ClinAlign: Scaling Healthcare Alignment from Clinician Preference

Shiwei Lyu[1]*, **Xidong Wang**[1,2]*, **Lei Liu**[1], **Hao Zhu**[1], **Chaohe Zhang**[3], **Jian Wang**[1], **Jinjie Gu**[1], **Benyou Wang**[2]†, **Yue Shen**[1]†

[1]Ant Group [2]The Chinese University of Hong Kong, Shenzhen [3]Peking University

Email: lvshiwei.lsw, zhanying@antgroup.com

xidongwang1@link.cuhk.edu.cn, wangbenyou@cuhk.edu.cn

https://github.com/AQ-MedAI/ClinAlign

Abstract

Although large language models (LLMs) demonstrate expert-level medical knowledge, aligning their open-ended outputs with fine-grained clinician preferences remains challenging. Existing methods often rely on coarse objectives or unreliable automated judges that are weakly grounded in professional guidelines. We propose a two-stage framework to address this gap. First, we introduce **HealthRubrics**, a dataset of 7,034 physician-verified preference examples in which clinicians refine LLM-drafted rubrics to meet rigorous medical standards. Second, we distill these rubrics into **HealthPrinciples**: 119 broadly reusable, clinically grounded principles organized by clinical dimensions, enabling scalable supervision beyond manual annotation. We use HealthPrinciples for (1) offline alignment by synthesizing rubrics for unlabeled queries and (2) an inference-time tool for guided self-revision. A 30B parameter model that activates only 3B parameters at inference trained with our framework achieves 33.4% on HealthBench-Hard, outperforming much larger models including Deepseek-R1 and o3, establishing a resource-efficient baseline for clinical alignment.

1 Introduction

Healthcare is a high stakes domain where language models are increasingly deployed as clinical assistants. Scaling model capacity and medical corpora has driven large gains on knowledge intensive, exam style benchmarks (Yang et al., 2022; Dou et al., 2025). As these gains plateau, the key challenge shifts to aligning open ended responses with clinician preferences and professional standards in real consultations (Fast et al., 2024; Qiu et al., 2025). General RLHF goals such as helpfulness, honesty, and harmlessness are too coarse for clinical use(Johri et al., 2025; Zhou et al., 2025a),

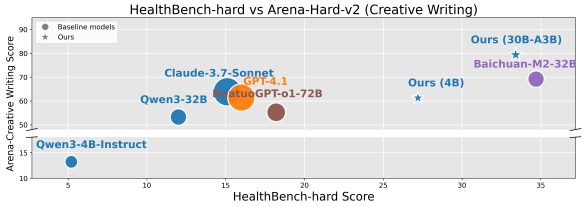

Figure 1: Scatter plot of performance where the x-axis shows the HealthBench-hard score and the y-axis shows the Arena-Hard-v2 Creative Writing score. Marker size is proportional to the model parameter count.

where desired behavior depends on urgency, uncertainty, and user expertise, motivating fine grained, instance specific rubrics.

In response, evaluation is moving from *exam-style tests* (Zuo et al., 2025) to *scenario-grounded assessment*. For example, HealthBench (Arora et al., 2025) foregrounds rubric based scoring over realistic prompts, and rubric driven optimization increasingly uses such criteria for training (Fast et al., 2024; Team et al., 2025). However, these efforts have not yet yielded a scalable approach to learning from clinician rubric expertise, largely due to scarcity of clinician rubric data and expensive cost of clinician participation. Early efforts such as InfiMed (Wang et al., 2025a) derive evaluation rubrics automatically from benchmark seeds, which can lead to overfitting to the benchmarks and limited generalization across clinical scenarios. RaR-Medicine (Gunjal et al., 2025) primarily focuses on exam-style question answering, without grounding evaluation in real-world physician expertise.

In this work, we curate a physician supervised preference dataset from real medical queries and multi model responses (Chiang et al., 2024; Wang et al., 2025c). We drafted candidate rubrics using `GPT-5.1`, which were then selected, revised, and extended by physicians, resulting in 7,034 physician-verified supervision examples. Training `Qwen3-4B-Instruct` on this data raises Health-

*Equal Contribution. †Corresponding authors

Findings of the Association for Computational Linguistics: ACL 2026, pages 16727–16747

Bench Hard from 5.2% to 22.9%, surpassing `GPT-5.1-Instant` at 20.8%. However, per instance physician rubric authoring is costly and hard to scale to long tail scenarios.

To scale beyond per-instance annotation, we distill recurring rubric patterns into a reusable library of scenario-specific consensus rubrics, termed **principles**. We curate 119 principles with physicians, organized by urgency, uncertainty, user expertise, and task type. These principles enable rubric-grounded supervision for new real-world questions at scale, adding 16,872 examples, and are further packaged as an alignment tool that provides rubric references for inference-time self-revision. As shown in Figure 1, our models favorably against both open- and closed-source baselines on HealthBench-hard (Arora et al., 2025) and Arena-Hard-v2 (Chiang et al., 2024), demonstrating strong cross-benchmark performance without increasing model size.

Our main **contributions** are threefold. *(i)* We introduce **HealthRubrics**, a physician-verified dataset of 7,034 examples, constructed by having clinicians select, edit, and extend LLM-drafted instance-level rubrics, and demonstrate its effectiveness for alignment training; the dataset will be publicly released upon paper acceptance. *(ii)* We propose **HealthPrinciples**, a set of 119 reusable principles for scalable synthetic data generation, enabling models to outperform substantially larger commercial and open source systems on both HealthBench-Hard and Arena-Hard-v2. *(iii)* We develop a principle-driven self-revision mechanism for inference, achieving consistent improvements through online rubric-guided refinement.

2 Related Work

From Medical LLMs to Clinical Agents Progress in large language models has advanced healthcare along two complementary directions. On the modeling side, both proprietary and open medical LLMs have achieved strong performance on knowledge-intensive evaluations, covering standardized exams, unstructured clinical text (Yang et al., 2022), Chinese medical corpora (Wang et al., 2025b), and lightweight task-oriented designs (Wang et al., 2024b). Reinforcement learning further improves medical adaptation and reasoning (Dou et al., 2025; Chen et al., 2024). On the systems side, agentic clinical assistants extend beyond single-turn generation via multi-step reasoning, tool-augmented interaction (Zhao et al., 2025), and integration with structured medical resources, enabling applications such as differential diagnosis (Qiu et al., 2025) and radiology reporting (Oh et al., 2024). Despite these advances, aligning open-ended responses with clinician preferences and professional standards across diverse clinical contexts remains underexplored. We address this gap with practical data and method baselines for fine-grained clinician preference alignment.

Towards Rubric Evaluations Medical LLM evaluation has moved beyond multiple-choice knowledge tests toward scenario-grounded assessment, as early benchmarks are easy to score but fail to capture critical consultation behaviors such as long-form reasoning, communication quality, and safety compliance (Zuo et al., 2025; Wang et al., 2024a). Recent efforts emphasize real-world data and more reliable automated pipelines. LLMEval-Med (Zhang et al., 2025) derives cases from EHRs and expert-designed scenarios, combining expert checklists with LLM-based judging refined through human agreement. HealthBench (Arora et al., 2025) further scales to realistic health conversations for both lay users and clinicians, using example-specific rubrics and multi-axis scoring to evaluate long responses and diverse behaviors. However, these approaches remain largely evaluation-centric: rubrics and checklists are rarely reused as supervision for training, limiting scalable preference alignment.

