Cura 1T: Specialized Model for Agentic Healthcare

actAVA AI Team

Abstract

Healthcare spans high-stakes communication, expert reasoning, and workflow execution, yet specialized LLMs that cover these use cases together remain limited. A healthcare model must handle patient consultation, clinical reasoning over text and images, interactive diagnosis, and electronic health record (EHR) tool use. These capabilities fail in different ways, and a narrow update for one task can degrade another. We present Cura 1T, a healthcare-specialized LLM trained through a human-gated self-evolution loop. In each evolution round, a training agent plans a target capability, trains the model, evaluates benchmark trajectories, and refines the data mixture from observed failures. This data-centered loop improves the model through targeted synthetic and curated examples rather than a single generic medical-data update. Across the healthcare evaluation suite, Cura 1T ranks at or near the top among frontier baselines, while remaining competitive on out-of-domain reasoning and agentic benchmarks.

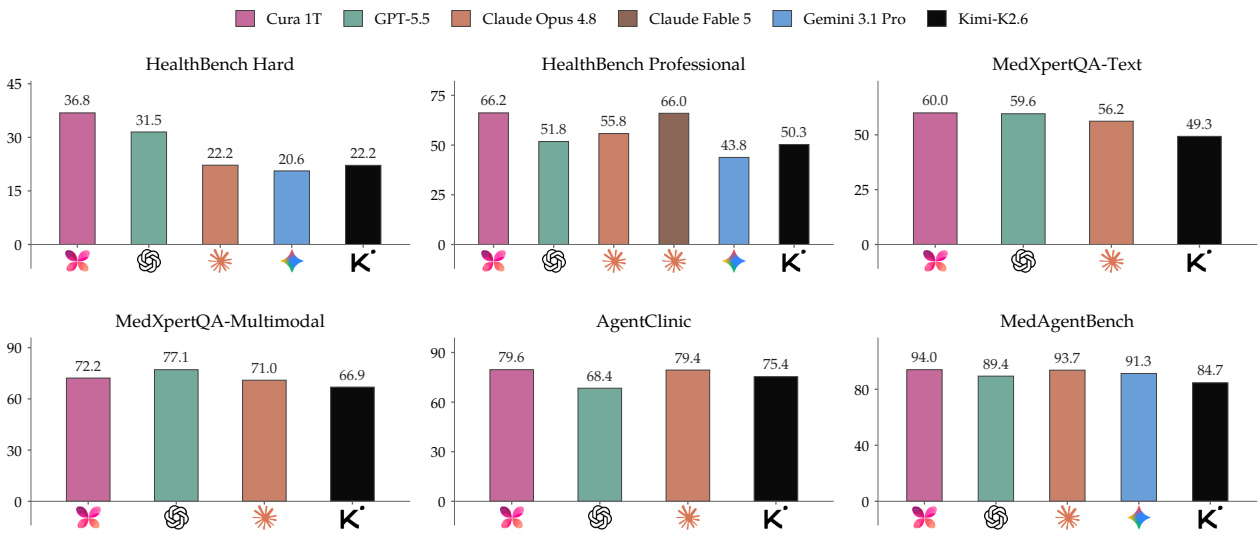

Figure 1 Performance of Cura 1T, frontier models, and the Kimi-K2.6 base across six Healthcare benchmark panels: MedAgentBench (Jiang et al., 2025), HealthBench Professional and Hard (Arora et al., 2025; OpenAI, 2026), MedXpertQA (Zuo et al., 2025), and AgentClinic (Schmidgall et al., 2024).

1 Introduction

Frontier language models now answer difficult clinical questions (Singhal et al., 2023), but healthcare deployment requires a broader and less forgiving capability profile. We focus on three use cases

that recur throughout this report. *Patient care* requires clinician- and patient-facing responses that follow physician guidelines (Singhal et al., 2023; Arora et al., 2025; OpenAI, 2026). *Clinical reasoning* requires solving expert medical questions across text and images (Jin et al., 2020; Pal et al., 2022; Singhal et al., 2023; Zuo et al., 2025). *Healthcare agentic tasks* require multi-turn diagnostic workups, electronic-health-record tool use, and long-horizon healthcare workflows under strict task protocols (Schmidgall et al., 2024; Jiang et al., 2025; Chen et al., 2026; Liu et al., 2026b,a; HL7 International, 2019). These tasks are related, but they fail in different ways: rubric omissions, missing facts, brittle reasoning, premature diagnoses, and malformed tool calls each require a different repair. Community progress has been strong on individual ingredients: medical reasoning benchmarks and multimodal expert QA (Jin et al., 2020; Pal et al., 2022; Zuo et al., 2025), rubric-graded clinical conversations (Arora et al., 2025; OpenAI, 2026), diagnosis and EHR-tool benchmarks (Schmidgall et al., 2024; Jiang et al., 2025), long-horizon healthcare-agent benchmarks (Chen et al., 2026; Liu et al., 2026b,a), and continuous-care model systems (Baichuan AI and THUBPM Group, Tsinghua University, 2026). A single specialized LLM trained to serve all three use cases remains comparatively underexplored.

Building such a model is a data-construction problem as much as a model-training problem. Compared with coding or mathematics, healthcare has less abundant training signal, and its sensitivity makes high-quality supervision harder to obtain. Useful examples are fragmented across guidelines, exams, patient interactions, images, and electronic-health-record workflows; many outcomes also cannot be checked by a simple executor or answer key. Because the failure modes are unevenly distributed across task categories, adding examples for one behavior can improve that behavior while eroding behavior that was already correct elsewhere. The central challenge is therefore not simply to choose a training algorithm or hyperparameter configuration, but to identify the missing data in the recipe, add publicly available or synthetic surrogates, and curate mixtures that transfer across tasks without unnecessary forgetting.

We build Cura 1T, a healthcare-specialized LLM to resolve this challenge. Cura 1T is trained with low-rank adapters (Hu et al., 2021) through a human-gated self-evolution loop. This design is also motivated by the recent shift toward recursive self-improvement and automated research workflows, enabled by stronger agentic models that can plan, execute, and inspect long technical tasks (autoresearch contributors, 2026; Yamada et al., 2025; Novikov et al., 2025). Manually running the full post-training cycle—planning a run, launching training, evaluating the result, reading trajectories, diagnosing failures, and rebuilding the data mixture—requires substantial expert time and slows iteration. In each experiment, an LLM agent defines the desired behavior and acceptance criteria, trains a candidate model, evaluates it under the relevant benchmarks, reads the graded trajectories, and refines the next data mixture from the observed failures. The agent uses specialized data-construction skills rather than applying a single generic "medical data" update. We present the detailed design and implementation of this self-evolution loop in Section 3.2.

