael Valdece Sousa Bastos, and Philip E Tetlock. Wisdom of the silicon crowd: Llm ensemble prediction capabilities rival human crowd accuracy. *Science Advances*, 10(45):eadp1528, 2024.

Wenqi Shi, Yuchen Zhuang, Yuanda Zhu, Henry Iwinski, Michael Wattenbarger, and May Dongmei Wang. Retrieval-augmented large language models for adolescent idiopathic scoliosis patients in shared decision-making. In *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–10, 2023.

Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. Biomegatron: larger biomedical domain language model. *arXiv preprint arXiv:2010.06060*, 2020.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8, 2025.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

Maojun Sun. Llamacare: A large medical language model for enhancing healthcare knowledge sharing. *arXiv preprint arXiv:2406.02350*, 2024.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624*, 2024.

Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *CoRR*, 2023.

A Venigalla, Jonathan Frankle, and M Carbin. Biomedlm: a domain-specific large language model for biomedical text. *MosaicML. Accessed: Dec*, 23(3):2, 2022.

Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. Clinicalgpt: large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv preprint arXiv:2306.09968*, 2023.

Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. Beyond direct diagnosis: Llm-based multi-specialist agent consultation for automatic diagnosis. *arXiv preprint arXiv:2401.16107*, 2024a.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*, 2024b.

Yizhong Wang, Ximing Yuan, Zheru Xiang, Hongwei Wang, Linjie Cai, Shuohang Weng, Jena Qiu, Bo Tseng, Cheng Chen, Yiran Sun, et al. Helpsteer2: Improving helpfulness and harmlessness for human preference fine-tuning. *arXiv preprint arXiv:2401.10997*, 2024c.

Yilin Wen, Zifeng Wang, and Jimeng Sun. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. *arXiv preprint arXiv:2308.09729*, 2023.

Martin Weyssow, Aton Kamanda, Xin Zhou, and Houari Sahraoui. Codeultrafeedback: An llm-as-a-judge dataset for aligning large language models to coding preferences. *arXiv preprint arXiv:2403.09032*, 2024.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*, 2(5):6, 2023.

Jiageng Wu, Xian Wu, and Jie Yang. Guiding clinical reasoning with large language models via knowledge seeds. *arXiv preprint arXiv:2403.06609*, 2024a.

Jiageng Wu, Xian Wu, Yefeng Zheng, and Jie Yang. Medkp: Medical dialogue with knowledge enhancement and clinical pathway encoding. *arXiv preprint arXiv:2403.06611*, 2024b.

Yuqi Wu, Guangya Wan, Jingjing Li, Shengming Zhao, Lingfeng Ma, Tianyi Ye, Mike Zhang, Ion Pop, Yanbo Zhang, and Jie Chen. Proai: Proactive multi-agent conversational ai with structured knowledge base for psychiatric diagnosis. *arXiv preprint arXiv:2502.20689*, 2025.

Chonghua Xue, Sahana S Kowshik, Diala Lteif, Shreyas Puducheri, Varuna H Jasodanand, Olivia T Zhou, Anika S Walia, Osman B Guney, J Diana Zhang, Serena T Pham, et al. Evaluation of large language model performance on the biomedical language understanding and reasoning benchmark. *medRxiv*, 2024. URL `https://www.medrxiv.org/content/10.1101/2024.05.17.24307411v1.full.pdf`.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.

Dingkang Yang, Jinjie Wei, Dongling Xiao, Shunli Wang, Tong Wu, Gang Li, Mingcheng Li, Shuaibing Wang, Jiawei Chen, Yue Jiang, et al. Pediatricsgpt: Large language models as chinese medical assistants for pediatric applications. *Advances in Neural Information Processing Systems*, 37:138632–138662, 2024b.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, et al. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540*, 2022.

Fan Zhang, Guangtao Zeng, Yinuo Zhang, Yanzhe Wang, Zonghai Yao, Guangya Wan, Jie Chen, Fei Wang, and Xiaoxiao Li. Evaluating large language models on clinical consent form generation. *arXiv*, abs/2504.00934, 2025. URL `https://arxiv.org/abs/2504.00934`.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*, 2023.

Rui Zhang, Yizhi Liang, Yi Fung, Xiangyu Peng, Jie Zhou, Juntao Li, Shuyi Wang, Xiaohuan Zhou, Yang Yan, Yixin Zhu, et al. Infobench: Evaluating instruction following ability in large language models. *ACL Anthology*, 2024. URL `https://aclanthology.org/2024.findings-acl.772.pdf`.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*, 2023.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Zijian Zhou, Miaojing Shi, Meng Wei, Oluwatosin Alabi, Zijie Yue, and Tom Vercauteren. Large model driven radiology report generation with clinical quality reinforcement learning. *arXiv preprint arXiv:2403.06728*, 2024.

Yinghao Zhu, Changyu Ren, Zixiang Wang, Xiaochen Zheng, Shiyun Xie, Junlan Feng, Xi Zhu, Zhoujun Li, Liantao Ma, and Chengwei Pan. Emerge: Integrating rag for improved multimodal ehr predictive modeling. *arXiv*, abs/2406.00036, 2024.