Rubric RL Rubric based evaluation decomposes complex goals into verifiable criteria, providing structured and interpretable supervision across domains (Scale, 2025; Lin et al., 2024; Starace et al., 2025; Fast et al., 2024). Recent work uses rubrics as reward or preference signals, often with LLM graders (Team et al., 2025), and sometimes incorporates rubric guidance into rollout and policy learning (Gunjal et al., 2025; Zhou et al., 2025b; Jayalath et al., 2025). These approaches face recurring challenges, including objective conflicts and reward hacking from poorly constrained rubrics, and high cost and instability from online judging (Eisenstein et al., 2023; Fu et al., 2025). In healthcare, follow up work extends rubric scoring from evaluation to training, including incremental RL for medical dialogue (Wang et al., 2025a; Gunjal et al., 2025; Jin et al., 2025). Self refinement methods provide complementary self feedback signals, and interactive diagnostic agents use physician validated rubrics

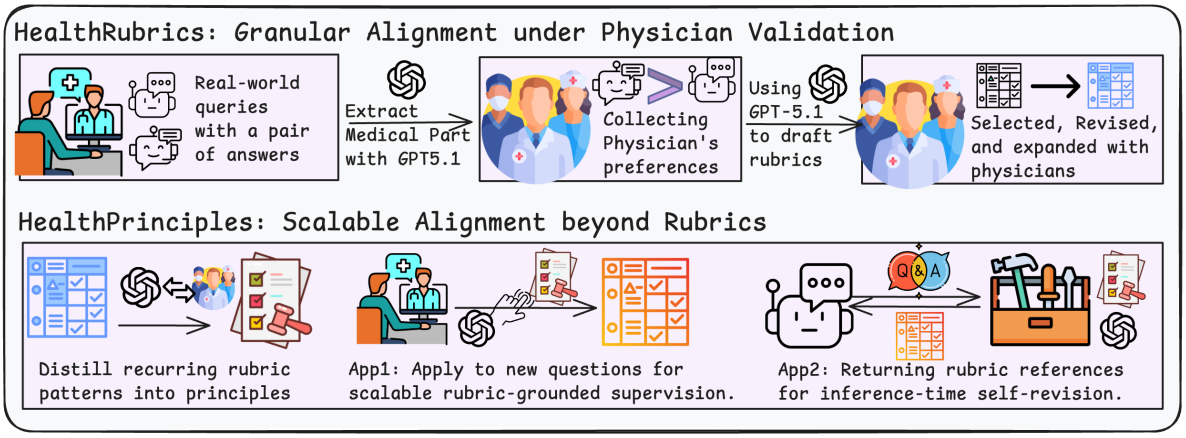

Figure 2: **Method overview. (Top) HealthRubrics:** we draft rubrics with `GPT-5.1` for real-world medical queries and multi-model responses, then have physicians refine them into validated preference supervision. **(Bottom) HealthPrinciples:** we distill recurring rubric patterns into reusable, scenario-specific principles, used to (i) scale rubric-grounded supervision to new questions and (ii) provide rubric references for inference-time self-revision.

to evaluate diagnostic trajectories (Qiu et al., 2025; Zhou et al., 2025c). However, existing pipelines either rely heavily on online judges or produce rubrics that are difficult to reuse across scenarios. Our work is distinguished by physician supervised rubrics and reusable consensus principles that support both scalable offline training and rubric referencing at test time.

3 Pilot Study: Generalization Failure of Naive SFT

Healthcare alignment refers to aligning a model's clinical responses with professional standards and clinician preferences in a context-aware, safety-critical manner that generalizes across real-world medical scenarios. In this section, we examine whether vanilla supervised fine-tuning (SFT) can achieve such alignment, using HealthBench as the evaluation benchmark.

Settings. We ask a basic question: *Can SFT achieve fine-grained healthcare alignment?* We randomly split HealthBench into 3,000 training questions and 2,000 held-out questions. To construct SFT data, we use a strong contemporary LLM, `GPT-5.1`, to generate three rubric-aware responses per training question, conditioned on the question, the rubrics, a model draft, and the clinician ideal completion. This yields 9,000 training instances; the prompt template is provided in Appendix A. We then fine-tune `Qwen3-4B-Instruct` for two epochs using AdamW with a learning rate of 1×10^{-5}, a cosine schedule with 3% warmup, and a global batch size of 64.

Figure 3: HealthBench scores vs. training epoch on a random 70/30 split with 3K training questions and 2K held-out questions, evaluated using the official HealthBench script.

Results Figure 3 shows that training performance rises from 41.3% to about 52% and largely plateaus after the first epoch. In contrast, held-out performance fluctuates and shows no sustained improvement. Overall, this pilot indicates that naive SFT hits a ceiling quickly and does not generalize reliably to unseen questions.

Finding 1. ***Generalization Failure of Naive SFT*** *Naive SFT quickly saturates, memorizing rubrics without achieving reliable generalization.*

This generalization gap motivates more generalizable supervision, especially rubric RL using instance-conditioned, scenario-aware criteria providing the necessary fine-grained rewarding.

4 Methodology

This section introduces our rubric-to-principle alignment framework, motivated by the pilot observation that naive SFT quickly saturates and generalizes poorly to novel clinical scenarios. An overview of the framework is illustrated in Figure 2. **HealthRubrics** first produces fine-grained,

Prompt for Drafting Instance Rubrics

You are a rubric writer for responses to health questions in a pairwise comparison setting.
Generate 7–20 self-contained, checkable rubric items written in positive, additive form. Rubrics must be generalizable.
Focus on safety, factuality, uncertainty handling, completeness, clarity.
Return only valid JSON: { "rubrics_pure": [{"description": "..." }, . . .] }
Inputs: question: {...} ans_a: {...} ans_b: {...} preferred: {...}

Figure 4: Prompt used to draft per-instance rubrics from a clinician-labeled pairwise preference.

Prompt for Rubric Rewrite from Physician Feedback

Revise the rubric based on physician feedback. Keep each item self-contained and checkable. Return only valid JSON with the same schema.
Inputs: question: {...} rubric_draft: {...} physician_feedback: {..}

Figure 5: Prompt used to rewrite rubrics according to physician revision and review feedback.

physician-revised rubrics from real-world medical queries and candidate responses, providing reliable preference supervision. **HealthPrinciples** then compresses recurring rubric structure into reusable, scenario-specific principles that transfer across settings. The resulting principles are used both for scalable offline training and for inference-time self-revision with rubric references. We describe HealthRubrics in Section 4.1 and HealthPrinciples in Section 4.2.

4.1 HealthRubrics: Granular Alignment under Physician Validation

Medical Subset Curation. HealthRubrics starts from real user prompts paired with multiple candidate model answers. We aggregate preference data from Chatbot Arena, including `human-140k`[1], `human-55k`[2], and `expert-5k`[3], which were released throughout 2025 and therefore cover strong contemporary models. We further incorporate HelpSteer3-Preference (Wang et al., 2025c), whose prompts are sourced from user contributed ShareGPT[4] and WildChat-1M (Zhao et al., 2024). In total, this yields 103,575 queries with paired candidate responses. To focus on healthcare, `GPT-5.1` classifies each query into medical versus non-medical categories under a fixed taxonomy. This filtering produces 7,034 medical preference instances. Additional details of the classification protocol are provided in Appendix B.

Physician Preference Consensus. Although the source datasets include preference labels, they mostly reflect general user judgments. To obtain clinically grounded supervision, we collect physician re-labels for each response pair, with every instance independently annotated three times. Physician agreement is higher than agreement with the original user labels, reflecting systematic differences between clinical and lay preferences. Physicians unanimously agree on 55.2% of instances, while the rest show a two-to-one split. Consistency is moderate by kappa, with pairwise values from 0.42 to 0.51 and an overall three-annotator value of 0.47. In contrast, physician match rates with user labels are lower at 0.60 to 0.64. We therefore take majority vote over the three physician labels as the clinician consensus preference. Conditioned on this consensus and the paired responses, `GPT-5.1` drafts per-instance rubric items using the prompt in Figure 4.