Figure 1 and Table 1 summarize the main results for Cura 1T. Cura 1T is the strongest model on five of the six healthcare benchmark panels and ranks second on the remaining MedXpertQA multimodal tasks. In Table 1, the same model consistently improves over the Kimi-K2.6 base across patient-care, clinical-reasoning, and healthcare-agentic evaluations. Section 4 gives the full benchmark setup, intermediate self-evolution results, and frontier-model comparisons. In addition, we report out-of-domain benchmark results in math, scientific reasoning, and agentic tasks. Cura 1T preserves strong performance on these benchmarks, indicating that the healthcare-focused training strategy does not obviously erode general reasoning or agentic capability under these evaluations.

Table 1 Improvement of Cura 1T from the Kimi-K2.6 base. Metrics differ by benchmark (rubric score for HealthBench, exact-letter pass@1 for MedXpertQA, task success for AgentClinic and MedAgentBench).

Benchmark	Base	Cura 1T	Δ
MedAgentBench	0.847	**0.940**	+0.093
HealthBench Professional	0.503	**0.662**	+0.159
HealthBench Hard	0.222	**0.368**	+0.146
MedXpertQA	0.569	**0.655**	+0.086
AgentClinic	0.754	**0.796**	+0.042

2 Related Work

Healthcare AI. Healthcare language-model work now spans patient care, clinical reasoning, and agentic clinical systems. Med-PaLM showed that large models can encode enough medical knowledge to answer licensing-exam questions (Singhal et al., 2023), while the Baichuan series, especially Baichuan-M4, extends medical modeling toward continuous care with tool use, patient memory, action constraints, evidence retrieval, and multimodal perception (Baichuan AI and THUBPM Group, Tsinghua University, 2026). Benchmark development follows the same broadening of scope. Patient-care benchmarks such as HealthBench grade open-ended clinical responses against physician-written rubrics (Arora et al., 2025), and HealthBench Professional extends this setting to real clinician chats (OpenAI, 2026). Clinical-reasoning benchmarks began with large-scale medical-exam question answering such as MedQA and MedMCQA (Jin et al., 2020; Pal et al., 2022); newer expert benchmarks such as MedXpertQA increase difficulty and add multimodal cases with real clinical images and context (Zuo et al., 2025). Agentic healthcare benchmarks then evaluate whether models can act over time: AgentClinic tests multi-turn diagnosis in a simulated clinic (Schmidgall et al., 2024), MedAgentBench evaluates clinically derived EHR tasks against a standards-compliant record (Jiang et al., 2025; HL7 International, 2019), and recent systems such as CHI-Bench, PhysicianBench, and HealthAgentBench extend evaluation toward long-horizon, role-composed, policy-driven healthcare workflows, realistic EHR environments, and unified agentic healthcare settings (Chen et al., 2026; Liu et al., 2026b,a). This report uses these benchmark families to evaluate Cura 1T across the three healthcare use cases introduced in Section 1.

LLM post-training. LLM post-training commonly combines supervised adaptation, reinforcement learning, self-training, and parameter-efficient updates. Supervised fine-tuning (SFT) adapts a pretrained model to task-specific demonstrations and is often used as the cold-start stage before more selective data construction. Reinforcement learning (RL) then improves the policy against task-level reward signals before later distillation or consolidation. Rejection sampling fine-tuning keeps a model's own highest-scoring generations and trains on them (Yuan et al., 2023; Touvron et al., 2023), while STaR bootstraps reasoning by sampling rationales, retaining those that reach the correct answer, and fine-tuning on the retained traces (Zelikman et al., 2022). Self-distillation fine-tuning (SDFT) uses the model's own samples as the student distribution and a teacher conditioned on additional context as the target, keeping updates closer to the base model's generation policy (Shenfeld et al., 2026). These methods are often combined with reasoning-aware data formats that preserve chain-of-thought behavior rather than replacing it with off-policy traces (Wei et al., 2022; DeepSeek-AI, 2025), and with low-rank adapters that update a small parameter-efficient module while leaving the base weights intact (Hu et al., 2021).

Self-evolution and auto-research. Self-Refine and Reflexion are early examples of language-model self-evolution. They use natural-language critique or task feedback to improve subsequent attempts, while OPRO, DSPy, and TextGrad treat prompts, programs, or LM-pipeline components as objects to optimize against a metric (Madaan et al., 2023; Shinn et al., 2023; Yang et al., 2023; Khattab et al., 2023; Yuksekgonul et al., 2024). The public autoresearch repository sharpened this framing around a coding agent that edits a real training loop, runs fixed-budget experiments, scores validation loss, and keeps or discards each change (autoresearch contributors, 2026). High-fidelity auto-research systems extend this pattern to scientific workflows and algorithm discovery, including AI Scientist, AI Scientist-v2, and AlphaEvolve, while ResearchGym and recent surveys emphasize that reliability, provenance, and reproducibility remain central bottlenecks (Lu et al., 2024; Yamada et al., 2025; Novikov et al., 2025; Garikaparthi et al., 2026; Tie et al., 2026). Cura follows the same closed-loop measurement discipline, but changes the optimized object: the loop does not primarily search prompts, code, or hyperparameters; it curates the post-training data mixture that is then trained into the model.

3 Cura 1T

Cura 1T is post-trained on top of Kimi-K2.6 (Kimi Team, 2025) for three healthcare use cases: *patient care*, *clinical reasoning*, and *healthcare agentic tasks*. The released model uses a 256K context window and native text + vision input. Patient care covers clinician- and patient-facing medical responses that must follow physician rubrics while remaining clear and safe. Clinical reasoning covers expert medical question answering across text and images, where the model must combine factual recall, differential reasoning, and answer selection. Healthcare agentic tasks cover multi-turn clinical workflows, where the model must gather evidence, use tools, and complete diagnosis or EHR tasks before producing a final answer. To efficiently train a one-trillion-parameter model, we build our training infrastructure around a lightweight adapter-training stack (see Section 3.1) and the *self-evolution* loop orchestrated by a training agent (see Section 3.2).