Appendix

A Limitations and Future Directions

Limitations and Future Work. While MedGUIDE is a comprehensive benchmark grounded in real clinical guidelines, there are natural opportunities for extension. Our current focus is on NCCN oncology guidelines, and future versions could expand to cover other specialties such as cardiology or endocrinology to further assess cross-domain generalization. Also, we explored guideline-grounded fine-tuning and context augmentation, future work may explore reinforcement learning or retrieval-enhanced methods to further align models with structured clinical reasoning. Overall, we see MedGUIDE as a solid foundation upon which richer diagnostic benchmarks can be built.

B Demographic Information of Annotators

Annotator ID	School & Program	Year	Age
01	Harvard Medical School	M3	27
02	Harvard Bioengineering & Department of Stem Cell and Regenerative Biology (HSCRB)	G5	27
03	Harvard Department of Stem Cell and Regenerative Biology (HSCRB)	N/A	24
04	Harvard Population Health Science - Statistical Genetics	incoming PhD	24
05	Harvard Population Health Science - Epidemiology	G1	25
06	MIT Biology	PhD	28

C Disease Abbreviations

Table 2 lists the abbreviations used in our dataset along with the corresponding full disease names derived from NCCN guidelines.

Abbreviation	Full Disease Name
`aml`	Acute myeloid leukemia
`bla`	Bladder cancer
`bon`	Bone cancer
`bre`	Breast cancer
`cer`	Cervical cancer
`cml`	Chronic myeloid leukemia
`cns`	Central nervous system cancer
`col`	Colon cancer
`eso`	Esophageal and esophagastric junction cancer
`gas`	Gastric cancer
`hcc`	Hepatocellular carcinoma
`nscl`	Non-small cell lung cancer
`pancreatic`	Pancreatic cancer
`prostate`	Prostate cancer
`sclc`	Small cell lung cancer
`thyroid`	Thyroid carcinoma
`vulvar`	Vulvar cancer

Table 2: Abbreviations used for disease categories in the MedGUIDE benchmark.

D Criteria for QA Data Quality Evaluation

Here we provide detailed criteria for both clinical and general aspects, along with their corresponding scoring rubrics.

D.1 Clinical Evaluation Criteria

1. Clinical Plausibility

Definition: Assess how realistic, accurate, and medically plausible the patient scenarios and clinical histories are.

- **Score 5 (Excellent):** Entirely realistic scenario; demographics, disease course, treatment history, and outcomes accurately reflect real-world clinical practice.
  Example: A 65-year-old male smoker with stage III NSCLC, weight loss, and persistent cough undergoes CT and biopsy confirming adenocarcinoma.
- **Score 4 (Good):** Mostly realistic with minor atypical or simplified details.
  Example: 30-year-old woman with hemoptysis and imaging-confirmed large lung mass, but no risk factors mentioned.
- **Score 3 (Fair):** Plausible but includes inconsistencies or missing details.
  Example: 75-year-old with metastatic prostate cancer but no urinary symptoms or PSA history.
- **Score 2 (Poor):** Multiple clinical inaccuracies.
  Example: 25-year-old female with thyroid cancer metastasizing to bones without prior neck symptoms.
- **Score 1 (Implausible):** Medically impossible or significantly flawed.
  Example: Localized thyroid cancer treated with first-line systemic chemotherapy.

2. Clinical Utility

Definition: Evaluate educational relevance and practical clinical usefulness of the scenario and question.

- **Score 5 (Excellent):** Highly relevant to key guideline-based decisions.
  Example: Choosing between surveillance vs. surgery for low-risk thyroid cancer.
- **Score 4 (Good):** Relevant but less frequent or less critical decision.
  Example: Choosing imaging modality for thyroid nodule staging.
- **Score 3 (Fair):** Basic or somewhat obvious decision.
  Example: Referral for suspected thyroid nodule.
- **Score 2 (Poor):** Trivial scenario.
  Example: Whether to evaluate a neck mass in a symptomatic adult.
- **Score 1 (Misleading):** Irrelevant or outdated guidance.
  Example: Surgical treatment for incidental 3 mm thyroid cyst.

3. Quality of Decision Path

Definition: Assess the clinical logic and fidelity to real-world guideline-based decision sequences.

- **Score 5 (Excellent):** Fully guideline-consistent and logically coherent.
- **Score 4 (Good):** Minor deviation, otherwise accurate.
- **Score 3 (Fair):** Small errors but overall direction intact.
- **Score 2 (Poor):** Multiple inconsistencies or unclear transitions.
- **Score 1 (Invalid):** Contradicts guidelines or nonsensical path.

4. Alignment to Decision Path

Definition: Evaluate if the patient profile matches the simulated guideline path.

- **Score 5 (Excellent):** All key nodes correctly included.
- **Score 4 (Good):** Minor omissions, overall aligned.
- **Score 3 (Fair):** Multiple inaccuracies but generally reasonable.

- **Score 2 (Poor):** Weak alignment; many steps wrong or missing.
- **Score 1 (Misaligned):** Entirely mismatched from the guideline.

5. Clinical Accuracy of Correct Answer

Definition: Confirm whether the correct answer is safe and guideline-aligned.

- **Score 5 (Excellent):** Fully guideline-supported.
  Example: RAI for iodine-avid thyroid cancer.
- **Score 4 (Good):** Minor ambiguity or edge-case.
  Example: RAI for borderline thyroid case.
- **Score 3 (Fair):** Acceptable but not ideal.
- **Score 2 (Poor):** Technically possible but incorrect.
- **Score 1 (Unsafe):** Clinically incorrect or harmful.

