WHERE CREDIT LANDS: VERIFIABLE STEP CREDIT FOR BOUNDED-MEMORY EHR AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) has enabled large language model (LLM) agents to solve clinical tasks over longitudinal electronic health records (EHRs), where an agent takes a sequence of retrieval and memory decisions over evidence scattered across years of visits. However, the usual optimizer, group relative policy optimization (GRPO), reduces each rollout to one score and gives every action the same advantage, so it cannot tell which action helped or hurt. Crediting each action needs its own score and a comparison with its alternatives, neither of which is readily available for current EHR agents. We propose **V**erifier-**A**nchored **P**rocess **A**dvantage (**VAPA**), a critic-free training method that pairs checks of individual actions with comparisons of alternative decisions, needs no learned reward model, and uses neither verifiers nor replay at test time.[1] First, deterministic checks score retrieval, memory, and stopping actions from the patient record, completed rollout, and task specification. Second, forked replay rewinds to the state before a selected action and samples alternative continuations, with weaker fallback groups where no fork reaches. Finally, a two-level advantage assigns local credit from these comparisons while preserving final-answer credit for every original action. Across MIMIC-IV calculation and retrieval tasks and the held-out MedAgentBench and EHRSHOT suites, VAPA has the highest mean among matched agent, memory, and RL baselines, with no further training for transfer. At the same sampled-token budget, replaying alternative decisions outperforms additional independent rollouts, and VAPA's lead over trajectory-level GRPO grows with decision-chain length and with a tighter memory bound.

1 INTRODUCTION

Large language model (LLM) agents are beginning to take on clinical tasks over longitudinal electronic health records (EHRs), from risk-score calculation to medication and procedure checks (Khandekar et al., 2024; Jiang et al., 2025b). A patient's record can span years, with the evidence for one task scattered across visits, orders, laboratory results, and diagnoses (Johnson et al., 2023; Wornow et al., 2023), so the agent often retrieves it through tools rather than reading the record at once. Each retrieval returns only events recorded before the task cutoff, consumes part of a fixed budget, and remains visible for one turn unless the agent stores it in a small memory. Solving a task therefore requires a sequence of decisions about what to retrieve, what to keep, and when to answer.

Because such tasks have verifiable final answers, reinforcement learning with verifiable rewards (RLVR) provides a direct training signal, and group relative policy optimization (GRPO) is a common way to learn from it (Shao et al., 2024; Guo et al., 2025; Zhang et al., 2025b; Lin et al., 2025). However, GRPO gives every action in a rollout the same advantage, so the final reward shows whether the answer was right, not which decision helped or hurt. In Figure 1A, the agent retrieves a creatinine value, later discards it to free a memory slot, and fails because the value is gone at the answer step. The failed rollout gives every action the same negative advantage, so the retrieval that found the value is penalized as much as the discard that lost it.

For an EHR agent to tell such decisions apart, each one needs a score of its own and a comparison with the alternatives available at the same point, but three limitations arise. **Unverified actions.**

[1] Code is available at https://anonymous.4open.science/r/VAPA

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.

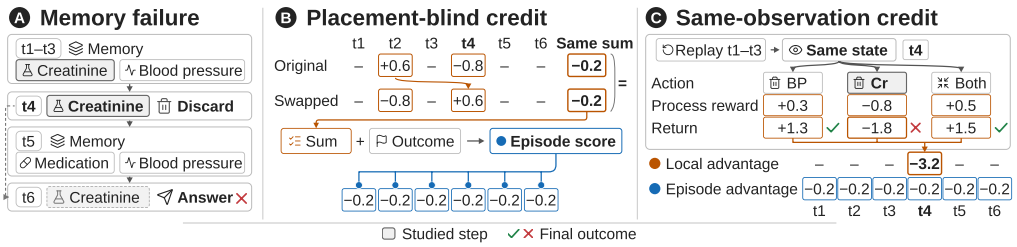


Figure 1: Credit assignment in bounded-memory EHR agents. (A) Discarding creatinine at t_4 frees a memory slot but leaves the value unavailable when needed at t_6. (B) Swapping the rewards at t_2 and t_4 leaves the episode score unchanged, so every action receives the same advantage. (C) The illustrated fork supplies a local advantage at t_4, while the episode advantage reaches all six actions.

Step-level rewards can supervise intermediate decisions (Uesato et al., 2022; Lightman et al., 2024; Cui et al., 2025), but retrieval, memory, and stopping actions do not come with per-action labels or existing verifiers. A learned reward model could supply these scores, but its estimation error would enter the credit itself. **Ungrouped steps.** A direct comparison requires alternative actions chosen with the same visible evidence and memory, but this pre-action state rarely recurs as rollouts diverge. A learned critic could estimate values without direct comparisons, but it must be fitted from the same rollouts it scores. **Single-level advantage.** An episode advantage scores every action by the rollout's final outcome, so it shows whether the rollout succeeded but not which action was responsible. For example, in Figure 1B, swapping the rewards of two steps leaves the episode score unchanged, and every action keeps the same advantage. A step-level (local) advantage does the reverse, comparing an action with alternatives sampled from the same state to show which action helped. However, it exists only where alternatives were sampled, and its outcome term is zero when all alternatives reach the same outcome. Neither level is sufficient alone.

In this work, we propose **V**erifier-**A**nchored **P**rocess **A**dvantage (**VAPA**), a critic-free method that addresses these limitations with three components, uses no learned reward model, and requires neither verifiers nor replay at test time. **Verifier rewards.** Deterministic checks use the cutoff-restricted record, completed rollout, and fixed task specification to score whether retrieval, memory, and stopping actions were grounded, timely, and needed. **Step groups.** Forked replay rewinds a logged rollout to the state before a selected action, samples alternatives under the frozen policy, and continues each branch to an answer, constructing same-state comparisons on demand. Canonical-observation collisions and a turn-index pool provide weaker fallbacks for steps not reached by a fork. **Two-level advantage.** In each original rollout, a local advantage compares decisions with available alternatives, while an episode advantage carries final-answer credit to every action, so each level covers what the other misses, as shown in Figure 1C. We train on MIMIC-IV, evaluate on its calculation and retrieval tasks, and carry the frozen policy to MedAgentBench and EHRSHOT (Khandekar et al., 2024; Jiang et al., 2025b; Wornow et al., 2023). VAPA has the highest mean of the matched RL baselines on every endpoint, including critic-based optimizers that fit a value head or a learned process reward model, reaches its answers in fewer actions, and keeps its lead over GRPO on a second backbone. Verifier rewards and step-level credit each help and help most together, replaying alternative decisions outperforms collecting more independent runs at the same budget, and the lead over trajectory-level training widens as decision chains grow longer.

2 Problem Formulation

We model longitudinal EHR reasoning as a partially observed interaction under a task-specific evidence cutoff, bounded patient memory, and a fixed cost budget. We then show why a trajectory-level reward cannot determine which action earned a process signal.

Instances and evidence. From task distribution $\mathcal{D}$, an instance $z = (E, x, \tau_c, y^\star)$ contains a timestamped patient record E, instruction x, prediction cutoff τ_c, and gold answer $y^\star$. For event e_i with timestamp τ_i, the agent and process verifiers access only $E_{\le \tau_c} = \{e_i \in E : \tau_i \le \tau_c\}$. Only the terminal scorer sees $y^\star$, which may describe a future outcome.

Agent state. At turn t, observation $o_t = (x, \tau_c, \mathcal{R}_t, M_t, h_t, b_t)$ contains the latest tool return $\mathcal{R}_t$, patient memory M_t, compact action history h_t, and remaining budget b_t. A return remains visible for one turn unless saved in M_t, which holds at most L items with values, status, scope, and evidence

pointers. The history records action metadata but no clinical values or reasoning text. Appendix A.2 gives the schemas and action signatures.

Execution. Policy π_θ with parameters θ samples action a_t from the grammar-masked legal set $\mathcal{A}_t$. Each non-answer action has cost $c(a) = 1$ and consumes one unit from initial budget $b_1 = C_{\max}$, while ANSWER costs zero and ends the episode with prediction $\hat{y}$ and evidence-pointer set P. Only ANSWER is permitted once the budget is exhausted or turn cap $T_{\max}$ is reached. A completed rollout is $\omega = (o_1, a_1, \ldots, o_T, a_T)$ with final turn T. The deterministic transition function Φ executes $o_{t+1} = \Phi(o_t, a_t; E_{\leq \tau_c})$, so replaying a logged prefix recovers the same state. Appendix B.2 specifies invalid-action handling.

Rewards. The terminal scorer $R_{\text{out}}(\hat{y}, y^\star)$ checks the committed answer and its format. Evidence pointers may enter process verification but not this score. The evaluation objective $J_{\text{eval}}(\theta)$ is its expectation over instances from $\mathcal{D}$ and rollouts from π_θ,

$$J_{\text{eval}}(\theta) = \mathbb{E}_{z\sim\mathcal{D}}\mathbb{E}_{\omega\sim\pi_\theta(\cdot|z)}[R_{\text{out}}(\hat{y}, y^\star)], \qquad \text{subject to} \quad \sum_{t=1}^{T} c(a_t) \leq C_{\max}. \tag{1}$$

Placement blindness. Trajectory-level training combines the terminal score, process rewards $r_t \in [-1, 1]$, and action costs into

$$R_{\text{traj}}(\omega) = R_{\text{out}}(\hat{y}, y^\star) + \lambda_p \sum_{t=1}^{T} r_t - \lambda_c \sum_{t=1}^{T} c(a_t), \qquad \lambda_p, \lambda_c \geq 0. \tag{2}$$

where λ_p and λ_c are nonnegative process-reward and action-cost weights. The process rewards are defined in Section 3.1. Trajectory-level GRPO converts R_{traj} into one group-relative advantage shared by every action token in the rollout (Shao et al., 2024; Guo et al., 2025).

Observation 1 (Trajectory-level placement blindness)**.** *Fix the sampled rollouts. Any two assignments of process rewards that preserve every rollout's R_{traj} yield identical trajectory-level advantages and policy updates.*

The advantage depends on process rewards only through their sum. Moving the discard penalty in Figure 1 to the answer therefore leaves the update unchanged. Shared prefixes can still provide implicit process credit (Sullivan & Koller, 2026), but this trajectory scalar does not preserve reward placement.

3 METHOD

VAPA turns deterministic checks into action-level training credit. Verifier rewards define what is scored, step groups provide alternatives for each action, and a two-level advantage combines the resulting local advantage with an episode advantage from the final answer (Figure 2).

3.1 VERIFIER REWARDS

Verifier rewards convert each check into a bounded process reward. During training, verifiers may inspect the completed rollout, the full cutoff-restricted record, and the frozen task specification that defines requested fields, windows, and evidence rules. They never inspect the gold terminal answer or post-cutoff events, and their privileged inputs never enter the policy observation.

Action scores. The map $\kappa(a)$ gives an action's type, so $\kappa_t = \kappa(a_t)$ and $\mathcal{V}_{\kappa_t}$ is its verifier set. Verifier ℓ has prespecified reliability weight $w_\ell > 0$, emits score $v_{\ell t} \in [-1, 1]$ when it fires, and otherwise abstains. Its firing indicator is $\delta_{\ell t} \in \{0, 1\}$. The process reward is

$$r_t = \frac{1}{W_{\kappa_t}} \sum_{\ell \in \mathcal{V}_{\kappa_t}} w_\ell \delta_{\ell t} v_{\ell t}, \qquad W_\kappa = \sum_{\ell \in \mathcal{V}_\kappa} w_\ell. \tag{3}$$

For a family with no process check, $r_t = 0$. The denominator sums over the whole family rather than the firing subset, so $r_t \in [-1, 1]$ and abstentions shrink it toward zero rather than amplifying whichever check fired. Appendix A.2 gives the fourteen predicates, their weights, and the set

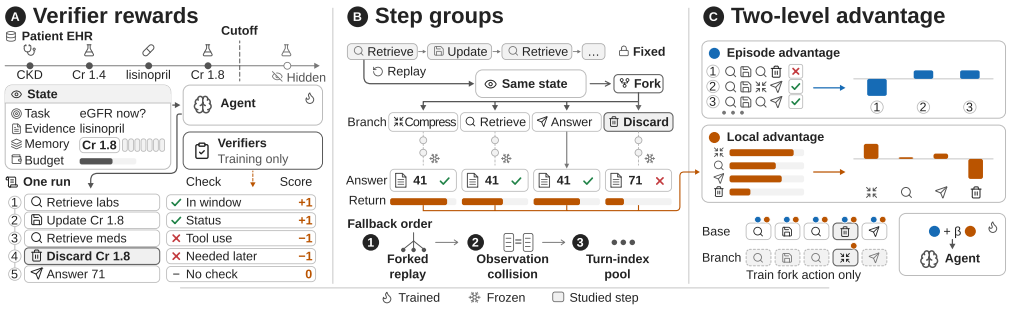


Figure 2: Overview of VAPA. (A) Deterministic verifiers score intermediate actions during training, while the outcome scorer evaluates terminal success. (B) Forked replay rebuilds one pre-action state and samples alternative actions, while canonical observation collisions and a within-instance turn-index pool extend local coverage. (C) Every base action receives the episode advantage and any available local advantage. The running example is the creatinine discard of Figure 1.

applicable to each action type. Cost stays separate because $c(a_t)$ measures resource use, while r_t distinguishes informative from avoidable actions of equal cost. A check may read the current turn or the completed rollout, and either way its score stays attached to the checked action.

Temporal check. WINDOWRESPECTED applies when an evidence-bearing action cites a pointer and the task specifies a time window. It returns $+1$ if every cited event is at or before the cutoff and within the window, and -1 otherwise. It abstains when inapplicable, for instance on a QUERYFIELD request that carries no pointer, or when required verifier inputs are unavailable.

Reward scale. During source training, the action cap bounds the largest process-and-cost difference between two trajectories by $2(C_{\max}+1)\lambda_p + C_{\max}\lambda_c$, allowing for the terminal answer's process checks. We select reward weights that keep this bound below the terminal success gap, so a successful rollout always has a higher raw training return than a failed one while process rewards distinguish how the answer was reached. Appendix B.2 gives the bound for the selected weights.

3.2 STEP GROUPS

Local advantages compare continuation returns within step groups. We prioritize forked replay at identical observations, then canonical observation collisions and a within-instance turn-index pool.

Forked replay. Let $\mathcal{B}(z) = \{\omega_i\}_{i=1}^{G_0}$ be G_0 independent base rollouts for instance z under policy $\pi_{\theta_{\text{old}}}$, with θ_{old} fixed during batch sampling. At selected turn (i,t), deterministic replay of the logged prefix, rather than resampling it, recovers $o_t^{(i)}$ exactly. We draw K fresh actions with replacement from $\pi_{\theta_{\text{old}}}(\cdot \mid o_t^{(i)})$ and retain duplicates. Each branch executes its action and continues under $\pi_{\theta_{\text{old}}}$ until termination. Prefix scores are fixed across siblings, while scores from turn t onward are recomputed for each branch. We call this tree-consistent scoring. In Figure 1C, the fork compares discarding creatinine with alternatives that retain it under the same policy. Appendix B.2 gives G_0, K, the fork schedule, selection rule, and statistical scope.

Collision groups. After fork-group assignment, eligible base and post-fork branch turns from the same instance form a collision group when their canonical-observation signatures agree. Canonicalization normalizes representation details the model-visible serialization can preserve, so this tier gives a recurrence baseline. Appendix A.2 specifies the signature and fork-anchor exclusion rule, and Appendix B.4 specifies branch-turn eligibility.

Turn-index pool. Still-unmatched base turns at the same turn index within the same instance form a pooled group whose members share a remaining budget but arrive along different prefixes. The pool provides a coarse return baseline, and its outcome component recovers an episode-style contrast over the same base returns (Appendix B.4). Because tiers are taken in order, the groups do not overlap.

3.3 TWO-LEVEL ADVANTAGE

Step groups carry answer information only through differences among their continuations, and step groups do not cover every base action. We keep the critic-free construction and add an episode advantage shared by every action in a base rollout.

Credit return. For completed base or branch rollout i, let T_i be its final turn, $R_{\text{out}}^{(i)}$ its terminal score, and $a_k^{(i)}$ and $r_k^{(i)}$ its action and process reward at turn k. At turn t, define

$$G_t^{(i)} = R_{\text{out}}^{(i)} - \lambda_c \sum_{k=t}^{T_i} c\Big(a_k^{(i)}\Big) + \lambda_p \sum_{k=t}^{T_i} \gamma^{\,k-t} r_k^{(i)}, \qquad \gamma \in (0,1]. \tag{4}$$

The answer score and realized costs are undiscounted, while γ controls how far process credit travels. Main step-level arms use $\gamma = 1$, giving $G_1^{(i)} = R_{\text{traj}}(\omega_i)$ under Equation 2. Any answer-format term is included in R_{out}, and setting $\lambda_p = 0$ yields the outcome-only step estimator.

Advantage terms. Let $g(i,t)$ be the strongest available group containing turn (i,t), where $|g(i,t)|$ is its number of members. For $|g(i,t)| > 1$, its leave-one-out local advantage is

$$A_{\text{loc}}(i,t) = G_t^{(i)} - \frac{1}{|g(i,t)| - 1} \sum_{(j,u) \in g(i,t) \setminus \{(i,t)\}} G_u^{(j)}, \tag{5}$$

and a singleton receives $A_{\text{loc}}(i,t) = 0$. At a fork, the K fresh siblings are conditionally independent draws from $\pi_{\theta_{\text{old}}}$ at the recovered observation, while the logged base action was drawn before anchor selection and its distribution can therefore be tilted by that selection. Collision and turn-index groups extend the comparison to their respective peers. Appendix B.2 states the selection bias and bounds it. The episode term compares final answers across the base rollouts of an instance,

$$A_{\text{ep}}^{(i)} = R_{\text{out}}^{(i)} - \frac{1}{G_0} \sum_{j=1}^{G_0} R_{\text{out}}^{(j)}, \qquad \omega_i \in \mathcal{B}(z). \tag{6}$$

Every base action in rollout i receives this term. Auxiliary branches are excluded because they start from selected intermediate states rather than the initial state.

Policy update. Let $\sigma_{\text{ep}}(z)$ be the standard deviation of episode advantages within $\mathcal{B}(z)$ and σ_{loc} the action-level standard deviation over complete non-singleton step groups in the batch. We use

$$\widehat{A}_{\text{ep}}^{(i)} = \frac{A_{\text{ep}}^{(i)}}{\sigma_{\text{ep}}(z) + \epsilon}, \qquad \widehat{A}_{\text{loc}}(i,t) = \frac{A_{\text{loc}}(i,t)}{\sigma_{\text{loc}} + \epsilon}, \qquad \widehat{A}_t^{(i)} = \widehat{A}_{\text{ep}}^{(i)} + \beta \widehat{A}_{\text{loc}}(i,t), \quad \beta \geq 0. \tag{7}$$

Here $\epsilon > 0$ is a numerical stabilizer and β scales the local advantage. Appendix D.2 gives the numerical settings and the masked objective in Equation 9. Base-action tokens receive $\widehat{A}_t^{(i)}$, while fresh fork-action tokens receive only $\beta \widehat{A}_{\text{loc}}$. Later branch turns receive a local term only inside a canonical collision group. Shared branch prefixes and other continuation turns are masked from the branch loss, while generated branch tokens remain charged to the sampling budget (Appendix D.2). The shared batch-level normalizer preserves differences in raw local scale across groups.

Local ordering. For two members (i,t) and (j,u) of a group g with $n = |g| > 1$, Equation 5 gives

$$A_{\text{loc}}(i,t) - A_{\text{loc}}(j,u) = \frac{n}{n-1}\left(G_t^{(i)} - G_u^{(j)}\right). \tag{8}$$

The local ordering follows the sampled continuation returns, which combine answer quality, action cost, and process rewards. When two continuations reach the same answer, process and cost differences can still distinguish preserving required evidence from retrieving it again later.