3.1 Training

The training protocol is implemented as a thin adapter-training stack. We add the training algorithms, benchmark-environment adapters, and loop orchestration needed for healthcare self-evolution.

Supervised fine-tuning (SFT) serves as the low-cost prerequisite step inside each round. Before launching a more time-consuming RL or SDFT run, the agent uses SFT to verify that the proposed data mixture and hyperparameters are well formed, train stably, and move the target metrics in the intended direction. Once this screen passes, the round proceeds through reinforcement learning (RL), followed by self-distillation fine-tuning (SDFT) (Shenfeld et al., 2026) as the final step.

SDFT trains the model from its own on-policy samples toward a teacher distribution conditioned on additional information. For a prompt x, privileged context c, and student sample $y \sim \pi_\theta(\cdot \mid x)$, we write this objective as

$$\mathcal{L}(\theta) = D_{\mathrm{KL}}(\pi_\theta(\cdot \mid x) \,\|\, \pi(\cdot \mid x, c)) = \mathbb{E}_{y \sim \pi_\theta(\cdot|x)} \left[\log \frac{\pi_\theta(y \mid x)}{\pi(y \mid x, c)} \right], \tag{1}$$

where $\pi_\theta(\cdot \mid x)$ is the student distribution and $\pi(\cdot \mid x, c)$ is the privileged-context teacher distribution. During the training of Cura, c is the extra information used to generate a clean teacher trajectory: an intervened or augmented trajectory, a reference behavior, or verified knowledge used to ground

teacher generation. The student sees only the original prompt after this context is removed. The practical effect is that the update is anchored to trajectories the model can itself produce, which is important for long medical reasoning traces where copying an off-policy chain of thought can damage the model's native reasoning behavior. Section 3.2 defines the data-construction actions that construct these contexts.

All experiments follow the same protocol. With optional human guidance, the training agent writes a plan, trains candidate adapters through the SFT-to-RL-to-SDFT stack, evaluates the model, and reports the evidence for a keep-or-revise decision. We run the loop separately for capabilities that need improvement. A kept round becomes the continuation point for the next iteration and contributes its validated data refinement to the growing mixture. A reverted round remains in the experimental record, but is not used as the basis for later improvement because the intervention caused severe side effects or was not favored. Cura 1T is trained from the consolidated mixture accumulated across completed capability loops, rather than from a single benchmark-specific update.

3.2 Self-evolution Loop

The self-evolution loop reflects a practical asymmetry in healthcare post-training: the limiting resource is not a fixed benchmark dataset with a nearly finished recipe, but scarce and uneven domain data for the use cases above. Recent self-evolution and automated-research workflows search over hypotheses, programs, skills, experiment trees, or training configurations (Yang et al., 2023; Khattab et al., 2023; Yuksekgonul et al., 2024; autoresearch contributors, 2026; Yamada et al., 2025; Novikov et al., 2025). Cura keeps the closed-loop discipline, but makes the data recipe the search object. The agent writes configurations and monitors training, but its central task is to analyze failures, synthesize targeted data, and curate a mixture that preserves previously solved behavior.

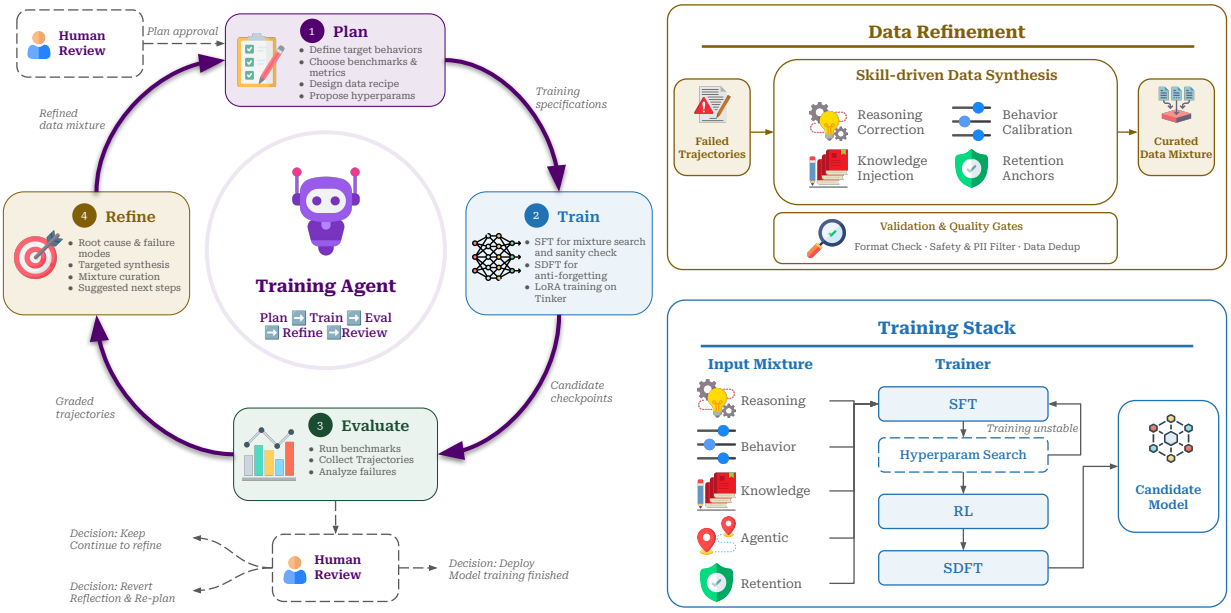

Figure 2 Left: Agent-managed self-evolution loop for Cura 1T. Human review gates the plan before training and the keep, revert, or deploy decision after evaluation. Right: Data refinement pipeline and training stack.