D.2 General Quality Criteria

1. Clarity and Detail Level

Definition: Evaluate if the clinical scenario is clear, unambiguous, and sufficiently detailed.

- **Score 5**: Clear, detailed, and precise medical language.
- **Score 4**: Mostly clear with minor ambiguities.
- **Score 3**: Some ambiguity or missing elements.
- **Score 2**: Lacks key clinical details.
- **Score 1**: Unclear or insufficient to interpret.

2. Consistency and Internal Logic

Definition: Assess logical consistency of the patient narrative.

- **Score 5**: Fully consistent with clear temporal logic.
- **Score 4**: Mostly consistent with minor issues.
- **Score 3**: Some contradictions.
- **Score 2**: Significant inconsistencies.
- **Score 1**: Illogical or medically impossible.

3. Safety and Toxicity

Definition: Ensure that scenarios and choices are medically safe and non-harmful.

- **Score 5**: Completely safe and appropriate.
- **Score 4**: Safe with minor issues.
- **Score 3**: Generally safe but questionable detail.
- **Score 2**: Contains safety concerns.
- **Score 1**: Unsafe or harmful.

4. Textual Quality and Professionalism

Definition: Assess grammar, readability, and tone.

- **Score 5**: Highly professional and grammatically correct.
- **Score 4**: Minor issues not affecting clarity.
- **Score 3**: Noticeable grammar/style flaws.
- **Score 2**: Hard to follow due to errors.
- **Score 1**: Poor quality and unprofessional.

5. Option Distinctiveness

Definition: Evaluate whether answer choices are clearly distinct.

- **Score 5**: All options clearly distinct.
- **Score 4**: Minor overlaps, still distinguishable.
- **Score 3**: Some overlap but manageable.
- **Score 2**: Significant overlap.
- **Score 1**: Indistinct or redundant choices.

E Reward Model Training

For each input sample x, we have a ground truth reward vector $[R_1(x), R_2(x), \ldots, R_5(x)]$ corresponding to the five evaluation dimensions. The multi-head reward model is trained to predict $[\hat{R}_1(x), \hat{R}_2(x), \ldots, \hat{R}_5(x)]$. This is formulated as a multi-label regression task, where each label is a 5-dimensional real-valued score vector.

We use mean squared error (MSE) as the loss function across all heads:

$$\mathcal{L}(x) = \frac{1}{5} \sum_{i=1}^{5} \left(\hat{R}_i(x) - R_i(x) \right)^2 \tag{5}$$

For training, we use a single NVIDIA H100 80GB GPU. The model is trained for 5 epochs with a learning rate of 2×10^{-5}. Hyperparameters and the backbone model are selected via grid search over five candidates:

- `Llama3.2-1B-Instruct`,
- `Llama3.2-3B-Instruct`,
- `Llama3.1-8B-Instruct`,
- `Qwen2.5-1.5B-Instruct`,
- `Qwen2.5-7B-Instruct`.

The search spans learning rates 5e-5, 2e-5, 1e-5, 5e-6 and epoch counts ranging from 1 to 16. The validation MSE is 0.28 at the end of training.

F MedGUIDE Evaluation Results

In Table 3, we present the detailed scores for both accuracy metrics across all 25 models evaluated on the MedGUIDE benchmark.

Table 3: Performance of 25 LLMs on MedGUIDE

Model	Accuracy	Weighted Accuracy
Claude-3-5-Haiku-20241022	0.4935	0.4755
Claude-3-7-Sonnet-20250219	0.5765	0.5562
ClinicalCamel-70B	0.4683	0.4360
Deepseek-V3	0.5957	0.5770
Deepseek-R1	0.6006	0.5812
GPT-4.1	0.6439	0.6254
GPT-4o-mini	0.5557	0.5342
O1	0.6239	0.6049
O4-mini	0.6075	0.5890
Llama-3.1-8B-Instruct	0.4955	0.4706
Llama-3.1-70B-Instruct	0.5654	0.5431
Llama-3.2-1B-Instruct	0.3121	0.2889
Llama-3.2-3B-Instruct	0.4377	0.4174
Medalpaca-7b	0.4116	0.3899
Medalpaca-13b	0.2721	0.2516
Meditron-7b	0.1701	0.1458
Meditron-70b	0.2530	0.2295
Mistral-7B-Instruct-v0.3	0.5034	0.4813
Mixtral-8x7B-Instruct-v0.1	0.5095	0.4915
Qwen2.5-1.5B-Instruct	0.4242	0.4044
Qwen2.5-7B-Instruct	0.5188	0.5007
Qwen3-4B	0.4789	0.4624
Qwen3-8B	0.4936	0.4733
Qwen3-14B	0.5554	0.5354
Qwen3-32B	0.5553	0.5366

G Correlation Between Benchmarks

G.1 Details of Benchmarks

IFEval. *Instruction-Following Evaluation (IFEval)* [Zhou et al., 2023] is a synthetic benchmark designed to probe an LLM's ability to follow *explicit, rule-verifiable constraints*. Each of the 541 prompts is paired with one to three constraints drawn from a catalogue of 25 rule types, covering simple length requirements ("write at least `N` words"), formatting rules ("output exactly two bullet points"), keyword constraints ("mention the string `AI` three times"), and lexical bans ("do not use the word `because`"), among others. Because every constraint can be checked automatically, evaluation is performed with *instruction-level loose accuracy*: a response is scored as correct if **all** constraints attached to that prompt are satisfied, allowing for innocuous preambles or markdown wrappers that do not violate any rule. (We follow the "loose" variant proposed by Zhou et al. [2023], which strips leading markdown headings and trailing boiler-plate before rule checking so as to avoid false negatives.) In the correlation study we treat each model's overall pass rate (%) on IFEval as one of the two score vectors.

