Prompt Optimization in Medical LLM Benchmarks

Ana Cunha, Geoffrey Gao, and Shahd Ibrahim

Technische Universität München, Munich, Germany

Abstract. Large Language Models (LLMs) are increasingly evaluated for healthcare applications, making reliable benchmarking essential. Yet benchmark outcomes may depend not only on the underlying model but also on the prompt and evaluation configuration. We investigate this issue on HealthBench, an open-ended benchmark of approximately 5,000 healthcare tasks evaluated with physician-developed rubrics. We compare five manually designed prompting strategies with Cost-Aware Prompt Optimization (CAPO) and additionally examine LLM-as-a-Judge robustness under different judge prompts and judge models. Our analysis focuses on how these experimental choices affect benchmark scores, model rankings, and evaluation consistency, with the goal of identifying practical considerations for more reproducible medical LLM benchmarking.

Keywords: medical LLMs · prompt optimization · HealthBench · LLM-as-a-Judge · benchmark robustness

1 Introduction

Large Language Models (LLMs) are increasingly used for medical question answering, patient communication, information retrieval, and decision-support tasks. As these systems move toward healthcare use, reliable evaluation is necessary for comparing models and selecting appropriate systems.

Traditional medical benchmarks such as MedQA and MedMCQA mainly assess factual knowledge through multiple-choice questions. HealthBench instead evaluates open-ended healthcare responses using physician-developed rubrics, allowing multiple clinically appropriate answers to receive credit. This setting better reflects realistic medical communication but also introduces additional experimental choices.

One such choice is the prompt. Changes in system instructions can alter response structure, detail, and safety behavior without changing model parameters. Benchmark scores may therefore reflect both model capability and prompt design. Prompt optimization offers a comparatively inexpensive way to explore this sensitivity, ranging from manually designed instructions to automated methods such as Cost-Aware Prompt Optimization (CAPO).

A second source of variability is the evaluator. HealthBench uses an LLM-as-a-Judge framework to score open-ended responses at scale. If scores change

with the judge prompt or judge model, benchmark conclusions may depend on the evaluation configuration itself. We therefore study both prompt optimization and judge stability within a unified benchmarking pipeline.

1.1 Contributions

- A reusable benchmarking pipeline supporting API-based and locally hosted LLMs.
- A comparison of five manual prompting strategies and CAPO on HealthBench.
- Judge stability analyses across alternative judge prompts and judge models.
- An analysis of the implications of prompt- and judge-sensitive evaluation for reproducible medical LLM benchmarking.

2 Related Work

Medical LLM evaluation has progressed from multiple-choice benchmarks such as MedQA and MedMCQA toward open-ended evaluation. HealthBench evaluates realistic healthcare tasks with physician-developed rubrics and LLM-based grading, enabling scalable assessment of free-text responses [2].

Prompt engineering can modify LLM behavior without retraining. Recent work has examined prompt optimization for medical benchmarks [1], while CAPO frames prompt selection as a cost-aware automated search problem [3]. In parallel, LLM-as-a-Judge evaluation has enabled large-scale scoring of open-ended outputs but raises questions about evaluator sensitivity. Our work connects these directions by examining prompt optimization together with judge prompt and judge model stability.

3 Methods

3.1 Benchmark and Pipeline

All experiments use HealthBench, which contains approximately 5,000 realistic healthcare tasks with structured evaluation rubrics. The pipeline separates response generation from scoring: a HealthBench question is combined with a selected prompt, sent to a generation model, and stored as JSON Lines (JSONL). Stored responses are then evaluated with the corresponding rubric by one or more LLM judges. Scores are aggregated for prompt comparison, category analysis, and judge-stability analysis. Separating generation and scoring allows outputs to be reused across evaluation configurations and supports interrupted runs through resume functionality.

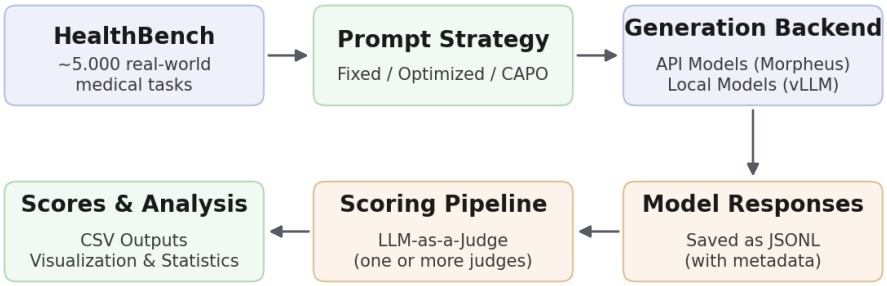

Fig. 1. Overview of the benchmarking pipeline from HealthBench input to prompt strategy, generation, LLM-based scoring, and analysis.