Credit scope. Fork comparisons credit actions through their sampled continuation returns. At $\gamma = 1$, redistributions preserving each member's suffix total leave the raw local advantage unchanged. Proposition 1 in Appendix D.2 gives the invariance conditions, including common prefix-to-suffix shifts, and discounted diagnostics, and narrows the blind region of Observation 1 to one post-fork suffix.

4 EXPERIMENTS

We evaluate four questions. Does VAPA improve over agent, memory, and RL baselines spanning critic-free and critic-based optimizers under matched caps (**RQ1**)? How does the frozen policy transfer to MedAgentBench and EHRSHOT (**RQ2**)? What do process rewards and the step-credit regime contribute, how does replay compare with extra independent rollouts at matched tokens, and how does the gain depend on the memory bound L (**RQ3**)? How does the advantage over trajectory-level credit vary with decision depth (**RQ4**)?

4.1 EXPERIMENTAL SETUP

Dataset and tasks. We construct training episodes from MIMIC-IV (Johnson et al., 2023). Source evaluation comprises 29 MedCalc-Bench calculators (Khandekar et al., 2024) and five structured-retrieval families. Verified calculation accuracy covers the 13 standard calculators and measures calculator selection and input binding, and Appendix C.4 reports the 16 proxies. Held-out evaluation uses 150 read-only MedAgentBench-Query episodes (Jiang et al., 2025b), eight EHRSHOT tasks on Stanford Medicine records (Wornow et al., 2023) that measure transfer across data sources and tasks for the frozen agents, and a chained suite of separately sampled depth-one through depth-six strata. Source validation determines model selection, and policy and inference settings are fixed for held-out evaluation. Episodes allow $L = 8$ memory items, $C_{\max} = 12$ costed actions, and $T_{\max} = 16$ turns in a 32,768-token context. Appendix A.1 gives construction rules and split statistics.

Design and baselines. The primary factorial uses Qwen3.5-9B (Qwen Team, 2026). All arms share an SFT checkpoint trained from one deterministic reference-program demonstration per episode, plus the curriculum, grammar, costs, optimizer schedule, and sampled-token budget. Table 2 crosses process-reward content, removed by setting $\lambda_p = 0$, with the complete step-credit regime of Section 3. We use five matched seeds for the factorial, no-fork comparisons, and the ten additional trained references, and three for the gpt-oss-20b replication (Agarwal et al., 2025). Qwen3.5-9B runs with thinking disabled and gpt-oss-20b at low reasoning (Appendix B.2). All controls use the shared interface and caps, and span prompted agents, bounded-memory agents, critic-based optimizers, and critic-free optimizers, as listed in Table 1. Target-trained clinical systems are reported separately as unmatched references. Appendix B.3 gives baseline adaptations and comparator scope.

No-fork control. We compare A2 and A4 with variants that allocate branch tokens to additional independent rollouts, retaining collision and turn-index groups and matching curriculum, reward content, estimator, sampled-token budget, and seed index. Their larger base groups change normalization, so the comparison evaluates the two sampling allocations (Appendix B.3).

Evaluation. We report verified calculation accuracy, grounded exact match, MedAgentBench task and cost-normalized success (CNS), and task-wise EHRSHOT AUPRC and AUROC, with metric-specific verifier checks run after policy execution. Self-consistency and oracle best-of-eight evaluate inference scaling and are reported in Appendix B.3, where oracle selection over eight samples exceeds every trained system on the source suites. Appendices C.1 and B.5 give statistical procedures and compute.

4.2 MAIN RESULTS (RQ1, RQ2)

Matched comparison. VAPA has the highest mean among the RL methods on all six endpoints in Table 1, with its largest margin over the strongest Qwen RL baseline on cost-normalized MedAgentBench success. Retrieval and both EHRSHOT contrasts pass the Holm adjustment against the pre-specified comparator and against the later-added Tree-GRPO with process rewards (Appendix C.1). Every RL arm improves on the shared SFT checkpoint.

Wider reference set. Breadth beyond the group-relative family does not close the remaining gap. A learned value head leaves PPO below GRPO on every endpoint, and a learned process reward model in place of the deterministic checks recovers part of the process-reward gain while raising

Table 1: Main performance (%). Agent backbones use Qwen3.5-9B unless noted. Every row shares the EHR interface, action caps, and sampled-token budget. Internal and inference-scaling controls are in Appendix B.3. Bold marks the best RL means. Underlining marks the strongest RL baseline. Subscripts give seed SDs for trained systems and patient-bootstrap SEs for fixed ones. ‡ marks Holm-adjusted $p < 0.05$ against the prespecified comparator (Appendix C.1).

		Calculation	Retrieval	MedAgentBench		EHRSHOT	
Family	Method	Verified accuracy	Grounded exact match	Task success	Cost-normalized success	Mean AUPRC	Mean AUROC
Prompted agents	ReAct (2022)	$77.05_{\pm 1.10}$	$65.83_{\pm 0.83}$	$22.00_{\pm 3.64}$	$18.17_{\pm 3.11}$	$36.39_{\pm 0.68}$	$66.51_{\pm 0.57}$
	Reflexion (2023)	$78.59_{\pm 1.07}$	$67.73_{\pm 0.85}$	$28.67_{\pm 3.80}$	$23.78_{\pm 3.18}$	$37.83_{\pm 0.72}$	$68.02_{\pm 0.58}$
	Adapted EHRAgent (2024)	$79.33_{\pm 1.02}$	$71.06_{\pm 0.79}$	$31.33_{\pm 3.78}$	$21.64_{\pm 2.86}$	$38.02_{\pm 0.71}$	$67.94_{\pm 0.59}$
Memory agents	Adapted Memory-R1 (2026b)	$79.86_{\pm 1.61}$	$72.19_{\pm 0.93}$	$33.87_{\pm 3.75}$	$26.05_{\pm 2.84}$	$38.86_{\pm 0.49}$	$68.77_{\pm 0.45}$
	Adapted Fine-Mem (2026)	$80.31_{\pm 1.57}$	$73.44_{\pm 0.88}$	$35.33_{\pm 3.61}$	$27.91_{\pm 2.77}$	$39.32_{\pm 0.47}$	$69.15_{\pm 0.43}$
Critic-based RL	PPO (2017)	$79.12_{\pm 1.83}$	$69.48_{\pm 1.09}$	$31.07_{\pm 4.35}$	$23.46_{\pm 3.41}$	$38.11_{\pm 0.54}$	$68.05_{\pm 0.52}$
	PPO + learned PRM (2024)	$80.72_{\pm 1.71}$	$72.85_{\pm 1.14}$	$35.60_{\pm 3.98}$	$28.14_{\pm 3.22}$	$39.44_{\pm 0.58}$	$69.31_{\pm 0.49}$
Critic-free RL	GRPO (2024)	$79.49_{\pm 1.54}$	$70.21_{\pm 0.68}$	$32.53_{\pm 3.93}$	$24.91_{\pm 2.98}$	$38.53_{\pm 0.36}$	$68.42_{\pm 0.39}$
	GiGPO (2025)	$80.94_{\pm 1.48}$	$73.66_{\pm 0.71}$	$36.13_{\pm 2.72}$	$28.87_{\pm 2.52}$	$38.97_{\pm 0.41}$	$69.70_{\pm 0.38}$
	Adapted LoGo-GRPO (2026a)	$81.24_{\pm 1.42}$	$74.58_{\pm 0.75}$	$37.87_{\pm 2.84}$	$30.42_{\pm 2.44}$	$39.95_{\pm 0.44}$	$70.13_{\pm 0.37}$
	Adapted VinePPO (2025)	$81.42_{\pm 1.36}$	$75.12_{\pm 0.68}$	$39.20_{\pm 2.56}$	$31.96_{\pm 2.29}$	$40.42_{\pm 0.43}$	$70.86_{\pm 0.36}$
	Tree-GRPO (2026)	$81.53_{\pm 1.44}$	$75.55_{\pm 0.79}$	$39.47_{\pm 2.68}$	$31.24_{\pm 2.71}$	$40.31_{\pm 0.42}$	$70.09_{\pm 0.35}$
	Adapted VPR (2026)	$81.81_{\pm 1.39}$	$75.30_{\pm 0.64}$	$38.93_{\pm 2.29}$	$32.68_{\pm 2.14}$	$40.66_{\pm 0.41}$	$71.12_{\pm 0.34}$
	Tree-GRPO + process	$\underline{82.18}_{\pm 1.36}$	$\underline{76.06}_{\pm 0.80}$	$\underline{40.53}_{\pm 2.60}$	$\underline{33.48}_{\pm 2.50}$	$\underline{41.08}_{\pm 0.39}$	$\underline{71.22}_{\pm 0.35}$
gpt-oss-20b (2025)	GRPO (2024)	$79.04_{\pm 1.79}$	$69.04_{\pm 0.71}$	$31.33_{\pm 4.06}$	$23.83_{\pm 5.95}$	$37.66_{\pm 0.67}$	$67.96_{\pm 0.69}$
	VAPA	$81.94_{\pm 1.55}$	$76.02_{\pm 1.26}$	$40.67_{\pm 3.06}$	$34.46_{\pm 4.25}$	$41.01_{\pm 0.61}$	$71.08_{\pm 0.63}$
Ours	VAPA	$\mathbf{82.63}_{\pm 1.31}$	$\mathbf{78.03}_{\pm 0.94}^{\ddagger}$	$\mathbf{42.67}_{\pm 3.46}$	$\mathbf{36.08}_{\pm 3.02}$	$\mathbf{41.86}_{\pm 0.33}^{\ddagger}$	$\mathbf{71.84}_{\pm 0.34}^{\ddagger}$

Table 2: Factorial ablation (%). The final rows give the joint difference, with its sample seed standard deviation, and the difference-in-differences interaction Δ_{int} in percentage points. Both use per-seed contrasts. † marks Holm-adjusted two-sided $p < 0.05$ within the five-contrast family. Appendix C.1 gives the interaction formula, paired intervals, and analysis protocol.

	Factors		Calculation	Retrieval	MedAgentBench		EHRSHOT	
Arm	Process reward	Step-credit regime	Verified accuracy	Grounded exact match	Task success	Cost-normalized success	Mean AUPRC	Mean AUROC
A1 GRPO	No	No	$79.49_{\pm 1.54}$	$70.21_{\pm 0.68}$	$32.53_{\pm 3.93}$	$24.91_{\pm 2.98}$	$38.53_{\pm 0.36}$	$68.42_{\pm 0.39}$
A2 Step	No	Yes	$80.59_{\pm 1.47}$	$71.94_{\pm 1.01}$	$35.07_{\pm 3.90}$	$27.28_{\pm 2.62}$	$40.11_{\pm 0.50}$	$69.53_{\pm 0.27}$
A3 Process	Yes	No	$81.00_{\pm 1.51}$	$72.76_{\pm 0.96}$	$34.67_{\pm 3.80}$	$28.62_{\pm 2.46}$	$39.21_{\pm 0.57}$	$69.28_{\pm 0.46}$
A4 VAPA	Yes	Yes	$\mathbf{82.63}_{\pm 1.31}$	$\mathbf{78.03}_{\pm 0.94}$	$\mathbf{42.67}_{\pm 3.46}$	$\mathbf{36.08}_{\pm 3.02}$	$\mathbf{41.86}_{\pm 0.33}$	$\mathbf{71.84}_{\pm 0.34}$
A4−A1 (pp)			$+3.14_{\pm 0.72}$	$+7.82_{\pm 0.28}$	$+10.13_{\pm 1.59}$	$+11.17_{\pm 0.58}$	$+3.33_{\pm 0.25}$	$+3.42_{\pm 0.24}$
Interaction Δ_{int} (pp)			$+0.53$	$+3.53^{\dagger}$	$+5.47^{\dagger}$	$+5.09$	$+1.07$	$+1.45$

the rate of rewarded-but-invalid actions (Appendix E.1). The bounded-memory references fall between the trajectory-level and step-level arms throughout, and every exploratory contrast in Appendix C.1 favors VAPA. Adapted EHRAgent leads the prompted systems on grounded retrieval, and its budget-free native setting matches VAPA there within noise while spending three times the actions (Appendix B.3).

Transfer. The frozen policy carries its lead to both held-out suites with no further training, and the contrast replicates on gpt-oss-20b, where VAPA leads GRPO on every endpoint. Target-trained clinical references still lead on EHRSHOT, but on labels the agent never sees. Read against the same encoder's few-shot curve, the zero-label policy matches roughly sixteen labeled examples for each task, and fitting one head on the agent's evidence under matched supervision closes about two thirds of the difference (Table 22).

4.3 Ablations and Controls (RQ3)

Factorial effects. Table 2 crosses process rewards with the complete step-credit regime. Each factor raises every endpoint mean alone, their combination is highest throughout, and the interactions are positive across endpoints, largest on MedAgentBench and smallest on verified calculation accuracy. Grounded retrieval, MedAgentBench task success, and the depth slope pass the five-contrast Holm adjustment. The step-credit factor includes grouping, replay, masking, and the update rule. Exact forks anchor about three percent of base turns, while turn-index pools cover roughly two thirds to three quarters. The interaction is also larger where more of the decision is left to the agent, on calculator proxies rather than standard calculators whose formulas the environment resolves, and on MedAgentBench tasks that aggregate several retrieved values rather than one (Appendix C.4).

Appendix C.2 reports reduced-budget ablations, in which uniform positive process rewards perform worse than removing any single tested verifier family, supporting the value of checks that distinguish useful actions.

Matched-budget replay. Replay improves calculation, retrieval, and both MedAgentBench endpoints under both reward settings, with all eight paired intervals excluding zero and larger mean gains under process rewards (Figure 3A). Appendices B.3 and B.4 give group composition and compute for both allocations.

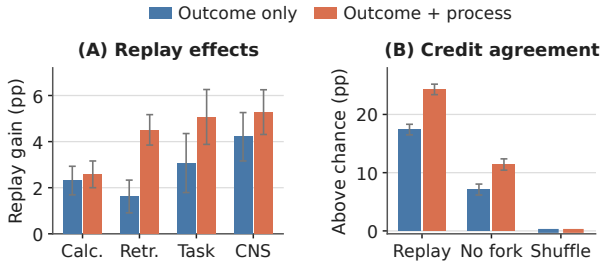


Figure 3: Replay effects and credit agreement. (A) Paired replay gains at matched sampled tokens. (B) Agreement above chance with the forced-action reference. Whiskers are 95% intervals. Exhaustive shuffling gives a zero reference.

Memory capacity. Repeating A1 and A4 at $L \in \{2, 4, 12, 16\}$ under the full budget, both arms improve with capacity while the VAPA margin declines monotonically on all three endpoints, with every interval excluding zero. Part of the advantage comes from the memory bound and part does not. Forced eviction nearly vanishes at the largest capacity, yet two thirds of the margin remains (Appendix C.3).

Credit agreement. At patient-disjoint states fixed before candidate sampling, we compare action orderings with independent forced-action estimates of outcome-and-cost value under the frozen SFT policy, holding observations and continuation policies fixed. Replay has the highest mean agreement under both reward settings (Figure 3B), and its paired advantage remains positive with more or fewer reference continuations, under a stricter reference-margin filter, and when action cost is omitted. The same probe ranks the PPO control's learned value head below both replay and the no-fork signal. Appendix D.1 gives the protocol, the composite-update comparison, and sensitivity results.

4.4 Long-Horizon Advantage (RQ4)

On the held-out chained suite, success declines for every arm as depth increases. The VAPA margin over GRPO more than doubles before leveling off, while the single-factor margins change less (Figure 4), and its prespecified seed-level slope excludes zero after Holm adjustment (Appendix C.1). Each chain passes a retrieved or calculated output to the next link, so later decisions depend on evidence produced earlier. Table 23 also reports cost-normalized success, whose margins compress on deeper chains as the required evidence-bearing actions consume more of the common budget. Depth strata are sampled separately, so the trend reflects their combined differences in length, composition, and difficulty.

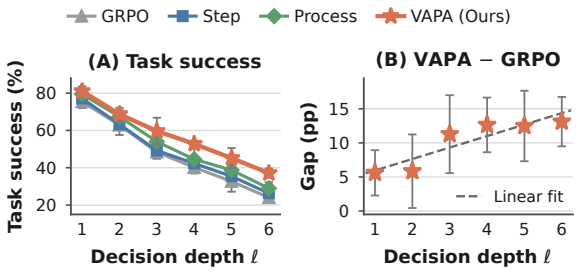


Figure 4: Horizon evaluation on frozen checkpoints. (A) Task success by depth. (B) The A4−A1 difference and its linear fit. Whiskers are five-seed 95% intervals.

4.5 Mechanism and Behavior

Case studies. Figure 5 illustrates the local and episode terms for one fork group. Appendix F presents two matched deployment pairs selected by frozen rules, one in which the replay-trained policy rechecks a stale value and one in which an extra query introduces a unit mismatch.

Behavioral audits. Under blinded annotation, process rewards reduce material trajectory errors in both credit regimes, with both paired intervals excluding zero, while paired changes in invalid

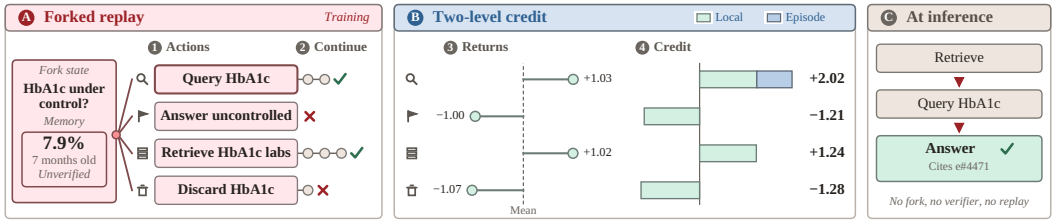


Figure 5: Local credit at one logged exact-fork state, chosen by a frozen rule. (A) Four candidates share the pre-action observation and continue under the frozen policy, with branch length as action cost. (B) Leave-one-out distances from the mean are local advantages (Equation 5, normalized by Equation 7), and only the base action adds the episode term. (C) Deployment uses one rollout.

verifier-positive actions span zero. Corrupting a required memory item lowers continuation utility for A2 through A4, sham reorderings have small effects, and the A4 margin widens with record volume. Appendix E gives the annotation, intervention, and stress-test protocols. The only scripted exploits that succeed against the deterministic catalog edit memory or status.

5 Related Work

Process supervision. Human labels, implicit rewards, and rule-based checks provide feedback on intermediate reasoning (Uesato et al., 2022; Lightman et al., 2024; Wang et al., 2024; Cui et al., 2025; Zhang et al., 2025a). Verifiable methods extend this feedback to agentic reasoning (Yuan et al., 2026; Pronesti et al., 2026) and tool use (Jin et al., 2025; Chen et al., 2025; Feng et al., 2026; Chang et al., 2026). Yet combining these scores into one trajectory reward can reinforce useful and unhelpful actions together, so credit assignment determines how intermediate feedback changes the policy (Li et al., 2026). VAPA uses record-grounded checks in local comparisons among alternative actions and their continuations.

Agent credit assignment. GiGPO compares actions at recurring observations (Feng et al., 2025), VinePPO uses sampled continuations (Kazemnejad et al., 2025), Tree-GRPO branches from shared prefixes (Ji et al., 2026), and iStar learns step rewards (Liu et al., 2026). The closest prior framework is Memory-R2, whose LoGo-GRPO method combines global credit with rerollouts from a shared memory state (Yan et al., 2026a) and learns session-level memory from question-answering rewards, whereas VAPA verifies individual EHR actions and extends coverage beyond selected forks. Our matched controls cover all of these regimes under a common token budget, together with critic-based optimizers that fit a value head or a learned process reward model, so the comparison is between complete credit regimes rather than isolated estimators.