MMLU–PROFESSIONAL-MEDICINE. *Massive Multitask Language Understanding (MMLU)* is a 57-subject multiple-choice benchmark that spans high-school, university, and professional-licensing curricula [Hendrycks et al., 2020]. The PROFESSIONAL-MEDICINE subset consists of 132 questions drawn from USMLE-style practice banks and medical-board preparatory materials. Questions demand factual recall of physiology, pathology and pharmacology as well as short clinical reasoning: each item provides four answer choices with exactly one correct option. Following the MMLU protocol, we report *five-shot accuracy*: every test question is preceded by five exemplars from the same sub-domain, and a model's answer is selected by its highest-log-probability choice. This accuracy forms the second score vector in our analysis.

Why these two benchmarks? IFEval isolates pure *instruction adherence* under transparent, rule-based constraints, while MMLU–PROFESSIONAL-MEDICINE measures *domain knowledge and*

clinical reasoning. By correlating MedGUIDE results with these two complementary axes we can disentangle whether strong guideline reasoning co-varies more with generic instruction-following skills, with medical knowledge, or with both.

G.2 Correlation Metrics for Benchmark Comparison

To quantify the relationship between model performance across different benchmarks, we use the following statistical correlation metrics:

Spearman's Rank Correlation Coefficient. Spearman's correlation measures the strength of a monotonic relationship between two sets of scores. It is defined as the Pearson correlation between the rank-transformed variables:

$$\rho_S = \frac{\sum_{i=1}^{n}(\mathrm{rank}(x_i) - \overline{r}_x)(\mathrm{rank}(y_i) - \overline{r}_y)}{\sqrt{\sum_{i=1}^{n}(\mathrm{rank}(x_i) - \overline{r}_x)^2}\sqrt{\sum_{i=1}^{n}(\mathrm{rank}(y_i) - \overline{r}_y)^2}},$$

where $\mathrm{rank}(x_i)$ and $\mathrm{rank}(y_i)$ are the ranks of x_i and y_i, and $\overline{r}_x$, $\overline{r}_y$ are their mean ranks.

Kendall's τ. Kendall's τ measures ordinal association by counting the number of concordant and discordant pairs:

$$\tau = \frac{C - D}{C + D},$$

where C is the number of concordant pairs and D is the number of discordant pairs among all $\binom{n}{2}$ possible pairs.

Pearson Correlation Coefficient. Pearson's correlation measures the strength of a linear relationship between two numerical variables:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

where $\bar{x}$ and $\bar{y}$ are the means of x and y, respectively.

These metrics allow us to assess whether strong performance on MedGUIDE correlates with performance on other medical reasoning benchmarks, either in terms of rank agreement or absolute score similarity.

Comparison of Metrics. Each correlation metric offers distinct interpretability and sensitivity characteristics. Below we summarize when each is most appropriate for comparing benchmark scores:

- **Pearson correlation (r)** asks: *Are the two score vectors linearly related?* It assumes interval-scale scores and is sensitive to outliers. This metric is most informative when the relationship between benchmarks is expected to be approximately linear (e.g., every 10-point gain on Benchmark A corresponds to an 8-point gain on Benchmark B).
- **Spearman correlation (ρ)** asks: *Do higher scores on one benchmark generally correspond to higher scores on the other?* It converts scores into ranks and measures monotonic relationships. This rank-based approach is robust to non-linear transformations and outliers, making it suitable when benchmarks differ in scale or exhibit curved relationships.
- **Kendall's τ** asks: *For each pair of models, does the higher-scoring model on one benchmark also rank higher on the other?* This purely ordinal metric counts concordant and discordant pairs, making it especially robust to noise and ties. It is most informative when N is small (as in our case with $N = 25$) and when a probabilistic interpretation of ranking agreement is desired.

By jointly considering these three metrics, we obtain a more comprehensive view of how two benchmarks align—capturing linear trends (Pearson), monotonicity (Spearman), and ordinal consistency (Kendall).

G.3 Correlation Results

Figure 6 shows the exact performance scores on three benchmarks. Figure 7 shows correlations between **MedGUIDE** and both of **IFEval** [Dong et al., 2024] and **MMLU-Professional Medicine** [Hendrycks et al., 2020] across 25 models, using Spearman's ρ, Kendall's τ, and Pearson's r metrics.

We observe strong positive correlations between MedGUIDE and both benchmarks. Interestingly, MedGUIDE exhibits even higher alignment with MMLU (e.g., $\rho = 0.85$, $r = 0.81$) than with IFEval ($\rho = 0.71$, $r = 0.75$), suggesting that despite its emphasis on structured decision-making, MedGUIDE retains a substantial factual and knowledge-intensive component. The relatively weaker Kendall's τ values (e.g., 0.56 for MedGUIDE–IFEval and 0.68 for MedGUIDE–MMLU) indicate moderate agreement in relative model ranking order.

Overall, the correlations highlight that MedGUIDE partially overlaps with both instruction-following and medical knowledge evaluation, while introducing unique challenges in decision-path compliance not fully captured by either benchmark alone.

