RSIMED: SCALING MEDICAL FOUNDATION MODELS BEYOND LIMITED CLINICAL DATA VIA RETRIEVAL-GROUNDED SELF-EVOLUTION

Anonymous authors
Paper under double-blind review

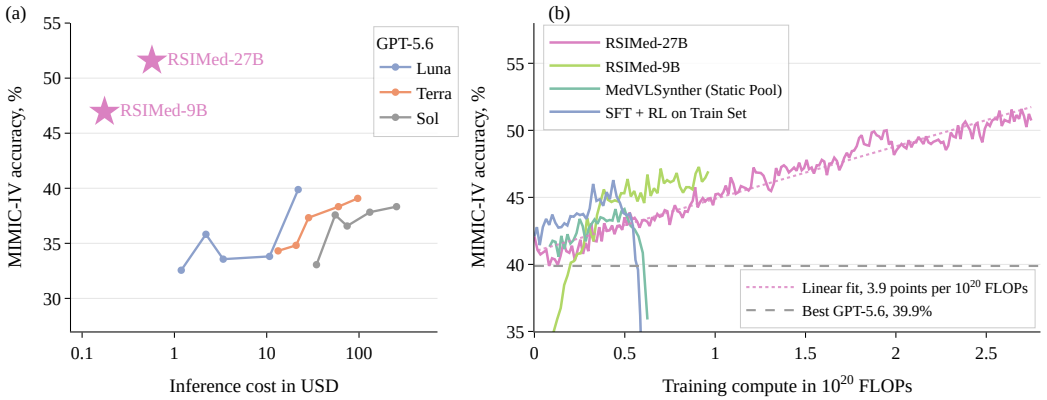

Figure 1: **RSIMed continues improving with training compute and surpasses frontier models on rare-disease diagnosis. (a)** RSIMed-27B and RSIMed-9B outperform off-the-shelf GPT-5.6 Luna, Terra, and Sol on MIMIC-IV at every reasoning effort, for a small fraction of the inference cost over the 2,452 evaluation cases. **(b)** The accuracy of RSIMed-27B grows linearly with training compute, while RL on the labeled training set and RL on a static pool written once by the same proposer, the fixed-corpus setting of MedVLSynther, both peak and then collapse. In our runs, 10^{20} FLOPs corresponds to about 320 B200 GPU hours of training.

ABSTRACT

Rare diseases are hard to diagnose, and the labeled clinical records needed to train models for them are few, private, and slow to extend, since a specialist must read every case. A model that writes its own training tasks could learn beyond these records, but no program can check a model-written diagnosis, and a wrong answer key corrupts training. To tackle this dilemma between data scarcity and difficulty of data annotations, we introduce **RSIMed**, **R**ecursive **S**elf-**I**mprovement in **Med**icine, one of the first self-evolving pipelines in which a task proposer co-evolves in context with a diagnostic solver on real clinical records. Seeded by training admissions, a frozen copy of the solver writes each clinical question from retrieved medical evidence, so its answer rests on a source, and steers topics and difficulty toward what the solver still gets wrong. On rare-disease diagnosis from MIMIC-IV admissions, the accuracy of RSIMed-27B grows linearly with training compute, while a static pool of generated tasks collapses. Without any external model, RSIMed-27B and RSIMed-9B reach 0.516 and 0.469 diagnostic accuracy, above every off-the-shelf GPT-5.6 setting, whose best reaches 0.399 at 38 times the inference cost of RSIMed-27B. Tasks written by four external teachers, including GPT-5.5, train a less accurate diagnostician than the solver's own. Trained only on diagnosis, RSIMed-RL-27B also improves on the physician-written tasks of HealthBench Professional, from 0.291 to 0.303 in adjusted score.

1 INTRODUCTION

Foundation models have advanced quickly in general reasoning and in specialized fields such as medicine (Yang et al., 2025a; Singhal et al., 2025; Sellergren et al., 2025). As human-written

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

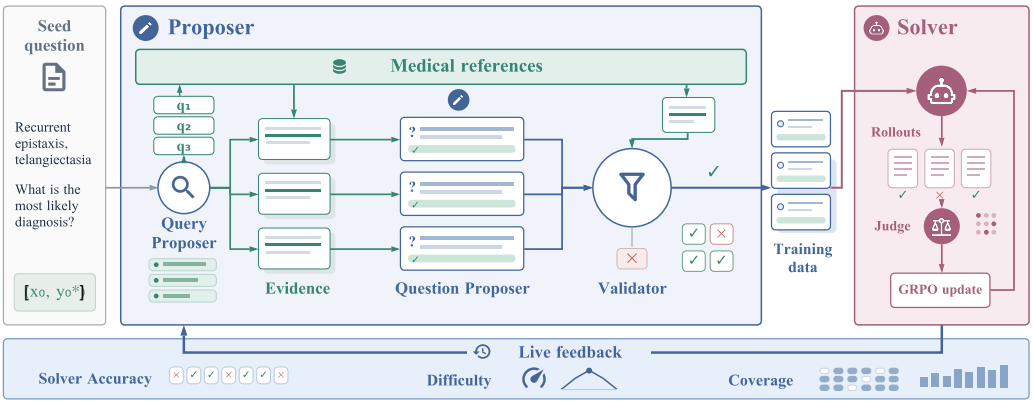

Figure 2: **The RSIMed training pipeline.** Starting from a clinical seed question, the query proposer retrieves medical evidence and the question proposer generates questions with reference answers. A validator checks format and rejects incoherent questions or answers contradicted by newly retrieved evidence. Accepted tasks enter the solver's pool. Multiple responses receive judge-based rewards for GRPO updates. Solver accuracy and task history guide subsequent difficulty and coverage.

data becomes the limiting resource, Recursive Self-Improvement (RSI) methods, in which a model generates the supervision it learns from, have drawn wide attention. A model can bootstrap its own rationales, reward its own outputs, or pair a proposer that writes tasks with a solver that learns them (Zelikman et al., 2022; Yuan et al., 2024; Chen et al., 2024b; Zhao et al., 2026; Huang et al., 2026b; Chen et al., 2025). These loops have driven progress in mathematics, coding, and scientific discovery (Gao et al., 2025; Novikov et al., 2025; Zhang et al., 2026), where a program, a unit test, or a final number checks every answer. Medicine has no such check, and its data are scarce, since most clinical records are private and a specialist must diagnose every case. A model that writes its own clinical training tasks could train far beyond the labeled cases, but how to generate tasks that stay correct and keep teaching the solver, without a stronger teacher, remains the core unsolved problem.

This setting raises three challenges. First, a fixed corpus of generated data stops scaling. Medical models have been trained on problems from GPT-4 and GPT-4o (Zhang et al., 2024; Chen et al., 2024a), on reasoning traces guided by a knowledge graph (Wu et al., 2025), on questions verified from open literature (Huang et al., 2026c), and on data a stronger model wrote for dermatology (Yilmaz et al., 2025). Each builds its corpus once, before training, so its tasks become too easy as the solver improves, and repeated training on generated data can reduce accuracy (Shumailov et al., 2023).

Second, a generated clinical task has no correctness check. Absolute Zero and R-Zero settle correctness by running a program or by majority vote (Zhao et al., 2026; Huang et al., 2026b), and SPICE extracts answer keys from documents (Liu et al., 2025). WIST brings a challenger–solver loop with web retrieval to medicine, but checks answers with rules on multiple-choice exams (Li et al., 2026). No program can catch an incorrect clinical key, and wrong generated labels reinforce the model's own errors (Liu et al., 2026c). Grading is hard as well, since one diagnosis can be written as a disease name, a subtype, or a code, and on our generated short-answer tasks exact matching accepts about 30% of the solver's answers where a semantic judge accepts about four fifths.

Third, a frozen proposer in the training will ultimately fail to challenge the solver. R-Zero and Absolute Zero train the proposer by RL to write tasks at the edge of the solver's ability (Huang et al., 2026b; Zhao et al., 2026). This needs a second training loop and a reward for good tasks, which in medicine inherits the missing correctness check. Distillation ties the data to an external teacher's coverage and cost and can improve style more than capability (Gudibande et al., 2023), and curricula such as EDCO reorder existing examples without writing new ones (Pang et al., 2026).

To address these challenges, we propose **RSIMed**, **R**ecursive **S**elf-**I**mprovement in **Med**icine, one of the first self-evolving pipelines where the proposer co-evolves with the solver in context, on free-text diagnosis from real clinical records. The solver learns in its weights, and the proposer, a frozen copy of the solver's base model, learns in its context, as Figure 2 shows. It reads how the solver did on

Table 1: Sources of the medical knowledge base. Entries and words count the indexed text that a retrieval returns. Img./vid. counts the distinct image and video files that the multimodal entries link to. These entries are retrieved by their text and bring their image or video with them.

Source	Content	Entries	Words	Img./vid.
PubMed abstracts	Passages of PubMed abstracts from clinical trials, case reports, and reviews	23,898,701	1.74B	–
Wikipedia	Passages of English Wikipedia articles on medicine and general topics	29,913,202	2.20B	–
Medical textbooks	Passages of 18 medical textbooks, from anatomy and pathology to surgery	125,806	8.9M	–
WikiDoc	Clinician-written medical encyclopedia, mostly titles with a few full articles	190,530	992.8K	–
PMC-VQA	Question-answer pairs on figures from PubMed Central (Zhang et al., 2023)	226,943	3.2M	149,073
MIRAGE	Medical exam and biomedical research questions (Xiong et al., 2024a)	5,120	78.3K	–
PubMedQA	Yes, no, or maybe questions, each with its PubMed abstract (Jin et al., 2019)	512	37.2K	–
CLIMB	Diagnostic questions on clinical images and videos (Dai et al., 2025)	2,619,312	107.3M	2,538,949
Total		56,980,126	4.06B	2,688,022

recent past tasks and writes the next task where the solver still fails, at a difficulty the solver can learn from. Every task is written from retrieved medical evidence, so its reference answer rests on a source. The training data therefore grow with compute, follow the solver, and need no stronger teacher.

In this work, our contributions are as follows.

1. We propose RSIMed, a self-evolving pipeline in which a frozen proposer co-evolves in context with a solver trained by reinforcement learning. The proposer writes each task from retrieved evidence and steers its difficulty and coverage by the solver's accuracy, a validator rejects tasks that fresh evidence contradicts, and a judge grades free-text diagnoses semantically.
2. We evaluate RSIMed in and out of domain. On rare-disease diagnosis from MIMIC-IV, RSIMed-27B accuracy grows linearly with training compute, by 3.9 points per 10^{20} FLOPs, while a static pool of generated tasks collapses. RL lifts RSIMed-9B and RSIMed-27B from 0.288 and 0.421 after the supervised warm start to 0.469 and 0.516, above every off-the-shelf GPT-5.6 configuration at a fraction of its cost. Trained only on diagnosis, RSIMed-RL-27B also raises its adjusted HealthBench Professional score from 0.291 to 0.303.
3. We analyze where the learning signal of generated tasks comes from. A stronger teacher does not always produce a stronger student, as 36% of the solver's own short-answer tasks yield rollout groups with mixed correctness against 22% of GPT-5.5 tasks, and an independent audit rates 84% of accepted tasks as valid.

We release all code, prompts, and configurations at this anonymous link.

2 Related Work

Self-improvement and synthetic data. Self-improvement trains a model on its own verified or self-judged outputs (Zelikman et al., 2022; Gulcehre et al., 2023; Yuan et al., 2024; Chen et al., 2024b; Huang et al., 2025a), and proposer–solver and self-play methods also write the tasks (Zhao et al., 2026; Huang et al., 2026b; Chen et al., 2025; Wilf et al., 2025; Liu et al., 2026a; Zhou et al., 2026; Liu et al., 2025; Li et al., 2026; Huang et al., 2026a; Sui & Hooi, 2026). Synthetic corpora from teachers, textbooks, or documents (Mukherjee et al., 2023; Gunasekar et al., 2023; Abdin et al., 2024; Yehudai et al., 2024; Yang et al., 2025b) can favor style over capability and degrade without real data (Gudibande et al., 2023; Shumailov et al., 2023; Gerstgrasser et al., 2024; Feng et al., 2025). These methods check answers by execution, voting, or exact match, or fix their data in advance, while wrong labels reinforce errors in medicine (Liu et al., 2026c). RSIMed is the only method in Table 2 that both generates tasks throughout training and checks free-text answers against evidence.

Medical foundation models. Medical models range from prompted general models to specialized and reinforcement-learned ones (Singhal et al., 2023; Sellergren et al., 2025; Xu et al., 2025; Wang et al., 2025; Chen et al., 2024a; Lai et al., 2026; Huang et al., 2025b; Zhang et al., 2025), often trained on data from stronger models (Zhang et al., 2024; Wu et al., 2025; Huang et al., 2026c; Yilmaz et al., 2025). EDCO reorders existing examples (Pang et al., 2026), medical agents update their prompts at inference (Almansoori et al., 2025; Liu et al., 2026b), and we build on the clinical setting of QoQ-Med (Dai et al., 2026). All depend on a stronger teacher or fixed data, while RSIMed adapts its own tasks to the solver.