Clinical and memory agents. EHRAgent executes code over multi-table EHRs and reuses successful cases (Shi et al., 2024), and we port it to the fixed patient tools and per-episode budget, also reporting its native code-execution setting. Memory-R1 trains separate memory-manager and answer agents (Yan et al., 2026b), and Fine-Mem adds auxiliary chunk questions with evidence-anchored attribution (Ma et al., 2026). Both are adapted to the same bounded memory and caps, as are ReAct and Reflexion. These designs change the architecture or the supervision channel, whereas VAPA verifies cutoff validity, units, status, and evidence consistency within one bounded-memory agent. Target-trained clinical predictors are reported separately, with a supervision-matched variant.

6 Conclusion

We introduced VAPA, a critic-free method turning record-grounded checks into local comparative credit while retaining final-answer credit. Its margin over trajectory-level GRPO widens with decision depth and with the tightness of the memory bound, and inference uses neither replay nor verifiers. Applying the method beyond EHRs requires reproducible states, task-grounded action checks, and fixed-policy continuations.

AI USE STATEMENT

We used large language models for drafting and editing, for parts of the experimental design and the analysis plan, for developing the conceptual framing, and for formulating and checking the mathematical claims, including the statement and proof of Proposition 1. This assistance also covered reporting templates and draft implementation specifications. The authors are responsible for verifying the experiments, code, and reported content.

ETHICS STATEMENT

We use only deidentified releases. MIMIC-IV and EHRSHOT are used under their credentialed data use agreements, and MedAgentBench under its released license. No direct identifiers or clinical free text are shown. Case studies use only deidentified structured values and relative dates exposed by released benchmark scorers. All record-derived data stay on institution-controlled compute. The systems answer questions about a record and do not diagnose, triage, or recommend treatment, so nothing here supports deployment without clinician review.

REPRODUCIBILITY STATEMENT

Appendices A and B specify task construction, splits, action and verifier semantics, training settings, and compute accounting. Appendices C.1 through E.3 give the statistical, audit, and intervention procedures. Reproduction requires the implementation, manifests, per-seed predictions, annotation records, and analysis scripts listed in Appendix B.6, together with independent credentialed access to the source data. Appendix G states the limitations of the reported evidence. Record-derived data are not redistributed. The anonymized artifact is at https://anonymous.4open.science/r/VAPA.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Qikai Chang, Zhenrong Zhang, Pengfei Hu, Jun Du, Jiefeng Ma, Yicheng Pan, Jianshu Zhang, Quan Liu, and Jianqing Gao. THOR: Tool-integrated hierarchical optimization via RL for mathematical reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0Af7UiJISU.

Mingyang Chen, Linzhuang Sun, Tianpeng Li, sunhaoze, ZhouYijie, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. Research: Learning to reason with search for LLMs via reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=OuGAwwAT8G.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. Retool: Reinforcement learning for strategic tool use in LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=tRk1nofSmz.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=QXEhBMNrCW.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Yuxiang Ji, Ziyu Ma, Yong Wang, Guanhua Chen, Xiangxiang Chu, and Liaoni Wu. Tree search for LLM agent reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ZpQwAFhU13.

Pengcheng Jiang, Cao Xiao, Adam Richard Cross, and Jimeng Sun. Graphcare: Enhancing healthcare predictions with personalized knowledge graphs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tVTN7Zs0ml.

Pengcheng Jiang, Cao Xiao, Minhao Jiang, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, and Jiawei Han. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=8fLgt7PQza.

Yixing Jiang, Kameron C Black, Gloria Geng, Danny Park, James Zou, Andrew Y Ng, and Jonathan H Chen. MedAgentBench: a virtual EHR environment to benchmark medical LLM agents. *NEJM AI*, 2(9):AIdbp2500144, 2025b.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Rwhi91ideu.

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.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. VinePPO: Refining credit assignment in RL training of LLMs. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Myx2kJFzAn.

Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, Soren Dunn, Serina S Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad W Safranek, Abid A Anwar, Andrew Zhang, et al. MedCalc-Bench: Evaluating large language models for medical calculations. *Advances in Neural Information Processing Systems*, 37:84730–84745, 2024.

Zhuofeng Li, Haoxiang Zhang, Seungju Han, Sheng Liu, Jianwen Xie, Yu Zhang, Yejin Choi, James Zou, and Pan Lu. In-the-flow agentic system optimization for effective planning and tool use. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Mf5AleTUVK.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024.

Jiacheng Lin, Zhenbang Wu, and Jimeng Sun. Training LLMs for EHR-based reasoning tasks via reinforcement learning. *arXiv preprint arXiv:2505.24105*, 2025.

Xiaoqian Liu, Ke Wang, Yuchuan Wu, Fei Huang, Yongbin Li, Jianbin Jiao, and Junge Zhang. Agentic reinforcement learning with implicit step rewards. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ooROvpmxMV.

Weitao Ma, Xiaocheng Feng, Lei Huang, Xiachong Feng, Zhanyu Ma, Jun Xu, Jiuchong Gao, Jinghua Hao, Renqing He, and Bing Qin. Fine-Mem: Fine-grained feedback alignment for long-horizon memory management. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19666–19684, 2026. doi: 10.18653/v1/2026.acl-long.900.

Massimiliano Pronesti, Anya Belz, and Yufang Hou. Beyond outcome verification: Verifiable process reward models for structured reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 32187–32202, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C Ho, Carl Yang, and May Dongmei Wang. EHRAgent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22315–22339, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

Michael Sullivan and Alexander Koller. GRPO is secretly a process reward model. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=nMGaOCVlDW.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Fries, and Nigam Shah. EHRSHOT: An EHR benchmark for few-shot evaluation of foundation models. *Advances in Neural Information Processing Systems*, 36:67125–67137, 2023.

Sikuan Yan, Ahmed Bahloul, Ercong Nie, Susanna Schwarzmann, Riccardo Trivisonno, Volker Tresp, and Yunpu Ma. Memory-R2: Fair credit assignment for long-horizon memory-augmented LLM agents. *arXiv preprint arXiv:2605.21768*, 2026a. URL https://arxiv.org/abs/2605.21768.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z Pan, et al. Memory-R1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12805–12825, 2026b.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Huining Yuan, Zelai Xu, Huaijie Wang, Xiangmin Yi, Jiaxuan Gao, Xiao-Ping Zhang, Yu Wang, Chao Yu, and Yi Wu. Verifiable process rewards for agentic reasoning. *arXiv preprint arXiv:2605.10325*, 2026.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-VL: Learning to reason with multimodal large language models via step-wise group relative policy optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1859–1869, October 2025a.

Sheng Zhang, Qianchu Liu, Guanghui Qin, Tristan Naumann, and Hoifung Poon. Med-RLVR: Emerging medical reasoning from a 3B base model via reinforcement learning. *arXiv preprint arXiv:2502.19655*, 2025b.

APPENDIX CONTENTS

A Tasks and Environment **15**
A.1 Tasks and Metrics . . . 15
A.2 Interface and Verifiers . . . 16

B Training and Baselines **19**
B.1 Training Algorithm . . . 19
B.2 Training Details . . . 20
B.3 Baselines and Controls . . . 22
B.4 Rollout Accounting . . . 25
B.5 Compute and Efficiency . . . 26
B.6 Governance and Release . . . 27

C Additional Results **28**
C.1 Statistical Analysis . . . 28
C.2 Design Ablations . . . 30
C.3 Memory Capacity . . . 31
C.4 Task-Level Results . . . 32
C.5 Target-Trained Clinical References . . . 34

D Credit and Optimization **36**
D.1 Credit Agreement . . . 36
D.2 Optimization Diagnostics . . . 38

E Behavior and Robustness **39**
E.1 Reward-Channel Audits . . . 39
E.2 Memory Interventions . . . 40
E.3 History Length . . . 41
E.4 Subgroup Breakdown . . . 42

F Case Studies **43**

G Limitations and Broader Impact **44**

A TASKS AND ENVIRONMENT

Conventions. Brackets denote 95% intervals. Appendix C.1 defines aggregation, seed dispersion, and patient-bootstrap uncertainty.

A.1 TASKS AND METRICS

Source-task construction. MIMIC-IV episodes pair a patient, instruction, cutoff, and executable reference program (Johnson et al., 2023). The manifest fixes input mappings, units, windows, aggregation, missingness, ties, and tolerances. Of 29 MedCalc-Bench definitions (Khandekar et al., 2024), 13 retain standard formulas and 16 are operationalized proxies, including four judgment-item proxies. Unsupported definitions are excluded. Standard-formula accuracy measures calculator selection and input binding. Retrieval covers latest laboratory values, medication activity, first pre-cutoff diagnosis, procedures within a window, and recorded-versus-missing status. Definitions and exclusions precede training.

Splits and sampling. A fixed patient hash partitions patients before episode construction. Duplicate patient-task-target-cutoff tuples are removed. Calculation cases need cutoff-valid inputs and finite outputs. Retrieval cases need a canonical answer or prespecified missingness state. Sampling is without replacement under a fixed seed. Caprini and APACHE II proxy training pools contain only 361 and 389 cases, yielding 21,550 of the 21,600 targeted training episodes. The targets are 11,600 training, 1,740 validation, and 3,480 test episodes over 29 calculation types, and 10,000, 1,500, and 3,000 over five retrieval families, with no borrowing across splits.

Horizon suite. The evaluation-only suite contains 250 held-out episodes at each depth from one through six. Every chain uses fixed retrieval families or standard calculators, with each output supplying an input to the next link and each link requiring an evidence-bearing action. Even the deepest chain leaves budget for memory management and recovery. All arms and seeds share the 1,500 episodes and greedy inference caps. No chained episode enters training. Comparisons are paired within episode, while separately constructed depths may differ in composition and difficulty.

Target tasks. MedAgentBench-Query contains the 150 released GET-only tasks (Jiang et al., 2025b). Its adapter maps calls and scorers to the shared grammar and budgets. EHRSHOT contains Stanford Medicine records (Wornow et al., 2023). The event-mapping specification preserves prediction cutoffs, concept identities, and unit semantics, drops events without a typed schema field, and marks unresolved units as ambiguous. Coverage accounting separates total, retained, unmapped, and unresolved-unit counts by task and patient. Agents use the source-task prompt, grammar, and budgets. Supervised references read the released files directly. We retain all three operational tasks and the three laboratory and two new-diagnosis tasks with the largest validation positive-label counts. Quotas precede ranking. Laboratory labels merge all non-normal categories into one positive class. Agents remain frozen, and supervised references use target training and validation splits with the same derived labels. The evaluation uses the selected tasks and instance caps below.

Frozen evaluation subsample. Before subsampling, the eight selected EHRSHOT-derived test tasks contain 190,517 label instances. Before model evaluation, each task is capped at 2,500 instances. Above the cap, round r admits the r-th eligible instance of each patient in a fixed, label-independent order until the cap is reached. Tasks below the cap are retained whole. Table 3 reports full and retained counts and prevalence. Metrics and patient-clustered intervals describe this retained sample, including its reduced weight on repeated instances.

Source metrics. Let $j(i)$ denote the calculator used by episode i, and let y_i and $\hat{y}_i$ be its reference and parsed terminal values. For numerical outputs, define the verification indicator ν_i by

$$\nu_i = \mathbf{1}[|\hat{y}_i - y_i| \le \epsilon_i], \qquad \epsilon_i = \max\left\{\epsilon_{\text{abs}}^{(j(i))}, 0.05|y_i|\right\}.$$

Here $\mathbf{1}[\cdot]$ is the indicator function and ϵ_i is the episode-specific tolerance. The absolute tolerance $\epsilon_{\text{abs}}^{(j)}$ is half the smallest unit of calculator j's frozen output convention, and is defined at zero and negative references. Categorical, rule-based, and date outputs need exact agreement after prespecified canonicalization. Verified accuracy averages ν_i within calculator, then weights calculators equally. For

Table 3. Selected EHRSHOT-derived endpoints and retained evaluation counts. Positive rates are percentages. Capped tasks retain the patient roster while reducing repeated instances.

			Test split			Frozen evaluation subsample		
Family	Task	Patients	Instances	Positive	Positive rate (%)	Instances	Positive	Positive rate (%)
Operational	Long length of stay	1,238	2,195	552	25.1	2,195	552	25.1
Operational	30-day readmission	1,190	2,189	260	11.9	2,189	260	11.9
Operational	ICU transfer	1,154	2,037	85	4.2	2,037	85	4.2
Laboratory	Anemia	2,002	58,155	39,970	68.7	2,500	1,676	67.0
Laboratory	Hyponatremia	1,956	67,028	19,274	28.8	2,500	720	28.8
Laboratory	Thrombocytopenia	1,998	56,338	19,137	34.0	2,500	874	35.0
Diagnosis	Hyperlipidemia	864	1,317	172	13.1	1,317	172	13.1
Diagnosis	Hypertension	755	1,258	159	12.6	1,258	159	12.6
All eight tasks			190,517			16,496		

retrieval, $m_i, g_i \in \{0, 1\}$ indicate canonical match and cutoff-valid grounding for the field and window. NOTRECORDED is grounded only when the tool return shows no qualifying event. Grounded exact match averages $m_i g_i$ within query family, then weights five families equally.

Target metrics. For the N MedAgentBench-Query tasks, let $s_i \in \{0, 1\}$ be official task success, $C_i = \sum_t c(a_{it})$ be realized action cost, and $C_{\max,i} > 0$ be the episode budget. The environment enforces $0 \le C_i \le C_{\max,i}$. We report task success (TS) and cost-normalized success (CNS),

$$\mathrm{TS} = 100\,\frac{1}{N}\sum_{i=1}^{N} s_i, \qquad \mathrm{CNS} = 100\,\frac{1}{N}\sum_{i=1}^{N} s_i\left(1 - \frac{C_i}{C_{\max,i}}\right).$$

Failures receive zero CNS. The adapter sets $C_{\max,i} = C_{\max} = 12$ for every MedAgentBench-Query, source, and horizon episode, so the subscript is kept only for generality. Direct context has $C_i = 0$, so its CNS equals task success.

Transfer readout. All selected EHRSHOT endpoints are binary. For an episode with a valid terminal ANSWER, let χ_i be the token prefix before its prediction field, and let $u_0 = \texttt{0}$ and $u_1 = \texttt{1}$ be the class identifiers. The manifest verifies both identifiers are single tokens under each backbone's tokenizer. For class index $k \in \{0, 1\}$, the frozen policy yields log score ℓ_{ik} and positive-class probability p_i,

$$\ell_{ik} = \log p_\theta(u_k \mid \chi_i), \qquad p_i = \frac{\exp(\ell_{i1})}{\exp(\ell_{i0}) + \exp(\ell_{i1})},$$

where $p_\theta(u_k \mid \chi_i)$ is the next-token probability of class identifier u_k under the frozen policy parameters θ.

Format imputation. Missing or malformed terminal answers receive the median valid score for the same system and task, or 0.5 if that system produces no valid score for the task. AUROC gives tied positive-negative pairs half credit, while AUPRC is non-interpolated average precision, treating a tied block at one threshold. These are whole-cohort metrics under a fixed imputation rule. Appendix C.4 reports failure rates and sensitivity to four block placements. Metrics are computed over retained label instances within each task and macro-averaged across tasks. Patient-clustered uncertainty follows Appendix C.1.

A.2 INTERFACE AND VERIFIERS

Record schema. Each cutoff-valid event $e_i = (\tau_i, d_i, f_i, v_i, \xi_i, \varsigma_i)$ records time τ_i, clinical domain d_i, field f_i, value v_i, optional text ξ_i, and source ς_i. Table 4 specifies access and memory operations.

Action grammar. Table 4 gives the action signatures. In order, q, w, and d denote a query, time window, and clinical domain. The remaining arguments are field f, calculator ϕ, binding vector $\mathbf{b}$, memory item m, item identifier j, status σ, scope ρ, justification η, item set S, prediction $\hat{y}$, and evidence-pointer set P. A syntactically valid action is legal only when its identifiers resolve in the

observation and its cost fits the budget. The environment never changes memory implicitly. UPDATEMEMORY initializes a new item or replaces an existing item's content and evidence pointers. Replacement preserves the existing status and scope, which only MARKSTATUS may change.

Table 4. Agent actions, arguments, and environment effects. Every nonterminal action costs one unit. Answering terminates the episode at zero tool cost.

Action	Arguments	Environment effect
RETRIEVE(q, w, d)	query, time window, domain	Returns ranked cutoff-valid evidence with stable pointers.
QUERYFIELD(f, w)	field, time window	Returns FOUND, NOTRECORDED, AMBIGUOUS, or OUTOFWINDOW.
CALCULATE$(\phi, \mathbf{b})$	calculator, input binding	Evaluates the frozen definition of ϕ on the bound inputs and returns a value under its output convention, or UNBOUND, UNRESOLVED, or AMBIGUOUS.
UPDATEMEMORY(m)	typed memory item	Creates a typed item, or replaces content and pointers while preserving its existing status and scope.
MARKSTATUS(j, σ, ρ, η)	item, status, scope, justification	Changes only the named item's status and validity scope.
COMPRESS(S)	nonempty item set	Replaces S with one summary retaining the union of required pointers.
DISCARD(j)	item identifier	Removes one item.
ANSWER$(\hat{y}, P)$	prediction, evidence chain	Terminates the episode at zero tool cost.

Prompt surfaces. Table 5 gives the shared prompt and a worked turn. Direct context replaces the action interface with a cutoff-restricted record and permits only ANSWER. Always retrieve uses a fixed action plan. EHRSHOT uses a binary task string and the prediction-field readout above. Tool returns enter the next user message. The artifact includes the exact chat templates and the four demonstrations used by the prompted controls.

Scaffold strings. ReAct fills the scaffold slot of Table 5 with `Think briefly about what the task still needs, then emit one action.` Reflexion fills it with `Before answering, state in one line whether the evidence in memory is sufficient and what is missing, then emit one action.` Both strings are frozen before training and their tokens count against the turn output cap.

Memory and retrieval semantics. Only UPDATEMEMORY, COMPRESS, and DISCARD change occupancy. Full-buffer writes fail without silent eviction. RETRIEVE filters cutoff-valid events by domain and window before ranking. QUERYFIELD applies normalized lookup, window, and tie rules. NOTRECORDED, OUTOFWINDOW, and AMBIGUOUS distinguish absence, evidence outside the window, and unresolved ties or units. The manifest fixes indexing, normalization, ranking, and tie handling.

Calculator resolution. CALCULATE$(\phi, \mathbf{b})$ requires a known calculator, complete bindings, and available budget. Bindings identify memory items or evidence pointers. The resolver applies the frozen units, windows, aggregation, formula, and output convention using only those inputs. Missing bindings, inadmissible or undefined inputs, and unresolved units or ties return UNBOUND, UNRESOLVED, or AMBIGUOUS. Both success and failure cost one unit. Missing inputs follow the manifest without implicit defaults. Tool-output rounding follows the frozen output convention. Terminal-answer tolerance follows Appendix A.1.

Canonical structured-state signature. For collision grouping only, we serialize the task and cutoff, tool return, structured memory, action window, remaining budget, and legal-action mask as canonical JSON with sorted keys. Timestamps, codes, units, numbers, and evidence pointers use normalized representations, and whitespace, rollout identifiers, and metadata are removed. Two observations collide only when these canonical byte strings are identical. Each group logs its serial-

Table 5. Shared prompt template and a worked retrieval turn. Angle brackets denote runner-filled slots. The footer explains serialization and is not emitted.