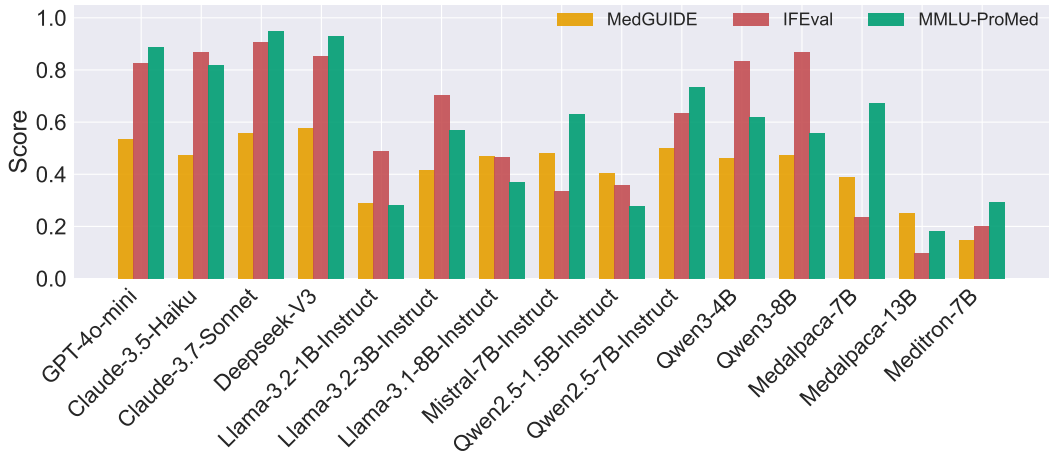


Figure 6: Performance scores of models across three benchmarks: MedGUIDE (Weighted Accuracy), IFEval, and MMLU-Professional Medicine.

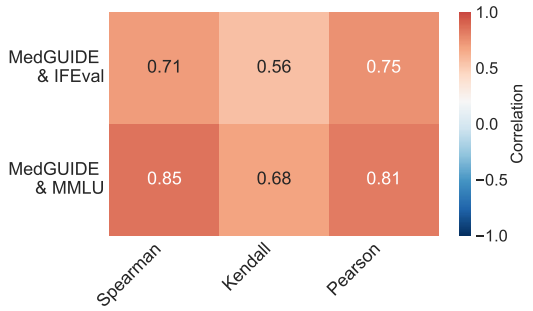


Figure 7: Pairwise correlations between MedGUIDE, IFEval, and MMLU-Professional Medicine across 25 models, evaluated via Spearman, Kendall, and Pearson metrics. MedGUIDE aligns more strongly with MMLU in absolute scoring, but shares notable ranking similarity with IFEval as well.

H Methods for Improving MedGUIDE Accuracy

H.1 Method I

Below in Table 4, we report the exact accuracy scores after applying Method I (including guideline in context), along with the corresponding relative improvements in percentage. These improvements are

also visualized in Figure 8. We then present a detailed case study demonstrating how Method I enables the model to follow the guideline and produce both the correct answer and a guideline-consistent explanation.

Table 4: LLM performance after applying clinical guideline by Method I, with relative improvement in percentages.

Model	**Accuracy**	**Weighted Accuracy**
Claude-3-5-Haiku-20241022	0.673 (36%)	0.658 (38%)
ClinicalCamel-70B	0.610 (30%)	0.583 (34%)
Deepseek-V3	0.601 (1%)	0.579 (0%)
GPT-4.1	0.775 (20%)	0.760 (21%)
O4-mini	0.783 (29%)	0.768 (30%)
Llama-3.1-8B-Instruct	0.671 (35%)	0.654 (39%)
Llama-3.1-70B-Instruct	0.769 (36%)	0.755 (39%)
Llama-3.2-1B-Instruct	0.354 (13%)	0.324 (12%)
Medalpaca-7b	0.407 (-1%)	0.376 (-4%)
Medalpaca-13b	0.291 (7%)	0.264 (5%)
Meditron-7b	0.292 (72%)	0.271 (86%)
Meditron-70b	0.490 (94%)	0.463 (102%)
Mixtral-8x7B-Instruct-v0.1	0.621 (22%)	0.599 (22%)
Qwen3-8B	0.667 (35%)	0.647 (37%)
Qwen3-4B	0.625 (31%)	0.605 (31%)
Qwen3-14B	0.740 (33%)	0.723 (35%)
Qwen3-32B	0.726 (31%)	0.711 (32%)

Figure 8: Improvement of accuracy and weighted accuracy for Method I

Case Study of Method I

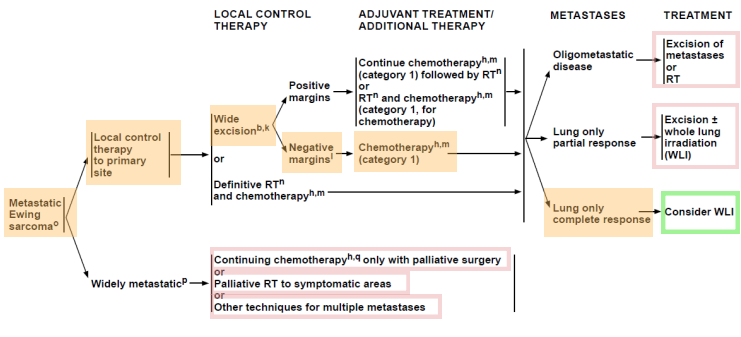


Figure 9: **NCCN Metastatic Ewing sarcoma Guideline.** The orange boxes illustrate the workflow through which the sample QA dataset is generated. Red and green annotations represent the correct and incorrect options.