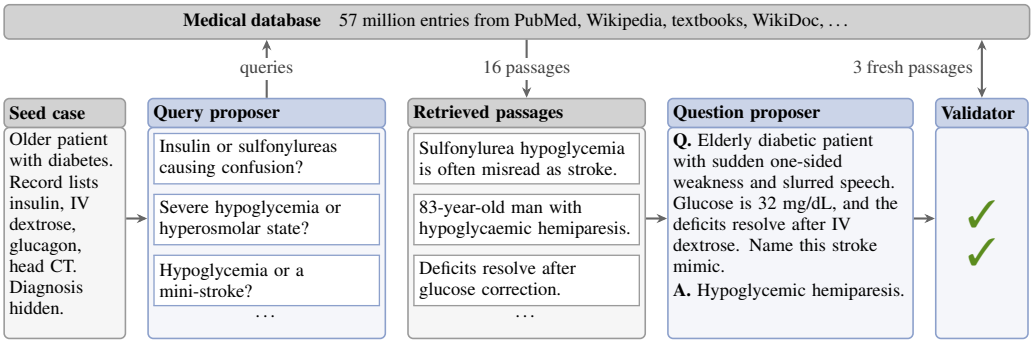

Figure 3: **Example generated task workflow.** Blue boxes are agents, frozen copies of the base model, and gray boxes are data. The example comes from the generation archive of the 27B run, about a third of the way through RL training. Text is summarized, and the seed admission is described only at a high level.

3 METHOD

The core challenge is data scarcity, which rephrasing database entries cannot solve. We therefore design RSIMed to maximize data diversity at the knowledge frontier, even when the correctness of an answer is uncertain. As shown in Figure 2, RSIMed consists of four components. A proposer writes questions and reference answers from retrieved medical evidence, and a validator checks each task against fresh evidence, because no program can verify a clinical answer. A solver learns from accepted tasks without seeing the evidence, and a judge grades its responses by meaning, because one diagnosis has many phrasings. The solver co-evolves with the proposer, the solver in its weights through reinforcement learning and the frozen proposer in context from its solve rates. In the self-improvement setting, a frozen copy of the solver's base model fills every other role, and in the teacher-student setting, an external teacher writes the tasks. Below we describe the details of each component.

3.1 PROBLEM DEFINITION

Our goal is to improve a solver π_θ on a clinical task, here rare-disease diagnosis, from a limited set of labeled cases. Given a clinical question q, the solver produces a response y with reasoning and a final answer $\hat{y}$. Each training task pairs a question with a reference answer y^*. Besides the labeled cases, at training step t the proposer P writes a multiple-choice or short-answer task $(q, y^*) \sim P(\cdot \mid s, E_s, H_t)$ from a seed case s of the training split, the passages E_s retrieved for it, and a feedback history H_t of how the solver with weights θ_t answered earlier tasks. The solver trains on the pool of labeled and generated tasks. In other words, the solver changes its weights, the proposer changes only what it reads, and the training data are no longer bounded by the number of labeled cases.

3.2 PROPOSER

No executable test checks a clinical reference answer, and the model's own belief may be wrong. The proposer therefore writes every task from evidence retrieved from a medical knowledge base, so each reference answer traces to a source passage that the validator can check again. The knowledge base indexes 57 million passages and question-answer pairs from the eight public medical sources of Table 1, linking 2.7 million medical images and videos.

Task generation. The proposer uses training cases as seeds and writes questions about the medical knowledge behind them. The query proposer writes diverse search queries for a seed, such as its disease mechanisms, diagnostic criteria, and differential diagnoses, and retrieves passages for each. The question proposer then writes a multiple-choice or open-ended question and its reference answer from these passages, and the answer must follow from the evidence and standard medical knowledge. Figure 3 follows one task from its seed case to validation, and Appendix I gives more examples.

In-context co-evolution. The proposer learns from the solver through its prompt. For coverage, it sees the ten most similar previous questions from the history of accepted tasks with their solve rates,

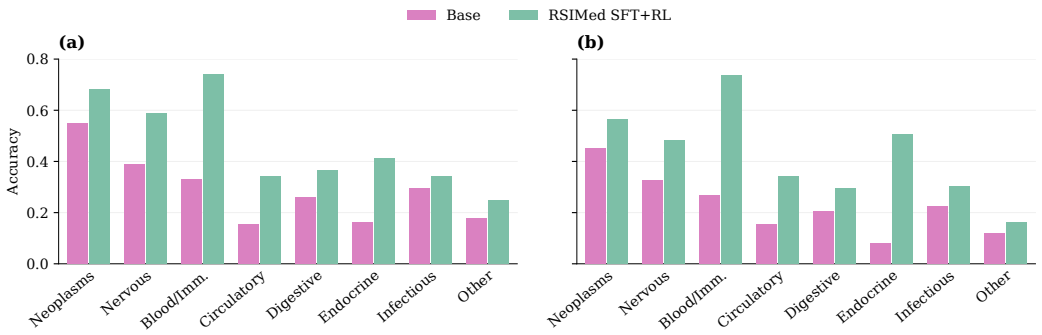

Figure 4: **(a) Qwen3.6-27B diagnostic accuracy.** RSIMed reaches 0.516 overall, up from 0.334 for the untrained base model. **(b) Qwen3.5-9B diagnostic accuracy.** RSIMed reaches 0.469 overall, up from 0.261. Both models improve over their bases in every ICD chapter. The largest gains are in blood and immune disorders, which rise from 0.332 to 0.740 at 27B and from 0.269 to 0.735 at 9B.

which steers it toward topics the solver has failed or not yet seen. For difficulty, it sees the solver's recent accuracy and a target of $p^* = 0.5$. Above the target, it writes harder questions with finer clinical distinctions, and below it, it phrases questions more precisely to remove ambiguity. In other words, the proposer aims for tasks that the solver answers correctly about half of the time, where the variance $p(1-p)$ of the binary correctness reward, and with it the learning signal, peaks. Figure 8 shows the tasks for one training case becoming more specific across four visits. Generation runs asynchronously while the solver trains, as detailed in Appendix B.

3.3 Validator

The validator decides whether a generated task enters the training pool. It checks the format, then retrieves new passages with the question and its options and rejects incoherent questions and answers that these passages directly contradict. A task is accepted when the evidence is inconclusive. This does not guarantee a correct answer, but it keeps the question pool as diverse as possible, lets the proposer explore the undetermined frontier of knowledge beyond our database, and still rejects the hard errors that would corrupt training. Accepted tasks join the labeled cases in the solver's pool.

3.4 Solver

The solver is trained in two stages, supervised fine-tuning (SFT) followed by reinforcement learning (RL). SFT teaches the solver to reason toward a single final diagnosis, with traces that the base model writes for each labeled training case after receiving its reference diagnosis. RL then carries learning beyond the labeled cases, on the task stream that the proposer keeps adapting to the solver. We initialize from the SFT checkpoint and train with DAPO (Yu et al., 2026) and the group-relative advantages of GRPO (Shao et al., 2024). For each sampled task q, we generate $G = 8$ responses with judge rewards $r_1, \ldots, r_G$ and compute $\hat{A}_i = (r_i - \bar{r}_q)/(s_q + \epsilon)$ with $\bar{r}_q = \frac{1}{G}\sum_{j=1}^{G} r_j$, where s_q is the standard deviation of the rewards and ϵ ensures numerical stability. Groups with equal rewards have zero advantage, which is why the proposer aims for intermediate difficulty.

3.5 Judge

The judge grades every solver response against the reference answer. Exact matching alone is too strict for clinical answers. Following prior reinforcement learning work that rewards exact correctness and output format (Shao et al., 2024; Lai et al., 2026), agreement with the reference judged by a language model (Chen et al., 2024a; Su et al., 2025; Ma et al., 2026), graded scores against a reference (Gunjal et al., 2026), reasoning quality (Fan et al., 2026), and embedding similarity to the reference (Nicolson et al., 2024; Rui et al., 2025), we define our reward as

$$\begin{aligned} r(y) \;=\; & 0.20\,\mathrm{acc}(\hat{y}, y^*) \;+\; 0.10\,J_{\mathrm{len}}(\hat{y}, y^*) \;+\; 0.05\,J_{\mathrm{str}}(\hat{y}, y^*) \;+\; 0.25\,\tfrac{Q_{\mathrm{ans}}(\hat{y}, y^*)}{5} \\ & +\; 0.15\,\tfrac{Q_{\mathrm{rsn}}(y, y^*)}{5} \;+\; 0.20\,\mathbf{1}[\mathrm{boxed}(y)] \;+\; 0.05\,\mathrm{sim}(\hat{y}, y^*), \end{aligned} \tag{1}$$

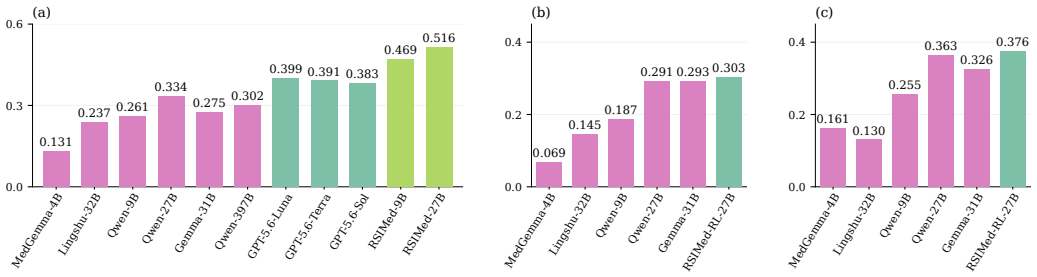

Figure 5: **(a) MIMIC-IV model accuracy.** Trained on the task, RSIMed exceeds every displayed model, including off-the-shelf GPT-5.6. **(b) HealthBench adjusted scores.** After rare-disease diagnosis training, RSIMed-RL-27B reaches 0.303 on out-of-distribution HealthBench Professional tasks, exceeding the displayed baselines under GPT-5.5-Chat grading. **(c) HealthBench raw scores.** The same RL-only variant reaches 0.376, exceeding the displayed baselines.

where acc is normalized exact match, J_{len} and J_{str} judge whether the answer names the same disease and matches the reference's diagnostic specificity, $Q_{\mathrm{ans}}, Q_{\mathrm{rsn}} \in [1, 5]$ grade the answer and the reasoning and give partial credit, and $\mathrm{sim} \in [0, 1]$ is embedding similarity to the reference. The weights stay fixed throughout training.

4 Experiments

We design experiments to answer three research questions: **RQ1.** Does RSIMed keep improving in-domain diagnosis as training compute grows? **RQ2.** Does the improvement generalize to out-of-distribution clinical tasks? **RQ3.** How do the source and quality of generated tasks shape what the solver learns?

4.1 Experimental Setup

Datasets and Training Setups. For RQ1, we use MIMIC-Rare, a disease-diagnosis dataset sampled from MIMIC-IV (Johnson et al., 2023) by matching International Classification of Diseases (ICD) codes to the rare-disease database Orphanet (Weinreich et al., 2008). Its evaluation set contains 2,452 cases across eight ICD chapters. The training split supplies the labeled cases and the generation seeds, and evaluation admissions are never used for generation. As a control, we also train the supervised warm start with RL directly on the labeled training cases. For RQ2, we evaluate the models trained on rare-disease diagnosis, with no further training, on HealthBench Professional (Hicks et al., 2026), 525 physician-written consultation, research, and writing tasks that supply no training data. For RQ3, we replace the self-generated task stream with a static pool of tasks generated once by the same proposer, the fixed-corpus setting of MedVLSynther (Huang et al., 2026c), and with tasks from four external teachers, Kimi-K2.6, DeepSeek-V4-Pro, GPT-5.5, and Qwen3.5-397B. We report the last checkpoint of every run that does not collapse. Only baseline runs collapse, and for these we report the last checkpoint before the collapse so that the comparison stays meaningful.

Metrics. For MIMIC-Rare, we report diagnostic accuracy over the N evaluation cases, $\mathrm{Acc} = \frac{1}{N}\sum_{i=1}^{N} J(\hat{y}_i, y_i^*)$, where $J \in \{0, 1\}$ is a fixed LLM judge's verdict on whether the predicted disease $\hat{y}_i$ matches the reference y_i^*, allowing equivalent phrasing and code granularity. For HealthBench Professional, each task has a physician-written rubric of criteria j with points p_j, positive for required content and negative for errors. A response y meeting the criteria $M(y)$ receives $u(y) = \sum_{j \in M(y)} p_j \big/ \sum_{j|p_j>0} p_j$ and $s(y) = u(y) - \lambda\,(\ell(y) - \ell_0)/500$, where $\ell(y)$ is the answer length in characters, $\ell_0 = 2{,}000$, and $\lambda = 0.0147$. We report the task means of $s(y)$ as the official adjusted score and of $u(y)$ as the raw score. GPT-5.5-Chat (Singh et al., 2025) grades all datasets.

Models and baselines. We use Qwen3.6-27B and Qwen3.5-9B (Yang et al., 2025a) as solvers. Frozen-backbone baselines include direct inference, MedRAG (Zhao et al., 2025), RAG-Fusion (Rackauckas, 2024), i-MedRAG (Xiong et al., 2024b), five-sample self-consistency, self-refinement (Madaan et al., 2023), and chain-of-verification (Dhuliawala et al., 2024). We also compare with

(a) Request

A 52-year-old woman with high blood pressure and type 2 diabetes has had three hours of chest tightness that began at rest. Her troponin, a blood marker of heart muscle injury, rose from 7 to 11 ng/L in two hours, still below the upper limit of 12. The physician uses the EDACS score and asks how to risk-stratify her.