Figure 2 consolidates the experiment-control loop and the two internal modules that matter for data-first self-evolution. *Plan* defines target behaviors, benchmarks, metrics, data recipe, and candidate hyperparameters, then passes through human plan approval. *Train* runs LoRA adapter training on the training stack, using SFT as the mixture and hyperparameter screen, RL as the reward-driven improvement stage, and SDFT as the final trainer of the round. *Evaluate* runs the benchmarks, collects trajectories, and analyzes failures. *Refine* uses the failed trajectories as evidence for root-cause analysis, converts the resulting failure modes into targeted data, curates the next mixture, and validates the candidate rows before proposing next steps. *Review* is the human gate around both the validated data refinement and the candidate model, producing a keep, revert, continue-to-refine, or deploy decision before the next round begins.

The refinement block aims to enhance the data mixture by synthesizing new trajectories. The agent first categorizes what went wrong, then synthesizes training data that targets the same missing capability while preserving retained behavior. Before those rows enter a new mixture, validation gates check format, safety and PII risk, duplicates, and coverage of the intended repair. Data synthesis is driven by the following agent skills:

- **Retention Anchor**: The agent adds a small set of examples for capabilities the current model already handles, so a targeted repair does not overwrite them.
- **Reasoning Correction**: The agent edits a failed reasoning trace with a frontier model while preserving the original problem shape. The corrected pattern then seeds new questions that require similar reasoning.
- **Knowledge Injection**: The agent identifies missing clinical knowledge, grounds it with retrieved documents, and synthesizes closed-book QA rows with explicit reasoning.
- **Behavior Calibration**: The agent compares desired and observed answer behavior, synthesizes rubrics that express the gap, and generates responses conditioned on those rubrics.
- **Other**: The agent handles task-specific failures outside the reusable reasoning, knowledge, and behavior categories, including agentic workflow errors.
- **Data Mixture Curation**: The agent merges synthesized correction rows with retention anchors, then deduplicates, caps, and reweights the candidate training set.

This design makes the experiment loop data-first. The agent reads a graded run, identifies the missing capability, and changes one mixture decision at a time: add retention anchors when a repair causes forgetting, increase targeted synthetic trajectories when a failure mode remains underrepresented, or remove rows that dilute the target distribution.

4 Evaluation

4.1 Setup

We evaluate Cura 1T on healthcare benchmarks that exercise patient-facing response quality, clinical reasoning, interactive diagnosis, and electronic-health-record tool use. The healthcare benchmarks include MedAgentBench, HealthBench Professional, HealthBench Hard, MedXpertQA, and AgentClinic. MedAgentBench measures task success on clinically derived FHIR tool-use tasks (Jiang et al., 2025). HealthBench Professional and HealthBench Hard score open-ended answers with physician-authored rubrics (Arora et al., 2025; OpenAI, 2026). MedXpertQA is exact-letter graded and reported as text, multimodal, and overall pass@1 (Zuo et al., 2025). AgentClinic evaluates inter-

Table 2 Components used by the self-evolution loop in Figure 2. The table summarizes how evaluation records are converted into refined data mixtures and how SFT, RL, and SDFT divide the training step.

Component	Role in the loop	Output
Plan	Define target behavior, metrics, data recipe, and candidate hyperparameters.	Training specifications
Train	Run SFT to verify the mixture and hyperparameters, RL to improve the policy against reward signals, and SDFT to consolidate the final adapter.	Candidate model
Evaluate	Run benchmark harnesses and collect graded trajectories and failure summaries.	Failed trajectories and metrics
Refine	Categorize failures, synthesize targeted data, curate the next mixture, validate candidate rows, and suggest next steps.	Validated data refinement
Reasoning Correction	Use a corrected reasoning trace to synthesize cases with similar phrasing or the same reasoning pattern.	Pattern-matched reasoning rows
Knowledge Injection	Use verified knowledge to synthesize nearby questions targeting the same missing concept, then emit closed-book rows.	Source-grounded knowledge rows
Behavior Calibration	Use rubric, fix guidance, or a reference response to synthesize nearby cases targeting the same omitted action or wrong policy.	Corrected behavior rows
Retention Anchors	Mark solved behavior that should be preserved when a repair is added.	Retention rows
Other	Cover task-specific failures outside the reusable reasoning, knowledge, and behavior categories.	Task-specific rows
Curated Data Mixture	Merge, deduplicate, and reweight correction rows with retention anchors.	Candidate training mixture
Validation Gates	Check format, safety and PII risk, duplicates, and repair coverage before training.	Validated training mixture
Human review	Approve the plan, then inspect the validated refinement and candidate result before the next round.	Keep, revert, refine, or deploy decision

active diagnosis in a simulated clinic; we use the tool-based harness detailed in Section 4.5 and report per subset and overall pass@1 (Schmidgall et al., 2024). All benchmarks are evaluated at T=1.0 except that the MedAgentBench development path in Figure 3 uses $T = 0.6$. The out-of-domain benchmarks in Section 4.6 include AIME, GPQA-Diamond, and τ^2-Bench.

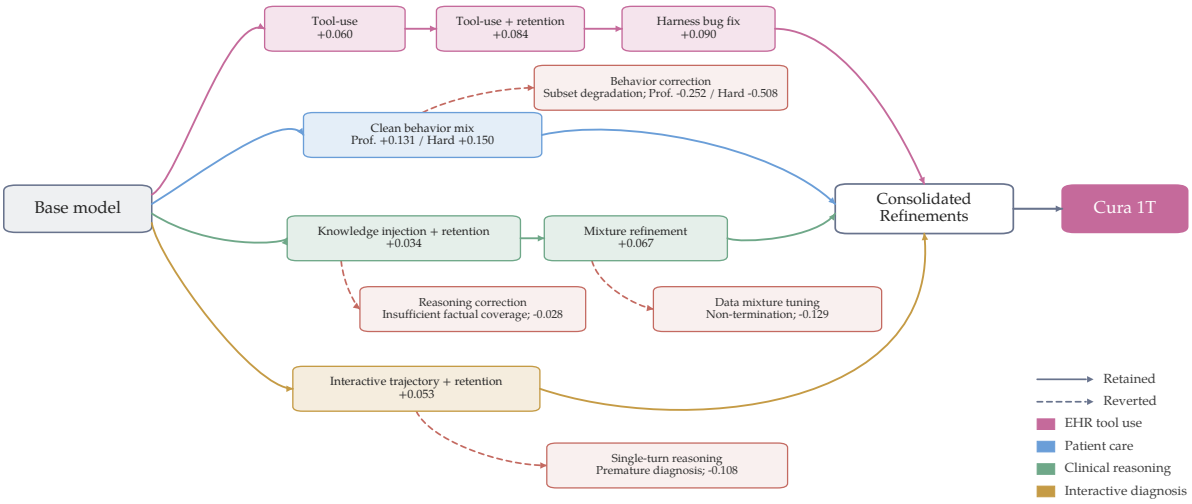

Figure 3 Evolution map from the base model to Cura 1T. Values are changes from benchmark-specific bases; solid and dashed red arrows mark retained and reverted interventions.