Prompt: A 16-year-old male was diagnosed with metastatic Ewing sarcoma following the evaluation of persistent pain in his right pelvis and a palpable mass. Initial imaging revealed a 5 cm tumor in the iliac wing with multiple lung metastases noted on a CT scan. He underwent neoadjuvant chemotherapy with vincristine, doxorubicin, and ifosfamide, followed by surgical resection of the primary tumor. Post-surgery, he received definitive radiation therapy to the primary site and additional chemotherapy for residual disease. Follow-up chest CT demonstrated complete resolution of the lung metastases, but there remains concern for micro-metastatic disease given the initial extensive pulmonary involvement. Given these findings, should we consider whole lung irradiation as a next step in managing his treatment?
Options:

(A) Excision of metastases or RT

(B) Consider WLI

(C) Continuing chemotherapy only with palliative surgery

(D) Excision ± whole lung irradiation (WLI)

(E) Palliative RT to symptomatic areas

(F) Other techniques for multiple metastases

Correct Answer: B.

Original Answer by GPT4.1: C
Original Explanation by GPT4.1: The correct management for this 16-year-old male with metastatic Ewing sarcoma is option C: Continuing chemotherapy only with palliative surgery. Despite the patient's comprehensive initial treatment regimen consisting of neoadjuvant chemotherapy, surgical resection of the primary iliac wing tumor, definitive radiation to the primary site, and additional chemotherapy, the history of extensive pulmonary metastases represents a significant prognostic challenge. While the follow-up chest CT demonstrates complete radiographic resolution of the visible lung metastases, patients with initial widespread metastatic Ewing sarcoma typically face a guarded prognosis. In this context, continuing systemic chemotherapy remains the backbone of treatment to control microscopic metastatic disease throughout the body, not just in the lungs. The palliative surgery component allows for addressing any local complications or symptomatic lesions that might develop, without subjecting the patient to the significant toxicity of whole lung irradiation, which carries substantial short and long-term side effects including pneumonitis, pulmonary fibrosis, and cardiac toxicity. This balanced approach prioritizes disease control while maintaining quality of life, recognizing that complete cure may be challenging in the setting of initially extensive metastatic disease. Systemic therapy offers the best chance of managing potential microscopic disease at all sites while minimizing treatment-related morbidity in this young patient who has already undergone intensive multimodal therapy.

Correct Answer by GPT 4.1 (using Method I): B
Correct Explanation by GPT 4.1 (using Method I): To understand the decision-making path from the clinical vignette to the final option B: Consider WLI, it is essential to follow the flowchart

step by step, tracing the structure logically from the root node to the leaf node. The patient is diagnosed with Metastatic Ewing Sarcoma, which represents the starting point at the root node of the flowchart. According to the guidelines, metastatic Ewing sarcoma is managed initially with Local Control Therapy to the primary site. In this case, the patient underwent surgical resection and radiation therapy, fitting the expected pathway for local control. After local control therapy, the flowchart directs the decision-making process to Adjuvant Treatment / Additional Therapy, which includes continued chemotherapy and radiation therapy. This step matches the patient's treatment history, where he received additional chemotherapy after surgery.
The next step is the Assessment of Metastatic Response, where the decision path diverges based on the response of lung metastases. The flowchart specifies three possible pathways: Oligometastatic Disease, which would prompt excision or radiation therapy; Lung Only Partial Response, suggesting incomplete control and potential excision ± whole lung irradiation (WLI); and Lung Only Complete Response. In this patient's case, imaging showed a complete response in the lung, guiding the decision towards Consider WLI as the next step to address potential micro-metastatic disease.
Following the path for Lung Only Complete Response, the guideline explicitly recommends Consider WLI for comprehensive management aimed at mitigating the risk of microscopic disease. This is where the decision path terminates, aligning precisely with the option B: Consider WLI in the clinical question. Thus, the stepwise traversal through the guideline flowchart follows a logical sequence: Metastatic Ewing Sarcoma → Local Control Therapy → Adjuvant Treatment → Lung Only Complete Response → Consider WLI, justifying B as the most clinically appropriate option.

Analysis: In this case study, we observe a clear improvement in both the reasoning and answer selection after incorporating the guideline into the prompt. In the original response (without guideline context), the model incorrectly selects option **C** and justifies its decision based on general considerations of prognosis and treatment-related toxicity. However, this explanation overlooks the patient's complete response in the lungs and fails to recognize the appropriate next step recommended in the guideline for such cases. In contrast, after including the guideline in the prompt (Method I), the model correctly identifies **B: Consider WLI** as the recommended next step. The explanation explicitly traces the decision path through the guideline, correctly mapping the patient's clinical course—metastatic disease, local control, adjuvant therapy, and lung-only complete response—to the corresponding recommendation. This demonstrates how incorporating the structured guideline enables the model to produce not only the correct answer but also a faithful and guideline-consistent rationale.

H.2 Method II

Below in Table 5, we report the exact accuracy scores after applying Method II (fine-tuning with guidelines), along with the corresponding relative improvements in percentage. These improvements are also visualized in Figure 10.