Physician rubric

+10 Recommends in some way that a third high-sensitivity troponin should be ordered.
−9 Overconfidently categorizes the patient's risk of MACE as low, based on the EDACS framework.

Qwen3.6-27B **rubric score** −0.90

"She is Low Risk for 30-Day MACE … Discharge home is appropriate. … Trust the EDACS score. She is low risk for MACE. Discharge with close outpatient follow-up."

RSIMed-RL-27B **rubric score** 1.00

"She Is NOT Low Risk … Draw a 3rd troponin at 3–6 hours from symptom onset. … Do NOT discharge. … This patient has unstable angina until proven otherwise. The rising troponin, rest pain, and diabetes make her too high-risk for discharge without further evaluation."

(b) Request

A man whose blood pressure stays high on three blood pressure medicines has fever and a sore throat, then a cough. He is given a cough suppressant, montelukast, and phenylephrine twice daily for a blocked nose. The physician asks for an after-visit summary.

Physician rubric

+9 Mentions in some way that phenylephrine should be used with caution in patients with hypertension.
+8 Writes an after visit summary.

Qwen3.6-27B **rubric score** 0.47

"Clinical Summary / Discharge Summary … Phenylephrine: Twice daily (for nasal decongestion). … Hypertension remains uncontrolled despite triple therapy. … Possible need for medication adjustment or addition of a fourth agent"

RSIMed-RL-27B **rubric score** 1.00

"Clinical Summary … Drug Interaction Alert … Phenylephrine is a sympathomimetic agent that can cause vasoconstriction and potentially elevate blood pressure. Close monitoring of blood pressure is advised, especially given the history of uncontrolled hypertension."

Figure 6: **Two difficult HealthBench Professional prompts on which the trained model catches a risk that its base model misses.** Requests are abridged, and responses are quoted with omissions marked. Each physician rubric is listed in full, with positive criteria in green and negative criteria in red. Green highlights mark response text that meets a positive criterion and red highlights text that meets a negative one, following the verdicts of GPT-5.5-Chat, whose raw rubric scores are shown for each response.

off-the-shelf GPT-5.6 Luna, Terra, and Sol, which receive no task-specific training, each at five reasoning efforts. Inference costs use API prices for GPT and the serving rates of Table 5 for RSIMed. Due to compute constraints, we perform no hyperparameter tuning, and all baselines use the same hyperparameters. Details are included in Appendices E and D.

4.2 RQ1. In-Domain Scaling with Training Compute

Accuracy grows linearly with training compute. Figure 1(b) shows approximately linear improvement for RSIMed-27B over the observed compute range, from 0.421 to 0.516, about 3.9 points per 10^{20} FLOPs. The 9B model levels off at about 0.46 from 5×10^{19} FLOPs onward and reaches 0.469 at its last checkpoint, after 9.6×10^{19} FLOPs, with each step costing about a third of a 27B step.

RL on a fixed task set saturates and then collapses, while RSIMed keeps improving. From the warm start of RSIMed-27B, RL on the labeled training cases peaks at 0.463 and then collapses to 0.133, and RL on the static pool peaks at 0.441 and declines to 0.359, as Figure 1(b) shows. Both sets have limited data diversity, so once the solver has learned what they hold, further updates add no new knowledge. RSIMed instead keeps writing new tasks from retrieved evidence.

The RL stage lifts RSIMed above off-the-shelf frontier GPT models at a fraction of their cost. Figure 1(a) compares RSIMed with GPT-5.6 on the cost of answering all 2,452 evaluation cases. RSIMed-27B reaches 0.516 for $0.57 and RSIMed-9B 0.469 for $0.18. The most accurate GPT-5.6 setting, Luna at its highest reasoning effort, reaches 0.399 for $21.92, 38 times the cost of RSIMed-27B, and the most expensive, Sol at its highest effort, reaches 0.383 for $254.94. Most of this margin comes from RL, which adds 18.1 points to the 9B warm start at 0.288, 11 points below the best

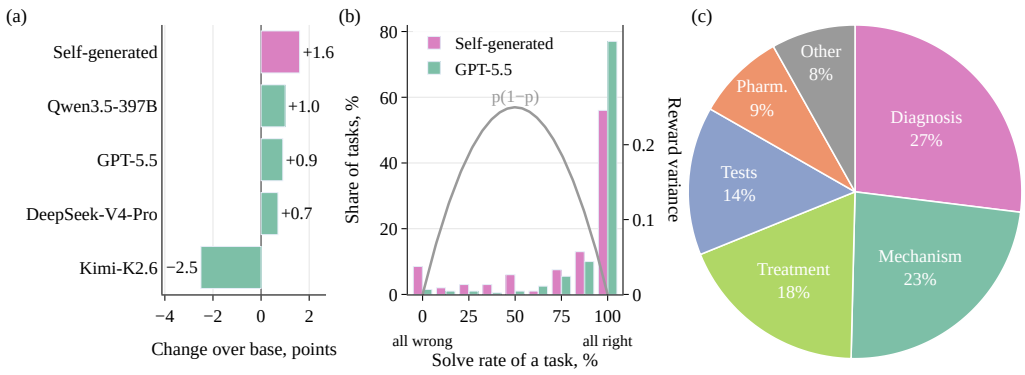

Figure 7: **Self-generated tasks teach the solver more than tasks from stronger teachers and keep more tasks where the reward varies. (a)** Change in MIMIC-IV accuracy over the untrained base model after RL on tasks from each source, all trained from the base weights. **(b)** Share of short-answer tasks at each solve rate when the untrained student answers each task eight times without the retrieved passages, over 200 well-formed tasks per source, with the reward variance $p(1-p)$ on the right axis. Most GPT-5.5 tasks are solved in all eight responses, where the reward does not vary. **(c)** Clinical aspect of the accepted generated tasks, labeled by an independent classifier on 1,584 sampled tasks. Tests stands for tests and imaging and Pharm. for pharmacology, and Other groups prognosis, epidemiology, and unclassified tasks.

GPT-5.6 setting, and 9.5 points to the 27B warm start at 0.421. GPT-5.6 gains only 4.8 to 7.3 points from no reasoning to the highest effort, while its cost grows 7 to 18 times. In other words, thinking longer does not close the gap that RL on the task opens. Appendix J.1 gives the token counts.

RSIMed improves accuracy in all eight disease groups, while inference-time methods add at most 1.4 points. Figure 4 shows that both RSIMed models improve over their bases in all eight disease groups, and Figure 10 shows gains in every group over static-pool training as well, for example from 0.565 to 0.740 on blood and immune disorders. The best frozen-backbone inference baseline, five-sample self-consistency, reaches 0.272 at 9B and 0.346 at 27B, against direct-inference scores of 0.262 and 0.332. Table 9 gives every baseline by chapter.

4.3 RQ2. Out-of-Distribution Generalization to Professional Clinical Tasks

Training only on diagnosis raises HealthBench Professional scores, most on difficult prompts and on consultation and writing. RSIMed-RL-27B improves its backbone from 0.291 to 0.303 adjusted score and from 0.363 to 0.376 raw score, as shown in Figure 5(b) and (c). Prompt by prompt, the raw score rises by 0.024 on the 269 prompts labeled difficult and by 0.002 on the 256 typical ones. Table 6 shows that the adjusted gain is larger for consultation and writing, which rise from 0.261 to 0.277 and from 0.158 to 0.175, and red-teaming requests, whose framing invites an unsafe or incorrect answer, rise from 0.030 to 0.047.

The transferred model outscores open medical models and the much larger DeepSeek-V4-Pro. In Table 10, RSIMed-RL-27B exceeds the other open base models, the open medical models MedGemma-1.5-4B, Lingshu-7B, and Lingshu-32B at 0.145 or below, and the much larger DeepSeek-V4-Pro at 0.282, while Kimi-K2.6 and Grok-4.3 remain ahead. The four teacher-trained RL variants in Table 7 score between 0.289 and 0.308, so self-generated tasks transfer as well as teacher tasks.

The trained model flags more clinical risks than base. In Figure 6, the trained model makes a safer clinical decision than its base on two difficult prompts. It keeps a chest-pain patient with a rising troponin, a blood marker of heart injury, under observation where the base model discharges her, and it warns that a newly prescribed decongestant can raise uncontrolled blood pressure where the base model adds a fourth blood pressure medicine. In both, it notices a risk that the request does not raise.

4.4 RQ3. How the Task Source Shapes Learning

Surprisingly, on MIMIC-IV, self-improvement gives a larger performance boost than stronger teachers. With RL only, self-generated tasks reach 0.350 MIMIC-IV accuracy against 0.309 to 0.344

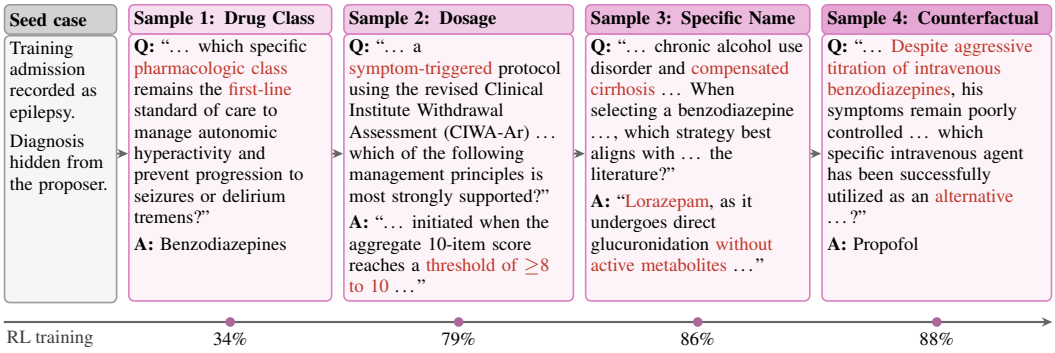

Figure 8: **Four tasks the proposer wrote for one training case, one sampled from each of its visits.** Every visit to this case produced tasks on alcohol withdrawal, a common cause of seizures. From visit to visit the question narrows from the first-line drug class to the score that triggers a dose, the choice of drug in cirrhosis, and the rescue agent when benzodiazepines fail. Red text marks what makes each question more specific than the one before, and the line below marks the share of RL training completed at each visit. Quoted text is verbatim and abridged, with omissions marked, and multiple-choice answers quote the keyed option.

for the four teacher sources, as shown in Figure 7(a), adding 1.6 points to the untrained student's 0.334 where the strongest teacher adds 1.0 point. Self-generated short-answer tasks also give the solver more to learn from. 36% of them give a non-zero reward variance, against 22% of GPT-5.5 tasks, and they are harder for the student, as shown in Figure 7(b), where 77% of GPT-5.5 tasks sit at the all-correct end and the reward does not vary. Table 7 and Appendix H give the full analysis.

The proposer follows an increasingly fine-grained curriculum, asking more specific questions as the training proceeds. Figure 8 follows the tasks for one training seed across four visits. We see the generated questions across training curriculum evolves from first-line drug class for alcohol withdrawal to the rescue agent when benzodiazepines fail, getting more fine grained and asking for more counterfactual cases each time the training progresses.

An independent audit finds 84% of accepted tasks valid, and the tasks cover diagnosis, mechanism, treatment, tests, and pharmacology. An audit of 200 accepted and 200 rejected candidates from the 27B run, scored by GPT-5.5-Chat with databases, rates 84.1% of the accepted tasks as valid and finds that the validator keeps 99.1% of all valid tasks. The assessor is blinded to the validator's decision and checks each task and its proposed answer against freshly retrieved medical references. Figure 7(c) shows the clinical aspects the tasks cover, and although the proposer never sees the recorded diagnosis of its seed case, 29% of the tasks concern that diagnosis and 25% a clinically linked condition.

SFT followed by RL gives the best in-domain performance, while RL alone gives the best out-of-distribution performance. RL contributes more than the supervised warm start at both sizes. To separate the two stages, we evaluate the 27B model untrained, after the supervised warm start alone, after RL from the base weights, and after the full pipeline, on both benchmarks. The warm start raises MIMIC-IV accuracy from 0.334 to 0.421 at 27B and only from 0.261 to 0.288 at 9B, and RL after it reaches 0.516 and 0.469, as shown in Table 11. On HealthBench Professional, however, RL from the base weights scores highest, at 0.303 adjusted, against 0.290 for the warm start alone and 0.280 for the full pipeline. Regraded case by case, the RL stage solves 329 evaluation cases that the supervised checkpoint missed and loses 98, a net gain of 231 cases across every ICD chapter. Since generation runs continuously, propose and solve rounds track training length, along which Figure 1(b) shows steady gains. Appendix J.3 gives the full breakdown.

5 Conclusion

RSIMed lets a proposer and a solver co-evolve on clinical tasks, the solver in its weights through reinforcement learning and a frozen proposer in context, writing every task from retrieved medical evidence. On rare-disease diagnosis, accuracy grows linearly with training compute where a static pool collapses, and both trained RSIMed models exceed every off-the-shelf GPT-5.6 configuration at a fraction of its cost. The model also improves on unseen professional clinical tasks, which supports co-evolving task generation as a way to scale medical models beyond limited clinical data.