3.2 Models and Implementation

The framework supports both remote API inference and locally hosted inference through a unified backend. We use the API backend for the larger generation models in the main prompt comparison, because these models are the primary systems under evaluation and are accessed through a shared inference service. Local models are served with vLLM (a high-throughput LLM serving engine) and are used mainly for judge-stability experiments, where running several smaller evaluators side by side is more practical than relying on a single remote judge. The implementation additionally supports reusable outputs, resume functionality, configurable judge prompts, multiple judge models, and parallel workers.

Table 1. Models used in the benchmarking pipeline, with backend and role.

Model	Backend	Role
Ministral-3-14B-Reasoning-2512	API	Generation
gemma-4-31B-it	API	Generation
Qwen3.6-27B	API	Generation
qwen-35-35b-coding	API	Generation / Judge
`qwen2_5_1_5b`	Local	Generation / Judge
`gemma_2_9b_it`	Local	Generation / Judge
`llama_3_1_8b`	Local	Generation / Judge
`qwen2_5_7b`	Local	Generation / Judge

3.3 Prompt Optimization

We evaluate five manual strategies chosen to probe distinct axes of medical response behavior rather than to exhaustively search prompt wording. Baseline provides an unmodified reference so that later gains or losses can be attributed to the added instruction. Safety First emphasizes cautious advice and escalation

when emergency care may be needed, reflecting a common priority in healthcare settings. Physician Role assigns a clinician persona to test whether role framing improves clinical framing and accuracy. Structured Output encourages an organized recommendation-then-reasoning format, testing whether explicit structure helps or displaces content rewarded by HealthBench rubrics. Direct/Concise prioritizes brevity to test whether shorter answers remain competitive when rubrics also reward safety disclaimers and thorough explanations.

We additionally evaluate CAPO as an automated prompt-optimization approach, because manual strategies alone cannot show whether search over candidate instructions yields further gains beyond hand-designed prompts. CAPO searches candidate prompts on a development set and refines promising candidates using mutation and crossover. For each generation model, we run 10 optimization steps on a 300-example development set drawn from HealthBench. Candidate scoring during optimization uses the same model for both generation and evaluation (self-grading), which reduces the evaluation budget but may introduce self-preference bias relative to an independent judge; we return to this point in the limitations. We keep the search relatively short and cost-aware (a limited number of optimization steps, block-wise evaluation, and a mild length penalty) to control evaluation budget while discouraging unnecessarily long prompts. Final evaluation is performed separately from the development search to limit optimization-specific overfitting.

3.4 LLM-based Evaluation and Judge Stability

For each generated response, the judge receives the original medical question, the model response, and the corresponding HealthBench rubric. We evaluate two forms of stability by changing one factor at a time, so that differences can be attributed to either the judge instruction or the judge model rather than to both jointly. For prompt stability, the judge model is fixed while the judge instruction is varied across alternative grading emphases (for example, stricter rubric adherence, clinical-reasoning focus, or semantic equivalence). For model stability, the judge prompt is fixed while the judge model is changed across the local evaluators in Table 1. Pearson correlation is used to compare score patterns between configurations because our question is whether rankings and relative score structure are preserved, not only whether mean scores remain similar.

4 Results

4.1 Prompt Strategy and Model Performance

Prompt choice changed HealthBench scores across the three models shown in the prompt-strategy comparison. Safety First improved scores relative to baseline for all three models, while Direct/Concise produced the lowest scores across all models in this comparison. We hypothesize that brevity-forcing instructions suppress safety disclaimers, referrals, and explanatory detail that HealthBench

rubrics reward, and that the safety instruction aligns model behavior more directly with the physician-developed rubric criteria.

CAPO-optimized prompts produced the highest scores shown in the figure, reaching approximately 0.58 for Qwen and Gemma and 0.47 for Ministral. Because CAPO candidates were scored using self-grading during optimization while the manual strategies were evaluated with an external judge, scores across these two configurations should be interpreted with care rather than compared as directly equivalent numbers. Nonetheless, the CAPO-optimized prompt outperformed every manual strategy under both evaluation settings.