SYSTEM

```
You answer one question about one patient record. Use only evidence recorded on or
before the cutoff. Emit exactly one action per turn.
Actions: Retrieve(query, window, domain) | QueryField(field, window) |
Calculate(calculator, bindings) | UpdateMemory(item) |
MarkStatus(id, status, scope, why) | Compress(ids) | Discard(id) |
Answer(prediction, evidence)
End your reply with the action call on one line, arguments in the order listed above
and comma-separated, and write nothing after that line.
window is a relative span such as last_90d, last_1y, or all. domain is one of lab,
med, dx, proc, note. bindings maps each required calculator input to a memory id or
an evidence pointer. evidence is a bracketed pointer list such as [e#4471, e#5210],
and a pointer is the ptr field of a tool return.
Memory holds at most <mem_cap> items, each with a value, a status of current | stale
| contradicted | uncertain, a validity scope, and an evidence pointer. A write to
full memory is rejected. Status changes require MarkStatus.
Every non-answer action costs one unit. Answer is free, ends the episode, and must
cite the pointers it relies on. At zero budget only Answer is legal.
<scaffold>
```

USER turn t

```
Cutoff <cutoff>  Budget <budget_left>/<budget_cap>  Memory <mem_used>/<mem_cap>
Task: <task>
Memory: <memory>  Last return: <tool_return>  History: <history>
```

ASSISTANT

```
QueryField(hba1c, last_90d)
```

TOOL

```
FOUND  value=6.4  unit=%  time=cutoff-21d  ptr=e#4471
```

USER turn $t+1$

```
Cutoff <cutoff>  Budget <budget_left>/<budget_cap>  Memory <mem_used>/<mem_cap>  ...
```

The footer is explanatory only. Slots correspond to the observation fields in Section 2. The scaffold is empty for the factorial arms and SFT only, and supplies the ReAct or Reflexion instruction for those controls. Body blocks are newline-separated. Wider inline gaps represent two spaces. The final ellipsis abbreviates repeated fields. The next user message includes the raw tool return.

ization version and content hash. Canonicalization removes field-order variation that remains visible in the policy prompt. Appendix E.2 tests sensitivity to that variation.

Collision-group scope. This tier groups recurring structured states and budgets across trajectories, allowing presentation details to vary. Exact-fork anchors retain their own tier, and later branch turns remain eligible for collision groups.

Verifier catalog. Each deterministic predicate decides applicability and returns a score in $[-1, 1]$, positive for satisfied, negative for violated, and zero for applicable neutral cases. Equation 3 averages firing scores using fixed reliability weights and a denominator that includes abstentions. The score is zero when all predicates abstain or are inapplicable. Predicates use action arguments, cutoff-restricted records, completed traces, and frozen manifest fields such as requested windows and reference-evidence rules. Gold terminal answers, post-cutoff events, and learned scores are excluded from their inputs. Table 6 organizes the predicates into seven families. For action type κ, $\mathcal{V}_\kappa$ contains its assigned predicates across these rows, and W_κ sums their fixed weights. The implementation specification in the artifact lists predicate identifiers, action membership, weights, required inputs, and scoring and abstention conditions.

Predicate weights. Table 7 names the fourteen predicates, their families, their fixed reliability weights, and the action types each one applies to. Six predicates that admit an exact decision against the record carry weight 2, and eight heuristic predicates carry weight 1. Weights are set before RL from this exactness distinction alone and are never tuned. The last column of the table gives $\mathcal{V}_\kappa$ for each action type, and the final row gives the resulting normalizer W_κ of Equation 3, so a reader can

Table 6. Deterministic verifier families and their positive and negative conditions. Table 7 gives the predicates of each family and the action types they apply to.

Family	Positive condition	Negative condition
Extraction	Field, value, unit, and time match the cited event	Any extracted component conflicts with its pointer
Grounding	Every asserted item resolves to supporting admissible evidence	Missing, unsupported, or mismatched pointer
Temporal validity	Evidence precedes the cutoff and satisfies the requested window	Post-cutoff or out-of-window evidence is used as support
Missingness	Recorded, absent, ambiguous, or out-of-window state is handled correctly	A value is invented or a qualifying value is ignored
Status and scope	Status and validity scope match the task and evidence	Status or scope is inconsistent with later use
Compression and discard	Required content and pointers are preserved, or an irrelevant item is removed	Required evidence is lost or provenance is broken
Stopping	The action advances an unresolved requirement, or the answer is supported	A redundant tool call is made, or the policy stops without sufficient evidence

evaluate r_t for any action from the predicate outputs alone. Because the denominator sums over the whole family rather than the firing subset, r_t stays in $[-1, 1]$ for every action type whatever the abstention pattern.

Table 7. The fourteen deterministic predicates. Weights are fixed before RL. R, Q, C, U, M, K, D, and A mark RETRIEVE, QUERYFIELD, CALCULATE, UPDATEMEMORY, MARKSTATUS, COMPRESS, DISCARD, and ANSWER. The final row gives W_κ, the sum of the weights applicable to each action type.

Predicate	Family	w_ℓ	R	Q	C	U	M	K	D	A
EXACTEXTRACTION	Extraction	2		✓	✓	✓				✓
UNITCONSISTENCY	Extraction	2			✓	✓				✓
POINTERRESOLUTION	Grounding	2			✓	✓				✓
SUPPORTGROUNDING	Grounding	2			✓	✓				✓
WINDOWRESPECTED	Temporal validity	2			✓	✓				✓
MISSINGNESSHANDLING	Missingness	2		✓	✓					✓
STATUSCORRECTNESS	Status and scope	1				✓	✓			✓
SCOPECORRECTNESS	Status and scope	1					✓			✓
CONTENTPRESERVATION	Compression and discard	1						✓		
PROVENANCEPRESERVATION	Compression and discard	1						✓	✓	
DISCARDNECESSITY	Compression and discard	1							✓	
CHAINCONSISTENCY	Grounding	1								✓
REDUNDANTCALL	Stopping	1	✓	✓	✓					
STOPPINGSUFFICIENCY	Stopping	1								✓
Normalizer W_κ			1	5	13	11	2	2	2	16

B TRAINING AND BASELINES

B.1 TRAINING ALGORITHM

Algorithm 1 combines replay sampling, tier assignment, local advantages, and branch-loss masking. Its inputs define θ_{SFT} as the supervised checkpoint, $\mathcal{V}$ as the verifier catalog, G_0 as the number of base rollouts, K as the number of fresh siblings for each fork, and B as the global sampled-token budget. During each update, θ_{old} is the frozen sampling copy of current actor parameters θ, θ_{ref} is the fixed KL reference, $\mathcal{T}$ is the accumulated trajectory set, N_{batch} is its newly generated token count, and *spent* is the cumulative generated-token count. Subsequent subsections specify these operations and their constants. The factorial evaluates this complete step-credit regime.

Algorithm 1. VAPA training with replay groups and action-level credit. All newly sampled tokens count toward the budget. The update uses eligible policy tokens.

Require: instance stream $\mathcal{D}$, SFT checkpoint θ_{SFT}, verifier catalog $\mathcal{V}$, base group G_0, siblings K, token budget B

$\theta \leftarrow \theta_{\mathrm{SFT}}$, reference adapter $\theta_{\mathrm{ref}} \leftarrow \theta_{\mathrm{SFT}}$, spent $\leftarrow 0$
while spent $< B$ **do**
  $\theta_{\mathrm{old}} \leftarrow \theta$ ▷ freeze the sampling policy for this batch
  $\mathcal{T} \leftarrow \emptyset$, $N_{\mathrm{batch}} \leftarrow 0$ ▷ accumulate one update batch
  while $N_{\mathrm{batch}} < 131{,}072$ and spent $< B$ **do**
    draw instance $z \sim \mathcal{D}$ under the curriculum of Appendix B.2
    roll out $\mathcal{B}(z) = \{\omega_i\}_{i=1}^{G_0}$ under $\pi_{\theta_{\mathrm{old}}}$ until episode termination
    select at most one replay-eligible state per stratum using pre-action weights over the realized candidate set
    for all selected turns (i, t) **do**
      replay the logged prefix to recover $o_t^{(i)}$ exactly
      sample K actions from $\pi_{\theta_{\mathrm{old}}}(\cdot \mid o_t^{(i)})$ and complete each branch
    end for
    score every turn of every completed rollout by Equation 3, sharing prefix scores across siblings and recomputing suffix scores per branch
    form each exact-fork group from its retained base action and fresh siblings
    group remaining eligible base and post-fork branch turns by canonical-observation collision
    pool still-unmatched base turns by turn index within the instance
    add all base rollouts and fresh branches to $\mathcal{T}$, keeping this instance group intact
    add $N_{\mathrm{generated}}(z)$ to both spent and N_{batch} ▷ all new tokens
  end while
  $G_t^{(i)} \leftarrow$ Equation 4 for every turn in $\mathcal{T}$
  $A_{\mathrm{loc}} \leftarrow$ Equation 5, leave-one-out within tier group, zero for singletons
  $A_{\mathrm{ep}} \leftarrow$ Equation 6 for every base action
  normalize advantages by Equation 7 under the rules of Appendix D.2
  assign $\widehat{A}_{\mathrm{ep}} + \beta \widehat{A}_{\mathrm{loc}}$ to base actions and $\beta \widehat{A}_{\mathrm{loc}}$ to eligible branch actions
  mask shared branch prefixes, post-branch turns outside a collision group, and malformed or retried generations
  if any eligible tokens remain, take one AdamW step on the masked surrogate plus KL to θ_{ref}
end while
return the final budget-matched θ
Inference runs one greedy trajectory from θ with no fork, no verifier, and no reference adapter.

Token counters. $N_{\mathrm{generated}}(z)$ is the complete group count defined in Appendix B.4. It enters the global and batch counters once per group.

B.2 Training Details

Models and initialization. Primary runs adapt the text-only language stack of Qwen3.5-9B (Qwen Team, 2026) with thinking disabled. Replication uses gpt-oss-20b (Agarwal et al., 2025) at low reasoning. Native templates and tokenizers share the system instruction and tool schema. Rank-16 LoRA with scaling factor $\alpha = 32$ and no dropout adapts attention, feedforward, and architecture-specific projections. Base weights, embeddings, and heads remain frozen. Weights and activations use BF16, with gpt-oss experts dequantized at loading. Optimizer state and statistics use FP32. Reasoning tokens are budgeted and omitted from later observations.

Supervised initialization. A deterministic controller executes each training reference program to produce one demonstration per episode. SFT trains action and answer tokens for one epoch while masking instructions, observations, and reasoning. All RL arms for a backbone share its final SFT checkpoint. During RL, generated reasoning belongs to its action and receives that action's advantage.

Table 8. Training settings shared across the primary experiments. Backbone-specific loading details are given in the text. Selected reward weights also transfer to the replication.

Setting	Value
Primary and replication	Qwen3.5-9B and gpt-oss-20b
Context and turn output	32,768 and 512 tokens
Memory and episode caps	$L = 8,\ C_{\max} = 12,\ T_{\max} = 16$
LoRA	rank 16, $\alpha = 32$, dropout 0
Precision and loading	Qwen3.5 BF16, gpt-oss-20b experts loaded in BF16, FP32 optimizer and statistics
SFT	1 epoch, learning rate 1×10^{-4}
RL	learning rate 5×10^{-6}, nominal 2^{25} sampled tokens
Base group and replay	$G_0 = 8$, at most 2 fork states, 3 siblings for each state
Return and local weight	$\gamma = 1,\ \beta = 1$
KL reference and weight	frozen SFT checkpoint, $\eta_{\mathrm{KL}} = 0.01$
Process and cost weight	$\lambda_p = 0.05,\ \lambda_c = 0.02$

Interface and curriculum. Table 8 gives the shared caps. Memory starts empty, and only explicit actions change it. Nonterminal actions cost one unit. A write to full memory is rejected with an error observation. Unparsable turns consume one turn and one available budget unit. A malformed answer, or a non-answer at zero budget, fails the episode. Since $C_{\max} = 12$ and answering is free, the 16-turn cap does not bind. RL samples the two source suites equally and task types uniformly within suite. The first 20% of tokens uses Retrieve, Calculate, UpdateMemory, and Answer on the lower two history quartiles. The next 30% adds QueryField, MarkStatus, and the third quartile. The final 50% enables all actions and the full training distribution. The schedule is fixed across arms.

Sampling and replay. Base and branch sampling use temperature 1.0, nucleus threshold $p = 1$, and no vocabulary truncation. Actor, behavior, and reference probabilities use that full-softmax distribution. Single-trajectory evaluation is greedy. Self-consistency and best-of-eight use the training sampler with fixed decoder seeds. An eligible replay state permits at least two legal actions. Remaining budget b_t defines the two selection strata

$$\mathcal{B}_{\mathrm{early}} = \{o_t \mid C_{\max} - b_t < C_{\max}/2\}\,, \qquad \mathcal{B}_{\mathrm{late}} = \{o_t \mid C_{\max} - b_t \geq C_{\max}/2\}\,.$$

Each base group contributes at most one anchor per stratum. Memory occupancy of at least 75% doubles its sampling weight. An empty stratum transfers its draw when another distinct eligible state exists. Weights use pre-action observations, and the candidate set reflects termination and eligibility along the completed base rollouts. The three fresh siblings are independent conditional on the anchor.

Anchor-selection bias. Let $S(i,t)$ be the event that turn (i,t) is chosen as an anchor. Because S is measurable with respect to the completed base rollouts, conditioning on it reweights the retained base action's law by $\Pr(S \mid a_t^{(i)}, o_t)/\Pr(S \mid o_t)$, and whenever that factor varies with $a_t^{(i)}$ the base member's expected local advantage carries a selection term that the sibling comparisons do not. Two properties bound what this can do. Leave-one-out advantages sum to zero inside a complete group, so a tilt on the base member appears as an equal and opposite shift spread over its siblings and cannot change the group's total credit. The tilt also acts through action identity alone rather than through the sampled continuation returns, so it shifts the base member's advantage without reordering the siblings among themselves. Measured over the reported fork groups, the base member's mean normalized local advantage is $+0.041$ $[0.028, 0.054]$ against a sibling mean of -0.014 $[-0.018, -0.009]$. A selection-free reading of the same groups, which drops the base member and compares only the K fresh siblings, leaves the sibling ordering unchanged and removes that shift. Appendix D.1 evaluates the composite ordering that results once episode credit is added to the base member. Duplicate sibling actions are retained, and branches terminate without further forks.

Optimizer and budget. We optimize an on-policy gradient surrogate with AdamW, one optimizer step for each accumulated batch. All arms use one on-policy update per sampled batch. A 3% linear warmup and cosine decay to 10% of the initial rate follow cumulative sampled token fraction, with weight decay 0.01 and gradient clipping at 1.0. Each nonfinal update accumulates complete base groups and branches to at least 131,072 sampled actor tokens, and no group is split across updates.

Sampling for the final update stops at the first intact group meeting the nominal global target, and paired runs differ by at most one group. Token accounting follows Appendix B.4, and advantage normalization follows Appendix D.2. We report the final budget-matched checkpoint.

Training trajectories. In Figure 6 the arms separate during the middle curriculum stage and retain their ordering over the remaining evaluated budget.

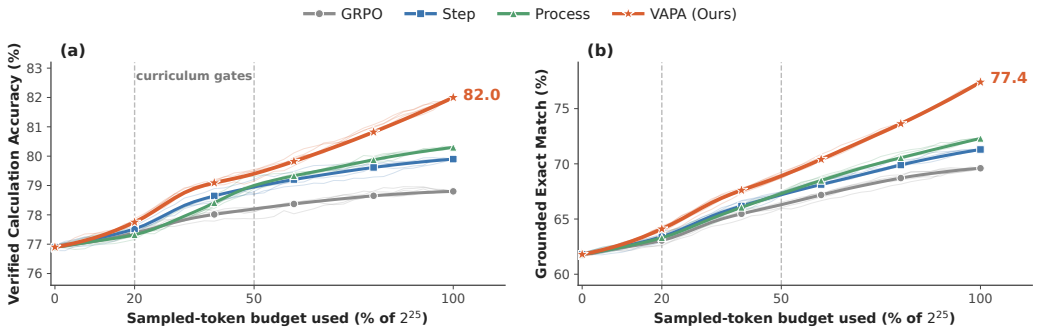


Figure 6. Source-validation learning curves across five policy seeds. Dashed vertical lines mark curriculum transitions. The horizontal axis is cumulative sampled actor tokens.

Rewards and selection. The terminal reward is $R_{\text{out}} = 2s - 1$, where s is answer correctness under the calculation or canonical retrieval scorer of Section 2. Retrieval grounding is evaluated separately and contributes through the process checks. The process reward $r_t \in [-1, 1]$ follows Equation 3, with frozen reliability weights. At $\lambda_p = 0.05$, $\lambda_c = 0.02$, and $\gamma = 1$, at most 13 scored turns and 12 costed actions bound the process-plus-cost difference between trajectories by

$$13(2\lambda_p) + 12\lambda_c = 1.54.$$

That bound is below the terminal success gap of 2 between $R_{\text{out}} = +1$ and $R_{\text{out}} = -1$, so a failed trajectory never outscores a successful one. The bound reaches 2 at $\lambda_p \approx 0.068$, which is why the VAPA λ_p grid in Table 10 stops at 0.05 although that value is selected. This bounds the unnormalized trajectory return. The episode term uses terminal rewards alone, and Section 3.3 rescales local advantages by a batch quantity. Setting $\lambda_p = 0$ produces the outcome-only arms without changing action costs. The joint source-validation grid in Table 10 selects λ_p, λ_c, and β. Appendix B.3 gives the pilot protocol and transfer settings.

B.3 Baselines and Controls

Comparison protocol. Matched RL systems share the actor, interface, reward inputs, caps, training configuration, and sampled-token budget of Section 4.1. Table 9 records how published methods are adapted to these constraints. Native EHRAgent and target-trained clinical predictors retain different resource or supervision settings and are reported separately as unmatched references.

Prompted controls. Direct context disables tools and memory and receives a cutoff-restricted record under a frozen temporal-stratified truncation rule. Its action cost is zero. Always retrieve executes a fixed, label-free query plan from the task schema and answers when the plan finishes or the action budget is reached. SFT only uses the shared initialization without RL. ReAct adds a rolling thought-action-observation scaffold (Yao et al., 2022). Reflexion adds bounded self-critique before answering (Shinn et al., 2023). Both receive four demonstrations chosen by a frozen nearest-task-family rule and run one trajectory under the same caps. They receive no gold answers, verifier output, or cross-episode memory, and reasoning counts toward the token cap. Their shared prompt differs only in the scaffold slot of Table 5.

RL baseline adaptations. GRPO is A1 in Table 2, and every row receives the same actor-token budget. Tree-GRPO + process retains its tree groups and tuned sibling weight but replaces the terminal-only return with Equation 4 under the shared verifier catalog. For trajectory i at turn t, adapted VPR instead normalizes $\lambda_p r_{i,t} - \lambda_c c(a_{i,t}) + \mathbf{1}[t = T_i] R_{\text{out}}^{(i)}$ across active trajectories (Yuan et al., 2026), where $\mathbf{1}[t = T_i]$ indicates the terminal turn. It uses the frozen global-batch fallback

Table 9. Matched RL controls. Columns give the advantage baseline, the process-reward channel, the local comparison, and forked sampling, all under the shared actor-token budget. *Verifier* marks the deterministic catalog of Appendix A.2.

Method	Advantage baseline	Process reward	Local comparison	Forks
A. CRITIC-BASED				
PPO (2017)	Learned value head	None	Token-level GAE	No
PPO + learned PRM (2024)	Learned value head	Learned	Token-level GAE	No
B. CRITIC-FREE				
GRPO (2024)	Group mean	None	None	No
GiGPO (2025)	Group mean	None	Natural state collisions	No
Adapted LoGo-GRPO (2026a)	Group mean	None	Shared memory state	Memory state
Tree-GRPO (2026)	Group mean	None	Sibling, inter-tree	Yes
Tree-GRPO + process (2026)	Group mean	Verifier	Sibling, inter-tree	Yes
Adapted VPR (2026)	Same-turn mean	Verifier	Same-turn normalization	No
Adapted VinePPO (2025)	Monte Carlo value	Verifier	Value at fork states	Resampled
VAPA	Group + step groups	Verifier	Fork, collision, turn	Yes

when fewer than four trajectories remain and carries neither delayed process credit nor an episode, state-matching, or replay term.