Model	Accuracy	Weighted Acc
Llama-3.1-8B-Instruct	0.5014 (1%)	0.4772 (1%)
Llama-3.2-1B-Instruct	0.3170 (2%)	0.2941 (2%)
Medalpaca-7b	0.4135 (0%)	0.3936 (1%)
Meditron-7b	0.3334 (96%)	0.3044 (109%)
Qwen3-4B	0.4844 (1%)	0.4661 (1%)
Qwen3-8B	0.4972 (1%)	0.4757 (1%)

Table 5: Performance of LLMs after fine-tuning by Method II, with improvements relative to baseline.

I Prompt Templates for Guideline-to-QA Generation

To generate high-quality medical QA data from NCCN guideline figures, we design a set of structured prompts to guide large language models through three stages: (1) converting clinical guideline diagrams into structured JSON; (2) extracting all clinical decision paths from the tree; and (3)

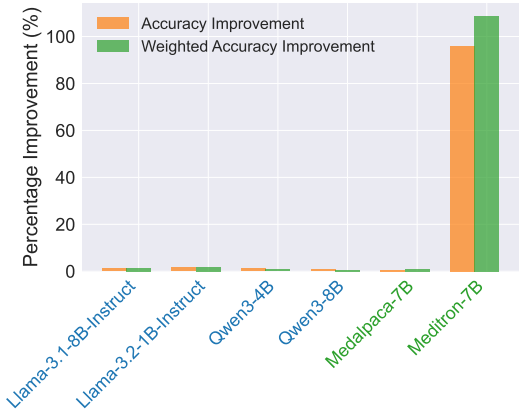


Figure 10: Improvement of accuracy and weighted accuracy for Method II

generating realistic patient vignettes and clinical questions based on each decision path. Below we provide a description and the exact text of each prompt used in this process.

I.1 Prompt: Convert Guideline Diagram to JSON Tree

This prompt asks the LLM to read a clinical decision tree (e.g., a screenshot of an NCCN guideline) and convert it into a machine-readable, hierarchically structured JSON format.

```
You are given a clinical decision tree diagram (as a
screenshot image) that outlines medical guidelines for
diagnosis or treatment.  Your task is to convert the decision
tree into a structured JSON format that precisely mirrors the
full logic and hierarchy presented in the figure.
* The JSON must accurately reflect all branching paths,
decision points, conditions, and outcomes.
* All treatment options, diagnostic steps, and relevant notes
(e.g.  footnotes or eligibility criteria) must be preserved.
* Maintain the exact hierarchical relationships and nesting,
so that the JSON could be used to programmatically reconstruct
the original tree.
* Use clear and descriptive key names based on the text in the
image.
* If treatments or conditions include multiple options or
logical conditions (e.g.  ‘‘A or B’’), represent them using
lists or nested structures as appropriate.
Example format:
{
"First relapse (morphologic or molecular)":  {
"Early relapse (<6 mo) after ATRA and arsenic trioxide (no
anthracycline)":  {
"Therapy":  [
"Anthracycline-based regimen as per APL-3",
"Gemtuzumab ozogamicin"
]
},
"Late relapse (≥6 mo) after arsenic trioxide-containing
regimen":  {
"Therapy":  [
"Arsenic trioxide ± ATRA ± (anthracycline or gemtuzumab
ozogamicin)"
]
```

```
}
},
"No remission":  {
"Next steps":  [
"Clinical trial",
"Matched sibling or alternative donor HCT"
]
}
}
```

I.2 Prompt: Extract All Decision Paths

This prompt asks the LLM to enumerate every possible decision path from the root to a leaf node in the structured guideline.

```
You are provided with an image of an NCCN clinical decision
tree guideline, outlining detailed medical instructions for
diagnosis and treatment.  Your task is to precisely list all
possible clinical decision paths depicted in the guideline.
Instructions:
* Represent each decision path as a Python list of strings.
* Each string must exactly match the text appearing in the
decision nodes, conditions, or treatment steps from the image,
without abbreviation, modification, or paraphrasing.
* Include every potential pathway from the initial decision
node down to each final leaf node.
Now, generate all possible paths.
```

I.3 Prompt: Generate a QA Sample from a Decision Path

This prompt instructs the model to generate a realistic patient vignette and a clinical question that follows the logic of a specific decision path.

```
You are provided with a clinical decision path derived from an
NCCN guideline, presented as an ordered list of decision nodes
(each node is a string).  Your task is to generate a realistic
patient vignette-a brief clinical case-including pertinent
medical history, timing of relapse, previous treatments,
test results, and clinical assessments required to match
precisely each node in the provided decision path from root
to leaf.  Conclude your vignette with a clinical question
asking explicitly about the appropriate next treatment step.
The correct answer must correspond exactly to the final node
of the provided path but should not be mentioned explicitly in
either the vignette or the question.

Decision Path:

  • "First relapse (morphologic or molecular)"
  • "Early relapse (<6 mo) after ATRA and arsenic trioxide (no
    anthracycline)"
  • "Therapy"

Formatting Instructions:

  • Present the entire vignette and concluding question as a
    single paragraph.
  • Do not reveal or imply the correct (leaf node) answer
    within the vignette or the question.
```

J Practical Implications for Clinical Use

MedGUIDE points to a pragmatic deployment recipe for medical LLMs as *clinician-in-the-loop, guideline-grounded decision support*:

- **Guideline Grounding.** Automatically retrieve the relevant decision tree at inference time and display the exact node-by-node path behind each recommendation, as demonstrated in our Method I results.
- **Transparent Rationale.** Explicitly link patient facts to decision nodes and output structured fields (e.g. recommended next step, contraindications, monitoring plans) suitable for integration into electronic health records (EHRs).
- **Uncertainty Gating.** When the model exhibits low confidence or ambiguous path alignment, it should defer surfacing missing information or escalating to clinician review rather than offering potentially unsafe recommendations.
- **Governance and Auditing.** Log the guideline version, model outputs, clinician overrides, and near-miss events to support ongoing safety monitoring, quality improvement, and regulatory compliance.