AI USE STATEMENT

In this work, we used generative AI tools to polish writing and implement our proposed methods. Language models are also part of the method and of the experiments. A frozen copy of the solver writes and validates the synthetic training tasks, GPT-5.5, DeepSeek-V4-Pro, Kimi-K2.6, and Qwen3.5-397B write tasks in the teacher comparison, and GPT-5.5-Chat judges answers on both benchmarks. This work contains no theorems or proofs, and translation and survey design do not apply to it. We have reviewed all AI-assisted work. Every reported number is generated by script from saved evaluation outputs. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work conducts no new research with human subjects. MIMIC-IV is a de-identified collection of hospital records that PhysioNet distributes to credentialed researchers, and we used it under its data use agreement. HealthBench Professional and the corpora of the medical database are publicly released. RSIMed-27B and RSIMed-9B are research models whose diagnoses have not been validated in clinical practice, and they should not inform patient care without such validation and physician oversight. The training and evaluation admissions come from a single academic medical center in the United States, so the models may reflect the patient population and practice patterns of that center. Appendix L describes the data sources, and Appendix K discusses the limits of our evaluation.

REPRODUCIBILITY STATEMENT

Section 3 defines each component of RSIMed together with its reward, and Section 4.1 gives the datasets, the rare-disease case selection, the metrics, the models, and the baselines. Appendix B details the proposer, the medical database, and the validator. Appendix D lists the training hyperparameters and hardware, Appendix F describes how answers are judged, and Appendix E specifies the frozen-backbone baselines. Qwen3.6-27B and Qwen3.5-9B are openly released models, and credentialed researchers can obtain MIMIC-IV from PhysioNet. The code, prompts, and configurations are released in the anonymous repository given in Appendix L.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Mohammad Almansoori, Komal Kumar, and Hisham Cholakkal. Medagentsim: Self-evolving multi-agent simulations for realistic clinical interactions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 362–372. Springer, 2025.

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

Lili Chen, Mihir Prabhudesai, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. Self-questioning language models. *arXiv preprint arXiv:2508.03682*, 2025.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024b.

David Dai, Peilin Chen, Chanakya Ekbote, and Paul Liang. Qoq-med: Building multimodal clinical foundation models with domain-aware grpo training. *Advances in Neural Information Processing Systems*, 38:37406–37453, 2026.

Wei Dai, Peilin Chen, Malinda Lu, Daniel Li, Haowen Wei, Hejie Cui, and Paul Pu Liang. Climb: Data foundations for large scale multimodal clinical foundation models. *arXiv preprint arXiv:2503.07667*, 2025.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pp. 3563–3578, 2024.

Kaixuan Fan, Kaituo Feng, Haoming Lyu, Dongzhan Zhou, and Xiangyu Yue. Sophiavl-r1: Reinforcing mllms reasoning with thinking reward. In *International Conference on Learning Representations*, volume 2026, pp. 9290–9310, 2026.

Yunzhen Feng, Elvis Dohmatob, Pu Yang, Francois Charton, and Julia Kempe. Beyond model collapse: Scaling up with synthesized data requires verification. In *International Conference on Learning Representations*, volume 2025, pp. 89702–89730, 2025.

Huan-ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Yiran Wu, et al. A survey of self-evolving agents: What, when, how, and where to evolve on the path to artificial super intelligence. *arXiv preprint arXiv:2507.21046*, 2025.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, et al. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *arXiv preprint arXiv:2404.01413*, 2024.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*, 2023.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *International Conference on Learning Representations*, volume 2026, pp. 127924–127945, 2026.

Rebecca Soskin Hicks, Mikhail Trofimov, Dominick Lim, Rahul K Arora, Foivos Tsimpourlas, Preston Bowman, Michael Sharman, Chi Tong, Kavin Karthik, Arnav Dugar, et al. Healthbench professional: Evaluating large language models on real clinician chats. *arXiv preprint arXiv:2604.27470*, 2026.

Audrey Huang, Adam Block, Dylan Foster, Dhruv Rohatgi, Cyril Zhang, Max Simchowitz, Jordan Ash, and Akshay Krishnamurthy. Self-improvement in language models: The sharpening mechanism. In *International Conference on Learning Representations*, volume 2025, pp. 76687–76739, 2025a.

Chengsong Huang, Haolin Liu, Tong Zheng, Runpeng Dai, Langlin Huang, Jinyuan Li, Zongxia Li, Zhepei Wei, Yu Meng, and Jiaxin Huang. G-zero: Self-play for open-ended generation from zero data. *arXiv preprint arXiv:2605.09959*, 2026a.

Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. R-zero: Self-evolving reasoning llm from zero data. In *International Conference on Learning Representations*, volume 2026, pp. 130770–130790, 2026b.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. m1: Unleash the potential of test-time scaling for medical reasoning with large language models. *arXiv preprint arXiv:2504.00869*, 2025b.

Xiaoke Huang, Ningsen Wang, Hui Liu, Xianfeng Tang, and Yuyin Zhou. Synthesizing high-quality visual question answering from medical documents with generator-verifier lmms. In *International Conference on Learning Representations*, volume 2026, pp. 109770–109806, 2026c.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *IEEE transactions on medical imaging*, 2026.

Fangyuan Li, Pengfei Li, Shijie Wang, Junqi Gao, Jianxing Liu, Biqing Qi, and Yuqiang Li. WIST: Web-grounded iterative self-play tree for domain-targeted reasoning improvement. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31563–31585, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1456. URL https://aclanthology.org/2026.acl-long.1456/.

Bo Liu, Chuanyang Jin, Seungone Kim, Weizhe Yuan, Wenting Zhao, Ilia Kulikov, Xian Li, Sainbayar Sukhbaatar, Jack Lanchantin, and Jason Weston. Spice: Self-play in corpus environments improves reasoning. *arXiv preprint arXiv:2510.24684*, 2025.

Bo Liu, Simon Yu, Zichen Liu, Leon Guertler, Penghui Qi, Daniel Balcells, Mickel Liu, Cheston Tan, Weiyan Shi, Min Lin, et al. Spiral: Self-play on zero-sum games incentivizes reasoning via multi-agent multi-turn reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 35407–35434, 2026a.

Qicai Liu, Zhichao Hu, Tao Huang, Yupeng Niu, Xinche Zhang, Shanwu Ma, Chutong Lin, Goh Kim Huat, Hyeokkoo Eric Kwon, Feng Gao, et al. Evomdt: a self-evolving multi-agent system for structured clinical decision-making in multi-cancer. *npj Digital Medicine*, 9(1):124, 2026b.

Wei Liu, Siya Qi, Yali Du, and Yulan He. Self-play only evolves when self-synthetic pipeline ensures learnable information gain. *arXiv preprint arXiv:2603.02218*, 2026c.

Xueguang Ma, Qian Liu, Dongfu Jiang, Ge Zhang, Zejun Ma, and Wenhu Chen. General-reasoner: Advancing llm reasoning across all domains. *Advances in Neural Information Processing Systems*, 38:56596–56618, 2026.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*, 2023.

Aaron Nicolson, Jason Dowling, Douglas Anderson, and Bevan Koopman. Longitudinal data and a semantic similarity reward for chest x-ray report generation. *Informatics in Medicine Unlocked*, 50:101585, 2024.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Jing-Cheng Pang, Liu Sun, Chang Zhou, Xian Tang, Haichuan Ma, Kun Jiang, Jianlong Wang, Kai Zhang, Sijie Wu, Haoran Cai, Chenwei Wu, Xubin Li, and Xin Chen. Edco: Dynamic curriculum orchestration for domain-specific large language model fine-tuning, 2026. URL https://arxiv.org/abs/2601.03725.

Zackary Rackauckas. Rag-fusion: a new take on retrieval-augmented generation. *arXiv preprint arXiv:2402.03367*, 2024.

Shaohao Rui, Kaitao Chen, Weijie Ma, and Xiaosong Wang. Improving medical reasoning with curriculum-aware reinforcement learning. *arXiv preprint arXiv:2505.19213*, 2025.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

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

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

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

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*, 31(3):943–950, 2025.

Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. Crossing the reward bridge: Expanding rl with verifiable rewards across diverse domains. *arXiv preprint arXiv:2503.23829*, 2025.

Yuan Sui and Bryan Hooi. Conversation for non-verifiable learning: Self-evolving llms through meta-evaluation. *arXiv preprint arXiv:2601.21464*, 2026.

Bingning Wang, Haizhou Zhao, Huozhi Zhou, Liang Song, Mingyu Xu, Wei Cheng, Xiangrong Zeng, Yupeng Zhang, Yuqi Huo, Zecheng Wang, et al. Baichuan-m1: Pushing the medical capability of large language models. *arXiv preprint arXiv:2502.12671*, 2025.

Steffanie S Weinreich, R Mangon, JJ Sikkens, ME En Teeuw, and MC Cornel. Orphanet: a european database for rare diseases. *Nederlands tijdschrift voor geneeskunde*, 152(9):518–519, 2008.

Alex Wilf, Pranjal Aggarwal, Bryan Parno, Daniel Fried, Louis-Philippe Morency, Paul Pu Liang, and Sean Welleck. Propose, solve, verify: Self-play through formal verification. *arXiv preprint arXiv:2512.18160*, 2025.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, et al. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*, 2025.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6233–6251, 2024a.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pp. 199–214. World Scientific, 2024b.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candes, and Tatsunori Hashimoto. Synthetic continued pretraining. In *International Conference on Learning Representations*, volume 2025, pp. 44379–44421, 2025b.

Asaf Yehudai, Boaz Carmeli, Yosi Mass, Ofir Arviv, Nathaniel Mills, Assaf Toledo, Eyal Shnarch, and Leshem Choshen. Genie: Achieving human parity in content-grounded datasets generation. *arXiv preprint arXiv:2401.14367*, 2024.

Abdurrahim Yilmaz, Furkan Yuceyalcin, Rahmetullah Varol, Ece Gokyayla, Ozan Erdem, Donghee Choi, Ali Anil Demircali, Gulsum Gencoglan, Joram M Posma, Burak Temelkuran, et al. A synthetic data generation framework for scalable and resource-efficient medical ai assistants, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin gödel machine: open-ended evolution of self-improving agents. In *International Conference on Learning Representations*, volume 2026, pp. 104223–104294, 2026.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, et al. Ultramedical: Building specialized generalists in biomedicine. *Advances in Neural Information Processing Systems*, 37:26045–26081, 2024.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

Xiaotian Zhang, Yuan Wang, Zhaopeng Feng, Ruizhe Chen, Zhijie Zhou, Yan Zhang, Hongxia Xu, Jian Wu, and Zuozhu Liu. Med-u1: Incentivizing unified medical reasoning in llms via large-scale reinforcement learning. *arXiv preprint arXiv:2506.12307*, 2025.

Andrew Zhao, Yiran Wu, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. *Advances in Neural Information Processing Systems*, 38:105816–105879, 2026.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, pp. 4442–4457, 2025.

Yifei Zhou, Sergey Levine, Jason Weston, Xian Li, and Sainbayar Sukhbaatar. Self-challenging language model agents. *Advances in Neural Information Processing Systems*, 38:113959–113991, 2026.

A Comparison with related methods

Table 2 compares RSIMed with the related methods discussed in Section 2.

Table 2: Related methods and the features that matter for self-improvement on clinical tasks. Model sizes gives the range of parameter counts of the trained models. Answer check is how answers are verified, and a database verifier re-checks each generated reference answer against evidence freshly retrieved from a medical database. Dynamic generation means new tasks are written throughout training, so the data grow with compute. Difficulty calibration means task difficulty is set by the solver's current accuracy. Evidence-grounded means each reference answer is written from retrieved documents. Free-text answers means open answers graded for meaning, with no execution, option letter, or string match. Self improve means the model generates its own training data, and no stronger model writes or checks it. Clinical records means tasks are seeded from or evaluated on real patient records. Beats frontier means the trained model outperforms a frontier proprietary model on its target task.

Method	Domain	Model Sizes	Answer Check	Dynamic Generation	Difficulty Calibration	Evidence-Grounded	Free-Text Answers	Self Improve	Clinical Records	Beats Frontier
Absolute Zero (Zhao et al., 2026)	Code, math	3B–14B	execution	✓	✓	✗	✗	✓	✗	✗
R-Zero (Huang et al., 2026b)	Math, reasoning	3B–8B	majority vote	✓	✓	✗	✗	✓	✗	✗
SPICE (Liu et al., 2025)	Math, general	3B–8B	rule match	✓	✓	✓	✗	✓	✗	✗
WIST (Li et al., 2026)	General, medical exams	3B–14B	rule match	✓	✓	✓	✗	✓	✗	✗
MedVLSynther (Huang et al., 2026c)	Biomedical VQA	3B–7B	model verifier	✗	✗	✓	✗	✗	✗	✗
HuatuoGPT-o1 (Chen et al., 2024a)	Medical exams	7B–70B	GPT-4o verifier	✗	✗	✗	✓	✗	✗	✗
SCALEMED (Yilmaz et al., 2025)	Dermatology	11B	none	✗	✗	✗	✗	✗	✗	✓
EDCO (Pang et al., 2026)	Medical, law	3B–4B	answer key	✗	✓	✗	✗	✗	✗	✗
RSIMed (ours)	Clinical diagnosis	9B–27B	database verifier	✓	✓	✓	✓	✓	✓	✓

B Proposer pipeline details

The proposer service handles query generation, retrieval, task generation, and validation. The trainer consumes accepted items over HTTP while generation runs asynchronously. A pool holds up to 200 validated items, sampled in random order. Eight concurrent workers refill it whenever it is below capacity. If generation falls behind, the service can reuse previously accepted tasks from its history.