Adapted VinePPO. Our VinePPO adaptation (Kazemnejad et al., 2025) uses the same shaped return and selects at most two states per base group. Three independent continuations from each replayed observation estimate the state value $\widehat{V}(o)$. Selected base turns receive $w_v(G_t - \widehat{V}(o_t))$, where w_v is the tuned value-baseline weight. Other turns receive the group-normalized trajectory return. All continuation tokens count toward the budget. For the same retained action and continuations, let G_{base} be the retained base-action return and G_k the return of fresh sibling k. VAPA's raw base-action local term is $G_{\text{base}} - K^{-1}\sum_{k=1}^{K} G_k$. The regimes additionally differ in normalization, episode credit, fallback groups, and auxiliary-action loss, so this comparison evaluates their combined design.

Critic-based controls. PPO adds a rank-16 scalar value-head adapter and uses action-token generalized advantage estimation (GAE) with trace-decay coefficient $\lambda_{\text{GAE}} = 0.95$, clip threshold 0.2, the shared terminal reward and costs, and no process rewards (Schulman et al., 2017). The head starts from SFT, trains jointly for one value epoch after each policy step, and uses batch-normalized advantages and the grid in Table 10. Appendix D.1 reports its held-out explained variance and induced action ordering.

Learned process reward. PPO + learned PRM keeps that optimizer and replaces the deterministic catalog with a process reward model, a rank-16 head on the same backbone, trained under a Bradley–Terry objective on 48,000 action-level preference pairs. Pairs are mined from the reserved pilot rollouts by outcome-conditioned contrast, which pairs two actions of the same type at the same turn index within one instance whose rollouts reach different terminal outcomes, and are split 90/10 by patient for early stopping. A $\tanh$ squashes the head's output into $[-1, 1]$, so it enters Equation 4 on the scale of Equation 3. Unlike the deterministic predicates it never abstains, so every legal action is scored. The head is frozen before RL and its weight is selected on the grid in Table 10, which matches the VPR grid. This is the learned-reward-model alternative discussed in Section 1, and Appendix E.1 audits its decisions and the behavior they induce.

Adapted LoGo-GRPO. Memory-R2's local-global objective uses the group-normalized trajectory return and three rerollouts at each of two recurring-memory anchors per base group (Yan et al., 2026a). Rerollout tokens count toward the shared budget. Unlike an exact VAPA fork, a shared-memory anchor can group turns with different tool returns, cutoffs, or remaining budgets.

Adapted Memory-R1. We retain separate memory-manager and answer adapters on the shared backbone and budget (Yan et al., 2026b). The manager controls memory actions, while the answer agent retrieves, calculates, and answers. Both train from the terminal outcome.

Adapted Fine-Mem. Four chunk questions per episode are generated before training from the cutoff-restricted record and scored by exact match (Ma et al., 2026). An equal-weight attribution term requires every retained memory item to carry a pointer cited by the answer. This supplies fine-grained supervision without comparing alternative actions at one state.

Adapted EHRAgent. EHRAgent's loop of code generation, execution, error feedback, and case reuse is retained (Shi et al., 2024). Under the matched interface, generated code may call only the episode's patient tools, each executed tool call costs one unit, and the case store is the same L-item memory rather than an unbounded cross-question bank. Four demonstrations follow the frozen nearest-task-family rule used for ReAct and Reflexion. The native configuration instead grants a cutoff-restricted multi-table database view, free-form execution with up to three repair attempts, and a case bank that persists across the suite, with no action budget, and Table 11 reports it as an unmatched reference. It averages 20.7 actions for each episode against 8.6 under the matched interface and 6.9 for VAPA, 14.3% of episodes need at least one repair, and 3.9% exhaust the repair allowance and are scored as failures. Its terminal answer still passes through the shared ANSWER grammar, so the transfer readout of Appendix A.1 applies unchanged and only the evidence-gathering path differs. Both configurations use Qwen3.5-9B, whereas the published system used a substantially larger proprietary model, so these rows measure the framework at this backbone.

Coefficient selection. Table 10 gives the grids and selected values. Each pilot uses the shared SFT checkpoint, one reserved seed, 2^{23} tokens, and the fractional curriculum, maximizing the mean of source validation calculation and retrieval. Final values are frozen before reported runs and transfer unchanged to the second backbone.

Table 10. Source-validation coefficients and selected values. Each reference tunes one coefficient on a five-point grid with its other settings fixed, either 0.25 to 4 in doublings for credit and value weights or 0.0125 to 0.2 in doublings for process weights. VAPA uses the joint grid shown.

Method	Coefficient	Selected
PPO	Value-loss coefficient	0.5
PPO + learned PRM	Process weight	0.05
GiGPO	State-credit weight	0.5
Adapted LoGo-GRPO	Local-global mixing weight	1
Adapted Memory-R1	Manager reward weight	0.5
Adapted Fine-Mem	Chunk-question weight	0.5
Tree-GRPO	Sibling-credit weight	1
Tree-GRPO + process	Sibling-credit weight	1
Adapted VPR	Process weight	0.05
Adapted VinePPO	Value-baseline weight	1
VAPA	λ_p, from $0.025, 0.05$	0.05
VAPA	λ_c, from $0.01, 0.02, 0.03$	0.02
VAPA	β, from $0.5, 1, 2$	1

Unmatched controls. Table 11 separates systems that do not share the matched action cost or sampled-token budget. Self-consistency and oracle best-of-eight use eight SFT trajectories (Wang et al., 2023; Cobbe et al., 2021). Self-consistency takes the plurality answer, breaking ties by length-normalized policy likelihood, and averages EHRSHOT class probabilities. Oracle best-of-eight selects the best hidden source-suite outcome and is not deployable.

No-fork control. The matched no-fork variants of A2 and A4 share initialization, instance order, curriculum, reward content, estimator, optimizer, caps, and seed index. Branch-token quota is reassigned to extra independent base rollouts on the same visited instances, in instance order, with allocation differing by at most one rollout within a stratum. This changes base-group size, normalization, and fallback availability. The control compares sampling allocations. Table 15 separately reports aggregate GPU-hours for the no-fork runs.

Table 11. References that share the task suites but not Table 1's action cost or sampled-token budget (%). Direct context answers with no tools at zero action cost, the inference controls spend eight trajectories for each episode, and native EHRAgent has no action budget, so its cost-normalized success is undefined. Subscripts are patient-bootstrap SEs. Oracle best-of-eight selects on the hidden outcome and is not deployable.

	Calculation	Retrieval	MedAgentBench		EHRSHOT	
System	Verified accuracy	Grounded exact match	Task success	Cost-normalized success	Mean AUPRC	Mean AUROC
A. INTERNAL CONTROLS						
Direct context	$75.19_{\pm 1.12}$	$50.30_{\pm 0.94}$	$16.00_{\pm 3.24}$	$16.00_{\pm 3.24}$	$36.22_{\pm 0.68}$	$65.94_{\pm 0.56}$
Always retrieve	$78.08_{\pm 1.03}$	$68.57_{\pm 0.84}$	$29.33_{\pm 4.12}$	$18.72_{\pm 2.48}$	$38.30_{\pm 0.74}$	$67.81_{\pm 0.58}$
SFT only	$77.56_{\pm 1.03}$	$62.37_{\pm 0.90}$	$24.67_{\pm 3.64}$	$19.89_{\pm 2.95}$	$36.95_{\pm 0.70}$	$67.18_{\pm 0.57}$
B. INFERENCE SCALING						
Self-consistency (2023)	$78.72_{\pm 1.06}$	$64.70_{\pm 0.89}$	$26.67_{\pm 3.73}$	$16.44_{\pm 2.44}$	$37.94_{\pm 0.73}$	$67.72_{\pm 0.58}$
Oracle best-of-eight (2021)	$85.32_{\pm 0.87}$	$84.53_{\pm 0.68}$	–	–	–	–
C. UNBUDGETED AGENT						
Native EHRAgent (2024)	$80.15_{\pm 0.98}$	$77.42_{\pm 0.74}$	$38.00_{\pm 3.96}$	–	$38.44_{\pm 0.70}$	$68.31_{\pm 0.60}$
VAPA	$82.63_{\pm 1.31}$	$78.03_{\pm 0.94}$	$42.67_{\pm 3.46}$	$36.08_{\pm 3.02}$	$41.86_{\pm 0.33}$	$71.84_{\pm 0.34}$

Table 12. Replay and no-fork means at matched sampled tokens (%), the eight within-setting contrasts summarized in Figure 3A. Gains are pp, with paired five-seed 95% t intervals. Appendix B.5 reports computation for both allocations.

Endpoint	Replay	No fork	Gain
A. OUTCOME ONLY (A2)			
Verified calculation accuracy	80.59	78.28	+2.31 [1.69, 2.93]
Grounded exact match	71.94	70.32	+1.62 [0.91, 2.33]
MedAgentBench task success	35.07	32.00	+3.07 [1.79, 4.35]
MedAgentBench cost-norm. success	27.28	23.07	+4.21 [3.16, 5.26]
B. OUTCOME AND PROCESS (A4)			
Verified calculation accuracy	82.63	80.05	+2.58 [2.00, 3.16]
Grounded exact match	78.03	73.52	+4.51 [3.85, 5.17]
MedAgentBench task success	42.67	37.60	+5.07 [3.88, 6.26]
MedAgentBench cost-norm. success	36.08	30.80	+5.28 [4.31, 6.25]

B.4 ROLLOUT ACCOUNTING

Group assignment. Algorithm 1 assigns each base turn to its highest eligible tier. Equal-return groups keep their tier and yield zero local credit. Appendix D.2 defines normalization and token eligibility.

Token accounting. $N_{\text{generated}}(z)$ includes every newly generated token in a complete base-plus-branch group, including masked, malformed, retried, and branch tokens. Cached or replayed prefixes are excluded. Prompt processing, tools, and verification are recorded separately. Complete groups remain intact across updates and can overshoot the nominal budget.

Realized coverage. Table 13 reports the highest available tier for each base turn, pooled over curriculum stages. At training temperature 1.0, base rollouts average 9.43 turns and 1,926 tokens under A2, and 7.62 turns and 1,567 tokens under A4. The residual tier contains turns without an eligible peer.

Fallback-tier properties. When a turn-index pool contains one turn from each of the G_0 base rollouts, its outcome component satisfies

$$A^{\text{out}}_{\text{loc}}(i,t) = R^{(i)}_{\text{out}} - \frac{1}{G_0 - 1} \sum_{\substack{j=1 \\ j \neq i}}^{G_0} R^{(j)}_{\text{out}} = \frac{G_0}{G_0 - 1} A^{(i)}_{\text{ep}}.$$

For a strict subset of base rollouts, the corresponding centering uses that subset's outcome mean. Process and cost suffixes add variation beyond this outcome contrast. When all outcomes agree, these auxiliary terms determine any nonzero local advantage. Appendix C.2 evaluates removal of this tier. Collision opportunities are concentrated at early turns because all base rollouts initially share o_1, while subsequent evidence and memory decisions can make their states diverge.

Fork-tier ceiling. With at most two fork states per eight-rollout base group, the reported mean lengths imply the exact-fork coverage of Table 13. Exact utilization requires integer selection counts. Fork branches account for 0.2750 and 0.2836 of generated tokens, and 12.2% and 15.8% of branch turns enter collision groups. Thus replay changes both fork-action training and collision coverage. The allocation control in Table 12 evaluates these effects together.

Table 13. Local-group coverage as a fraction of base-rollout turns. Pooled fractions weight arms by their realized turn counts. Nonzero credit is reported separately.

Tier	A2	A4	Pooled
Exact-fork group	0.0265	0.0328	0.0296
Canonical-observation collision	0.2172	0.2953	0.2559
Within-instance turn index	0.7503	0.6531	0.7021
No available group	0.0060	0.0188	0.0123
Nonzero local term	0.8430	0.8334	0.8382

Realized budget. Under the reported ledger, A2 closes 1,580 groups averaging 21,250 tokens and takes 238 updates, with 20,568 tokens above the nominal 2^{25} budget. A4 closes 1,921 groups averaging 17,475 tokens and takes 240 updates, with 15,043 tokens above nominal. Nonfinal updates close on the first whole group reaching 131,072 tokens and average 141,026 and 140,058 tokens respectively. Shorter trajectories allow more instance exposure at a matched token budget.

B.5 Compute and Efficiency

Hardware and accounting. Each run uses one 96 GB NVIDIA RTX PRO 6000 Blackwell GPU, with four jobs per node and hours measured as device-hours. A colocated paged-attention server refreshes the actor adapter after each update. Separate actor and reference adapters share one backbone through sequential forwards, while checkpointing and accumulation keep training on one device. The actor-token budget includes branches. Training hours include generation, tools, verifiers, reference probabilities, optimization, and checkpoints, but exclude loading and index construction. Table 14 reports measured hours and peak allocated memory, with additive-model throughput estimates for the Qwen factorial.

Table 14. Training and source-test inference cost. GPU hours show reported per-run means and seed SDs. Qwen factorial throughput entries are additive-model estimates. Peak memory and inference costs are reported separately.

		Training			Single trajectory inference		
Backbone	Checkpoint	GPU hours for each run	Sampled tokens per s	Peak memory in GiB	Actions for each episode	Generated tokens for each episode	Latency in s for each episode
Qwen3.5-9B	SFT	1.5	NA	52.4	9.8	2,090	57.0
	A1	8.69 ± 0.17	1,072	71.8	9.0	1,870	51.2
	A2	9.75 ± 0.23	956	76.3	8.1	1,660	45.5
	A3	9.17 ± 0.33	1,016	73.1	7.9	1,600	43.8
	A4	10.39 ± 0.35	897	77.4	6.9	1,350	37.2
gpt-oss-20b	SFT	2.3	NA	64.7	10.2	2,290	45.8
	A1	14.02 ± 0.42	665	81.2	9.4	2,010	40.5
	A4	16.89 ± 0.86	552	82.6	7.3	1,490	30.2

Evaluation compute. Pilots use 2^{23} sampled tokens and final RL runs use 2^{25}. Evaluation batches 48 episodes per GPU, reaching 1,909 generated tokens/s for Qwen3.5-9B and 2,493 for gpt-oss-20b. Table 15 reconstructs evaluation cost from suite sizes, generated tokens, and executed passes. Shared self-consistency and oracle trajectories are charged once.

Table 15. Reproduction compute by component. Panel A covers fixed-configuration runs and retains the full evaluation allowance. Panel B separates search and design screening. RL actor-token counts are nominal, and aggregate evaluation hours are reconstructed from suite sizes and throughput.

Component	Runs	GPU hours	RL tokens in billions
A. FIXED-CONFIGURATION REPRODUCTION			
SFT initialization for both backbones	2	4	NA
Primary Qwen3.5-9B factorial with 4 arms, each with 5 seeds	20	190	0.671
Controls without forks with 2 arms, each with 5 seeds	10	97	0.336
RL reference systems with 10 systems, each with 5 seeds	50	539	1.678
Memory-capacity sweep with 2 arms, each with 4 capacities and 5 seeds	40	382	1.342
gpt-oss-20b replication with 2 arms, each with 3 seeds	6	93	0.201
Credit assignment probe and memory interventions	NA	8	NA
Evaluation across suites and inference controls	NA	319	NA
Reproduction allowance	128	1,632	4.228
B. SEARCH AND DESIGN SCREENING			
Hyperparameter pilot for Qwen3.5-9B	18	47	0.151
Baseline coefficient pilot with 10 systems, each with 5 settings	50	120	0.419
Design ablations with 16 configurations, each with 3 seeds	48	125	0.403
Search and screen subtotal	116	292	0.973

Reproduction budget. Fixed-configuration reproduction uses 1,632 GPU-hours, including 1,305 for training and 327 for probes and evaluation. Search and design screening are separate. Clinical-reference refits, the fitted head, native EHRAgent, loading, index construction, and development add 21 GPU-hours and no actor tokens. Whole-group overshoot raises the two-panel total from 5.201 to at most 5.207 billion actor tokens. Table 14 separately reports single-trajectory source-test actions, tokens, and latency.

B.6 GOVERNANCE AND RELEASE

Pinned versions. The study uses Qwen3.5-9B and gpt-oss-20b with MIMIC-IV v3.1, selected EHRSHOT tasks, MedAgentBench-Query, and corrected MedCalc-Bench definitions. MIMIC-IV v3.1 has DOI `10.13026/kpb9-mt58`. The companion manifest records checkpoint revisions, dataset and label checksums, the adapter version, and the environment digest.

Access and credentialing. Record access requires compliance with the source data-use agreements and institutional authorization. The release record documents the applicable institutional determination. Record-derived episodes are not redistributed. Examples use relative time offsets and deidentified identifiers, and reconstruction requires independent access to the source databases.

Processing and release. Record-derived prompts and trajectories are processed locally on institution-controlled compute with downloaded open-weight checkpoints. The release manifest identifies the version, license, and checksum of each dependency, dataset access conditions continue to govern record-derived artifacts, and the task adapter and reconstruction code are distributed separately.

Reproduction records. Reconstruction requires the task adapter, construction and split manifests, trainer configuration, annotation records, and per-seed predictions. The companion specification and recomputation scripts define their required fields.

C Additional Results

C.1 Statistical Analysis

Endpoints and aggregation. RL systems, GraphCare, and KARE report means and sample standard deviations across seeds. Fixed systems report patient-clustered bootstrap standard errors. Each backbone shares one SFT checkpoint across RL runs. Calculation evaluates thirteen standard calculators separately from sixteen operationalized proxies. MedAgentBench-Query averages its 150 instances. EHRSHOT metrics are task-macro averages on the retained sample.

Factorial contrasts. Let $m_{a,s}^{(j)}$ denote endpoint j for arm $a \in \{1,2,3,4\}$ and matched seed $s \in \{1,\dots,S\}$, where S is the number of seeds. Let $\bar{m}_a^{(j)} = S^{-1}\sum_{s=1}^{S} m_{a,s}^{(j)}$. We report three prespecified contrasts

$$\begin{aligned}\widehat{\Delta}_{\text{proc|step}}^{(j)} &= \bar{m}_4^{(j)} - \bar{m}_2^{(j)},\\ \widehat{\Delta}_{\text{step|proc}}^{(j)} &= \bar{m}_4^{(j)} - \bar{m}_3^{(j)},\\ \widehat{\Delta}_{\text{int}}^{(j)} &= (\bar{m}_4^{(j)} - \bar{m}_3^{(j)}) - (\bar{m}_2^{(j)} - \bar{m}_1^{(j)}).\end{aligned}$$

The first contrast measures process-reward content within the step-credit regime, the second the complete regime with process reward present. That regime jointly changes group construction, replay availability, branch-token masking, and the two-level advantage. The interaction tests whether the gain is larger under that regime than under trajectory-level assignment. The overall $\bar{m}_4^{(j)} - \bar{m}_1^{(j)}$ difference is descriptive. The two simple effects are differences of Table 2 cells and every one of their paired intervals excludes zero. All contrasts are in percentage points.

Seed-level uncertainty. For within-seed contrast d_s, the paired interval is

$$\bar{d} \pm t_{0.975,S-1}\frac{s_d}{\sqrt{S}},$$

where $\bar{d}$ and s_d are the mean and sample standard deviation of the S seed-specific differences d_s, and $t_{0.975,S-1}$ is the 97.5th percentile of Student's t distribution with $S-1$ degrees of freedom. Contrasts use unrounded within-seed values, allowing a one-unit rounding difference in the last displayed digit. Three-seed backbone comparisons provide secondary replication. Intervals condition on fixed evaluation sets, use training seeds as the replication unit, and accompany two-sided tests.

Patient-level resampling. Fixed-system and patient-sampling analyses use 10,000 paired bootstrap replicates and percentile intervals. MedAgentBench tasks sharing a named patient form a cluster, with unnamed tasks as singletons. EHRSHOT resamples patients jointly from the union of task cohorts before macro-averaging. Hierarchical audit and stress-test resampling first selects matched seed tuples, then patient clusters. Trained-system endpoint contrasts instead use seed-level paired t intervals.