Figure 3 summarizes the evolution path of Cura 1T. The consolidated Cura 1T model is trained from the consolidated mixture after each evolution path saturates.

4.2 MedAgentBench

MedAgentBench tests whether a model can execute clinically derived EHR workflows through FHIR tool calls, which makes it a direct check on healthcare agentic reliability. We evaluate the model as a native tool-caller against a running FHIR server, using renderer-native function calling with read, write, and finish actions. The grader checks whether each tool call satisfies the required resource schema and clinical content.

The principal failure mode is brittle execution of EHR writes. The base model often identifies the intended clinical action but produces a resource that cannot be accepted as correct because a required field is missing, a coded value is wrong, or the payload is inconsistent with the clinical request. *Tool-use* targets these errors with synthetic trajectories that exercise the same FHIR write patterns in new patient contexts. *Tool-use + retention* adds successful trajectories from surrounding tool behaviors to prevent a narrow repair from degrading the rest of the workflow. *Harness bug fix* corrects the evaluation configuration and restores the intended prompt, allowing the refined model to be measured under the proper protocol.

Table 3 MedAgentBench task success. Bold marks the best score and underline marks the second-best distinct score.

Model	Intervention	Overall	Decision
Kimi-K2.6	–	0.883	–
Round 1	Tool-use	0.943	keep
Round 2	Tool-use + retention	0.967	keep
Round 3	Harness bug fix	<u>0.973</u>	keep
Cura 1T	Consolidated data mixture	0.940	release
Claude Opus 4.8	–	**0.977**	–
GPT-5.5	–	0.894	–
Gemini 3.1 Pro	–	0.913	–

Table 3 shows that the self-evolution rounds raise task success from 0.883 to 0.973. Cura 1T reaches 0.940 after consolidation, while the strongest frontier reference reaches 0.977. The same development path is summarized in Figure 3, and a matched task-level trace is provided in Appendix A.

4.3 HealthBench

HealthBench evaluates whether a model gives clinically appropriate open-ended answers rather than only selecting a multiple-choice option. The base model scores 0.503 on Professional and 0.222 on Hard. Failures are dominated by omitted required points rather than explicit penalty violations, and rejection sampling saturates quickly: the best score among multiple rollouts is not substantially higher than the average, suggesting that the model behavior requires correction. *Behavior correction* uses a broad behavior-fix mixture for omitted rubric points, following the ClinAlign rubric-and-principle synthesis pattern (Lyu et al., 2026); this first version is reverted after substantial degradation on a subset of tasks. *Clean behavior mix* keeps the same target while removing chart-template boilerplate, adding cleaner principle-rubric examples, and shortening training.

Table 4 HealthBench rubric scores at T=1.0. Bold marks the best score and underline marks the second-best distinct score in each column.

Model	Intervention	Professional	Hard	Decision
Kimi-K2.6	–	0.503	0.222	–
Round 1	Behavior correction	0.601	0.332	revert
Round 2	Clean behavior mix	<u>0.634</u>	**0.372**	keep
Cura 1T	Consolidated data mixture	**0.662**	<u>0.368</u>	release

Table 4 shows why HealthBench required a cleaner refinement rather than a broader one. Behavior correction raises the full-set scores to 0.601 on Professional and 0.332 on Hard, but the degraded subset falls by 0.252 on Professional and 0.508 on Hard relative to its base-model performance. The round is therefore reverted despite its aggregate gains. Clean behavior mix raises Professional to 0.634 and Hard to 0.372; after cross-benchmark consolidation, Cura 1T reaches the strongest Professional score at 0.662 while remaining close to the HealthBench-specific round on Hard at 0.368.

4.4 MedXpertQA

MedXpertQA tests expert medical reasoning across text and multimodal questions. The base model shows both recoverable reasoning gaps and missing clinical knowledge. Reasoning-only correction helps some local reasoning patterns but reduces overall accuracy, so retention becomes central to the evolution record. *Reasoning correction* and *Reasoning correction, extended* add corrected rationales without enough factual coverage and are reverted. *Knowledge injection + retention* adds closed-book clinical knowledge while mixing examples the model already answers correctly. *Mixture refinement* further balances the retained examples, while *Data mixture tuning* is reverted because overlong traces cause non-termination.

Table 5 MedXpertQA split and overall pass@1 at T=1.0. The overall score weights 2,450 text questions and 2,000 multimodal questions. Bold marks the best score and underline marks the second-best distinct score in each column.

Model	Intervention	Text	Multimodal	Overall	Decision
Kimi-K2.6	–	0.484	0.672	0.569	–
Round 1	Reasoning correction	0.447	0.656	0.541	revert
Round 1L	Reasoning correction, extended	0.454	0.657	0.545	revert
Round 2	Knowledge injection + retention	0.521	0.703	0.603	keep
Round 3	Mixture refinement	0.560	<u>0.728</u>	0.636	keep
Round 4	Data mixture tuning	–	–	0.440	revert
Cura 1T	Consolidated data mixture	**0.600**	0.722	<u>0.655</u>	release
Claude Opus 4.8	–	0.562	0.710	0.628	–
GPT-5.5	–	<u>0.596</u>	**0.771**	**0.675**	–

Table 5 shows that the reverted reasoning-only rounds trail the base model, while Knowledge injection + retention raises overall pass@1 to 0.603 and Mixture refinement raises it to 0.636. The consolidated Cura 1T row reaches 0.655 overall pass@1, improving over the base and Claude Opus 4.8 while remaining second to GPT-5.5 on the overall metric.

4.5 AgentClinic

AgentClinic evaluates whether a model can gather evidence and diagnose through a simulated multi-turn clinical encounter. We adapt the harness into a tool-native format with explicit structured tools for clinical interaction, so the model acts through structured calls rather than a text-only dialogue protocol. All rows in Table 6 use this harness at T=1.0 on MedQA, MedQA-Ext, NEJM, and NEJM-Ext.