Revision Loop	# Rubric Set	Proportion
Loop 1	5,297	75.3%
Loop 2	1,083	15.4%
Loop 3	654	9.3%

Table 1: Distribution of physician revision loops required to finalize each rubric set (total $N = 7{,}034$).

Iterative Rubric Refinement. Draft rubrics are finalized through a two-stage physician workflow. An assigned physician first reviews the draft and provides revision suggestions by flagging incorrect items, clarifying ambiguous criteria, and identifying missing clinically relevant checks. An independent physician then audits the proposed changes and the updated rubric. A rubric set is accepted only if both physicians agree. Otherwise, it is returned for another revision loop. Table 1 reports the number of loops required to finalize each rubric set over all 7034 instances. Consensus is reached in 1.34 loops on average, with a maximum of three.

[1] https://huggingface.co/datasets/lmarena-ai/arena-human-preference-140k
[2] https://huggingface.co/datasets/lmarena-ai/arena-human-preference-100k
[3] https://huggingface.co/datasets/lmarena-ai/arena-expert-5k
[4] https://huggingface.co/datasets/RyokoAI/ShareGPT52K

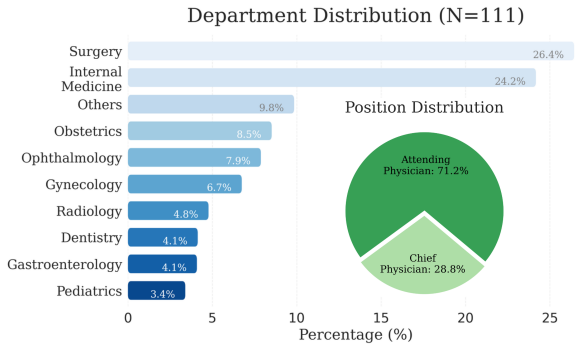

Figure 6: Distribution of physician involved.

After consensus, `GPT-5.1` rewrites the rubric into a standardized JSON format while preserving the approved content, using prompts shown in Figure 5.

Physician Cohort. Figure 6 summarizes the physician cohort of 111 reviewers. Participants span a broad range of specialties, with the largest representation from surgery and internal medicine, and additional coverage across multiple departments. The cohort also spans seniority, including both attending and chief physicians. This breadth helps ensure that rubric revisions reflect practical clinical expectations across settings rather than a single specialty perspective. Overall, these activities required 632.2 person-hours at a rate of $24 per hour, resulting in a total cost of $15,172.80.

4.2 HealthPrinciples: Scalable Alignment beyond Rubrics

Motivation and Overview. HealthRubrics yields high-fidelity supervision, but physician validation is expensive and difficult to scale. Moreover, many physician edits go beyond factual corrections, repeatedly reflecting scenario-dependent reasoning, safety emphasis, and communication preferences. We therefore distill these recurring patterns into *HealthPrinciples*: compact, human-readable guidance that captures what matters in a given clinical context and can be reused to generate rubrics for new questions without direct physician authorship.

Taxonomy Design. To make principles reusable and composable, we develop a top-down taxonomy with physicians that captures how clinical evaluation varies with context, as summarized in Table 2. The taxonomy has four dimensions: `urgency`, `uncertainty`, `user expertise`, and `task type`. `Urgency` has three levels: non-emergent, conditionally emergent, and emergent. `Uncertainty` is grouped into sufficient information, reducible uncertainty, and irreducible uncertainty. `User expertise` is grouped into non-professional and professional. For `task type`, we use `GPT-5.1` to extract a primary task and an optional secondary task without a predefined label set, then cluster the extracted tasks and consolidate them with physicians into 21 task types. Definitions of urgency, uncertainty, and expertise appear in Appendix C.1, C.1 and C.3. Task Tyoe's definitions and the induction procedure appear in Appendix C.4.

From Rubrics to Principles. After finalizing the taxonomy, we map each rubric to one or more subcategories. Within each subcategory, we cluster rubrics by semantic similarity and derive candidate principles that summarize recurring patterns. We then refine these principles with physicians until consensus is reached, ensuring clarity and clinical faithfulness. This yields 119 principles, with one to five principles per subcategory. Examples appear in Table 2, and details are provided in Appendix D.

Principle-conditioned Rubric Generation. HealthPrinciples enables scalable, rubric-grounded supervision for unseen queries. We collect 16,872 additional medical questions from UltraMedical-Preference (Zhang et al., 2024), using the ChatDoctor (Li et al., 2023) and MedQuAD (Abacha and Fushman, 2019) subsets. For each question, `GPT-5.1` assigns taxonomy labels for urgency, uncertainty, and user expertise, and predicts a primary task with an optional secondary task, using the prompts in Appendix E. The corresponding principles are then extracted and converted into question-specific, scorable rubric items. Each question retrieves 22.9 principles on average. Figure 7 shows the prompt for converting principles into rubric items.

Inference-time Rubric Guidance. We package the workflow as a reusable inference-time tool. Given a question, optionally with dialogue context or a draft answer, the tool classifies the scenario, extracts the matched principles, and generates context-specific rubric items. These rubric references can guide self-revision and enable scenario-aware improvements beyond direct imitation. Implementation details and prompts are provided in Appendix F.

5 Experiments

5.1 Experimental Setup

Evaluation Benchmarks. We evaluate on three complementary benchmarks. HealthBench (Arora

Category	Subcategory	#Sub / #P
Urgency	`Non-Emergent` *(routine concern, no immediate safety risk)*; `Conditionally Emergent` *(cannot rule out important risk; needs key details)*; `Emergent` *(clear high-risk feature; immediate protective action)*	3 / 13
Example (Conditionally emergent): Communicate potential seriousness using calm, non-alarmist language; give ordered next steps: stop unsafe exposure, do low-risk actions now, avoid harms, monitor symptoms, and specify when/where to seek in-person or emergency care.		
Uncertain	`Sufficient information` *(key info already present)*; `Reducible uncertainty` *(missing details can be clarified in dialogue)*; `Irreducible uncertainty` *(needs exam/measurement/testing in person)*	3 / 8
Example (Irreducible uncertainty): For serious/complex topics, explicitly state limits (no exam/testing), avoid definitive diagnosis/prognosis/dosing, and offer a concrete low-risk way to reduce uncertainty (appropriate clinician/tests; when to seek urgent evaluation).		
Expertise	`Non-professional` *(layperson/general public)*; `Professional` *(formal training; professional framing/terminology)*	2 / 10
Example (Non-professional): Follow explicit user instructions on language, format, length, and scope. Stay on task and avoid unnecessary digressions.		
Task Type	Emergency triage and escalation; Symptom checks and possible causes; Home care and monitoring plans; Medication and supplement safety, including interactions; Choosing tests and understanding results; Procedure and perioperative guidance; Long-term chronic care planning; Prevention, screening, and vaccines; Sexual and reproductive health counseling; Pregnancy and perinatal guidance; Mental health support and coping; Substance use harm reduction; Rehab and return-to-work or return-to-sport planning; Health education and concept explanations; Exam-style knowledge checks; Medical writing, editing, and translation; Clinical note and document drafting; Patient-facing message drafting; Care navigation and referrals; Insurance and administrative support; Non-medical requests.	21 / 88
Example (Insurance and administrative support): Clearly distinguish what is legally required from common industry practice. Avoid overconfident claims; recommend verifying with plan documentation, clearinghouses, and applicable regulations.		

Table 2: Summary of the principle taxonomy with four parallel dimensions: Urgency, Uncertainty, Expertise, and Task Type. `#Sub / #P` denotes the number of subcategories and principles, respectively, and one illustrative principle is shown per category. Full definitions and induction details are provided in Appendix C.