Notably, all three CAPO-optimized prompts independently converged on safety-first language — emphasizing patient safety, epistemic limits, and appropriate referrals — despite starting from different seed prompts and being optimized separately per model. This convergence suggests that CAPO is learning to align with HealthBench rubric criteria rather than discovering arbitrary wording, and it provides a behavioral explanation for why safety-oriented instructions consistently outperform other manual strategies.

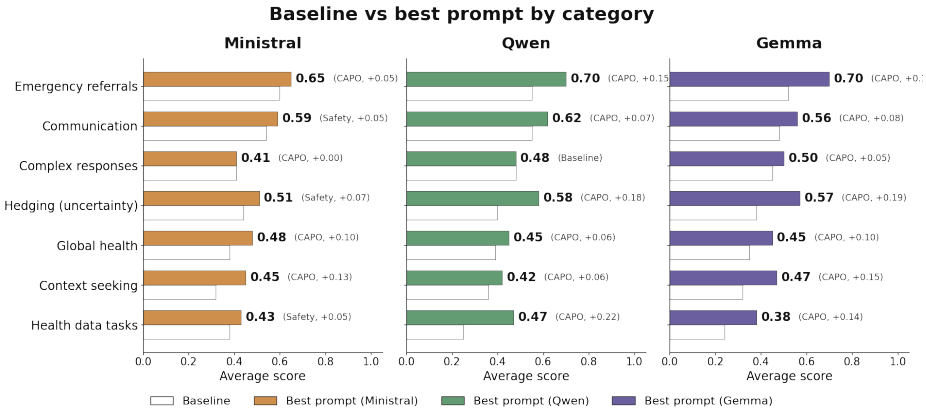

Fig. 2. HealthBench score by prompt strategy and generation model. Values are reported for the evaluated configurations shown.

4.2 Effect on Model Ranking and Categories

The relative ordering of models changed across prompt configurations. In the baseline comparison, Ministral scored highest among the three displayed models, whereas the CAPO-optimized configuration shown in Fig. 2 placed Qwen and Gemma above Ministral. Category-wise results further show that prompt effects are heterogeneous: the best prompt differs by model and category, and some categories show only small gains over baseline (Fig. 3). We hypothesize that categories with smaller gains are those where the model's baseline already

matches the rubric priorities, so additional prompting mainly rearranges wording rather than adding credited content.

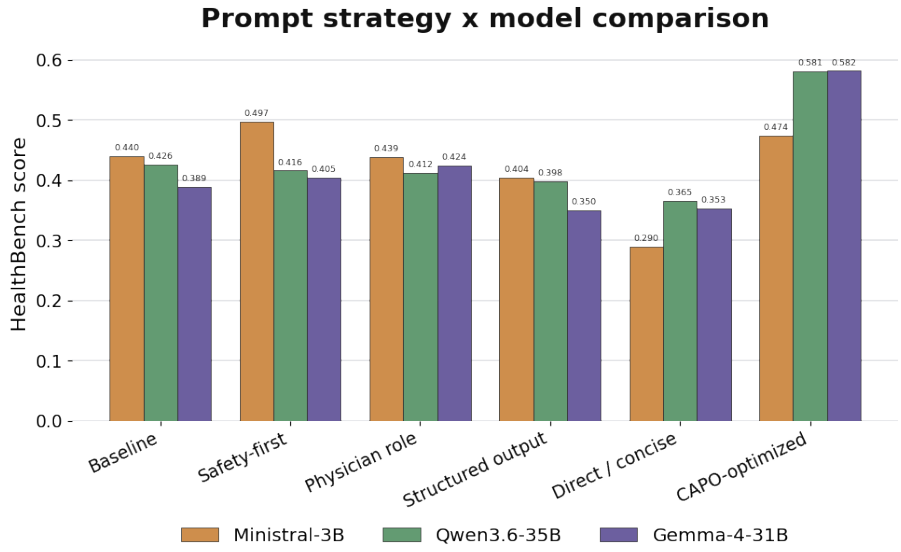

Fig. 3. Baseline versus the best evaluated prompt by HealthBench category for each model.

4.3 Judge Prompt Stability

Judge-prompt correlations were consistently positive and relatively high. Across the five prompt variants, pairwise Pearson correlations ranged from approximately 0.71 to 0.80 (Fig. 4, left). Thus, changing the judge instruction affected individual scores, but the overall score patterns remained broadly similar.