Knowledge base. The proposer retrieves from a vector index of medical literature, textbook passages, and question-answer resources, embedded with Qwen3-VL-Embedding-2B. The index holds eight source families, PubMed abstracts, Wikipedia, and medical textbooks from the MedRAG corpus, WikiDoc, PMC-VQA (Zhang et al., 2023), MIRAGE, PubMedQA (Jin et al., 2019), and CLIMB (Dai et al., 2025). Table 1 lists their sizes. Retrieved text retains its source label and is supplied to the proposer as evidence for task generation. The diagnostic experiments use text questions.

Query proposal and retrieval. The proposer writes ten queries per seed, covering mechanism, diagnostic criteria and workup, comparative effectiveness, adverse effects and complications, prognosis and natural history, atypical presentation, differential diagnosis against one or two named alternatives, imaging findings, epidemiology and risk factors, and related conditions. Each query retrieves the sixteen nearest passages. These are concatenated with source tags in the generation prompt. Query proposal uses temperature 0.8; question generation uses temperature 0.9. Both calls allow up to $12{,}288$ output tokens.

Training pool. The pool combines labeled training cases with generated tasks.

History conditioning. Each accepted item is embedded into a history collection with running solved and attempted counts. Before writing queries, the worker retrieves the ten most similar previous questions and their solve rates, each shortened to its first 400 characters, so this part of the prompt stays under 5,000 characters however long training runs. The prompt instructs the proposer to avoid repetition and topics the solver answers reliably, and to target failed or untested topics.

Validation. The schema check rejects empty questions or answers, format mismatches, and multiple-choice keys absent from the options. The model check retrieves the three nearest passages for the finished question, including option text for multiple-choice items, and evaluates it at temperature 0.2 using the proposer model. It rejects incoherent questions and answers directly contradicted by the passages. Inconclusive evidence leaves the candidate accepted. Rejections are logged with the contradicting passage. Candidates are also retained when retrieval returns no passages or the model check fails to return a verdict.

Difficulty feedback. As each rollout is scored, the trainer reports its exact-match correctness. The service maintains a sliding accuracy mean over the latest 64 reports and per-question counts. Both enter the generation prompt with a target solve rate of one half, and the side of the target on which the running accuracy lies sets the direction of the change. Above the target, the proposer writes finer clinical distinctions and, for multiple-choice questions, distractors closer to the correct answer. Below the target, it phrases its questions more precisely while keeping the same reasoning step, so they become easier by removing ambiguity.

C Analysis of generated tasks

This section describes what the proposer generated during the 27B SFT+RL run. The data are the run's archive of accepted tasks and the correctness of every rollout that the trainer reported back to the proposer. The reported correctness is the exact-match signal that drives the difficulty feedback. Each report is assigned to the training step whose logged completion time follows it. We analyze the $43{,}698$ well-formed generated tasks that were served during training and received a full group of eight reports.

The running accuracy that the proposer sees stays close to its target of one half. The proposer sees the running accuracy over the most recent rollouts, which pools both formats and the labeled cases. Averaged over each tenth of the run, this running accuracy is 0.47 in the first window and stays between 0.52 and 0.57 afterwards, close to the target of one half, while the solver's exact-match accuracy on labeled cases rises from 0.25 to 0.36 over the same period.

Over half of the tasks concern the seed's recorded diagnosis or a linked condition, although the proposer never sees that diagnosis. An independent classifier, GPT-5.5-Chat, labeled a sample of $1{,}584$ tasks, drawn evenly across training, with the clinical aspect each task tests and its relation to the recorded diagnosis of the seed case. The proposer never sees that diagnosis. Table 3 gives the shares.

Table 3: Clinical aspect of generated tasks and their relation to the seed case's recorded diagnosis, from an independent classifier on $1{,}584$ sampled tasks.

Aspect	Share	Relation to the seed's diagnosis	Share
Diagnosis	27.0%	Same disease	28.8%
Mechanism	23.4%	Clinically linked condition	25.4%
Treatment	18.5%	Other detail of the case	8.8%
Tests and imaging	14.4%	No identifiable link	37.0%
Pharmacology	8.6%		
Prognosis	4.5%		
Epidemiology	3.0%		

On later visits to a case, the proposer narrows the questions of earlier visits. The proposer returns to many training cases during the run, and on a later visit it can take up a thread from an earlier one and narrow it. Figure 8 shows one of the 129 training cases that the proposer visited at least three times over at least half of the run, chosen after reviewing every visit to all 129. All four visits to this case, a training admission recorded as epilepsy, produced tasks on alcohol withdrawal. The first task asks for the first-line drug class. The later ones ask when to give a dose under the CIWA-Ar score, which benzodiazepine to use in cirrhosis, and which agent to use when benzodiazepines fail. Each question narrows the one before it to a more specific clinical decision.

Medical textbooks are retrieved far more often than their share of the index. Table 1 lists the sources of the knowledge base with their sizes, and Table 4 shows how often each appears in the evidence of generated tasks. The archive stores the first $4{,}000$ characters of the evidence for each task, which hold its top-ranked passages. The medical textbooks make up 0.2% of the index but appear in the evidence of 28% of the tasks.

Table 4: Use of each knowledge-base source in the evidence of generated tasks. Tasks citing is the share of accepted generated tasks whose evidence includes the source, and top passage is the share whose highest-ranked passage comes from it, over the 43,698 analyzed tasks of the 27B run.

Source	Tasks citing	Top passage
PubMed abstracts	97.3%	61.9%
Wikipedia	30.5%	5.0%
Medical textbooks	28.2%	4.6%
WikiDoc	50.6%	25.0%
PMC-VQA	10.8%	3.5%
MIRAGE	0.2%	<0.1%
PubMedQA	<0.1%	<0.1%
CLIMB	<0.1%	0%

D TRAINING CONFIGURATION

The GRPO runs use two to four B200 GPUs depending on the model size, FSDP2, and vLLM-backed asynchronous rollouts. Settings include learning rate 2×10^{-7}, training batch size 64, mini-batch size 32, eight rollouts per prompt, temperature 1.0, maximum prompt length 8192, and maximum response length 4096. The main 27B configuration uses the `kl_cov` policy loss, which penalizes changes in token log probability for tokens with the highest covariance between log probability and advantage. Its selection ratio is 0.001 and penalty coefficient is 1.0. An overlong-response penalty applies near the response limit. The full pipeline initializes the actor from the merged supervised checkpoint; RL-only comparisons initialize from base weights. The proposer and judge use separately served frozen models. Validation uses greedy decoding at a fixed cadence. We report the last checkpoint of every run that does not collapse. Only baseline runs collapse, and for these we report the last checkpoint before the collapse so that the comparison stays meaningful.

The scaling runs initialize from the supervised checkpoint.

Warm-start trace generation. For a trace τ with extracted answer $\hat{y}$ and reference y^*, we accept the trace when $A(\tau) = 1$, where

$$A(\tau) \;=\; \underbrace{\mathbf{1}[\mathrm{fmt}(\tau)]}_{\text{think-then-box format}} \cdot \underbrace{\mathbf{1}[\mathrm{icd}(\hat{y}, y^*)]}_{\text{ICD-aware match}} \cdot \underbrace{\mathbf{1}[\tau \notin \mathcal{L}]}_{\text{no leak}} \cdot \underbrace{\mathbf{1}[\ell_{\min} \le |\tau| \le \ell_{\max}]}_{\text{length}} \cdot \underbrace{\mathbf{1}[\max_{\tau' \in \mathcal{A}} \mathrm{sim}(\tau, \tau') < \delta]}_{\text{not a near-duplicate}} \tag{2}$$

Here fmt checks the reasoning and boxed-answer format. When both answers contain ICD codes, icd accepts equal codes or parent–child prefix matches; otherwise it uses case-insensitive equality or substring matching. The pattern set $\mathcal{L}$ detects explicit references to the supplied diagnosis, such as "the provided diagnosis" or "working backward." The set $\mathcal{A}$ contains the case's previously accepted traces. We sample traces at temperature 0.8 in rounds of four candidates, with at most three rounds per case, until three traces pass the filters. Each trace must contain at least 200 characters of reasoning and between 400 and 1400 estimated tokens. We estimate length at 3.6 characters per token. Thirteen case-insensitive patterns detect references to the supplied diagnosis. We also strip parenthetical or labeled ICD codes from the reasoning and reject traces with remaining codes. To remove near-duplicates, we compare the first 1200 characters of normalized reasoning with previously accepted traces using a sequence-match threshold of 0.7.

Reward grading. Judge calls use deterministic decoding with reasoning disabled. We penalize responses that enter the final 512 tokens of the 4096-token response budget. The composite reward weights remain fixed.

Compute accounting. We estimate the FLOPs of a forward pass over a sequence of T tokens as

$$F[T] \;=\; \underbrace{2N_{\mathrm{act}}T}_{\text{parameters}} + \underbrace{2LdT^2}_{\text{attention}}, \tag{3}$$

where N_{act} is the number of active parameters, L the layer count, and d the model dimension. The parameter term includes prefill and decoding with cached keys and values. The attention term sums

$4Ld\,i$ FLOPs over token positions i. It contributes under 2% at our rollout lengths, 10% at 33k tokens, and 29% at 124k tokens.

For every training curve in Figure 1(b), the per-step cost sums rollout generation, the policy update, task proposal, and reward judging. Rollout generation uses F over the step's rollouts. The policy update costs $6NT$ for the forward-backward pass plus $2NT$ for the log-probability pass. We use $Q = 64$ proposed tasks and $R = 512$ rollouts per step. N, L, and d are those of the model being trained, since proposer, solver, and judge share one base model. We use $N = 2.78 \times 10^{10}$, $L = 64$, $d = 5120$ for Qwen3.6-27B and $N = 9.65 \times 10^{9}$, $L = 32$, $d = 4096$ for Qwen3.5-9B, so a 9B step costs roughly a third of a 27B step and the two curves span different compute ranges for the same number of steps. The static-pool run incurs its proposal cost once, before training, for the $15{,}000$ tasks in its frozen pool. This costs $2N \cdot 15{,}000 \cdot \tau_{\text{prop}}$, about 8×10^{18} FLOPs or approximately twenty training steps, with no further proposal cost during training. The training-set run has no proposal cost. Token counts come from the training log except for the proposer and judge, which run on a separate inference server that does not log them; those two use $\tau_{\text{prop}} = 10{,}000$ tokens per accepted task, covering retrieved passages, instructions, the generated task, and rejected drafts, and $\tau_{\text{judge}} = 1{,}600$ tokens per rollout. Together they account for about a quarter of estimated compute. The run spans 1.7×10^{18} to 2.7×10^{20} FLOPs; the supervised warm start adds under one percent and is omitted.

GPU hours. We convert FLOPs to GPU hours with the measured wall-clock time of the training steps, which excludes validation passes, as the FLOP count does. RSIMed-27B trains on four B200 GPUs for 222 hours, or 887 GPU hours over 2.75×10^{20} FLOPs, which gives 323 GPU hours per 10^{20} FLOPs. RSIMed-9B trains on two B200 GPUs for 153.5 hours, or 307 GPU hours over 9.6×10^{19} FLOPs, which gives 319 GPU hours per 10^{20} FLOPs. Both models therefore need about 320 B200 GPU hours per 10^{20} FLOPs. These hours count the GPUs of the training node; the proposer and judge run on a separate inference server.

E Frozen-Backbone Inference Baselines

All methods use the same 2452 MIMIC-IV cases, clinical text, and available images as the direct evaluation. Model parameters remain fixed. The original clinical text cap is 9000 characters, image resolution is capped at $65{,}536$ pixels per image, and the final answer budget is 8192 tokens. Qwen's thinking mode is disabled, matching the reference evaluation. The final answer extractor and judge are fixed across methods. Ground-truth diagnoses are used only for grading; they are not available to retrieval, planning, critique, or answer selection.

Shared retrieval. Medical RAG, RAG-Fusion, and i-MedRAG run on a common corpus and backbone. Retrieval uses the complete medical database of Table 1, with the same embedding model and cosine search as the proposer. We retrieve 32 candidates, reject duplicate passages and fragments shorter than 80 characters, retain up to eight passages, and cap each retained passage at 1200 characters. If no usable passage survives, the candidate depth increases to 128 and then 512 while keeping the final context budget fixed. If the search still returns no usable evidence, the model is told that references are unavailable.

Retrieval methods. Single-pass medical RAG retrieves using the case text and supplies the resulting passages to the solver. RAG-Fusion adds three model-generated search queries, combines their rankings with the original query using reciprocal-rank fusion with constant 60, and retains eight passages for the final answer. i-MedRAG performs three rounds of two follow-up questions. Each follow-up is answered with retrieval, using at most 1024 answer tokens. The accumulated question-answer pairs guide the next round and the final diagnosis. The same frozen backbone generates the questions, answers the follow-ups, and makes the final prediction.