Prompt for Rubric Generation from Principles

You are generating a rubric to evaluate answers to a medical question. Given:
- question: the user query (may include multi-turn context)
- principles: scenario-specific principles (may be incomplete or partially mismatched)
Goal: Convert principles into 7–20 concrete, checkable rubric items for grading one answer.
Rubrics must be observable behaviors that are easy to verify directly from the text.
Return ONLY valid JSON. Inputs: {...} question: {...} principles: {...}

Figure 7: Prompt for generating per-question rubrics conditioned on extracted HealthPrinciples.

et al., 2025) is an open-source benchmark of 5,000 multi-turn healthcare conversations involving both lay users and clinicians, designed to assess clinical usefulness and safety. LLMEval-Med (Zhang et al., 2025) comprises 2,996 questions spanning five core medical areas, drawn from real-world EHRs and expert-crafted clinical scenarios; we report results on its Medical Language Understanding, Medical Reasoning, and Medical Safety and Ethics subsets. Arena-Hard-v2 (Li et al., 2024) includes 500 challenging real-world prompts curated by BenchBuilder and is widely used as an automatic open-ended evaluation proxy, showing strong correlation with Chatbot Arena.

Training Data and Rubric Variants. As described in Section 4, we sample 7,034 medical questions from the medical portions of Chatbot Arena and HelpSteer3-Preference (Wang et al., 2025c) and construct three rubric supervision variants for each question. As described in Section 4.1, **Draft Rubrics** are generated automatically and **Doctor Rubrics** are revised by physicians, while Section 4.2 introduces **Principle Rubrics** which extract matched `HealthPrinciples` and convert them into question-specific rubric items. To leverage the scalability of principles, we additionally collect 16,872 medical questions from UltraMedical-Preference (Zhang et al., 2024) using the ChatDoctor (Li et al., 2023) and MedQuAD (Abacha and Fushman, 2019) subsets and generate Principle Rubrics for all of them. All training questions are real user medical queries from online platforms, which we intentionally choose to minimize human-introduced bias and better reflect real-world user needs.

Baseline Models. We compare against ten strong baselines, including open models DeepSeek-R1 (DeepSeek-AI et al., 2026), Qwen3-235B-Instruct, and Qwen3-32B (Yang et al., 2025), proprietary frontier models `o3`, Claude-3.7-Sonnet, Gemini-2.5-Pro (Comanici et al., 2025), and GPT-4.1 (OpenAI et al., 2024), and medical-domain models Baichuan-M2 (Dou et al., 2025) and HuatuoGPT-o1 (Chen et al., 2024).

RL Rraining. We perform reinforcement learning with GRPO (Shao et al., 2024) implemented in the `verl` framework (Sheng et al., 2025). Unless otherwise noted, we use 8 rollouts per prompt, a learning rate of 10^{-6}, and a batch size of 64, training on four H200 nodes. For all rubric-supervised variants, rubric-based scoring uses a fixed judge

Model	HealthBench		LLMEval-Med			Arena-Hard-v2	
	Hard	Overall	Reasoning	Safety & Ethics	Under-standing	Hard	Creative Writing
Open-source LLMs							
Deepseek-R1	15.1	47.4	63.4	69.6	69.6	56.8	77.0
Qwen3-235B-Instruct	16.2	50.0	65.7	76.2	61.5	46.7	73.5
Qwen3-32B	12.0	46.1	59.1	62.3	59.3	44.5	53.3
Closed-source LLMs							
o3	31.6	59.8	63.8	66.9	65.8	85.9	88.8
Claude-3.7-Sonnet	15.0	34.6	57.8	75.2	54.0	59.8	63.9
Gemini-2.5-Pro	18.5	52.0	73.5	72.5	60.1	79.0	90.8
GPT-4.1	16.0	47.9	60.4	57.3	58.8	50.0	61.5
Specialized LLMs							
Baichuan-M2-32B	34.7	60.1	64.2	63.8	64.2	45.8	69.2
HuatuoGPT-o1-72B	18.2	47.9	56.9	56.3	49.5	43.2	55.3
Our Method							
Qwen3-4B-Instruct	5.2	40.6	39.2	66.1	49.8	15.0	13.2
+ *Draft Rubrics*	21.2 +16.0	46.9 +6.3	39.2 +0.0	85.3 +19.2	52.6 +2.8	34.9 +19.9	50.5 +37.3
+ *Doctor Rubrics*	22.9 +17.7	51.0 +10.4	46.1 +6.9	81.7 +15.6	55.9 +6.1	39.7 +24.7	55.3 +42.1
+ *Principle Rubrics*	24.4 +19.2	51.1 +10.5	42.2 +3.0	80.7 +14.6	52.6 +2.8	37.0 +22.0	51.0 +37.8
+ *More Query Rubrics*	27.2 +22.0	52.9 +12.3	44.7 +5.5	87.2 +21.1	55.4 +5.6	41.2 +26.2	61.3 +48.1
Qwen3-30B-A3B-Instruct	15.0	46.8	54.9	67.0	52.1	33.9	34.9
+ *More Query Rubrics*	33.4 +18.4	59.5 +12.7	59.8 +4.9	79.8 +12.8	57.8 +5.7	74.6 +40.7	79.4 +44.5

Table 3: Model results on HealthBench, LLMEval-Med, and Arena-Hard-v2. *Draft Rubrics* are auto-generated rubrics; *Doctor Rubrics* are physician-revised; *Principle Rubrics* are derived from matched `HealthPrinciples`; *More Query Rubrics* scales supervision with additional medical queries.

model, Qwen3-32B. We stop training when performance saturates, which typically occurs within 20 optimization steps in our main runs.

5.2 Main Results

Table 3 summarizes the main results on HealthBench, LLMEval-Med, and Arena-Hard-v2.

Expert-verfied Rubrics Rubric conditioned RL remains effective under automatically generated supervision. Training with **Draft Rubrics** yields strong gains over models across benchmarks, consistent with the pilot results in Section 3 and showing that rubric based rewards provide a learnable and generalizable signal. **Doctor Rubrics** further improve over Draft Rubrics, with the largest gains on HealthBench in both Hard and Overall settings and on Arena-Hard-v2. This gap suggests that physician edits remove ambiguity and fix mis specified criteria, producing higher fidelity rewards and better policy updates. Doctor supervision also makes safety critical behaviors such as triage and escalation more consistent and reduces reward hacking driven by vague rubric items, which together explains the stronger overall gains.

From Rubrics to Principles **Principle Rubrics** are competitive with physician edited rubrics even without additional scaling. Trained on the same 7,034 questions, they match and sometimes slightly exceed Doctor Rubrics, suggesting that the extracted *HealthPrinciples* capture transferable evaluation structure rather than dataset specific artifacts. A key factor is coverage. Principle retrieval pools recurring requirements across semantically similar questions, so the rubric for a given query typically spans more facets such as risk assessment, uncertainty handling, and actionable next steps than a purely question local rubric. Scaling principles to more questions further improves results. **More Query Rubrics** add principle generated rubrics for 16,872 real user medical queries and deliver broad gains on HealthBench and Arena-Hard-v2. This pattern is consistent with greater robustness from diversified and scenario matched supervision.

From Medical Alignment to General Usefulness Although training optimizes medical rubric rewards, gains transfer to the open ended Arena-Hard-v2 setting. This suggests that rubric grounding strengthens general instruction following behaviors such as tracking user intent, organizing

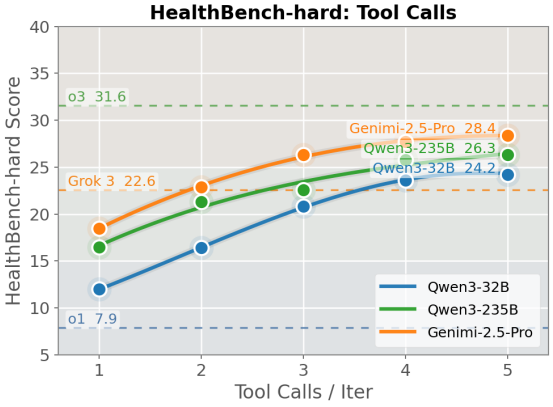

Figure 8: Tool-call Scaling on HealthBench-hard.