Multiplicity and display. The factorial family contains four interactions and the horizon slope. Baseline comparisons contain six endpoints per reference family. Panel A is exploratory because its references were selected using the original three seeds. Panel B uses adapted VinePPO, the prespecified comparator, and defines the main-table ‡ markers. Panel C gives the exploratory Tree-GRPO with process rewards comparison. Panels D through F in Table 18 are exploratory for the same reason. Holm adjustment uses all six displayed raw p values in each panel. Simple effects and diagnostics retain nominal intervals.

Freeze order. Analysis choices were fixed in the following order, each before the data it governs were seen. Task construction, splits, verifier catalog, and reward weights were fixed before any RL run. The factorial design, the five confirmatory contrasts of Table 16, the five-seed count, and the horizon suite were fixed after the reserved pilot seed and before the five reported seeds were drawn.

Adapted VinePPO was named the ‡ comparator at that point, which is why Panel B of Table 17 is confirmatory and Panel A is not. Panel A reports references selected during a three-seed exploratory round that preceded the reported runs and used a different reference for each endpoint, so its per-endpoint choice is itself a selection step and its intervals are nominal. Tree-GRPO with process rewards, and later the families of Table 18, were added after the reported seeds existed and are exploratory throughout. No earlier result was recomputed after a reference was added.

Confirmatory family. Table 16 reports five confirmatory contrasts using paired t statistics on four degrees of freedom. The horizon row tests the mean of five seed-specific A4–A1 slopes across six depths. The two rows below the rule report secondary comparisons with nominal intervals.

Table 16. Confirmatory and secondary contrasts with paired five-seed 95% t intervals. Holm adjustment covers the first five rows. Effects are pp, or pp per depth for the slope.

Endpoint	$\widehat{\Delta}_{\text{int}}$	95% interval	Raw p	Holm p
Verified calculation accuracy	+0.53	[-1.02, 2.07]	0.3994	0.3994
Grounded exact match	+3.53	[2.37, 4.70]	0.0011	0.0055
Horizon slope, A4−A1	+1.68	[0.90, 2.46]	0.0039	0.0156
MedAgentBench task success	+5.47	[2.89, 8.04]	0.0042	0.0156
EHRSHOT mean AUPRC	+1.07	[0.20, 1.95]	0.0271	0.0542
MedAgentBench cost-norm. success	+5.09	[2.44, 7.74]	0.0060	NA
EHRSHOT mean AUROC	+1.45	[0.91, 1.99]	0.0018	NA

Baseline contrasts. Table 17 separates the three critic-free reference sets of the original design.

Table 17. Baseline differences (pp) with paired five-seed 95% t intervals. Panels A, B, and C use the earlier selected references, adapted VinePPO, and Tree-GRPO with process rewards. Holm adjustment covers all six endpoints within each panel.

Endpoint	Reference	$\widehat{\Delta}$	95% interval	Raw p	Holm p
A. Exploratory comparisons with previously selected references					
Verified calculation accuracy	VPR	+0.82	[-0.45, 2.09]	0.1466	0.2314
Grounded exact match	Tree-GRPO	+2.48	[1.45, 3.51]	0.0026	0.0156
MedAgentBench task success	Tree-GRPO	+3.20	[-1.23, 7.63]	0.1157	0.2314
MedAgentBench cost-norm. success	VPR	+3.40	[0.44, 6.36]	0.0331	0.0993
EHRSHOT mean AUPRC	VPR	+1.20	[0.65, 1.75]	0.0037	0.0185
EHRSHOT mean AUROC	VPR	+0.72	[0.35, 1.09]	0.0058	0.0232
B. A4 minus adapted VinePPO					
Verified calculation accuracy	VinePPO	+1.21	[-0.09, 2.51]	0.0615	0.1230
Grounded exact match	VinePPO	+2.91	[1.85, 3.97]	0.0016	0.0096
MedAgentBench task success	VinePPO	+3.47	[-1.85, 8.78]	0.1443	0.1443
MedAgentBench cost-norm. success	VinePPO	+4.12	[1.20, 7.04]	0.0172	0.0516
EHRSHOT mean AUPRC	VinePPO	+1.44	[0.88, 2.00]	0.0020	0.0100
EHRSHOT mean AUROC	VinePPO	+0.98	[0.60, 1.36]	0.0021	0.0100
C. A4 minus Tree-GRPO + process					
Verified calculation accuracy	Tree-GRPO+P	+0.45	[-0.77, 1.67]	0.3626	0.6282
Grounded exact match	Tree-GRPO+P	+1.97	[1.09, 2.85]	0.0034	0.0204
MedAgentBench task success	Tree-GRPO+P	+2.13	[-3.02, 7.28]	0.3141	0.6282
MedAgentBench cost-norm. success	Tree-GRPO+P	+2.60	[0.05, 5.15]	0.0471	0.1413
EHRSHOT mean AUPRC	Tree-GRPO+P	+0.78	[0.33, 1.23]	0.0084	0.0336
EHRSHOT mean AUROC	Tree-GRPO+P	+0.62	[0.32, 0.92]	0.0045	0.0225

Extended reference contrasts. Table 18 compares A4 with the strongest member of each family added after the original reference set was frozen, namely adapted LoGo-GRPO among the critic-free memory methods, PPO with a learned process reward model among the critic-based methods, and adapted Fine-Mem among the memory agents. All eighteen differences are positive, and sixteen

survive Holm adjustment within their panel. The two that do not are the calculation and task-success contrasts against adapted LoGo-GRPO. The one endpoint where the intervals are widest is MedAgentBench task success, as in Panels A through C, because its 150 episodes give the noisiest seed-level estimate. Adapted EHRAgent is a fixed system, so it is compared by patient-clustered bootstrap instead. A4 exceeds its matched configuration by $+6.97$ grounded exact-match points $[5.42, 8.51]$ and $+14.44$ cost-normalized points $[10.98, 17.85]$, and exceeds its budget-free native configuration by $+0.61$ grounded exact-match points $[-0.93, 2.16]$ and $+4.67$ task-success points $[-0.84, 10.12]$.

Table 18. Differences from the reference families added after the original set was frozen (pp), with paired five-seed 95% t intervals. Holm adjustment covers the six endpoints of each panel. All three panels are exploratory.

Endpoint	Reference	$\widehat{\Delta}$	95% interval	Raw p	Holm p
D. A4 MINUS ADAPTED LOGO-GRPO					
Verified calculation accuracy	LoGo-GRPO	$+1.39$	[0.20, 2.58]	0.0317	0.0635
Grounded exact match	LoGo-GRPO	$+3.45$	[2.57, 4.33]	0.0004	0.0024
MedAgentBench task success	LoGo-GRPO	$+4.80$	[0.43, 9.17]	0.0381	0.0635
MedAgentBench cost-norm. success	LoGo-GRPO	$+5.66$	[3.05, 8.27]	0.0038	0.0115
EHRSHOT mean AUPRC	LoGo-GRPO	$+1.91$	[1.26, 2.56]	0.0012	0.0048
EHRSHOT mean AUROC	LoGo-GRPO	$+1.71$	[1.16, 2.26]	0.0010	0.0048
E. A4 MINUS PPO WITH A LEARNED PROCESS REWARD MODEL					
Verified calculation accuracy	PPO+PRM	$+1.91$	[0.52, 3.30]	0.0189	0.0189
Grounded exact match	PPO+PRM	$+5.18$	[4.01, 6.35]	0.0002	0.0015
MedAgentBench task success	PPO+PRM	$+7.07$	[3.12, 11.02]	0.0076	0.0153
MedAgentBench cost-norm. success	PPO+PRM	$+7.94$	[4.91, 10.97]	0.0019	0.0057
EHRSHOT mean AUPRC	PPO+PRM	$+2.42$	[1.70, 3.14]	0.0007	0.0029
EHRSHOT mean AUROC	PPO+PRM	$+2.53$	[1.95, 3.11]	0.0003	0.0015
F. A4 MINUS ADAPTED FINE-MEM					
Verified calculation accuracy	Fine-Mem	$+2.32$	[1.02, 3.62]	0.0078	0.0115
Grounded exact match	Fine-Mem	$+4.59$	[3.56, 5.62]	0.0002	0.0012
MedAgentBench task success	Fine-Mem	$+7.34$	[3.55, 11.13]	0.0058	0.0115
MedAgentBench cost-norm. success	Fine-Mem	$+8.17$	[5.30, 11.04]	0.0014	0.0041
EHRSHOT mean AUPRC	Fine-Mem	$+2.54$	[1.92, 3.16]	0.0003	0.0014
EHRSHOT mean AUROC	Fine-Mem	$+2.69$	[2.13, 3.25]	0.0002	0.0011

C.2 DESIGN ABLATIONS

Protocol. Each design ablation changes one element of A4 and uses three matched seeds at 2^{23} sampled tokens under the fractional curriculum. The reference uses the same reduced budget and seeds. Evaluation covers source tasks and MedAgentBench-Query. Figure 7 reports within-seed differences and paired t intervals. These exploratory comparisons are separate from the full-budget factorial. The 48 runs are included in the compute ledger.

Row definitions. Predicate rows remove one verifier family and recompute rewards over the reduced catalog. The uniform row gives every legal action process reward $+1$. Estimator rows remove one advantage term. The masking row also trains shared prefixes and unmatched branch continuations on local credit. Pool removal leaves unmatched base turns with zero local advantage. Group-size rows vary K or G_0 at fixed generated tokens, and the memory row halves L. Other settings remain fixed.

What the screen shows. Uniform positive process rewards perform worse than removing any single tested verifier family. Removing either advantage term produces larger losses than the individual predicate removals. Predicate effects vary across endpoints, with grounding and missingness more relevant to retrieval and stopping checks more relevant to MedAgentBench success (Figure 7).

Sibling count and masking. The $K = 7$ versus $K = 3$ intervals span zero on all endpoints. Training shared prefixes and unmatched continuations on local credit produces small mean changes,

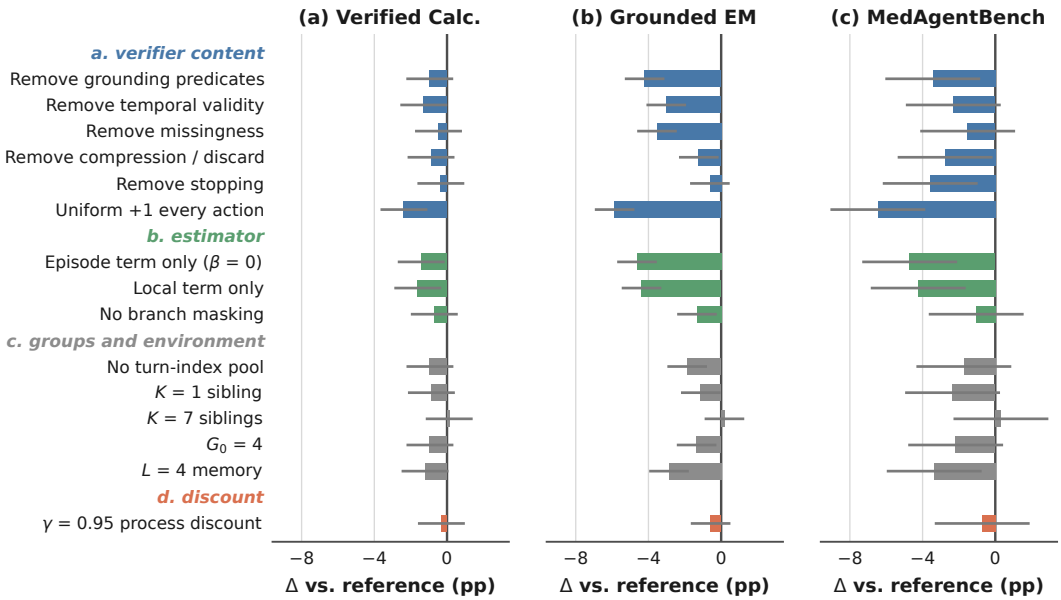


Figure 7. Changes from the reduced-budget VAPA reference (pp). Each variant uses three matched seeds and 2^{23} tokens. Whiskers are paired 95% t intervals.

with two intervals spanning zero. These estimates use three matched seeds and a quarter of the main training budget (Sullivan & Koller, 2026).

Pool removal. Pool removal lowers the endpoint means, with intervals spanning zero for calculation and MedAgentBench. The removed pool combines an episode-style outcome contrast with process and cost suffix information, so this ablation estimates their joint contribution at the reduced budget. The memory row halves the main capacity to $L = 4$ as a single tight-memory contrast, and Appendix C.3 runs the sweep at full budget.

Discount. A fifteenth variant, row d of Figure 7, uses $\gamma = 0.95$ at the same reduced budget and three seeds. Changes are -0.31 points with interval $[-1.60, 0.98]$ for calculation, -0.58 with $[-1.67, 0.51]$ for retrieval, and -0.72 with $[-3.33, 1.89]$ for MedAgentBench success. All three intervals span zero. Discounting reweights future process rewards. The main configuration uses $\gamma = 1$.

C.3 Memory Capacity

Protocol. The bounded memory is the constraint the method is designed for, so we vary it at full budget rather than in the reduced-budget screen. We repeat A1 and A4 at $L \in \{2, 4, 12, 16\}$ with five matched seeds, the shared SFT checkpoint, and 2^{25} sampled tokens, holding $C_{\max} = 12$, $T_{\max} = 16$, reward weights, β, and replay settings at their selected values. The $L = 8$ column reproduces the main runs. Task construction is unchanged, so an episode's evidence requirement is fixed and only the number of items the agent may retain varies. The 40 additional runs are in Table 15.

Results. Table 19 and Figure 8 give the sweep. Both arms improve monotonically with capacity and both flatten above $L = 12$, while the A4 margin moves in the opposite direction on all three endpoints and stays positive with intervals excluding zero at every capacity. Slopes come from an ordinary least-squares fit of the five per-capacity margins on $\log_2 L$, with intervals from the residual standard error on three degrees of freedom.

Reading the trend. Table 19, panel D, gives the mechanism behind the declining part of the margin. Peak occupancy rises with capacity but plateaus well below the largest bound, so at the wide settings the bound rarely binds and forced eviction, the decision the local advantage is built to score, nearly disappears. Repeat retrievals follow the same curve, and at every capacity A4 evicts and re-retrieves less than A1 at equal budget, which is the behavior the eviction-sensitive checks

reward. The margin declines far more slowly than the eviction rate, however. Evictions fall by two orders of magnitude while the MedAgentBench margin falls by about a third and remains clearly positive at the widest setting. The memory bound therefore accounts for part of the advantage and not for most of it. The residual is consistent with the horizon result of Section 4.4, where retrieval, calculation, and stopping decisions constrain later ones through the action budget rather than through memory, and those decisions are scored at every capacity.

Table 19. Memory-capacity sweep at full budget (%). Entries are five-seed means with SDs, and Δ is the paired A4−A1 margin with its 95% t interval. Slopes are points per doubling of L. The $L = 8$ column reproduces Tables 1 and 2.

	$L = 2$	$L = 4$	$L = 8$	$L = 12$	$L = 16$	Slope
A. VERIFIED CALCULATION ACCURACY						
A1	$77.21_{\pm 1.70}$	$78.46_{\pm 1.62}$	$79.49_{\pm 1.54}$	$80.18_{\pm 1.48}$	$80.44_{\pm 1.45}$	+1.09
A4	$80.94_{\pm 1.52}$	$81.83_{\pm 1.41}$	$82.63_{\pm 1.31}$	$82.99_{\pm 1.27}$	$83.12_{\pm 1.25}$	+0.74
Δ	+3.73	+3.37	+3.14	+2.81	+2.68	−0.35
95% CI	[2.70, 4.76]	[2.40, 4.34]	[2.25, 4.03]	[1.94, 3.68]	[1.82, 3.54]	[-0.48, -0.22]
B. GROUNDED EXACT MATCH						
A1	$64.85_{\pm 0.82}$	$67.62_{\pm 0.74}$	$70.21_{\pm 0.68}$	$71.84_{\pm 0.66}$	$72.40_{\pm 0.65}$	+2.57
A4	$74.11_{\pm 1.06}$	$76.03_{\pm 1.01}$	$78.03_{\pm 0.94}$	$78.92_{\pm 0.90}$	$79.15_{\pm 0.88}$	+1.75
Δ	+9.26	+8.41	+7.82	+7.08	+6.75	−0.83
95% CI	[8.75, 9.77]	[7.98, 8.84]	[7.47, 8.17]	[6.70, 7.46]	[6.34, 7.16]	[-0.99, -0.66]
C. MEDAGENTBENCH TASK SUCCESS						
A1	$24.53_{\pm 4.02}$	$28.80_{\pm 3.88}$	$32.53_{\pm 3.93}$	$34.67_{\pm 3.61}$	$35.47_{\pm 3.55}$	+3.71
A4	$36.80_{\pm 3.72}$	$40.53_{\pm 3.55}$	$42.67_{\pm 3.46}$	$43.47_{\pm 3.33}$	$43.73_{\pm 3.29}$	+2.30
Δ	+12.27	+11.73	+10.13	+8.80	+8.26	−1.41
95% CI	[9.86, 14.68]	[9.48, 13.98]	[8.17, 12.11]	[6.91, 10.69]	[6.43, 10.09]	[-1.97, -0.85]
D. REALIZED MEMORY BEHAVIOR FOR EACH EPISODE (A1 / A4)						
Peak items held	2.00 / 1.97	3.81 / 3.62	5.94 / 5.31	6.71 / 5.94	6.90 / 6.08	–
Evictions	3.88 / 3.41	1.95 / 1.52	0.62 / 0.34	0.14 / 0.06	0.04 / 0.01	–
Repeat retrievals	2.41 / 1.86	1.37 / 0.94	0.58 / 0.31	0.27 / 0.12	0.19 / 0.09	–

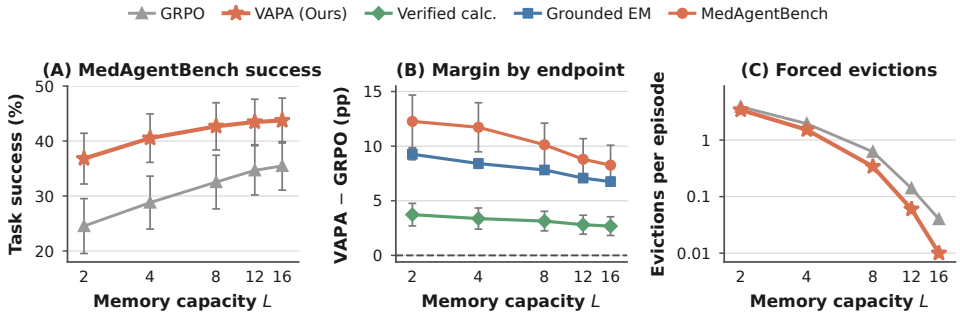


Figure 8. Memory capacity at full budget. (A) MedAgentBench task success, with five-seed 95% intervals. (B) The A4−A1 margin on three endpoints, with paired intervals. (C) Forced evictions for each episode, log scale.

C.4 Task-Level Results

Reporting. Task-level results disaggregate the factorial without additional significance tests, with uncertainty defined in Appendix C.1.

Source tasks. Figure 9 reports gains over A1 and factorial interactions for 29 calculators and five retrieval families, using 120 episodes per calculator and 600 per retrieval family. Standard calculators and operationalized proxies are displayed separately. Averaged over the sixteen proxies the four arms reach 60.18, 62.54, 63.62, and 70.83, against the standard-calculator column of Table 2. The artifact contains per-task levels and seed records.