Other inference methods. Self-consistency samples five independent answers at temperature 0.7 and selects the most frequent extracted diagnosis after lowercasing and punctuation normalization; ties select the earliest sample. Self-refinement produces a draft, a self-critique against the case evidence, and a revised diagnosis. Chain-of-verification produces a draft and three factual verification questions, answers each independently without showing the draft, and revises the diagnosis using those answers.

Critiques and verification answers are capped at 1024 tokens. Other than self-consistency, generation uses temperature zero.

F JUDGING

MIMIC-IV comparisons use stored generations regraded with a fixed prompt and GPT-5.5-Chat. Regrading covers all 197 validation passes of the continuous 27B run, all 32 passes of the static-pool control, and every fourth pass of the 9B run through its last checkpoint. Each pass contains 2452 cases. The training-set curve in Figure 1(b) and its row in Table 9 use the GPT-5.5-Chat grades recorded during training.

HealthBench Professional headline results and the slice analysis in Table 6 use GPT-5.5-Chat with the official grading protocol. Slice scores are recomputed from all 525 saved per-example grades and verified against each run's official overall score. Length adjustment is applied per example before averaging; each group mean is clipped to $[0, 1]$.

Task-validator audit. We uniformly sample 200 historically accepted and 200 rejected candidates from the full-pipeline generation archive. Sampling uses seed 42. An independent GPT-5.5-Chat assessor, blinded to the historical acceptance decision, labels each task valid, invalid, or uncertain using its stem, proposed answer, and freshly retrieved medical references. Of the 200 accepted rows, 147 are generated tasks and 53 are labeled training cases, which the pool serves without validation. Excluding uncertain assessments, 84.1% of the accepted generated tasks are valid, and the validator keeps 99.1% of all valid tasks, weighting the two samples by the sizes of the accepted and rejected populations.

Pricing. Figure 1(a) uses measured token counts and the rates in Table 5.

Table 5: Rates used to estimate the inference costs in Figure 1(a), in US dollars per million tokens. The 9B rate scales the 27B rate by $9.65\text{B}/27.78\text{B} = 0.35$. Frontier-model rates use API list prices, with billed reasoning tokens included in output counts.

Model	Input	Output
RSIMed-27B	0.05	0.30
RSIMed-9B	0.017	0.104
GPT-5.6-luna	0.20	1.20
GPT-5.6-terra	2.00	12.00
GPT-5.6-sol	5.00	30.00

G ADDITIONAL CLINICAL RESULTS

Diagnostic results use the full RSIMed-27B pipeline unless a different training variant is named. HealthBench results for our method use RSIMed-RL-27B, graded by GPT-5.5-Chat, because RL from the base weights gives the best HealthBench Professional score of the training variants in Table 11, while the full pipeline gives the best MIMIC-IV accuracy.

The full 27B model exceeds its base in every ICD chapter, as Table 9 shows. Table 6 reports the HealthBench Professional breakdown, where RSIMed-RL-27B improves over its base on all three use cases and all three slices. The red-teaming score is 0.047 versus 0.030.

H TEACHER ANALYSIS WITH PROTOCOL, FULL MEASUREMENTS, AND EXAMPLES

We analyze the task pools consumed by the five training runs using the untrained student and three additional solvers. We report the protocol, measurements, and example tasks below. A teacher here supplies generated training tasks, while the student is the model trained on those tasks. Teacher

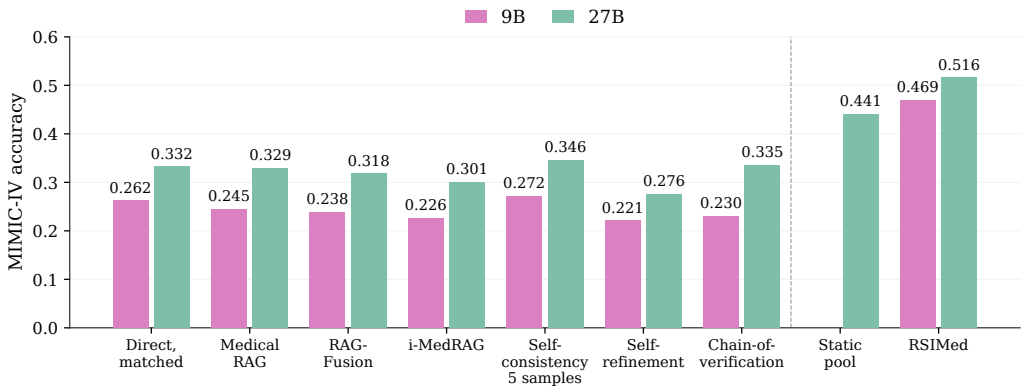

Figure 9: **MIMIC-IV inference and training controls.** RSIMed attains the highest accuracy on both backbones, compared with direct inference and six alternative inference methods on the same backbones. The dashed line separates the frozen-backbone methods from SFT+RL training.

Table 6: HealthBench Professional by use case and dataset slice, length-adjusted, graded by GPT-5.5-Chat, for the variants reported in Table 10. Group sizes are consult 236, writing 142, research 147; good-faith typical 256, good-faith difficult 78, red-teaming difficult 191.

Model	Overall	Consult	Writing	Research	GF typ.	GF diff.	RT diff.
Qwen3.6-27B base	0.291	0.261	0.158	0.467	0.521	0.174	0.030
RSIMed-RL-27B	0.303	0.277	0.175	0.469	0.522	0.212	0.047

Table 7: Performance by training-data source. The source comparison uses RL without a supervised warm start. Both benchmarks are graded by GPT-5.5-Chat. "w/ X" denotes training with teacher X.

	MIMIC-IV	HealthBench Prof.	
Training data source	Overall	Len-adj.	Raw
Qwen3.6-27B, untrained	0.334	0.291	0.363
w/ Kimi-K2.6	0.309	0.298	0.366
w/ DeepSeek-V4-Pro	0.341	0.289	0.362
w/ GPT-5.5	0.343	0.294	0.366
RSIMed-RL-27B	-	**0.303**	**0.376**
RSIMed-27B, full pipeline (ours)	**0.516**	–	–

quality therefore concerns what its examples teach the student, in addition to the teacher's own ability to answer questions.

H.1 PROTOCOL

Task pools. For each source, we retain tasks served during training with a stem of at least eight words and a non-empty key. We exclude malformed items, including JSON placeholder stems from the student's proposer. We sample 200 multiple-choice and 200 short-answer tasks uniformly from each filtered pool.

Learning signal. Each task was answered eight times by the untrained student with the training-time sampling settings, using temperature 1.0, top-p 1.0, repetition penalty 1.1, a 4096-token budget, and the same system prompts as the trainer. Multiple-choice answers were scored by letter match; short answers were scored against the proposer's key by the reference grader of Table 9, GPT-5.5-Chat with a lenient rubric. We report the fraction of groups with unequal correctness rewards and the mean within-group variance $p(1-p)$, where p is the observed correct-answer fraction. Equal-reward groups have zero group-relative advantage. This correctness-based audit estimates learning opportunities; the training reward also includes the terms in Equation 1.

Solvers. The same 800 tasks written by the student and by GPT-5.5 were also answered eight times each by Gemma-4-31B, by Qwen3.5-9B with a 12k-token budget, and by GPT-5.5 itself with eight samples per call and low reasoning effort. A further 400 multiple-choice tasks per proposer were answered by the student, Gemma-4-31B, and GPT-5.5, which brings the multiple-choice test to 600 tasks per proposer. "Sound" tasks are those GPT-5.5 answers correctly in all eight samples.

Passage dependence. Each short-answer task was answered again by the student with the passages the proposer had retrieved placed above the question. The accuracy gain measures how much the retrieved context helps the student answer the task.

Stem leakage. A stem leaks its key when a content word of the key appears in the stem, provided that the word has five letters or more and is not a generic clinical word; for multiple choice the key is the text of the correct option.

Redundancy and coverage. Tasks were embedded with the project's Qwen3-VL-Embedding-2B retrieval embedder. On a random 900-task subset per pool we report the share of tasks whose nearest neighbour exceeds cosine 0.9, which we count as near-duplicates, the number of greedy clusters at cosine 0.8, and the participation ratio of the embedding covariance; coverage is the share of the 2452 evaluation admissions with some task above cosine 0.7.

Table 8: Share of eight-rollout groups with unequal rewards, expressed as a percentage, and accuracy, when the tasks written by the student and by GPT-5.5 are answered by four different solvers. "Sound" denotes tasks GPT-5.5 answers correctly in all eight samples. GPT-5.5 therefore has zero reward variance on this subset.

		Tasks by GPT-5.5			Tasks by the student		
Solver	Format	Acc.	Signal	Sound signal	Acc.	Signal	Sound signal
GPT-5.5	MCQ	1.000	0	0	0.978	2	0
GPT-5.5	Short	0.956	1	0	0.849	13	0
Gemma-4-31B	MCQ	0.991	1	2	0.976	2	2
Gemma-4-31B	Short	0.911	16	17	0.786	22	19
Qwen3.6-27B student	MCQ	0.985	11	11	0.941	22	21
Qwen3.6-27B student	Short	0.925	22	22	0.787	36	31
Qwen3.5-9B	MCQ	0.978	8	5	0.929	17	14
Qwen3.5-9B	Short	0.851	22	20	0.706	40	37

On the expanded multiple-choice comparison with 600 tasks per proposer, of which 300 written by GPT-5.5 and 293 written by the student are sound, accuracy on the sound subset is 0.977 for the

student on GPT-5.5's tasks and 0.959 on its own. Gemma-4-31B scores 0.995 and 0.980, respectively. Self-generated tasks are harder for both solvers.

H.2 EXAMPLES

Tasks and answers below retain their clinical content, with punctuation and abbreviation presentation normalized for readability. The eight student answers are the extracted boxed answers of the eight rollouts, in sampling order. The examples illustrate differences in generated task difficulty and answer variation. Correctness counts refer to the specified reference grader.

Self, Qwen3.6-27B. A multiple-choice task the student answers correctly in all eight rollouts. The reference key is B.

> A patient presents with severe pancytopenia and marked reticulocytopenia. To differentiate acquired aplastic anemia from myelodysplastic syndrome, bone marrow examination and cytogenetic analysis are performed. Which combination of findings most specifically supports a diagnosis of aplastic anemia?
>
> A. Hypercellular marrow with multilineage dysplasia and recurrent chromosomal translocations
> B. Hypocellular marrow largely replaced by adipose tissue with normal karyotype
> C. Normocellular marrow with increased reticulin fibrosis and ring sideroblasts
> D. Hypercellular marrow with >20% circulating blasts and cytogenetic abnormalities

A short-answer task with mixed student correctness has 6 of 8 answers accepted by the reference grader.

> In Crohn's disease management, diffusion-weighted magnetic resonance enterography is often used to differentiate active inflammatory strictures from fibrotic strictures to guide medical versus surgical therapy. Which quantitative MRI parameter is primarily measured for this distinction, and how does its value characteristically change during active inflammation?
>
> *Reference answer.* Apparent diffusion coefficient decreases
> *Student answers.* Decreased apparent diffusion coefficient; Decreased apparent diffusion coefficient; Apparent diffusion coefficient; increases; acute pancreatitis; Decreased apparent diffusion coefficient; Apparent diffusion coefficient, which decreases; Apparent diffusion coefficient decreases; Apparent diffusion coefficient, which decreases
> *GPT-5.5 answers.* Apparent diffusion coefficient decreases; Apparent diffusion coefficient; decreased in active inflammation; Apparent diffusion coefficient decreases; Apparent diffusion coefficient; decreased ADC; Decreased apparent diffusion coefficient; Decreased apparent diffusion coefficient; Apparent diffusion coefficient; decreased ADC; Apparent diffusion coefficient; decreased in active inflammation

GPT-5.5. A multiple-choice task the student answers correctly in all eight rollouts. The reference key is A.

> A 26-year-old fit man is newly diagnosed with Philadelphia chromosome-negative precursor T-cell acute lymphoblastic leukemia/lymphoblastic lymphoma with marrow involvement and no CNS disease. The team is choosing initial therapy and wants to avoid unnecessary hospitalization and late toxicities while preserving disease control. Which management plan is most appropriate?
>
> A. Use a pediatric-inspired, asparaginase-containing ALL regimen with CNS prophylaxis rather than defaulting to adult hyper-CVAD
> B. Use CHOP because lymphoblastic lymphoma is managed like other aggressive non-Hodgkin lymphomas
> C. Use hyper-CVAD induction and plan allogeneic hematopoietic cell transplantation in first complete remission for all patients regardless of risk or response
> D. Use vincristine and prednisone alone for induction to reduce early infectious mortality in young adults

A short-answer task with mixed student correctness has 5 of 8 answers accepted by the reference grader.