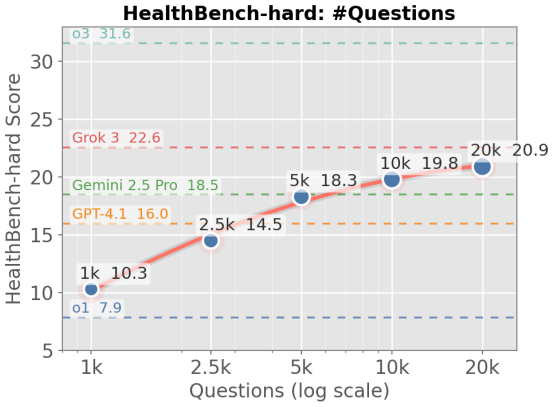

Figure 9: Question scaling on HealthBench-hard under a fixed training-FLOPs budget with Qwen3-4B-Instruct.

actionable guidance, and communicating limitations appropriately. In contrast, explicitly reasoning heavy metrics such as LLMEval-Med *Reasoning* change little, indicating that rubric aligned training mainly improves *helpfulness and safety* rather than problem solving depth. A promising next step is to combine rubric based RL with objectives that more directly train multi step reasoning, with the goal of improving both alignment and reasoning.

Inference-time Scaling Figure 8 shows that allowing multiple inference-time calls to our rubric-guidance tool consistently improves HealthBench-hard performance across backbones, even without any specific training, indicating that extracted principles and generated rubrics provide actionable revision signals at test time. Performance increases with more iterations but gradually saturates after a few calls, suggesting diminishing returns once major rubric mismatches are corrected and the remaining errors are harder to fix through further rubric feedback alone.

6 Analysis

Question Scaling with Fixed Budget Under a fixed compute budget, this subsection evaluates how performance scales with question coverage. Following Section 5.1, we randomly subsample 1k, 2.5k, 5k, 10k, and 20k questions from the full training pool. Training FLOPs are held constant by matching each run to training on 20k questions for two epochs, so smaller datasets are trained for proportionally more epochs, such as four for 10k and eight for 5k. For each setting, we evaluate the best checkpoint on HealthBench-hard.

Figure 9 shows a monotonic improvement as the number of distinct questions increases. The gains are largest from 1k to 5k and remain positive up to 20k, with diminishing returns at larger scales. This indicates that rubric-based RL benefits more from supervision diversity than from additional epochs on a narrow prompt set, likely because broader coverage better spans clinical intents, risk profiles, and failure modes. Overall, scaling is most effective by increasing the number of training questions.

Rubrics Scoring Model Choice To select an efficient and accurate rubric-satisfaction scorer, we randomly sample 1,000 HealthBench questions and evaluate model answers from both training and evaluation, yielding 11,446 rubric judgments. Treating the HealthBench protocol with GPT-4.1 as the reference, we measure Qwen3 scoring accuracy, which increases with model scale: Qwen3-4B, 14B, 32B, and 235B achieve 76.4%, 80.2%, 87.6%, and 87.9%, respectively. Given the negligible gain from 32B to 235B, we use Qwen3-32B as our default scorer for the best accuracy–efficiency trade-off.

7 Conclusion

In this work, we addressed the challenge of aligning medical language models with fine-grained standard. We introduced a robust framework anchored by **HealthRubrics**, a physician-verified dataset, and **HealthPrinciples**, a reusable taxonomy that distills expert consensus for scalable supervision. Through three progressive strategies: learning on validated data, scaling via principle-synthesized rubrics, and inference-time guidance, we demonstrated that such an alignment enables models to surpass frontier proprietary systems. These findings suggest that incorporating structured clinical logic is as effective as scaling model parameters for specialized tasks. By releasing our data, principles, and tool, we provide a practical, resource-efficient foundation to accelerate future research into safe and reliable healthcare AI.

Limitations

Despite the overall gains, we identify two key limitations. First, the proposed method does not yield a substantial improvement in intrinsic reasoning in our experiments. Although it helps the model produce better answers in some cases, the gains on multi-step reasoning tasks are not consistently significant. This suggests that the method primarily improves response quality by leveraging external signals more effectively, rather than fundamentally strengthening internal reasoning, and leaves room for further refinement through more targeted supervision and tighter integration between reasoning and tool use.

Second, as a inference-time scaling strategy, the benefits saturate quickly. Increasing tool usage or sampling beyond a certain point produces diminishing returns, indicating that simply scaling inference-time tool calls is unlikely to deliver sustained improvements. A promising direction is to train a more agentic model that can plan, decide when to invoke tools, and coordinate tool use more effectively, enabling better credit assignment and more efficient inference-time computation.

References

Asma Ben Abacha and Dina Demner Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinform.*, 20(1):511:1–511:23.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, and 1 others. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2026. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, and 1 others. 2025. Baichuan-m2: Scaling medical capability with large verifier system. *arXiv preprint arXiv:2509.02208*.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, and 1 others. 2023. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. *arXiv preprint arXiv:2312.09244*.

Dennis Fast, Lisa C Adams, Felix Busch, Conor Fallon, Marc Huppertz, Robert Siepmann, Philipp Prucker, Nadine Bayerl, Daniel Truhn, Marcus Makowski, and 1 others. 2024. Autonomous medical evaluation for guideline adherence of large language models. *NPJ Digital Medicine*, 7(1):358.

Jiayi Fu, Xuandong Zhao, Chengyuan Yao, Heng Wang, Qi Han, and Yanghua Xiao. 2025. Reward shaping to mitigate reward hacking in rlhf. *arXiv preprint arXiv:2502.18770*.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. 2025. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*.

Dulhan Jayalath, Shashwat Goel, Thomas Foster, Parag Jain, Suchin Gururangan, Cheng Zhang, Anirudh Goyal, and Alan Schelten. 2025. Compute as teacher: Turning inference compute into reference-free supervision. *arXiv preprint arXiv:2509.14234*.

Yongnan Jin, Xurui Li, Feng Cao, Liucun Gao, and Juanjuan Yao. 2025. Multidimensional rubric-oriented reward model learning via geometric projection reference constraints. *arXiv preprint arXiv:2511.16139*.

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Leandra A Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M Van Allen, David Kim, and 1 others. 2025. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature medicine*, 31(1):77–86.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *Preprint*, arXiv:2406.11939.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Preprint*, arXiv:2303.14070.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2024. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. *arXiv preprint arXiv:2406.04770*.

Yujin Oh, Sangjoon Park, Hwa Kyung Byun, Yeona Cho, Ik Jae Lee, Jin Sung Kim, and Jong Chul Ye. 2024. Llm-driven multimodal target volume contouring in radiation oncology. *Nature Communications*, 15(1):9186.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Pengcheng Qiu, Chaoyi Wu, Junwei Liu, Qiaoyu Zheng, Yusheng Liao, Haowen Wang, Yun Yue, Qianrui Fan, Shuai Zhen, Jian Wang, and 1 others. 2025. Evolving diagnostic agents in a virtual clinical environment. *arXiv preprint arXiv:2510.24654*.