Overall, MedGUIDE is meant to *augment*, not replace, clinical judgment, and the benchmark's metrics (adherence, transparency, safe deferral) give concrete targets for trustworthy deployment.

K Extended Discussion and Design Considerations

K.1 Scope and Generalization

This version of MedGUIDE focuses on oncology mainly because of the availability and clarity of NCCN decision trees, which make it easier to map structured logic to model behavior. Besides, the overall pipeline (extracting decision logic, generating vignettes, and scoring model responses) does not assume anything oncology-specific. Most clinical specialties use similar decision trees, so we see MedGUIDE as a generalizable framework that can be extended to other domains like cardiology or endocrinology without major changes.

K.2 Use of Synthetic Vignettes

We chose to generate patient vignettes synthetically via prompting rather than extracting from EHRs. Our main considerations are: (1) privacy concerns, (2) the need to precisely control the match between vignettes and decision paths, and (3) the desire to cover the full decision space of each guideline. While synthetic text may lack some of the nuance of real-world clinical notes, we mitigate this by prompting from multiple models and applying strict quality filtering. The goal of our current work is not to perfectly mimic clinical language, but to create plausible, decision-relevant scenarios tied to ground-truth logic.

K.3 Robustness of Quality Filtering

To make the data as clean and consistent as possible, we apply a two-stage filtering process: a trained reward model checks for clinical validity, and a small ensemble of LLM-as-a-judge prompts scores textual clarity, internal logic, and safety. This hybrid approach helps us catch both subtle medical inconsistencies and broader language issues. While it is impossible to guarantee perfect quality at scale, this setup gives us a practical way to balance coverage and rigor without requiring full manual review of 8K samples.

K.4 Design Choice: Multiple-Choice Format

We stick with multiple-choice questions in this first release because under this setting, rigorous evaluation and quantification are possible. Moreover, our design of multiple-choice correctness based on the decision paths is tightly aligned with the structure of clinical decision trees and consistent with existing medical LLM applications. We fully recognize that real-world clinical decisions can be

more open-ended, but designing a reliable benchmark for open-ended guideline following is much harder: it is extremely challenging to define a metric to measure how well a model's open-ended answer follows the guideline. We see our current setup as a starting point, and plan to explore more flexible formats (e.g., chain-of-thought or free-text reasoning) in future versions of MedGUIDE.

K.5 Simple Methods for Guideline Grounding

We deliberately chose two lightweight methods: guideline-in-context prompting and fine-tuning, because these methods are straightforward. These are not meant to be state-of-the-art techniques, but rather easy baselines that help test whether models can actually benefit from structured guideline information. We expect more complex methods (such as structured planning, tool use, or retrieval-augmented generation) to show further improvements, but they are outside the scope of this paper and will be explored in future work.

K.6 Ethical and Release Considerations

All examples in MedGUIDE are synthetically generated, with no real patient data involved. We include filtering criteria for professionalism, internal consistency, and safety, and instruct models to avoid harmful or stigmatizing content during generation. The dataset is intended purely for research and benchmarking, not for clinical deployment. In the public release, we will also provide a best-practices usage document to help ensure the dataset is used responsibly and in alignment with its intended purpose.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's contributions: introducing the MedGUIDE benchmark, evaluating a diverse set of LLMs, analyzing correlations with existing benchmarks, and proposing improvement methods. These claims are consistently supported throughout the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section 6, acknowledging scope (cancer-specific guidelines), model coverage, and inference-time assumptions without overstating them.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper is empirical and does not include theoretical results or proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe model evaluation procedures, selection thresholds, quality criteria, and hyperparameters for fine-tuning. Code and data are made available.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The MedGUIDE dataset and code are publicly released via Hugging Face and anonymous repository links (see footnote in Section 2).

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Fine-tuning settings, learning rates, epochs, and dataset filtering procedures are described in Sections 3.3 and 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Due to the deterministic nature of most evaluated models (especially APIs) with temperature 0, we report accuracy without error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe API models used (inference only), and for fine-tuning we report model names, epochs, learning rates, and note resource constraints (Section 5).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All data is synthetic or derived from published clinical guidelines; we ensure no patient data or personal health information is used.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 6 briefly touches on clinical relevance and risks of model misuse. The broader impact of guideline adherence is implicit throughout the paper.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: MedGUIDE is derived from publicly available clinical protocols, and we only release model-generated synthetic data with no patient information.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All models are either public or commercial APIs used per their terms; citations are included for all datasets and models.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Human annotators (clinical experts) were involved in quality rating. Annotation details and demographics are provided in Appendix B.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Human annotators (clinical experts) were involved in quality rating. Annotation details and demographics are provided in Appendix B.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: All annotators were consenting medical professionals. No patient data or interventions were used. We obtained verbal or written agreement to participate in this research task in compliance with institutional research policy.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs were not used in the development of the core methods or experiments in this work. They were only used for non-scientific purposes such as writing assistance and figure plotting.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.