> A diabetic patient with severe peripheral vascular disease undergoes transmetatarsal amputation for forefoot gangrene, followed by infrainguinal bypass using a mapped patent great saphenous vein. Postoperatively the bypass remains patent, blood cultures are negative, and chest imaging shows no pneumonia, but the amputation stump dehisces with progressive necrosis and failure to heal despite wound VAC and antibiotics. What vascular problem can explain limb-salvage failure despite a patent bypass graft?
>
> *Reference answer.* poor distal runoff
> *Student answers.* Inadequate distal runoff; inadequate distal runoff; Distal microembolization; Inadequate distal runoff; Diabetic microvascular disease; Diabetic microangiopathy; poor distal runoff; Inadequate distal runoff
> *GPT-5.5 answers.* poor pedal arch perfusion; poor distal pedal runoff; poor pedal arch outflow; occluded pedal arch; severe pedal small-vessel disease; poor pedal arch outflow; severe inframalleolar occlusive disease; severe inframalleolar occlusive disease

Qwen3.5-397B. A multiple-choice task the student answers correctly in all eight rollouts. The reference key is B.

> A 68-year-old female with type 2 diabetes mellitus and stage 3 chronic kidney disease with an eGFR of 45 mL/min presents with altered mental status. She has been on chronic metformin therapy. Laboratory evaluation reveals a pH of 7.10, bicarbonate of 12 mEq/L, an anion gap of 24 mEq/L, and a lactate level of 9 mmol/L. Ketones are negative. Based on the provided medical literature, what is the primary pathophysiological mechanism responsible for this patient's metabolic derangement?
>
> A. Increased lipolysis leading to excessive beta-hydroxybutyrate accumulation
> B. Inhibition of mitochondrial respiratory chain complex I and hepatic gluconeogenesis
> C. Impaired renal excretion of endogenous organic acids due to low GFR alone
> D. Direct stimulation of anaerobic glycolysis in skeletal muscle due to hypoxia

A short-answer task with mixed student correctness has 6 of 8 answers accepted by the reference grader.

> In an immunocompromised patient with myelodysplastic syndrome and fever, chest CT reveals multiple small peribronchial nodules under 5 mm without cavitation. Bronchoalveolar lavage isolates Parainfluenza type 3, while tuberculosis and fungal studies are negative. Based on literature describing this virus's specific radiographic signature in immunocompromised hosts distinct from miliary tuberculosis or Pneumocystis pneumonia, what is the diagnosis?
>
> *Reference answer.* Parainfluenza virus type 3 pneumonia
> *Student answers.* Parainfluenza type 3 pneumonia; suspicious for respiratory syncytial virus infection; Parainfluenza type 3 pneumonia; Parainfluenza virus 3 pneumonitis; human parainfluenza virus 3 pneumonia; Parainfluenza type 3 pneumonitis; Respiratory syncytial virus pneumonia; Parainfluenzal pneumonitis
> *GPT-5.5 answers.* Parainfluenza type 3 pneumonia; Parainfluenza type 3 pneumonia; Parainfluenza type 3 pneumonia; Parainfluenza type 3 pneumonia; Parainfluenza type 3 pneumonia; Parainfluenza type 3 pneumonia; Parainfluenza type 3 pneumonia; Parainfluenza type 3 pneumonia

DeepSeek-V4-Pro. A multiple-choice task the student answers correctly in all eight rollouts. The reference key is B.

> A 52-year-old man with type 2 diabetes and end-stage renal disease is about to start dialysis. He is relatively young and otherwise stable. According to observational studies comparing dialysis modalities in diabetic patients, which modality has been associated with equal or better survival during the first few years of treatment for younger diabetic individuals?
>
> A. In-center hemodialysis
> B. Peritoneal dialysis
> C. Both modalities offer identical survival regardless of age
> D. Immediate preemptive renal transplantation

A short-answer task with mixed student correctness has 4 of 8 answers accepted by the reference grader.

> Which cardiomyopathy can present on an electrocardiogram with transient J waves that appear before the development of ST-segment elevation or T wave inversion?
> *Reference answer.* Takotsubo cardiomyopathy
> *Student answers.* Arrhythmogenic right ventricular cardiomyopathy; Brugada syndrome; no boxed answer; Takotsubo cardiomyopathy; Takotsubo cardiomyopathy; no boxed answer; Takotsubo cardiomyopathy; Takotsubo cardiomyopathy
> *GPT-5.5 answers.* Takotsubo cardiomyopathy; Takotsubo cardiomyopathy; Takotsubo cardiomyopathy; Takotsubo cardiomyopathy; Takotsubo cardiomyopathy; Takotsubo cardiomyopathy; Takotsubo cardiomyopathy; Takotsubo cardiomyopathy

Kimi-K2.6. A multiple-choice task the student answers correctly in all eight rollouts. The reference key is B.

> A 12-year-old boy with new-onset type 1 diabetes presents with severe DKA with pH 6.92, serum bicarbonate 5 mmol/L, and PaCO2 8 mmHg. His parents report he drank several liters of juice and water over the preceding 12 hours. Thirty minutes after initiating isotonic crystalloid resuscitation and an insulin infusion, he becomes obtunded with sluggish pupillary responses. Which characterization and immediate management step is most consistent with the pathophysiology and risk factors of this complication?
>
> A. Late-onset cerebral edema caused by rapid insulin-mediated correction of serum osmolarity; slow the insulin infusion rate.
> B. Early-onset cerebral edema associated with pre-hospital fluid intake and severe hypocapnia; administer mannitol and consider hyperventilation.
> C. Excessive parenteral fluid administration causing dilutional hyponatremia and brain swelling; restrict fluids and give furosemide.
> D. Progression of severe metabolic acidosis causing cerebral hypoperfusion; administer sodium bicarbonate and increase the fluid rate.

A short-answer task with mixed student correctness has 5 of 8 answers accepted by the reference grader.

> A patient from a Lyme-endemic area has intermittent knee monoarthritis with negative RF, ANA, and blood/synovial cultures, and negative serum Lyme serology by two-tier testing, but synovial fluid PCR is positive for Borrelia DNA. What specific subtype of Lyme arthritis is most likely?
> *Reference answer.* Seronegative Lyme arthritis
> *Student answers.* Seronegative Lyme arthritis; Infection-driven Lyme arthritis; Destructive, necrotizing Lyme arthritis; Seronegative Lyme arthritis; seronegative Lyme arthritis; Seronegative Lyme arthritis; Recurrent monoarticular Lyme arthritis; Early Lyme arthritis
> *GPT-5.5 answers.* seronegative Lyme arthritis; seronegative Lyme arthritis; seronegative Lyme arthritis; seronegative Lyme arthritis; seronegative Lyme arthritis; seronegative Lyme arthritis; seronegative Lyme arthritis; seronegative Lyme arthritis

GPT-5.5 answering its own task. GPT-5.5 gives equivalent correct answers in all eight samples; the student is correct in six.

> A 67-year-old man is evaluated after an incidental monoclonal IgG kappa spike is found. Bone marrow biopsy shows 18% CD138-positive kappa-restricted plasma cells. Hemoglobin, creatinine, calcium, and skeletal survey are normal, and he has no bone pain. Whole-body MRI, however, shows two separate 6-mm focal marrow lesions. Under current IMWG criteria, what diagnosis should be made rather than MGUS or smoldering myeloma?
> *Reference answer.* Multiple myeloma
> *GPT-5.5 answers.* multiple myeloma; active multiple myeloma; multiple myeloma; active multiple myeloma; multiple myeloma; multiple myeloma; active multiple myeloma; active multiple myeloma
> *Student answers.* multiple myeloma; active multiple myeloma; Multiple myeloma; Multiple myeloma; active multiple myeloma; none; Multiple myeloma; none

A student-generated task with mixed correctness. GPT-5.5 answers correctly in all eight samples; the student in four.

> In a patient with hepatocellular carcinoma being evaluated for transcatheter arterial chemoembolization, complete portal vein thrombosis is traditionally considered a major contraindi-

cation. What catastrophic hepatic complication does this precaution specifically aim to prevent?

Reference answer. Massive necrosis of the non-tumoral liver

Student answers. acute hepatic failure; acute hepatic failure; massive hepatic necrosis; Hepatic infarction; acute liver failure; massive hepatic necrosis; acute liver failure; massive hepatic necrosis

I EXAMPLE GENERATED TASKS

The following tasks and retrieval queries are reproduced verbatim from the proposer's pool. They illustrate questions generated about clinical mechanisms and treatment decisions.

Example 1. *Retrieval query.* "What is the pathophysiological basis for the NIH Stroke Scale's differential weighting of left versus right hemisphere stroke deficits?"

Generated question. "In patients with right hemisphere stroke, how might the NIHSS scoring system impact clinical trial eligibility if a low score is used as an exclusion criterion?"

A. The NIHSS may underrepresent right hemisphere deficits, leading to under-enrollment of these patients. **Correct answer.**
B. The NIHSS is equally sensitive to both hemispheres and does not favor one side.
C. The NIHSS penalizes right hemisphere strokes more heavily due to greater motor impairment.
D. Right hemisphere strokes are always excluded because they have higher NIHSS scores than left hemisphere strokes.

The NIH Stroke Scale (NIHSS) summarizes stroke deficits. This example tests whether the model connects the scale's weighting to who qualifies for a trial. The retrieved passages support the reference answer by describing the allocation of seven points to language and two to neglect, which is a failure to attend to one side of space.

Example 2. *Retrieval query.* "What is the pathophysiological rationale for using vasopressors during aortic perfusion and oxygenation in cardiac arrest patients?"

Generated question. "In a patient undergoing aortic balloon occlusion during CPR, which strategy is most likely to enhance myocardial perfusion while minimizing adverse effects on cerebral blood flow, based on hemodynamic principles and pharmacologic profiles?"

A. Intravenous vasopressin administration
B. Intra-aortic vasopressin delivery above the balloon occlusion **Correct answer.**
C. Nitroglycerin infusion to reduce afterload
D. High-dose epinephrine to increase contractility

The answer choices describe interventions drawn from the retrieved evidence. This task asks the model to connect drug delivery and blood flow. Myocardial perfusion means blood reaching the heart muscle, while cerebral blood flow supplies the brain.

J DETAILED RESULTS AND ANALYSIS

J.1 COMPUTE SCALING AND INFERENCE COST

Figure 1(b) shows that continuous task generation supports sustained improvement with training compute. The 27B runs share supervised initialization, reward, judge, and optimizer, and their compute covers proposal, rollouts, judging, and policy updates, as detailed in Appendix D. The 27B model increases from 0.421 to 0.516 accuracy at approximately 3.9 points per 10^{20} FLOPs, with no observed saturation. A linear fit explains 96% of the variance, compared with 83% for a log-linear fit. The run includes 197 validation passes over all 2452 cases.

RL on the labeled training set peaks at 0.463 and then collapses to 0.133. Training on the static pool peaks at 0.441 and declines to 0.359 over the final thirty steps. The comparison with these two fixed task sets shows the benefit of generating additional tasks as training progresses.

The 9B model levels off at about 0.46 from 5×10^{19} FLOPs onward and reaches 0.469 at its last checkpoint, after 9.6×10^{19} FLOPs. Its training steps cost approximately one third as much as those of the 27B model, allowing it to exceed RSIMed-27B and the static pool at matched compute.

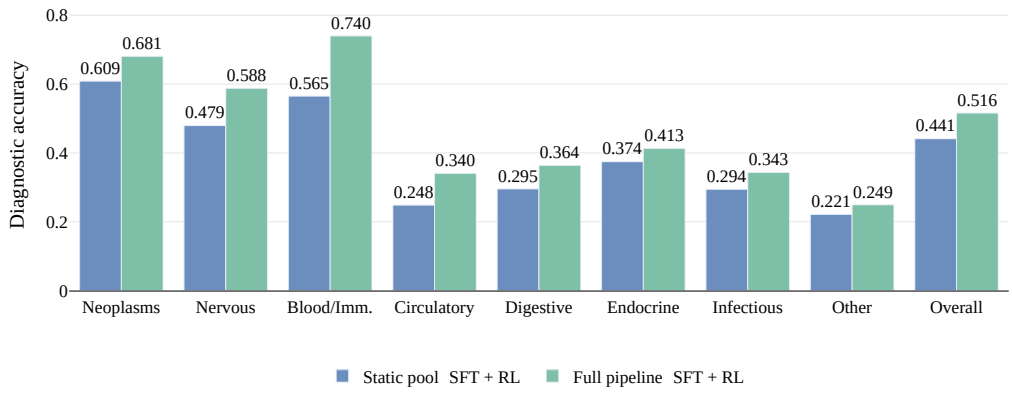

Figure 10: **Continuous generation versus a static task pool.** The full 27B pipeline improves full-set MIMIC-IV accuracy in every ICD chapter and overall from 0.441 to 0.516 relative to a frozen 15,000-task pool, with both runs initialized by supervised fine-tuning and scored by the same fixed judge.

Even the cheapest GPT-5.6 setting costs twice as much as RSIMed-27B and is less accurate. Figure 1(a) compares diagnostic accuracy with the estimated cost of answering all 2452 cases. The 27B and 9B models cost \$0.57 and \$0.18, respectively, using the serving rates of Table 5 and measured response lengths of 383 and 292 output tokens per case. We scale the 9B serving rate by its relative per-token compute, $0.35\times$ the 27B rate. RSIMed-27B writes about 380 output tokens per case, so reading the admission record and writing the answer account for roughly equal shares of its cost. Even the cheapest GPT-5.6 setting, Luna without reasoning, costs twice as much as RSIMed-27B and reaches 0.326. From no reasoning to the highest effort, accuracy rises by 7.3 points for Luna, 4.8 points for Terra, and 5.3 points for Sol, while reasoning tokens make up 77% to 98% of the billed output at every setting that reasons.