AI Scale. 2025. Vista: Visual-language understanding leaderboard, 2025. *URL https://scale. com/leaderboard/visual_language_understanding. Accessed*, 3.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, page 1279–1297. ACM.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, and 1 others. 2025. Paperbench: Evaluating ai's ability to replicate ai research. *arXiv preprint arXiv:2504.01848*.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Pengkai Wang, Pengwei Liu, Zhijie Sang, Congkai Xie, Hongxia Yang, and 1 others. 2025a. Infimed-orbit: Aligning llms on open-ended complex tasks via rubric-based incremental training. *arXiv preprint arXiv:2510.15859*.

Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, and 1 others. 2024a. Cmb: A comprehensive medical benchmark in chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6184–6205.

Xidong Wang, Nuo Chen, Junyin Chen, Yidong Wang, Guorui Zhen, Chunxian Zhang, Xiangbo Wu, Yan Hu, Anningzhe Gao, Xiang Wan, and 1 others. 2024b. Apollo: A lightweight multilingual medical llm towards democratizing medical ai to 6b people. *arXiv preprint arXiv:2403.03640*.

Xidong Wang, Jianquan Li, Shunian Chen, Yuxuan Zhu, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Junying Chen, Jie Fu, Xiang Wan, and 1 others. 2025b. Huatuo-26m, a large-scale chinese medical qa dataset. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3828–3848.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. 2025c. Helpsteer3-preference: Open human-annotated preference data across diverse tasks and languages. *Preprint*, arXiv:2505.11475.

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

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, and 1 others. 2022. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540*.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, Xingtai Lv, Hu Jinfang, Zhiyuan Liu, and Bowen Zhou. 2024. Ultramedical: Building specialized generalists in biomedicine. *Preprint*, arXiv:2406.03949.

Ming Zhang, Yujiong Shen, Zelin Li, Huayu Sha, Binze Hu, Yuhui Wang, Chenhao Huang, Shichun Liu, Jingqi Tong, Changhao Jiang, and 1 others. 2025. Llmeval-med: A real-world clinical benchmark for medical llms with physician validation. *arXiv preprint arXiv:2506.04078*.

Weike Zhao, Chaoyi Wu, Yanjie Fan, Xiaoman Zhang, Pengcheng Qiu, Yuze Sun, Xiao Zhou, Yanfeng Wang, Xin Sun, Ya Zhang, and 1 others. 2025. An agentic system for rare disease diagnosis with traceable reasoning. *arXiv preprint arXiv:2506.20430*.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *Preprint*, arXiv:2405.01470.

Juexiao Zhou, Haoyang Li, Siyuan Chen, Zhangtianyi Chen, Zhongyi Han, and Xin Gao. 2025a. Large language models in biomedicine and healthcare. *npj Artificial Intelligence*, 1(1):44.

Yang Zhou, Sunzhu Li, Shunyu Liu, Wenkai Fang, Kongcheng Zhang, Jiale Zhao, Jingwen Yang, Yihe Zhou, Jianwei Lv, Tongya Zheng, and 1 others. 2025b. Breaking the exploration bottleneck: Rubric-scaffolded reinforcement learning for general llm reasoning. *arXiv preprint arXiv:2508.16949*.

Yuxuan Zhou, Yubin Wang, Bin Wang, Chen Ning, Xien Liu, Ji Wu, and Jianye Hao. 2025c. Enhancing the medical context-awareness ability of llms via multifaceted self-refinement learning. *arXiv preprint arXiv:2511.10067*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

A SFT Data Generation Prompt

To reduce brittleness from imitating a single reference completion, we expand each training question into multiple rubric-aware targets. For each question, we provide `GPT-5.1` with the original question, the official HealthBench rubrics, a current model draft response, and a clinician-written ideal completion. The draft and ideal serve as complementary anchors, highlighting common failure modes and desired coverage, while the rubrics define the scoring contract the generated answer should satisfy. Using the template in Figure 10, we sample three independent answers per question to create stylistic and structural variants that aim to meet the same rubric requirements. This produces a 9,000-instance SFT dataset from 3,000 training questions, used in the pilot SFT experiments in Section 3.

B Medical Query Classification for HealthRubrics

To extract a high precision medical subset from the pooled preference corpora we use `GPT 5.1` to classify each conversation as medical related or not using a fixed guideline A conversation is labeled MEDICAL if a responsible response would require clinical or biomedical knowledge including diagnosis treatment medication use prognosis safety risks or interpretation of medical information Conversations that are mainly administrative or only loosely health related without clinical reasoning are labeled NON MEDICAL We use deterministic decoding for consistency and manually spot check a random sample to confirm precision The prompt is shown in Figure 11

C Subcategory Definitions and Task Type Induction

This appendix provides the full definitions for the HealthPrinciples taxonomy used in Section 4.2. Table 2 in the main text presents a compact view; here we detail every subcategory and describe how the task-type inventory was induced from data with physician input.

C.1 Urgency

- **Non-emergent.** No immediate safety threat is suggested; the query concerns routine health issues or general information and typically does not require time-sensitive action.
- **Conditionally emergent.** Time-sensitive risk cannot be ruled out from the given information; urgency depends on missing key details. The response should ask targeted clarifying questions and provide conditional escalation guidance with clear red flags.
- **Emergent.** The query indicates high-risk features or an immediate safety threat and warrants prompt protective action and escalation guidance.

C.2 Uncertainty

- **Sufficient information.** The question includes the key details needed to give safe and effective guidance; no essential clarifications are required to proceed.
- **Reducible uncertainty.** Important details are missing or ambiguous but can be obtained through follow-up questions in dialogue.
- **Irreducible uncertainty.** The uncertainty cannot be resolved remotely and requires examination, objective measurement, diagnostic testing, or in-person professional evaluation.

C.3 User expertise

- **Non-professional.** The user is a layperson or general public; responses should prioritize plain language and explicit actionability.
- **Professional.** The user indicates relevant formal training or uses professional framing and terminology; responses may use technical language while maintaining safety.

C.4 Task Type Inventory

Each query is assigned a primary task and an optional secondary task. Task types are designed to be readable labels aligned with Table 2; we avoid underscored identifiers in the paper text.

Task type induction. We do not pre-specify a task taxonomy. Instead, we first ask `GPT-5.1` to freely summarize the *primary* and *secondary* task implied by each query in short natural language. We then cluster these task summaries using `GPT-5.1` to propose multiple candidate clusterings at different granularities. Finally, physicians review the candidates, reconcile ambiguous boundaries, and select a 21-category inventory that best matches clinical practice and dataset coverage. During this process, physicians additionally introduced

the task family *Procedure and perioperative guidance*, which was under-represented in early model-generated clusterings but frequently appeared in real patient questions and clinician feedback.

Free-form task extraction prompt. Figure 12 shows the template used to extract primary and secondary tasks before clustering.

Final task families and definitions. Below are the 21 task families used in this work.

- **Emergency triage and escalation.** Assess whether an emergency or time-sensitive risk is present and whether immediate escalation is required, and specify clear red flags and time windows for seeking care. Response considerations: prioritize high-yield red flags and explicit timelines; avoid undue alarm or false reassurance; default to conservative in-person evaluation when key information is missing.

- **Symptom assessment and differential considerations.** Provide an initial assessment based on symptoms and history, outline plausible etiologies, and identify missing information that would change risk or management. Response considerations: distinguish common versus high-risk causes and state rationale; ask targeted clarifying questions; avoid presenting possibilities as a definitive diagnosis.

- **Self-care, monitoring, and follow-up.** Offer home-care and self-management guidance and define thresholds and time windows for follow-up, reassessment, strategy changes, or escalation. Response considerations: provide actionable steps and measurable targets; specify stop rules and escalation triggers; avoid high-risk procedures.

- **Medication and supplement safety.** Provide general guidance on medication or supplement use with emphasis on interactions, adverse effects, contraindications, and when professional confirmation is needed. Response considerations: account for comorbidities, pregnancy, renal or hepatic impairment, and concomitant drugs; avoid prescribing-level dosing or substitution directives.