The main failure mode is premature diagnosis before completing the clinical workup, especially on NEJM-style subsets. *Single-turn reasoning* is reverted because isolated answer rationales do not preserve clinical conduct. *Interactive trajectory + retention* instead teaches the model to elicit evidence before committing to a diagnosis, while mixing synthesized examples seeded from correctly answered cases to preserve existing behavior.

Table 6 AgentClinic pass@1 by subset under the tool-native protocol. Bold marks the best score and underline marks the second-best distinct score in each column.

Model	Intervention	MedQA	MedQA Ext	NEJM	NEJM Ext	Overall	Decision
Kimi-K2.6	–	<u>0.869</u>	0.827	0.400	0.567	0.754	–
Round 2	Interactive trajectory + retention	0.841	0.841	**0.800**	**0.717**	**0.807**	keep
Cura 1T	Consolidated data mixture	**0.879**	<u>0.850</u>	**0.800**	<u>0.625</u>	<u>0.796</u>	release
Claude Opus 4.8	–	0.841	**0.874**	**0.800**	0.608	0.794	–
GPT-5.5	–	0.832	0.808	<u>0.467</u>	0.358	0.684	–

Table 6 shows that Interactive trajectory + retention improves overall pass@1 from 0.754 to 0.807, with the largest gains on NEJM and NEJM-Ext. Cura 1T preserves most of this gain after consolidation, reaching 0.796 overall pass@1 and matching the best reported NEJM score under the same tool-native harness.

4.6 Out-of-domain Evaluation

We also evaluate several out-of-domain benchmarks to check whether Cura 1T preserves general capabilities after healthcare specialization. Figure 4 groups the benchmarks into two families: *Reasoning* covers AIME 2025, AIME 2026, and GPQA-Diamond (Mathematical Association of America, 2026; Rein et al., 2023), while *Agentic* covers the airline, retail, and telecom domains in τ^2-Bench (Barres et al., 2025). Cura 1T is on par with frontier comparators on each reasoning benchmark and τ^2-airline. On τ^2-retail and τ^2-telecom, Cura 1T surpasses the models with publicly reported scores. These checks indicate that Cura 1T preserves general reasoning and agentic capability in adjacent domains.

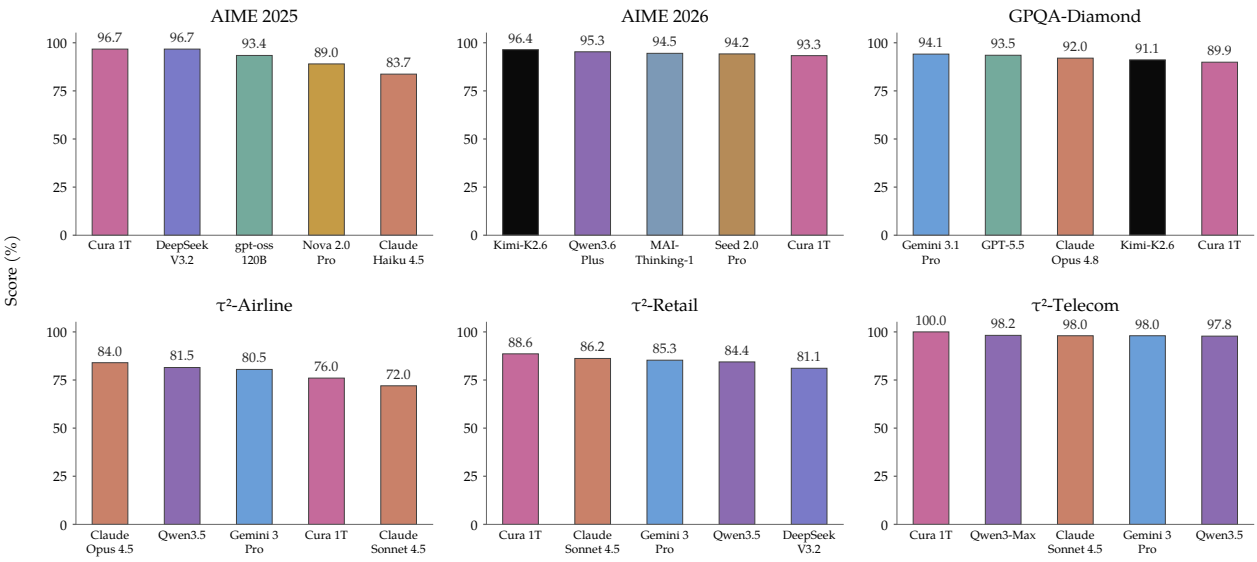

Figure 4 Out-of-domain evaluation results for the Cura 1T.

5 Discussion

The central lesson from our experiments is that healthcare specialization is a data-mixture problem before it is a hyperparameter-search problem. The self-evolution loop was useful because it treated evaluation failures as evidence for what data was missing, then used SFT to screen candidate mixtures, RL to improve the policy on the selected objective, and SDFT to consolidate the final round. This matters when available training data is sparse and uneven, for example healthcare: a model may need more rubric-following behavior for patient care, more factual coverage for expert questions, or more workflow-specific examples for agentic tool use. Searching the training stack alone would not identify those missing pieces.

The benchmark traces also show why a single generic medical-data update is insufficient. HealthBench failures were dominated by omitted rubric items and verbosity-driven grading failures, so the useful repair was behavior calibration with cleaner principle-rubric data. MedXpertQA mixed reasoning-pattern failures with missing knowledge, which made reasoning correction alone unstable until knowledge injection and retention were added. MedAgentBench exposed brittle FHIR writes in which the intended clinical action did not reliably translate into an executable resource, while AgentClinic required task-specific interactive trajectories that changed how the model gathered evidence before answering. These differences support the design choice in Section 3.2: the refinement step should categorize failures before synthesizing data.

6 Conclusion

Cura 1T shows how a healthcare-specialized LLM can be built through a self-evolution loop. Starting from Kimi-K2.6, the loop turns graded benchmark failures into targeted data construction: the agent plans a round, trains candidate adapters, evaluates the model, analyzes failed trajectories, and curates the next mixture. The main contribution is not a new benchmark-specific prompt or a single medical dataset, but a repeatable training loop that treats data mixture construction as the primary object of search.