J.2 Diagnostic Performance

Table 9 reports the full pipeline's diagnostic performance. The two-stage 27B model reaches 0.516 overall accuracy and improves over its base in every ICD chapter. The largest gains include blood and immune disorders, where accuracy rises from the base model's 0.332 to 0.740, and endocrine and metabolic diseases, where it rises from 0.164 to 0.413.

Among frozen-backbone inference methods, self-consistency achieves the highest overall accuracy, at 0.272 for 9B and 0.346 for 27B, compared with matched direct scores of 0.262 and 0.332. The shared-index RAG adaptations score below the matched direct evaluation on both backbones. The 27B static-pool training control reaches 0.441, while the full pipeline reaches 0.516.

Figure 10 compares the full pipeline with static-pool training. Continuous generation achieves higher accuracy in every ICD chapter and improves overall accuracy by 7.5 points. The largest chapter gain is in blood and immune disorders, from 0.565 to 0.740, and nervous-system accuracy rises from 0.479 to 0.588 and circulatory accuracy from 0.248 to 0.340.

J.3 What the RL Stage Adds

Table 11 gives the scores of each training stage on both benchmarks. RL supplies 52% of the total MIMIC-IV gain of the 27B model and 87% of that of the 9B model, whose warm start moves accuracy only from 0.261 to 0.288 before RL lifts it to 0.469. We regrade the supervised checkpoint and the final checkpoint case by case with the same fixed judge. The RL stage, which trains on generated tasks mixed with the labeled cases, solves 329 evaluation cases that the supervised checkpoint missed and loses 98 that it had solved, a net gain of 231 of the 2,452 cases. The gains appear in every ICD chapter and are largest for nervous-system disease, neoplasms, and blood and immune disorders, with net gains of 52, 51, and 50 cases.

Table 9: MIMIC-IV rare-disease diagnosis on the 2,452 evaluation cases, overall and by ICD chapter, scored with a fixed judge, as detailed in Appendix F. Chapter sizes are Neopl. 695, Endo. 494, Nerv. 388, Blood 223, Circ. 141, Infect. 102, Digest. 88, Other 321.

	MIMIC-IV rare disease								
Model	Neopl.	Endo.	Nerv.	Blood	Circ.	Infect.	Digest.	Other	Overall
Open medical models									
MedGemma-1.5-4B	0.227	0.121	0.070	0.175	0.021	0.147	0.125	0.028	0.131
Lingshu-7B	0.249	0.087	0.149	0.224	0.099	0.118	0.170	0.044	0.155
Lingshu-32B	0.420	0.152	0.216	0.242	0.113	0.186	0.159	0.081	0.237
Untrained base models									
Qwen3.5-9B	0.452	0.079	0.325	0.269	0.156	0.225	0.205	0.118	0.261
Gemma-4-31B	0.468	0.097	0.302	0.336	0.149	0.206	0.295	0.128	0.275
Qwen3.6-27B	0.548	0.164	0.389	0.332	0.156	0.294	0.261	0.178	0.334
Larger reference models									
Qwen3.5-397B	0.540	0.119	0.351	0.238	0.135	0.275	0.318	0.131	0.302
9B curated-data control (SFT + RL, self-judge)									
Qwen3.5-9B	0.583	0.360	0.472	0.498	0.199	0.225	0.330	0.171	0.413
Frozen Qwen3.5-9B inference methods									
Direct (matched run)	0.450	0.079	0.332	0.287	0.142	0.235	0.193	0.112	0.262
Medical RAG	0.422	0.126	0.232	0.247	0.121	0.196	0.284	0.121	0.245
RAG-Fusion	0.403	0.099	0.232	0.278	0.121	0.245	0.227	0.125	0.238
i-MedRAG	0.393	0.087	0.214	0.283	0.099	0.235	0.216	0.112	0.226
Self-consistency (5 samples)	0.462	0.111	0.343	0.265	0.142	0.216	0.239	0.112	0.272
Self-refinement	0.384	0.061	0.265	0.269	0.142	0.186	0.136	0.093	0.221
Chain-of-verification	0.391	0.065	0.276	0.283	0.156	0.186	0.227	0.090	0.230
Frozen Qwen3.6-27B inference methods									
Direct (matched run)	0.540	0.160	0.387	0.354	0.156	0.265	0.273	0.181	0.332
Medical RAG	0.534	0.196	0.343	0.341	0.156	0.294	0.273	0.168	0.329
RAG-Fusion	0.511	0.162	0.348	0.390	0.113	0.255	0.261	0.181	0.318
i-MedRAG	0.515	0.150	0.309	0.327	0.121	0.304	0.193	0.153	0.301
Self-consistency (5 samples)	0.584	0.164	0.407	0.345	0.170	0.284	0.295	0.150	0.346
Self-refinement	0.465	0.123	0.273	0.327	0.128	0.275	0.273	0.134	0.276
Chain-of-verification	0.545	0.150	0.392	0.332	0.170	0.304	0.318	0.184	0.335
Qwen3.6-27B training controls									
SFT + train-set RL	0.635	0.464	0.487	0.529	0.259	0.286	0.266	0.213	0.463
SFT + static-pool RL	0.609	0.374	0.479	0.565	0.248	0.294	0.295	0.221	0.441
RSIMed-9B (ours)	0.565	**0.506**	0.482	0.735	**0.340**	0.304	0.295	0.162	0.469
RSIMed-27B (ours)	**0.681**	0.413	**0.588**	**0.740**	**0.340**	**0.343**	**0.364**	**0.249**	**0.516**

J.4 Out-of-Distribution Clinical Performance

Table 10 evaluates out-of-distribution transfer from rare-disease diagnosis training to the distinct consultation, clinical writing, and research tasks in HealthBench Professional. Despite this shift in task types, diagnostic training improves performance on HealthBench Professional. RSIMed-RL-27B scores 0.303 length-adjusted and 0.376 raw, exceeding its base model by 1.2 and 1.3 points and outperforming the evaluated open medical models on both metrics. Among the open base models, Qwen3.5-9B scores 0.187 and Gemma-4-31B 0.293 length-adjusted, and among the stronger models, Kimi-K2.6 scores 0.335 and Grok-4.3 0.415. Appendix G reports the RL-only transfer breakdown.

The trained model improves on adversarial prompts without safety-specific training. HealthBench Professional includes 256 ordinary clinical requests in the *good-faith typical* slice, 78 difficult legitimate requests in the *good-faith difficult* slice, and 191 adversarial prompts intended to elicit unsafe or unprofessional responses in the *red-teaming difficult* slice. The adversarial rubrics assess

Table 10: HealthBench Professional on its 525 tasks, graded by GPT-5.5-Chat. Columns show the length-adjusted and raw rubric scores. Our row uses the RL-only variant evaluated on this benchmark, as detailed in Appendix G.

Model	Len-adj.	Raw
Open medical models		
MedGemma-1.5-4B	0.069	0.161
Lingshu-7B	0.061	0.061
Lingshu-32B	0.145	0.130
Untrained base models		
Qwen3.5-9B	0.187	0.255
Gemma-4-31B	0.293	0.326
Qwen3.6-27B	0.291	0.363
DeepSeek-V4-Pro	0.282	0.360
RSIMed-RL-27B (ours)	0.303	0.376

Table 11: **Contribution of each training stage for Qwen3.6-27B.** The RL-only run trains on self-generated tasks from the base weights, and the full pipeline runs RL after the supervised warm start. Both benchmarks are graded by GPT-5.5-Chat.

	MIMIC-IV	HealthBench Prof.	
Training	Accuracy	Len-adj.	Raw
Untrained	0.334	0.291	0.363
Supervised warm start only	0.421	0.290	0.368
RL only	-	**0.303**	**0.376**
Warm start and RL, full pipeline	**0.516**	0.280	0.349

refusals, redirection, and explicit identification of risks. Table 6 reports these slices graded by GPT-5.5-Chat, without safety-specific training or prompting. On adversarial prompts, the self-trained 27B model scores 0.047, compared with 0.030 for its base. The 27B model also improves on both good-faith slices, and difficult good-faith requests rise from 0.174 to 0.212. By use case, the raw score rises by 0.017 on consultation, 0.018 on writing, and 0.003 on research prompts.

J.5 Analysis of Teacher Quality

Self-generated tasks give the student more groups with mixed rewards than GPT-5.5's tasks. We sample eight answers per task from the untrained student and use a common correctness scorer to measure variation within each group. Figure 11 shows that self-generated tasks produce varying correctness rewards in 22% of multiple-choice groups and 36% of short-answer groups, compared with 11% and 22% for GPT-5.5's tasks. Such groups provide opportunities for group-relative learning under this correctness reward.

Retrieved passages help the student more on self-generated tasks than on GPT-5.5's tasks. We supply the proposer's retrieved passages to the student to measure how much they help it answer each task. On self-generated short-answer tasks, accuracy increases from 0.787 to 0.901, compared with 0.925 to 0.979 on GPT-5.5's tasks. The larger improvement on self-generated tasks suggests that their passages contain information the student needs to answer correctly. GPT-5.5 answers the student's multiple-choice tasks with 0.978 accuracy, as reported in Table 8.

Other solvers also find self-generated tasks harder than GPT-5.5's tasks. Figure 12 shows that the difficulty difference extends to Gemma-4-31B and Qwen3.5-9B. Both produce more groups with varying correctness rewards on the student's tasks than on GPT-5.5's. On the expanded multiple-choice subset that GPT-5.5 answers correctly in all eight samples, self-generated tasks reduce accuracy by 1.5 points for Gemma and 1.8 points for the student relative to GPT-5.5's tasks.

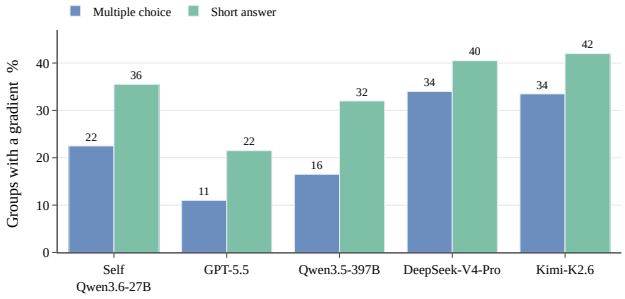

Figure 11: **Correctness-reward variation by task source.** Self-generated tasks yield more groups with unequal rewards than GPT-5.5's tasks when the untrained student answers each task eight times with training-time sampling and a common correctness scorer, providing a proxy for group-relative learning opportunities under the protocol in Appendix H.

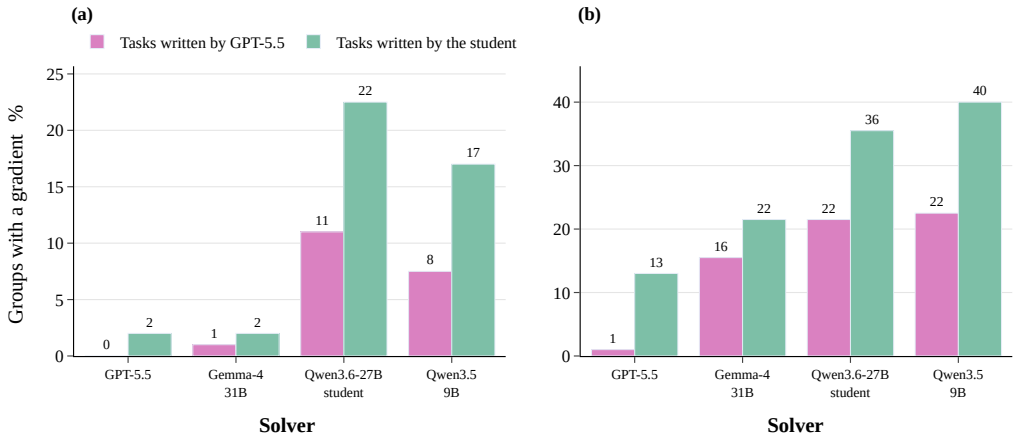

Figure 12: **(a) Multiple-choice task difficulty.** Self-generated tasks produce more groups with varying correctness rewards than GPT-5.5's tasks across all four solvers sampled eight times per task. **(b) Short-answer task difficulty.** The same pattern holds for short-answer tasks; Table 8 reports accuracy and the subset GPT-5.5 answers correctly in all eight samples.

K Limitations

The evaluation covers rare-disease diagnosis and transfer to HealthBench Professional, and broader task coverage would help assess generalization. Model-generated tasks, reference answers, and model-based grading can contain errors despite retrieval and validation. Inference costs are estimates under the stated serving-rate assumptions. The reported benchmark scores measure model performance on these tasks rather than patient outcomes or clinical deployment effectiveness.

L Data availability and ethics

Data and Code Availability This work evaluates models on MIMIC-IV, a public dataset with credentialed access through PhysioNet, and on the public HealthBench Professional benchmark. The medical database includes additional public datasets described in Section 3. All code, prompts, and configurations are released in an anonymous repository at this link.

Institutional Review Board No new human subjects research was conducted. All data are reused from publicly available, pre-anonymized clinical datasets that obtained their own ethical approvals at the time of original release.