- **Diagnostics and test interpretation.** Discuss test selection and interpretation, address limitations, and provide results-contingent next steps such as repeat testing, additional workup, referral, or observation. Response considerations: explain reference ranges and false positives or negatives; avoid definitive conclusions from isolated values.

- **Procedure and perioperative guidance.** Provide peri-procedural and perioperative guidance, including preparation, recovery, complication recognition, and relevant follow-up milestones. Response considerations: emphasize warning signs and return precautions; avoid clinician-specific instructions that require the procedural or anesthesia team; reinforce individualized discharge instructions.

- **Chronic disease management.** Describe a long-term management framework including goals, monitoring cadence, lifestyle and pharmacologic strategies, complication prevention, and self-management. Response considerations: present a structured plan; emphasize adherence and longitudinal follow-up; avoid one-size-fits-all targets or unsupervised medication changes.

- **Prevention, screening, and vaccines.** Provide prevention and risk-reduction guidance including vaccination, screening, health behaviors, and post-exposure actions. Response considerations: specify eligibility, contraindications, and time windows; note guideline and local variation; avoid absolute guarantees.

- **Sexual and reproductive health counseling.** Provide education and counseling on sexual health, contraception, sexual function, and common concerns. Response considerations: use nonjudgmental language; address STI risk and testing; include escalation criteria for severe pain, heavy bleeding, and safety concerns.

- **Pregnancy and perinatal care.** Address pregnancy and perinatal health issues including symptom evaluation, antenatal pathways, medication and lifestyle considerations, and postpartum recovery. Response considerations: use conservative safety thresholds; flag medications needing obstetric confirmation; state key triggers such as bleeding and hypertensive symptoms.

- **Mental health support.** Provide mental health information and support, risk identification, and guidance on when to seek professional or emergency help. Response considerations: prioritize self-harm or violence risk screening and crisis pathways; avoid substituting for professional diagnosis.

- **Substance use harm reduction.** Provide counseling using a harm-reduction approach including withdrawal and relapse considerations and referral indications. Response considerations: avoid actionable details that facilitate misuse; highlight overdose and withdrawal warning signs; maintain supportive language.

- **Rehabilitation and return to activity.** Support rehabilitation and functional recovery with graded plans, restrictions, milestones, and reassessment points for return to sport, work, or school. Response considerations: define stop rules; provide measurable progression; specify contraindications.

- **Health education and general explanations.** Provide accessible explanations of biomedical concepts and terminology. Response considerations: use layered explanations; communicate uncertainty; avoid drifting into individualized clinical advice.

- **Knowledge check and exam preparation.** Support clinician- or exam-oriented questions with structured takeaways and workflows. Response considerations: note regional guideline variability; avoid fabricating guideline details when uncertain.

- **Medical writing, editing, and translation.** Support medical or scientific writing and translation with terminology consistency and appropriate register. Response considerations: preserve meaning without inventing results or citations; flag ambiguous source phrasing.

- **Clinical documentation drafting.** Draft professional clinical records such as SOAP notes, referral letters, or discharge summaries. Response considerations: prioritize objective, time-anchored facts and traceable assessment and plan; protect privacy.

- **Patient-facing communication and wording.** Develop patient- or public-facing materials emphasizing clarity, tone, and actionable guidance. Response considerations: use plain language; communicate uncertainty; avoid absolutes.

- **Care navigation and referrals.** Provide guidance on which service to contact, how to prepare for visits, and follow-up planning. Response considerations: provide concrete checklists; avoid presenting navigation guidance as diagnostic conclusions.

- **Insurance and administrative processes.** Address insurance, certification, and compliance processes where the deliverable is procedural guidance rather than clinical decision-making. Response considerations: defer to institutional requirements; avoid facilitating misrepresentation; flag privacy and compliance risks.

- **Non-medical.** Queries that do not require clinical reasoning. Response considerations: respond without unnecessary healthcare framing; if a hidden health concern emerges, suggest reframing into an appropriate medical task type.

D Principle Extraction from Rubrics

This appendix describes how we derive *HealthPrinciples* from physician-validated rubric items. The extraction pipeline has two stages: (i) Taxonomy-based routing of rubric sets into subcategories and (ii) Iterative clustering and compression within each subcategory to produce candidate principles for physician refinement.

D.1 Stage I: Mapping rubrics to taxonomy subcategories

Each validated rubric set is assigned to one or more taxonomy subcategories. We formulate this as a multi-label classification problem over the full set of subcategories across the four taxonomy axes. We use `GPT-5.1` with a constrained label set and a strict JSON output format to ensure routable, consistent outputs.

D.2 Stage II: Iterative clustering and compression within each subcategory

After routing, each subcategory contains a corpus of rubric items. To summarize recurring

patterns at scale, we apply semantic clustering with iterative compression. We compute rubric-item embeddings using `Qwen3-Embedding-8B` (L2-normalized), cluster them with `MiniBatchKMeans`, and select representative items by cosine similarity to each cluster centroid. We then ask `GPT-5.1` to summarize each cluster into a single representative sentence. The resulting candidates are reclustered and re-summarized iteratively until a compact set is produced for physician review.

Compression schedule. We use a fixed compression ratio of 60: each iteration compresses roughly 60 candidate items into 1 representative candidate. We run up to three iterations, enabling hierarchical coverage of up to $60 \times 60 \times 60$ items. When the remaining candidate count drops below 100, we directly summarize them into 5 candidates to form a manageable set for physician refinement.

Summarization constraints. All summarization prompts are in English and require strict JSON outputs, as shown in Figure 14. Candidates are single-sentence, observable, and checkable criteria (i.e., rubric-like statements rather than abstract advice). Physicians subsequently edit and consolidate these candidates into the final HealthPrinciples inventory.

E Question Classification Prompts

This appendix presents the prompts used for taxonomy prediction and rubric construction. Given a user question (possibly multi-turn), `GPT-5.1` predicts taxonomy labels for `urgency`, `uncertainty`, and `expertise`, and assigns exactly one primary task family with up to two optional secondary task families. All predictions are restricted to the predefined label set (Appendix C) and returned as strict JSON for downstream retrieval. The question-classification prompt is shown in Figure 15, and the rubric-generation prompt is shown in Figure 17.

Notes. In the prompt, {subcategory description} is replaced with the label definitions for each axis from Appendix C, and the model is required to choose labels exactly as defined there. We validate that returned labels are in-vocabulary and that the JSON schema is well-formed before extracting principles.

F Inference-time Rubric Guidance Tool

This appendix lists the prompts used by our inference-time tool. Given a question—optionally accompanied by prior dialogue context and/or a draft answer—the tool first classifies the scenario and then converts the retrieved principles into context-specific rubric items. In both prompts, subcategory description is instantiated with the taxonomy definitions in Appendix C. Importantly, all classifications and rubric generation are performed by the tool's underlying model itself (i.e., without invoking any external models). The tool validates that predicted labels are in-vocabulary before extracting principles.

Prompt for Rubric-Aware Answer Generation with GPT-5.1

You are a careful clinician-style medical assistant. You will write an answer to a patient-facing health question.
Your goal is to maximize the HealthBench rubric score.

Task
Given a question and its rubrics, produce a high-quality answer that satisfies as many rubric items as possible.
You will also see a model draft answer and a clinician ideal completion. Use them as reference only.
Do not copy sentences verbatim. Write a fresh answer.

Hard requirements
- Write in ENGLISH.
- Follow the rubrics closely and include key safety guidance and triage when applicable.
- Do not invent facts. If needed, ask clarifying questions.
- Be specific and actionable; use short paragraphs and bullet points when helpful.
- Avoid policy or rubric talk. Do not mention rubrics, scoring, evaluation, or the prompt.
- Do not mention the draft or the ideal completion.
- Output only the final answer text.