Across the clinical suite, Cura 1T improves over the served base on patient-care, clinical-reasoning, and healthcare-agentic evaluations. The strongest gains come from different data actions in different settings: tool-use trajectories for MedAgentBench, cleaner behavior data for HealthBench, knowledge and reasoning repair with retention for MedXpertQA, and task-specific interactive trajectories for AgentClinic. The out-of-domain results on AIME, GPQA-Diamond, and τ^2-Bench indicate that these healthcare gains can be obtained without an obvious loss of general reasoning or agentic ability, suggesting that the healthcare specialization did not collapse the base model's capabilities.

Limitation Due to restriction of compute, data, and other resource, our result is still bounded by the current training regime. Future development of Cura should broaden clinical behavior coverage, aim for long-horizon agentic work, and move beyond LoRA updates when it scales.

Safety and clinical use Cura 1T is a research model, but not a medical service. It should not be used as a substitute for a clinician. The benchmarks in this report measure narrow competencies under fixed harnesses, and strong scores on those benchmarks do not establish safety for unsupervised clinical use. The numbers should be read as evidence about Cura 1T under the stated evaluations, instead of a guarantee of clinical behavior in deployment.

Acknowledgments

We thank the actAVA research and infrastructure teams, and the maintainers of the open benchmarks and open-weight models this work is built on.

References

Rahul K. Arora et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

autoresearch contributors. autoresearch: AI agents running research on single-GPU nanochat training automatically. GitHub repository, 2026.

Baichuan AI and THUBPM Group, Tsinghua University. Baichuan-m4: A clinical-grade medical agent system for continuous care. *arXiv preprint arXiv:2606.08982*, 2026.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-bench: Evaluating conversational agents in a dual-control environment. *arXiv preprint arXiv:2506.07982*, 2025.

Haolin Chen, Deon Metelski, Leon Qi, Tao Xia, Joonyul Lee, Steve Brown, Kevin Riley, Frank Wang, T. Y. Alvin Liu, Hank Capps, et al. χ-Bench: Can AI agents automate end-to-end, long-horizon, policy-rich healthcare workflows? *arXiv preprint arXiv:2605.16679*, 2026.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Aniketh Garikaparthi, Manasi Patwardhan, and Arman Cohan. ResearchGym: Evaluating language model agents on real-world ai research. *arXiv preprint arXiv:2602.15112*, 2026.

HL7 International. Fast healthcare interoperability resources (fhir). https://www.hl7.org/fhir/, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. MedAgentBench: A realistic virtual EHR environment to benchmark medical LLM agents. *arXiv preprint arXiv:2501.14654*, 2025.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. DSPy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2023.

Kimi Team. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Qianchu Liu, Sheng Zhang, Guanghui Qin, Jeya Maria Jose Valanarasu, Maximilian Rokuss, Mingyu Lu, Timothy Ossowski, Juan Manuel Zambrano Chaves, Cliff Wong, Peniel Argaw, et al. HealthAgentBench: A unified benchmark suite of realistic agentic healthcare environments for challenging frontier AI agents. *arXiv preprint arXiv:2606.31179*, 2026a.

Ruoqi Liu, Imran Q. Mohiuddin, Austin J. Schoeffler, Kavita Renduchintala, Ashwin Nayak, Prasantha L. Vemu, Shivam C. Vedak, Kameron C. Black, John L. Havlik, Isaac Ogunmola, et al. PhysicianBench: Evaluating LLM agents in real-world EHR environments. *arXiv preprint arXiv:2605.02240*, 2026b.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Shiwei Lyu, Xidong Wang, Lei Liu, Hao Zhu, Chaohe Zhang, Jian Wang, Jinjie Gu, Benyou Wang, and Yue Shen. Clinalign: Scaling healthcare alignment from clinician preference. *arXiv preprint arXiv:2602.09653*, 2026.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023.

Mathematical Association of America. American invitational mathematics examination (aime). https://maa.org/maa-invitational-competitions/, 2026.

Alexander Novikov, Ngan Vu, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

OpenAI. Healthbench professional: Evaluating large language models on real clinician chats. *arXiv preprint arXiv:2604.27470*, 2026.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. *arXiv preprint arXiv:2203.14371*, 2022.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: A multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, et al. Large language models encode clinical knowledge. *Nature*, 620:172–180, 2023.

Guiyao Tie, Jiawen Shi, Dingjie Song, Yixiao Huang, Ziji Sheng, Xueyang Zhou, Daizong Liu, Pan Zhou, Yongchao Chen, Ran Xu, Lifang He, Qingsong Wen, Manling Li, Cong Lu, Shuai Li, Pengtao Xie, Yixuan Yuan, Rui Meng, Lei Xing, Lichao Sun, Caiming Xiong, Philip S. Yu, and Jianfeng Gao. AutoResearch AI: Towards ai-powered research automation for scientific discovery. *arXiv preprint arXiv:2605.23204*, 2026.

Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*, 2023.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*, 2023.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. TextGrad: Automatic "differentiation" via text. *arXiv preprint arXiv:2406.07496*, 2024.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. Star: Bootstrapping reasoning with reasoning. *arXiv preprint arXiv:2203.14465*, 2022.

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

A Appendix

A.1 Training Hyperparameters

Table 7 reports the final training hyperparameter of Cura 1T. Hyperparameters for these screening runs vary by capability; the table reports the final consolidation rather than capability-specific screening runs or mixture proportions.

Table 7 Training hyperparameters for Cura 1T.

Configuration	Cura 1T
Algorithm	SFT $\rightarrow$ RL $\rightarrow$ SDFT
Base model	Kimi-K2.6
LoRA rank	32
Learning rate	3×10^{-4}, linear schedule
Batch size	128
Sequence limit	256K tokens
Training length	6 epochs with early stopping

A.2 Harness Implementations

AgentClinic. We replace the original free-text control protocol with two native tools. The doctor calls `order_test(test_name)` to obtain an objective finding and `submit_diagnosis(diagnosis)` to terminate the encounter. History taking remains ordinary assistant dialogue, which is passed to the patient simulator and returned as a user turn. A test result can enter the trajectory only through an executed `order_test` call, removing the free-text router's ambiguity between a requested and a reported finding. NEJM images are attached once in the opening turn, while MedQA encounters remain text-only. The upstream diagnosis comparator grades the terminal diagnosis, and all models in Table 6 use this interface.