4.4 Judge Model Stability

Changing the judge model produced substantially lower agreement. Off-diagonal Pearson correlations in the model comparison ranged from approximately 0.12 to 0.60 (Fig. 4, right). The weakest agreement occurred between `qwen2_5_1_5b` and `llama_3_1_8b` (0.12), while the strongest shown was between `gemma_2_9b_it` and `qwen2_5_7b` (0.60). This variation is markedly larger than that observed across judge prompts. We hypothesize that judge models differ more in how they interpret open-ended medical criteria—especially borderline or partially satisfied rubrics—than alternative phrasings of the same judging instructions do.

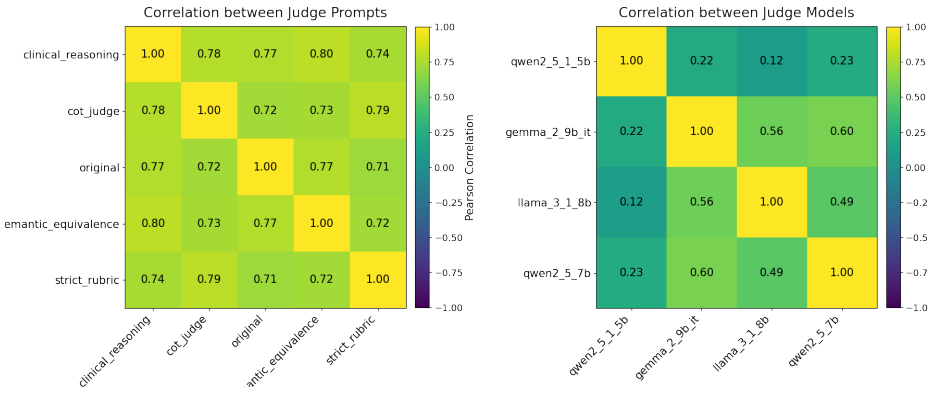

Fig. 4. Judge stability. Left: Pearson correlation between judge-prompt variants. Right: Pearson correlation between judge models.

5 Discussion

The experiments show that prompt choice is a meaningful experimental variable in medical LLM benchmarking. The same model can receive different HealthBench scores under different instructions, and the best manual strategy is not identical across models. The observed ranking change across prompt configurations is especially relevant when benchmark scores are used to compare candidate models: a single fixed prompt may favor one model's instruction-following behavior over another's. In particular, we hypothesize that CAPO improves scores by discovering instructions that better align response style with HealthBench rubric criteria, rather than by changing underlying medical knowledge.

The judge-stability results add a second source of uncertainty. Judge prompts produced correlations of roughly 0.71–0.80, whereas judge-model correlations ranged from about 0.12 to 0.60. Within the evaluated settings, judge-model choice therefore had a larger effect on score consistency than judge-prompt wording. For medical benchmarks, where small score differences may influence conclusions about model quality, reporting the judge configuration and testing robustness across evaluators is important.

These findings do not imply that prompt optimization changes a model's underlying medical knowledge or that the highest benchmark score necessarily identifies the safest model for deployment. Rather, they show that reported performance is conditional on the prompt and evaluator used. Benchmark studies should therefore document these settings and avoid treating a single score as an absolute measure of clinical capability.

5.1 Limitations and Future Work

The study is limited to HealthBench and to the generation and judge models evaluated in this project, so the observed patterns should not be generalized

to all medical benchmarks or model families. Evaluation is LLM-based rather than a new human-expert assessment, and benchmark improvements should not be interpreted as evidence of improved clinical outcomes. The CAPO experiments use self-grading during candidate scoring, which may inflate scores relative to an independent evaluator; a cross-grader validation would strengthen these results. Future work should validate the findings on additional medical benchmarks, include a broader range of judge models, investigate multi-judge or ensemble evaluation to reduce dependence on a single evaluator, and re-evaluate CAPO-optimized prompts using the same external judge as the manual strategy comparison.

6 Conclusion

We investigated prompt optimization and judge stability in HealthBench. Prompt strategy changed model scores and relative rankings, while judge-model choice produced substantially greater score variation than judge-prompt wording in the evaluated configurations. These results support transparent reporting of prompts and judge settings, together with robustness analyses, when using open-ended medical LLM benchmarks.

References

1. Aali, A., et al.: Prompt optimization improves robustness of language model benchmarks for medical tasks. In: ML4H 2025 Findings Track (2025)
2. Arora, R.K., et al.: HealthBench: Evaluating large language models towards improved human health. arXiv preprint arXiv:2505.08775 (2025)
3. Zehle, T., Schlager, M., Heiß, T., Feurer, M.: CAPO: Cost-aware prompt optimization. In: Proc. 4th Int. Conf. Automated Machine Learning. PMLR, vol. 293, pp. 18/1–45 (2025)