Inputs
Question
{question}

Rubrics
{rubrics}

Model draft answer
{draft}

Clinician ideal completion
{ideal}

Output
Write one final answer that best satisfies the rubrics.

Figure 10: Prompt template for generating rubric-aware SFT targets with `GPT-5.1`. We generate three answers per training question to increase coverage and reduce overfitting to a single phrasing.

Prompt for Medical Query Classification full

You are a medical text classifier Given the following conversation determine if it is medical related. You must only output a single JSON object. The JSON must have exactly one key ïs_medical·. The value must be either true or false in lowercase. Do not output any explanation or additional text.Conversation: {conversation}

Figure 11: Prompt used to filter medical related conversations prior to rubric construction

Prompt for Free-form Task Extraction

You are analyzing a medical user question.
Identify:
1. the primary task the assistant is being asked to perform
2. secondary tasks if present, otherwise None
Write tasks as short natural-language phrases, not as labels from a fixed list.
Do not give medical advice. Only describe the task(s).
question: {question}

Figure 12: Prompt used to extract free-form descriptions of tasks types prior to inducing the task taxonomy.

Prompt for Mapping a Rubric Set to Taxonomy Subcategories (template)

You are assigning taxonomy labels to a rubric set for medical answer evaluation.

You will be given:
- question: the original user question
- rubrics: a list of rubric items for evaluating answers to the question
- taxonomy: the full set of allowed subcategory labels for four axes:
Urgency ({subcategories}), Uncertainty ({subcategories})
Expertise ({subcategories}), Task Type ({21 families})

Task:
Select ALL subcategories that the rubrics reflect.
This is multi-label: a rubric set may map to multiple task types and may include guidance relevant to more than one axis.
Choose only from the provided label lists.
If uncertain, choose the minimal set that still covers the rubrics.

Output format (STRICT JSON only):
{
"urgency": ["A1" | "A2" | "A3"],
"uncertainty": ["B1" | "B2" | "B3"],
"expertise": ["C1" | "C2"],
"task_type": ["<Task Family Name>", ...],
"rationale": "one short sentence"
}

Inputs:
question:
{question}

rubrics:
{rubrics}

taxonomy (allowed labels):
{taxonomy}

Figure 13: Prompt used to route each rubric set into one or more taxonomy subcategories prior to clustering.

Prompt for Cluster Summarization into Representative Rubric Candidates (template)

You are an expert rubric editor for evaluating responses to health/medical questions.

Task:
Rewrite the input rubric sentences into exactly {k} representative rubric sentences in ENGLISH.

Hard requirements:
- Output EXACTLY {k} items.
- Return ONLY a JSON array of exactly {k} strings (no extra text).
- Each item MUST be a single sentence, self-contained, specific, and checkable by a non-expert.
- Each item MUST capture a distinct recurring requirement; avoid duplicates/overlap.
- Each item MUST be written as an observable criterion (e.g., "Mentions X", "Advises Y when Z"), not an abstract principle (avoid "be empathetic/clear/accurate" unless tied to concrete observable elements).
- Keep each sentence short (preferably ≤ 25 words); remove hedging and avoid multi-clause lists unless essential.
- Do NOT mention "the list", "above", "these items", files, prompts, or any process.
- Do NOT introduce new medical requirements not present in the input.

Input rubrics:
{joined}

Output format (STRICT):
Return ONLY a JSON array of exactly {k} strings.

Figure 14: Prompt used to summarize each rubric-item cluster into exactly k representative, rubric-like candidates under strict constraints.

Prompt for Taxonomy Classification (template)

```
r"""You are a medical conversation scenario classifier.

Input:
- question: the user's full question (may include multiple turns).

Task:
Classify the conversation along four independent axes and output labels with one-sentence definitions. Use plain language suitable for non-experts.

Urgency: {subcategory description}
Uncertainty: {subcategory description}
Expertise: {subcategory description}
Task Type: {subcategory description}
- Assign exactly one primary_task_family capturing the main deliverable/intent, and optionally 0–2 secondary_task_families for important supporting needs.

Output format (STRICT):
Return ONLY valid JSON with the following exact top-level schema (no extra keys, no markdown, no surrounding text):
{
"scenario": {
"urgency": {"label": "...", "definition": "..."},
"uncertainty": {"label": "...", "definition": "..."},
"expertise": {"label": "...", "definition": "..."},
"primary_task_family": {"label": "...", "definition": "..."},
"secondary_task_families": [{"label": "...", "definition": "..."}],
"description": "..."
}
}

Now generate the output for:
question: {{question}}
""".strip()
```

Figure 15: Prompt used to classify each question into taxonomy labels for principle retrieval and rubric generation.

Prompt 1: Scenario Classification (with optional dialogue context and draft answer)

```
r"""You are a medical conversation scenario classifier.

Inputs (some may be empty):
- question: the user's question
- dialogue_context: prior messages (optional)
- draft_answer: a candidate answer (optional)

Task:
Use {question} as the primary signal. If dialogue_context is provided, use it to clarify intent, constraints, and user expertise.
If draft_answer is provided, use it only as weak evidence about what the assistant is attempting to do.

Classify the scenario along four independent axes and output labels with one-sentence definitions. Use plain language suitable for non-experts.

Urgency: {subcategory description}
Uncertainty: {subcategory description}
Expertise: {subcategory description}
Task Type: {subcategory description}
- Assign exactly one primary_task_family capturing the main deliverable/intent.
- Optionally assign 0–2 secondary_task_families for important supporting needs.

Output format (STRICT):
Return ONLY valid JSON with the following exact schema (no extra keys, no markdown, no surrounding text):
{
"scenario": {
"urgency": {"label": "...", "definition": "..."},
"uncertainty": {"label": "...", "definition": "..."},
"expertise": {"label": "...", "definition": "..."},
"primary_task_family": {"label": "...", "definition": "..."},
"secondary_task_families": [{"label": "...", "definition": "..."}],
"description": "..."
}
}

Now generate the output for:
question: {{question}}
dialogue_context: { {dialogue context} }
draft_answer: {{draft answer}}
""".strip()
```

Figure 16: Prompt used to classify a question when optional dialogue context and/or a draft answer is available.

Prompt 2: Convert Retrieved Principles into Context-specific Rubric Items

```
r"""You are generating scorable rubric items for evaluating a medical answer.

Inputs:
- question: the user's question
- dialogue_context: prior messages (optional)
- draft_answer: a candidate answer (optional)
- scenario: predicted taxonomy labels
- principles: retrieved HealthPrinciples matched to the scenario

Task:
Convert the retrieved principles into rubric items that are tailored to this specific question and context.
If draft_answer is provided, write rubric items that directly help diagnose and revise likely failures in the draft.

Hard requirements:
- Output rubric items in ENGLISH.
- Each rubric item MUST be one sentence, specific, and checkable by a non-expert.
- Avoid vague style-only advice.
- Do NOT add new medical requirements beyond the provided principles and the user context.
- Prefer safety-critical items first when urgency is high or uncertainty is irreducible.

Output format (STRICT):
Return ONLY valid JSON with the following exact schema (no extra keys, no markdown, no surrounding text):
{
"rubric_items": ["...", "...", "..."],
"principle_coverage": [
{"principle_id": "...", "used_in_items": [0, 2]}
]
}

Now generate the output for:
question: {{question}}
dialogue_context: {{dialogue_context}}
draft_answer: {{draft_answer}}
scenario: {{scenario_json}}
principles: {{principles_json}}
""".strip()
```

Figure 17: Prompt used to convert retrieved HealthPrinciples into question-specific rubric items at inference time.