MedAgentBench. We expose three renderer-native operations—FHIR read, FHIR write, and final submission—through `fhir_get`, `fhir_post`, and `finish`. Reads are executed against the live FHIR server. Writes are recorded but do not mutate the server; the grader instead inspects the URL and resource payload in the tool-call history. The final-submission tool stores the ordered answer values and terminates the episode. The environment permits up to eight turns, preserves parse failures for recovery, and scores the completed trajectory with the benchmark's task-specific FHIR graders.

A.3 Case Studies

A.3.1 MedAgentBench: Referral ServiceRequest

Task and criterion. Task `task8_1` asks for an orthopedic surgery referral ServiceRequest for synthetic patient S2016972 at timestamp 2023-11-13T10:15:00+00:00. The grader requires SNOMED 306181000000106, `active` status, `order` intent, `stat` priority, the exact subject reference, and an SBAR note.

Stage	Response excerpt	Result
Base	`fhir_post(ServiceRequest, priority=routine, code=306181000000106, display=...)`	Reward 0
Round 1 (Tool-use)	`fhir_post(ServiceRequest, priority=stat, code=306181000000106)`	Reward 1
Round 2 (Tool-use + retention)	`priority=stat`; exact SNOMED coding, subject, timestamp, and SBAR note are preserved.	Reward 1
Round 3 (Harness bug fix)	Under the corrected harness: `priority=stat` with exact coding and `finish([])`.	Reward 1

Change. The base model understands the referral intent but misses the exact FHIR contract. Later responses satisfy the schema and content constraints, and the harness bug fix confirms the correction under the intended evaluation protocol.

A.3.2 HealthBench Professional: EROSTIN 10MG

Task and criterion. The prompt asks, "what medication is EROSTIN 10MG". The criterion requires identifying ebastine as the active ingredient and avoiding confusion with estradiol or Elestrin.

Stage	Response excerpt	Result
Base	"I cannot definitively identify what drug 'Erostin 10 mg' is based on the brand name alone."	Reward 0
Behavior correction	"EROSTIN 10 mg is an antihistamine medication containing ebastine." The task is correct, although the round is reverted for subset-level degradation.	Reward 1
Retention check	"EROSTIN 10 mg is an antihistamine medication that contains ebastine as its active component."	Reward 1
Cura 1T	"EROSTIN 10 mg is an antihistamine medication that contains ebastine as its active ingredient."	Reward 1

Change. This is a drug-identification failure rather than a reasoning failure. Behavior correction fixes this example locally but is not retained globally; the intermediate retention check and Cura 1T preserve the corrected active ingredient. A matched Clean behavior mix trace is not available for this example, so it is not shown.

A.3.3 HealthBench Professional: Sciatica Coding Regression

Task and criterion. The prompt asks for the current ICD-10-CM coding for sciatica. A correct answer must distinguish the billable laterality codes: `M54.30` for unspecified side, `M54.31` for right, and `M54.32` for left. The parent category `M54.3` is not the complete unspecified-side code.

Stage	Response excerpt	Result
Base	"`M54.30`—sciatica, unspecified side; `M54.31`—right; `M54.32`—left."	Reward 1.000
Behavior correction	"`M54.3`—sciatica, unspecified side," while retaining `M54.31` and `M54.32`.	Reward 0.000
Cura 1T	Restores `M54.30`, `M54.31`, and `M54.32`, and distinguishes ICD-10-CM from the WHO category.	Reward 0.973

Change. Behavior correction preserves the correct code family but drops the final digit for unspecified laterality, turning a correct base response into an incomplete coding recommendation. Cura 1T restores the billable code and the laterality distinction. This matched trace illustrates the subset-level degradation that caused Round 1 to be reverted despite its higher overall score.

A.3.4 MedXpertQA Text-197: DCIS Radiotherapy Benefit

Task and criterion. The question asks which ductal carcinoma in situ patient gains the greatest local-control benefit from radiotherapy. The gold answer is D: a 0.9 cm grade-3 DCIS after lumpectomy with a negative margin.

Stage	Response excerpt	Result
Base	"A positive surgical margin signifies likely residual disease and the highest baseline recurrence risk … F."	Reward 0
Round 1 (Reasoning correction)	"Option F combines … grade 3 histology and a positive margin. F."	Reward 0
Round 1L (extended)	"The highest-risk breast-conservation patient … F."	Reward 0
Round 2 (Knowledge injection + retention)	"Positive-margin cases … are not rescued by RT alone … Option D … D."	Reward 1
Round 3 (Mixture refinement)	"Positive-margin cases … are not rescued by RT alone … Option D … D."	Reward 1
Cura 1T	"Positive margins … require further surgery … Option D … D."	Reward 1

Change. The missing piece is a clinical constraint: positive margins change the management problem instead of simply increasing radiotherapy benefit. The retained rounds correct that constraint and preserve the answer through consolidation.

A.3.5 AgentClinic NEJM-Ext 21: Desquamative Interstitial Pneumonia

Task and criterion. The case describes a 34-year-old woman with tobacco use, dyspnea and dry cough, diffuse ground-glass opacities with peripheral consolidation, nondiagnostic bronchoalveolar lavage, and biopsy showing pigment-laden macrophages. The gold diagnosis is desquamative interstitial pneumonia.

Stage	Response excerpt	Result
Base	"The most likely diagnosis is Chronic Eosinophilic Pneumonia."	Reward 0
Reverted diagnostic attempt	"DIAGNOSIS READY: Pulmonary alveolar proteinosis," despite normal anti-GM-CSF and nondiagnostic BAL/PAS results.	Reward 0
Interactive trajectory + retention	"The diffuse ground-glass opacities … and macrophages in the airspaces … [are] classic for Desquamative Interstitial Pneumonia."	Reward 1
Cura 1T	Requests biopsy review, receives "Extensive alveolar filling with pigment-laden macrophages," and submits desquamative interstitial pneumonia.	Reward 1

Change. This multi-turn case depends on gathering and using discriminating evidence before diagnosing. The retained interactive trajectory teaches the workup behavior, and Cura 1T preserves that evidence-led diagnosis.