	A2−A1	A3−A1	A4−A1	$\widehat{\Delta}_{\text{int}}$
Standard calculators (13)				
Delta gap	+0.83	+3.50	+4.83	+0.50
CKD-EPI GFR	+2.50	+1.33	+4.50	+0.67
Free water deficit	+1.50	+2.17	+4.50	+0.83
Corrected calcium	+1.33	+2.17	+4.17	+0.67
Maintenance fluids	+1.17	+2.17	+4.00	+0.67
Body mass index	+0.33	+2.50	+3.00	+0.17
Body surface area	+1.17	+0.83	+2.67	+0.67
Anion gap	+0.50	+1.50	+2.50	+0.50
Corrected sodium (Hillier)	+0.83	+1.33	+2.50	+0.33
Creatinine clearance (C-G)	+1.83	+0.33	+2.33	+0.17
Fractional excretion of Na	+1.50	+0.50	+2.17	+0.17
Mean arterial pressure	+1.17	-0.17	+1.83	+0.83
Serum osmolality	-0.33	+1.50	+1.83	+0.67
Operationalized proxies (16)				
PERC rule	+3.83	+4.00	+16.33	+8.50
Caprini	+2.00	+3.00	+12.67	+7.67
HEART	+2.50	+3.00	+12.67	+7.17
SOFA	+2.50	+3.67	+11.83	+5.67
Glasgow coma scale	+1.83	+4.00	+11.50	+5.67
Wells DVT	+5.00	+4.00	+11.50	+2.50
MELD-Na	+2.00	+2.00	+11.33	+7.33
Padua prediction	+2.67	+4.17	+10.67	+3.83
CURB-65	+1.33	+4.33	+10.67	+5.00
Wells PE	+3.67	+2.83	+10.33	+3.83
CHA2DS2-VASc	+2.50	+4.33	+10.00	+3.17
HAS-BLED	+1.83	+2.50	+9.33	+5.00
Revised cardiac risk index	+1.50	+3.17	+9.17	+4.50
Child-Pugh	+1.17	+3.33	+7.83	+3.33
APACHE II	+1.83	+3.67	+7.33	+1.83
qSOFA	+1.67	+3.17	+7.33	+2.50
Retrieval families (5)				
Medication activity at time	+3.13	+3.57	+12.20	+5.50
Procedure within a window	+1.83	+1.63	+8.57	+5.10
Recorded vs. missing status	+1.83	+2.90	+6.27	+1.53
First pre-cutoff diagnosis	+1.43	+3.13	+6.20	+1.63
Latest laboratory value	+0.43	+1.53	+5.87	+3.90

Figure 9. Per-task joint gains (A4–A1) and factorial interactions (pp). Tasks are sorted within suite by the joint gain.

Calculators and proxies. Joint gains and interactions are larger on proxies, whose task definitions and input-composition requirements add sources of difficulty. Standard-calculator accuracy is the primary endpoint.

MedAgentBench-Query. Table 20 divides the read-only tasks by categories fixed before evaluation. Information-retrieval tasks require one reference GET, while data-aggregation tasks combine values from multiple retrievals. The overall row weights every instance equally. The interaction is larger on data aggregation for both task success and cost-normalized success.

EHRSHOT transfer. Table 21 gives task-level results on the retained cohorts in Table 3. A4 leads the agent macro means but has lower ICU-transfer AUPRC than the other arms. With target-label training, CLMBR-T-base exceeds the frozen A4 policy on six of eight tasks in each panel. Table 22 reports the unmatched clinical-reference means.

Table 20. MedAgentBench-Query success and cost-normalized success by task category (%). Entries show five-seed means and standard deviations under the shared inference budget.

Category	Instances	A1	A2	A3	A4
A. TASK SUCCESS					
Information retrieval	96	$34.58_{\pm 2.98}$	$37.71_{\pm 4.85}$	$36.88_{\pm 4.39}$	$44.79_{\pm 2.85}$
Data aggregation	54	$28.89_{\pm 6.09}$	$30.37_{\pm 4.26}$	$30.74_{\pm 4.65}$	$38.89_{\pm 5.40}$
Overall	150	$32.53_{\pm 3.93}$	$35.07_{\pm 3.90}$	$34.67_{\pm 3.80}$	$42.67_{\pm 3.46}$
B. COST-NORMALIZED SUCCESS					
Information retrieval	96	$27.26_{\pm 2.20}$	$29.77_{\pm 2.69}$	$31.56_{\pm 2.44}$	$38.47_{\pm 2.86}$
Data aggregation	54	$20.74_{\pm 5.60}$	$22.84_{\pm 4.70}$	$23.40_{\pm 4.30}$	$31.82_{\pm 5.35}$
Overall	150	$24.91_{\pm 2.98}$	$27.28_{\pm 2.62}$	$28.62_{\pm 2.46}$	$36.08_{\pm 3.02}$

Table 21. Task-level EHRSHOT-derived transfer results (%). Agent entries show seed standard deviations. CLMBR-T-base uses target-task training and shows patient-bootstrap standard errors.

Task	CLMBR-T-base	A1	A2	A3	A4
A. AUPRC					
Operational					
Long length of stay	$58.57_{\pm 2.10}$	$42.38_{\pm 1.08}$	$43.56_{\pm 0.73}$	$44.60_{\pm 0.85}$	$49.70_{\pm 0.89}$
30-day readmission	$37.76_{\pm 3.11}$	$22.95_{\pm 0.30}$	$25.86_{\pm 0.59}$	$25.23_{\pm 0.16}$	$27.97_{\pm 0.77}$
ICU transfer	$25.63_{\pm 4.58}$	$9.17_{\pm 0.82}$	$13.45_{\pm 0.48}$	$9.28_{\pm 0.64}$	$8.80_{\pm 1.02}$
Laboratory					
Anemia	$96.40_{\pm 0.28}$	$82.48_{\pm 0.47}$	$82.98_{\pm 0.58}$	$82.66_{\pm 0.63}$	$84.17_{\pm 0.31}$
Hyponatremia	$53.53_{\pm 2.02}$	$48.66_{\pm 0.59}$	$49.69_{\pm 1.12}$	$46.66_{\pm 0.97}$	$51.28_{\pm 0.85}$
Thrombocytopenia	$74.73_{\pm 1.52}$	$53.91_{\pm 0.84}$	$56.31_{\pm 0.97}$	$55.22_{\pm 0.94}$	$59.23_{\pm 0.80}$
Diagnosis					
Hyperlipidemia	$22.83_{\pm 3.04}$	$24.96_{\pm 1.21}$	$25.03_{\pm 1.56}$	$26.65_{\pm 1.62}$	$26.97_{\pm 0.54}$
Hypertension	$25.87_{\pm 3.75}$	$23.75_{\pm 0.54}$	$23.99_{\pm 0.26}$	$23.39_{\pm 0.79}$	$26.77_{\pm 0.91}$
Macro mean	$49.41_{\pm 0.90}$	$38.53_{\pm 0.36}$	$40.11_{\pm 0.50}$	$39.21_{\pm 0.57}$	$41.86_{\pm 0.33}$
B. AUROC					
Operational					
Long length of stay	$79.53_{\pm 1.08}$	$69.53_{\pm 0.70}$	$70.90_{\pm 0.58}$	$70.82_{\pm 0.61}$	$74.02_{\pm 0.33}$
30-day readmission	$78.87_{\pm 1.45}$	$67.68_{\pm 0.79}$	$69.14_{\pm 0.82}$	$69.09_{\pm 0.36}$	$72.28_{\pm 0.49}$
ICU transfer	$83.39_{\pm 2.16}$	$64.74_{\pm 0.75}$	$66.01_{\pm 1.24}$	$66.20_{\pm 1.27}$	$69.00_{\pm 1.29}$
Laboratory					
Anemia	$93.27_{\pm 0.45}$	$70.81_{\pm 0.75}$	$71.68_{\pm 0.73}$	$71.21_{\pm 0.84}$	$73.37_{\pm 0.39}$
Hyponatremia	$73.66_{\pm 1.13}$	$69.16_{\pm 0.63}$	$70.31_{\pm 0.86}$	$69.86_{\pm 0.73}$	$72.28_{\pm 0.73}$
Thrombocytopenia	$83.96_{\pm 0.90}$	$69.71_{\pm 0.72}$	$70.70_{\pm 0.33}$	$70.24_{\pm 0.59}$	$72.71_{\pm 0.63}$
Diagnosis					
Hyperlipidemia	$66.07_{\pm 2.29}$	$68.05_{\pm 0.73}$	$69.14_{\pm 0.60}$	$68.89_{\pm 0.70}$	$71.19_{\pm 0.60}$
Hypertension	$69.66_{\pm 2.30}$	$67.68_{\pm 0.85}$	$68.36_{\pm 0.42}$	$67.93_{\pm 0.45}$	$69.87_{\pm 0.72}$
Macro mean	$78.55_{\pm 0.52}$	$68.42_{\pm 0.39}$	$69.53_{\pm 0.27}$	$69.28_{\pm 0.46}$	$71.84_{\pm 0.34}$

Imputation sensitivity. We record 5,472 format failures across systems, with failure rates ranging from zero for supervised references to 2.49% for self-consistency. The primary and valid-output metrics differ by at most 0.18 AUPRC points and 0.31 AUROC points. Agent-arm ordering is reported unchanged across the four imputation placements. The artifact specifies those placements, per-system failure records, failure-label composition, and all resulting metric levels.

C.5 Target-Trained Clinical References

Purpose and protocol. These target-trained predictors show what task-specific EHRSHOT labels provide and are not matched RL baselines. CLMBR-T-base uses its pretrained coded-event encoder (Wornow et al., 2023). GraphCare retains personalized knowledge graphs and bi-attention prediction (Jiang et al., 2024), while KARE uses community retrieval with separate reasoning and label objectives under Qwen3.5-9B (Jiang et al., 2025a). All use the same tasks, binary labels, cutoffs, and

retained test instances as Table 21. Training-only patients determine vocabularies, graph statistics, exemplar indexes, and validation-selected settings. Reporting uses five fresh seeds for GraphCare and KARE, while CLMBR follows the fixed-system convention of Appendix C.1.

Label-budget controls. CLMBR-T-base is refit at $k \in \{8, 16, 32, 128\}$ labeled instances per task using five fixed, patient-disjoint draws that preserve prevalence. The frozen agent consumes no EHRSHOT labels. The supervision-matched VAPA variant keeps the policy, decoding, and interface frozen and fits one ℓ_2-regularized logistic head per task to its terminal answer probability, retained-memory concepts, recency and status flags, cited pointers, and action cost. The head sees no field the agent did not retrieve. Validation selects the regularization strength from $\{10^{-3}, 10^{-2}, 10^{-1}, 1\}$, and five patient-level resamples provide dispersion.

Model selection. Target validation maximizes the mean of AUPRC and AUROC. The learning-rate grids are $\{10^{-5}, 10^{-4}, 3\times10^{-4}, 10^{-3}\}$ for GraphCare and $\{10^{-6}, 5\times10^{-6}, 10^{-5}\}$ for KARE, with the published GraphCare rate included. KARE retrieves one positive and one negative training example by pre-cutoff similarity while excluding the query patient. Missing scores follow Appendix A.1.

Reading the comparison. Full-supervision CLMBR leads the frozen agent by 7.55 AUPRC and 6.71 AUROC points. Without target labels, VAPA lies between the 16- and 32-label CLMBR rows on AUPRC and between the 8- and 16-label rows on AUROC. Fitting a task head to the agent's own evidence recovers about two thirds of the full-supervision gap and exceeds both knowledge-graph references on AUPRC. These results locate the frozen policy on the supervision axis without treating target-trained predictors as matched baselines.

Table 22. Target-trained clinical references on eight EHRSHOT-derived tasks (%). These are not matched RL baselines. The table states the EHRSHOT labels each system consumes and ends with the label-free frozen agent. Subscripts are seed SDs, except patient-bootstrap SEs for supervised CLMBR and label-draw SDs for its few-shot rows.

System	EHRSHOT labels	Evaluation mode	AUPRC	AUROC
CLMBR-T-base (2023)	Full partition	Predictor	$\mathbf{49.41}_{\pm 0.90}$	$\mathbf{78.55}_{\pm 0.52}$
at $k = 128$ for each task	128 per task	Predictor	$46.18_{\pm 1.04}$	$76.37_{\pm 0.66}$
at $k = 32$ for each task	32 per task	Predictor	$43.09_{\pm 1.37}$	$74.18_{\pm 0.88}$
at $k = 16$ for each task	16 per task	Predictor	$41.32_{\pm 1.62}$	$72.63_{\pm 1.05}$
at $k = 8$ for each task	8 per task	Predictor	$38.94_{\pm 1.95}$	$70.31_{\pm 1.34}$
Adapted GraphCare (2024)	Full partition	Predictor	$43.63_{\pm 1.08}$	$74.77_{\pm 0.93}$
Adapted KARE (2025a)	Full partition	Predictor	$45.65_{\pm 1.32}$	$76.30_{\pm 1.16}$
VAPA + fitted head	Full partition	Frozen agent, fitted head	$46.93_{\pm 0.78}$	$76.42_{\pm 0.63}$
VAPA	None	Frozen agent	$41.86_{\pm 0.33}$	$71.84_{\pm 0.34}$

Horizon results. Table 23 expands Figure 4 with cost-normalized success. Depths use separate task strata, with methods paired within episodes. The A4–A1 success slope belongs to the confirmatory family. Other slopes and interactions are descriptive, and cost-normalized margins compress on deeper chains.

Table 23. Task success and cost-normalized success by decision depth (%). Entries show five-seed means and standard deviations. Slope units are pp per depth.

Depth	A1	A2	A3	A4	A4−A1	$\widehat{\Delta}_{\text{int}}$
A. TASK SUCCESS						
1	$75.44_{\pm 2.78}$	$77.12_{\pm 3.27}$	$79.44_{\pm 3.24}$	$81.04_{\pm 1.95}$	$+5.60_{\pm 2.68}$	$-0.08_{\pm 4.22}$
2	$62.88_{\pm 4.28}$	$63.28_{\pm 1.66}$	$67.36_{\pm 2.84}$	$68.72_{\pm 2.76}$	$+5.84_{\pm 4.35}$	$+0.96_{\pm 5.46}$
3	$48.40_{\pm 2.87}$	$49.28_{\pm 3.33}$	$54.00_{\pm 3.16}$	$59.68_{\pm 5.74}$	$+11.28_{\pm 4.61}$	$+4.80_{\pm 9.65}$
4	$40.16_{\pm 2.60}$	$42.32_{\pm 2.09}$	$44.56_{\pm 5.49}$	$52.80_{\pm 1.36}$	$+12.64_{\pm 3.23}$	$+6.08_{\pm 4.61}$
5	$32.72_{\pm 4.48}$	$35.44_{\pm 2.88}$	$38.80_{\pm 2.24}$	$45.20_{\pm 4.30}$	$+12.48_{\pm 4.17}$	$+3.68_{\pm 8.41}$
6	$24.00_{\pm 1.98}$	$26.32_{\pm 1.51}$	$28.96_{\pm 2.63}$	$37.12_{\pm 1.95}$	$+13.12_{\pm 2.92}$	$+5.84_{\pm 4.69}$
Slope	$-10.17_{\pm 0.81}$	$-9.84_{\pm 0.25}$	$-9.93_{\pm 0.31}$	$-8.49_{\pm 0.40}$	$+1.68_{\pm 0.63}{}^{\dagger}$	$+1.12_{\pm 1.28}$
B. COST-NORMALIZED SUCCESS						
1	$59.73_{\pm 2.36}$	$61.13_{\pm 2.56}$	$62.81_{\pm 2.69}$	$66.21_{\pm 1.60}$	$+6.48_{\pm 2.24}$	$+2.01_{\pm 3.59}$
2	$42.52_{\pm 3.07}$	$43.31_{\pm 1.38}$	$46.07_{\pm 1.97}$	$49.06_{\pm 2.09}$	$+6.55_{\pm 3.31}$	$+2.20_{\pm 3.80}$
3	$27.38_{\pm 1.83}$	$28.40_{\pm 2.01}$	$31.11_{\pm 1.77}$	$36.91_{\pm 3.62}$	$+9.53_{\pm 2.88}$	$+4.78_{\pm 6.05}$
4	$18.15_{\pm 1.12}$	$19.69_{\pm 1.01}$	$20.91_{\pm 2.63}$	$27.46_{\pm 0.74}$	$+9.31_{\pm 1.42}$	$+5.01_{\pm 2.53}$
5	$11.09_{\pm 1.53}$	$12.60_{\pm 1.08}$	$14.13_{\pm 0.92}$	$19.17_{\pm 1.89}$	$+8.08_{\pm 1.70}$	$+3.53_{\pm 3.24}$
6	$5.49_{\pm 0.54}$	$6.57_{\pm 0.58}$	$7.17_{\pm 0.84}$	$12.05_{\pm 0.58}$	$+6.56_{\pm 0.81}$	$+3.80_{\pm 1.34}$
Slope	$-10.71_{\pm 0.53}$	$-10.68_{\pm 0.34}$	$-10.98_{\pm 0.35}$	$-10.57_{\pm 0.32}$	$+0.14_{\pm 0.45}$	$+0.38_{\pm 0.75}$

Backbone replication. We repeat the A1 versus A4 contrast with gpt-oss-20b using three matched seeds, its SFT checkpoint, $G_0 = 8$, and $B = 2^{25}$, transferring reward and replay settings without retuning. VAPA leads GRPO by $+2.91$, $+6.98$, $+9.33$, $+10.63$, $+3.35$, and $+3.12$ points across the six Table 1 endpoints. Tokenizer, reasoning, and compute details follow Appendices B.2 and B.5.

D Credit and Optimization

D.1 Credit Agreement

Probe construction. Before evaluating trained arms, we freeze 1,200 patient-disjoint observations from shared-SFT source trajectories, with one per patient and equal allocation across source-suite and spent-budget strata (Appendix B.2). At each observation, the frozen SFT policy samples one base and three sibling actions using the RL sampler. Duplicate canonical actions remain in the group but are excluded from pairwise agreement. One forced-action continuation scores each candidate, and $K_{\text{ref}} = 8$ independent continuations estimate its reference value. Candidates and suffix seeds are shared across reward definitions and credit signals, with disjoint credit and reference suffixes. Replay scores candidates as an exact-fork group. No-fork candidates enter separately, preventing probe siblings from becoming collision peers while retaining canonical-collision and turn-index rules. Shuffling averages all $4! = 24$ score permutations. With half credit for ties, every retained pair has exactly 50% expected agreement, so this reference has no confidence interval. Local advantages are evaluated before multiplication by β. For observation o, forced action a, and continuation index k, let $\tau_{o,a,k}$ be the sampled finite-horizon continuation and let $C(\tau_{o,a,k})$ be its total action cost. The reference estimate is

$$\widehat{Q}_{\text{ref}}(o,a) = \frac{1}{K_{\text{ref}}} \sum_{k=1}^{K_{\text{ref}}} \Big[R_{\text{out}}(\tau_{o,a,k}) - \lambda_c C(\tau_{o,a,k}) \Big],$$

where $\widehat{Q}_{\text{ref}}(o,a)$ averages the K_{ref} outcome-and-cost continuation returns. Action a is forced at o and the remainder is sampled from the frozen SFT policy. The reference excludes process rewards but keeps the shared action-cost term. Suffixes never enter training, grouping, checkpoint selection, or candidate-credit computation.

Paired inference. For credit method m, let $c_m(o, a)$ be its scalar credit score for action a at observation o. For an unordered pair of distinct actions, Δc_m and ΔQ_{ref} are the first-minus-second differences in method credit and reference value. Four candidates give six pairs per state and 7,200 across the 1,200 states. We exclude 964 pairs with identical canonical actions and 586 with zero reference difference, leaving 5,650. Each retained pair receives agreement score ℓ_m,

$$\ell_m = \begin{cases} 1, & \Delta c_m \Delta Q_{\text{ref}} > 0, \\ \frac{1}{2}, & \Delta c_m = 0, \\ 0, & \Delta c_m \Delta Q_{\text{ref}} < 0. \end{cases}$$

We average ℓ_m within state, giving patients equal weight. Figure 10 reports the number of states retaining at least one pair. Of the 1,200 states, 281 lose no pairs, 417 lose one, 373 lose two, and 129 lose three. Only the strict reference-margin analysis excludes entire states. Figure 3B shows marginal agreement, and Figure 10 shows paired replay-minus-no-fork differences using the same candidates and suffixes. The primary difference is $+10.30$ pp with interval $[8.97, 11.63]$ for outcome-only reward and $+12.90$ with $[11.68, 14.16]$ for outcome and process. All sensitivity intervals exclude zero on the positive side. With common observations and weights, paired estimates equal differences of marginal agreements. Patient-clustered bootstrap resampling recomputes each paired difference.

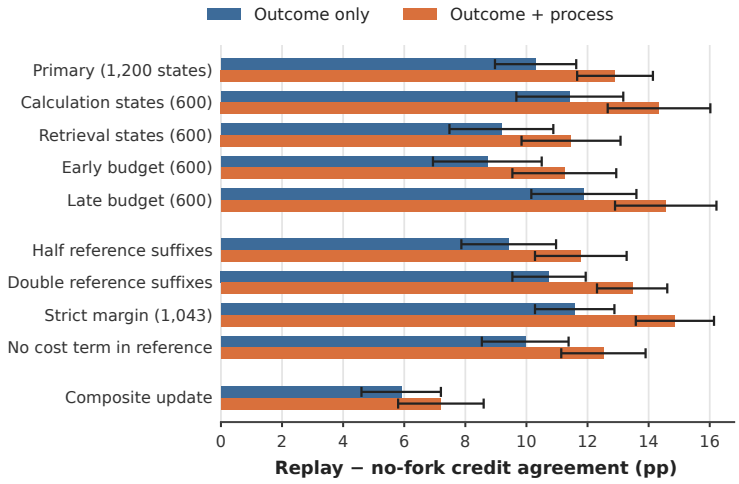


Figure 10. Replay-minus-no-fork credit agreement (pp), with patient-bootstrap 95% intervals. Composite rows use the asymmetric base-versus-branch credit assignment described in the text.

Learned value baseline. The probe also scores the PPO control's fitted value head, using the same observations, candidates, and reference suffixes. On states sampled from the frozen SFT policy, the head reaches held-out explained variance 0.31 $[0.24, 0.38]$ at the end of training. Because the head is trained without process rewards, the matched comparison is with the outcome-only signals. Its induced action ordering agrees with the forced-action reference on 58.7% $[57.6, 59.8]$ of retained pairs, against 57.1% for the outcome-only no-fork signal and 67.4% for outcome-only replay, with 50% as chance. A learned critic at this budget therefore supplies a baseline no better than the group mean it would replace and clearly worse than replay, which is what the matched comparison in Table 1 shows at the endpoint level.

Sensitivity and scope. Sensitivity analyses split source suite and spent-budget stratum, vary K_{ref}, require a reference margin above one Monte Carlo standard error, and omit action cost. Independent candidates and suffixes at fixed anchors estimate outcome-and-cost values under the frozen SFT policy when $\lambda_p = 0$. Process rewards add a separate action-quality target. Training uses trajectory-selected anchors and broader fallback groups, as specified in Appendix B.2.

Composite update. The composite probe adds normalized episode credit to the base member. Its replay-minus-no-fork agreement is smaller than the local-only difference, while sibling-only pairs are unchanged. The comparison measures action ordering at the fixed probe states.

D.2 Optimization Diagnostics

Normalization and batching. Episode normalization uses the G_0 base outcomes for one instance. Local normalization pools action-level advantages from complete nonsingleton groups across all tiers and instances in the accumulated batch. Singletons receive zero local credit and are excluded from this pool. Groups and scales are fixed before the optimizer step. The numerical specification uses population standard deviations for both advantage scales and $\epsilon = 10^{-8}$ in FP32. An empty local pool contributes zero local credit.

Loss and masking. Let $\mathcal{I}$ contain eligible policy-token indices in an accumulated batch. Token x_n has preceding token history h_n and receives its action's assigned coefficient $\widetilde{A}_n$ from Equation 7. Shared branch prefixes, unmatched continuation turns, and malformed or retried generations are excluded. All generated tokens remain budgeted. Token averaging weights longer actions more heavily. Returns, groups, scales, and coefficients stay fixed during differentiation, and empty eligible sets produce no update.

Optimizer specification. Starting at the sampling policy θ_{old}, take one AdamW step on actor parameters θ to minimize $\mathcal{L}_{\text{PG}} + \eta_{\text{KL}}\mathcal{L}_{\text{KL}}$. Here $\mathcal{L}_{\text{PG}}$ is the policy-gradient loss, $\mathcal{L}_{\text{KL}}$ is the forward-KL penalty, and $\eta_{\text{KL}} = 0.01$ is its weight,

$$\begin{aligned} \mathcal{L}_{\text{PG}} &= -\frac{1}{|\mathcal{I}|}\sum_{n\in\mathcal{I}} \text{stopgrad}(\widetilde{A}_n)\log \pi_\theta(x_n \mid h_n), \\ \mathcal{L}_{\text{KL}} &= \frac{1}{|\mathcal{I}|}\sum_{n\in\mathcal{I}}\sum_{v\in\mathcal{V}(h_n)} \pi_\theta(v \mid h_n)\log\frac{\pi_\theta(v \mid h_n)}{\pi_{\text{ref}}(v \mid h_n)}. \end{aligned} \tag{9}$$

Here stopgrad treats its argument as constant during differentiation, $\mathcal{V}(h_n)$ is the legal next-token vocabulary, v indexes a token in that vocabulary, and π_{ref} is the fixed SFT reference policy. Actor and reference probabilities use the same grammar mask and renormalization. The forward KL sums over that vocabulary at sampled eligible histories and is evaluated in token chunks. The AdamW first- and second-moment coefficients are $(\beta_1, \beta_2) = (0.9, 0.999)$ and its numerical epsilon is 10^{-8}, together with the learning rate, weight decay, and clipping in Appendix B.2.

Fork-group asymmetry. Local action-level coefficients sum to zero within a complete group. Adding episode credit to the base member changes this sum. Each coefficient multiplies its own token gradients, so the resulting update also depends on action lengths and histories. The episode term can reinforce or oppose local credit. Appendix D.1 evaluates the composite ordering.

Proposition 1 (Fork-group placement invariance at $\gamma = 1$)**.** *Fix a sampled fork group anchored at turn t, with fixed trajectories, terminal rewards, and costs, and tree-consistent scores on its shared prefix. At $\gamma = 1$, every raw fork-group advantage in Equation 5 is unchanged by any redistribution, within each branch, of process rewards among turns $k \geq t$ that preserves that branch's suffix total. It is also unchanged when rewards before t are moved into the suffixes by the same total for every group member.*

Proof of Proposition 1. At $\gamma = 1$, Equation 4 depends on process rewards only through each member's suffix sum. Preserving these sums leaves all returns and local advantages unchanged. If every return instead shifts by the same d, each leave-one-out mean also shifts by d, which cancels in Equation 5. Both identities hold for any fixed grouping tier. Tree-consistent prefix scoring makes transferring the same prefix total into every fork-group suffix induce this common shift. □

Scope conditions. Fixed grouping and suffix totals suffice for the raw local-credit identity. Shared prefixes are needed for the common prefix-to-suffix shift, and normalized credit also requires an unchanged batch scale. Moving rewards across a suffix boundary can change its total. At $\gamma < 1$, later rewards receive less weight.

Discounted relocation diagnostic. Logged groups are rescored at diagnostic discount $\gamma_{\text{diag}} = 0.95$ after moving post-anchor process rewards to the terminal turn while preserving suffix totals. Table 24 gives ordering preservation, Kendall's rank correlation τ, and $\Delta\widehat{A}_{\text{loc}}$, the mean absolute change in normalized local advantage. No model update is performed. Displacement is larger in the early-budget stratum.

Fork-group census. The worked four-distinct-action example selects the base-member local advantage closest to the subset median, breaking ties by episode identifier and policy seed.

Table 24. Suffix-preserving terminal relocation in the reported groups. Brackets give 95% intervals for discounted diagnostics. The $\gamma = 1$ row follows algebraically, with Kendall's $\tau = 1$ wherever the statistic is defined.

Stratum	Groups	Order preserved	Kendall τ	$\Delta\widehat{A}_{\mathrm{loc}}$
$\gamma = 1$, all groups	19,160	1.000	1.000	0.0000
$\gamma_{\mathrm{diag}} = 0.95$, all groups	19,160	0.949 [0.933, 0.965]	0.981 [0.974, 0.988]	0.0058 [0.0055, 0.0061]
$\mathcal{B}_{\mathrm{early}}$	9,580	0.942 [0.917, 0.965]	0.978 [0.968, 0.987]	0.0090 [0.0085, 0.0096]
$\mathcal{B}_{\mathrm{late}}$	9,580	0.956 [0.933, 0.977]	0.984 [0.974, 0.992]	0.0025 [0.0022, 0.0027]

E Behavior and Robustness

E.1 Reward-Channel Audits

Common audit protocol. Two clinically qualified annotators independently inspect cutoff-valid records, traces, memory transitions, and evidence chains under a frozen codebook, followed by blinded adjudication. Arm, seed, rewards, advantages, verifier outputs, and outcomes are hidden. Six hours of training and a 150-item calibration round precede annotation. Balanced audits require $\kappa \geq 0.70$. The penalty audit instead requires raw agreement ≥ 0.80 because its skewed deserved-penalty marginal deflates κ. All audits follow policy and verifier freezing.

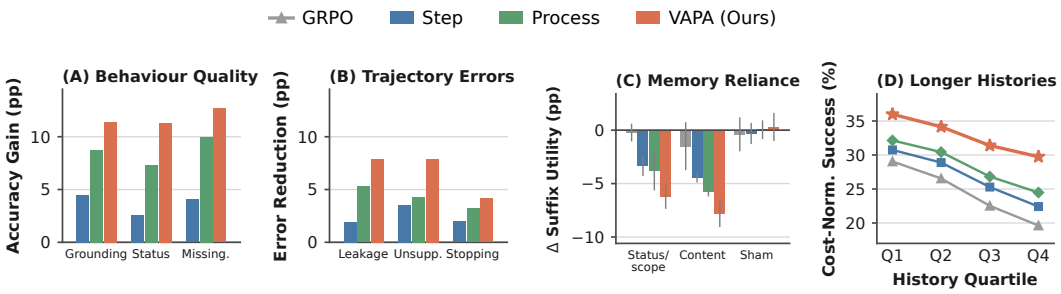


Figure 11. Behavioral diagnostics. (A,B) Predicate gains and trajectory-error reductions relative to A1. (C,D) Memory-intervention effects and CNS by history quartile. Whiskers give 95% intervals.

Verifier-decision audit. We sample 1,700 action–verifier opportunities from 340 held-out patients across evidence, temporal/status, memory, and stopping families. A positive label means that the predicate warrants credit. Table 25 reports agreement, confusion counts, precision, and recall.

Table 25. Verifier-decision audit by family. N counts action–verifier opportunities and κ measures annotator agreement. TP, FP, FN, and TN give the confusion counts. Precision and recall are percentages.

Family	N	κ	TP	FP	FN	TN	Prec.	Rec.
Evidence	450	0.81	267	6	15	162	97.80	94.68
Temporal/status	420	0.84	245	7	12	156	97.22	95.33
Memory	430	0.82	252	12	33	133	95.45	88.42
Stopping	400	0.75	212	13	19	156	94.22	91.77

Applied-penalty audit. Family quotas select 1,200 of 33,200 penalties across 140 patients. Precision is the deserved-penalty rate, MinP the lowest predicate-level precision, Agr raw annotator agreement, and a sign flip an unwarranted penalty on an action deserving positive credit. Estimates use quota weights and 4,000 patient-clustered bootstrap replicates. Raw agreement remains high even where skewed marginals depress Cohen's κ, so Table 26 reports both agreement measures and adjudicated counts.

Table 26. Applied-penalty audit by verifier family. Precision and minimum predicate precision are percentages. Brackets are patient-bootstrap 95% intervals, with raw agreement and Cohen's κ.

Family	Penalties	Precision	MinP	Unwarranted	Sign flips	Agr	κ
Evidence	330	95.15 [92.51, 97.31]	93.64	16	4	0.891	0.39
Temporal/status	300	89.67 [86.07, 92.99]	89.00	31	1	0.933	0.72
Memory	310	84.52 [80.54, 88.17]	76.70	48	11	0.881	0.62
Stopping	260	88.85 [84.65, 92.55]	86.15	29	3	0.835	0.46

Trajectory and rewarded-action audits. Randomized annotation covers 960 trajectories from 240 held-out cases, balanced by suite and seed block. A material error must affect the answer, evidence, or budget through inadmissible evidence, incorrect status or scope, lost evidence, missingness handling, or premature stopping. Wrong answers alone do not qualify. Separately, positive-score actions are stratified by arm, seed, verifier family, and score quartile, with up to 1,200 per arm and inverse inclusion weights. Invalidity includes unsupported credit and redundant or irrelevant behavior. Table 27 preserves seed blocks and patients in paired comparisons.

Table 27. Material trajectory errors and invalid rewarded actions (%). Arm contrasts have paired hierarchical-bootstrap 95% intervals. Cohen's κ describes pre-adjudication agreement.

	Material trajectory error	Invalid verifier-positive action
A1	37.92	4.34
A3	29.58	4.54
$A3 - A1$	-8.33 [-11.99, -4.68]	$+0.20$ [-2.38, 2.78]
A2	30.83	3.41
A4	22.08	3.89
$A4 - A2$	-8.75 [-12.59, -4.91]	$+0.48$ [-0.86, 1.82]
Cohen's κ	0.79	0.73

Figure endpoints. Figure 11A covers 973 grounding, 684 memory-status, and 379 missingness actions. Panel B partitions material errors by first affected family, with an unplotted residual completing Table 27. Resampling preserves seed blocks, cases, and each case's actions.

Stress tests. One prespecified state from each of 300 held-out patients receives three label-preserving and three label-flipping transformations. Six scripted strategies test duplicate retrieval, repeated field queries, redundant or irrelevant writes, indiscriminate status changes, and unsupported pointer reuse. A scripted attempt succeeds when it is invalid under the codebook, raises the auxiliary process-minus-cost score

$$R_{\text{aux}} = \sum_t (\lambda_p r_t - \lambda_c c(a_t))$$

relative to the same-observation control, and does not improve R_{out}. The sum defining R_{aux} runs over executed turns t. A third test samples 32 legal continuations per state and blindly audits the highest-R_{aux} and a uniform member. Table 28 reports all attempts. Successful scripted exploits edit memory or status, while the selection effect has an interval spanning zero. Resampling preserves seed blocks and patients.

Learned reward model. Applying the same 1,700-opportunity audit to PPO + learned PRM gives positive-decision precision 78.41% $[74.06, 82.32]$, against 94.22–97.80% for the deterministic families. Rewarded-but-invalid actions reach 9.26% $[7.81, 10.78]$, versus 3.89% for A4, a paired difference of $+5.37$ points $[3.88, 6.86]$. Scripted exploitation succeeds for five of six learned-model strategies, including duplicate retrieval and repeated field queries, while only unsupported value insertion behind a valid pointer fails against both reward channels. This audit motivates deterministic checks despite their narrower coverage.

E.2 Memory Interventions

Probe states. Patient-disjoint shared-SFT trajectories supply states containing a task-required, cutoff-grounded memory item that is absent elsewhere in view and legally reacquirable. Edits pre-

Table 28. Reward-exploitation stress tests with patient and attempt counts. Estimates are percentages except the paired difference (pp). Brackets give clustered 95% intervals. No interval is estimated for zero-event rows.

Audit	Reported endpoint	Patients	Attempts	Estimate (%)	95% interval	κ
Metamorphic	Label-preserving violations	300	900	1.22	[0.22, 2.56]	
Metamorphic	Label-flipping violations	300	900	3.11	[0.89, 5.67]	
Scripted	Duplicate RETRIEVE	300	300	0.00	*n/a*	0.82
Scripted	Repeated QUERYFIELD	300	300	0.00	*n/a*	0.80
Scripted	Redundant UPDATEMEMORY	300	300	2.00	[0.00, 5.00]	0.74
Scripted	Task-irrelevant UPDATEMEMORY	300	300	1.33	[0.00, 3.33]	0.76
Scripted	Indiscriminate MARKSTATUS	300	300	0.67	[0.00, 2.00]	0.78
Scripted	Unsupported value, valid pointer	300	300	0.00	*n/a*	0.83
Selection	Max-R_{aux} continuation	300	300	2.67	[0.33, 6.00]	0.78
Selection	Uniform continuation	300	300	1.00	[0.00, 3.00]	0.75
Selection	Paired difference (pp)	300	300	+1.67	[-1.33, 5.00]	

serve type and length. Of 660 observations from 540 patients, 243 lack valid edits, 19 exceed the two-state cap, and 78 follow roster completion. The 320 retained states include 80 patients with two states and 160 with one. Selection precedes outcomes.

Interventions. Edits change only M_t, preserving the remaining observation, budget, and initial legal-action mask. Sham reorders exchangeable JSON fields. Status/scope corruption inserts a type-valid but task-incompatible label. Content corruption inserts a plausible atomic value that conflicts with the cited record and changes the reference result. Deterministic corruption maps are independent of policy continuations.

Frozen-policy suffixes. Each state resumes under five final checkpoints and two decoder seeds shared across arms and conditions, using temperature 1.0, nucleus threshold $p = 1$, no vocabulary truncation, and standard caps. Conditions share tools and cutoff-valid records. Policies may repair memory through ordinary costed actions, without verifier guidance. For arm a, state i, policy seed j, decoder seed k, and condition q, let $\tau^{(q)}_{aijk}$ be the sampled suffix trajectory, t_i its intervention turn, $T^{(q)}_{aijk}$ its final turn, $s(\tau) \in \{0, 1\}$ its terminal success, and a_u its action at turn u. Define

$$U^{(q)}_{aijk} = s\left(\tau^{(q)}_{aijk}\right) - \frac{\lambda_c}{2} \sum_{u=t_i}^{T^{(q)}_{aijk}} c(a_u).$$

This outcome-only utility is an affine normalization of $R_{\text{out}} - \lambda_c C$, where $C = \sum_{u=t_i}^{T^{(q)}_{aijk}} c(a_u)$ is suffix action cost, because $R_{\text{out}} = 2s - 1$. Cost includes every executed action, including failed termination. Suffixes without a valid ANSWER have $s = 0$. The paired effect is

$$d^{(q)}_{aijk} = 100 \left(U^{(q)}_{aijk} - U^{(\text{intact})}_{aijk}\right), \qquad q \in \{\text{sham}, \text{status/scope}, \text{content}\},$$

with negative values indicating lower utility. Averaging decoder seeds within state and states within patient gives patients equal weight. Figure 11C reports arm means and two-sided five-seed t intervals. The A1 corruption intervals include zero.

Interpretation. Sham effects are small under the tested memory reorderings. Content and status/scope corruption reduce utility in the reported comparisons, with intervals excluding zero for A2 through A4. The corruption effects reflect both reliance on the edited memory and the policy's ability to repair it.

E.3 HISTORY LENGTH

History measure. History length $H_i = |E^{(i)}_{\leq \tau_{c,i}}|$ counts canonical pre-cutoff events and is identical across methods. Capped trajectories remain terminal failures. Quartiles are fixed separately within calculation and retrieval over unique patient-cutoff pairs, with ties kept together. The respective cut points are 236, 428, and 803 events, and 205, 371, and 718 events.

Quartile analysis. All four arms evaluate each case with five final matched checkpoints. Let a index the arm, i the case, and j the policy seed. For terminal success s_{aij}, action cost C_{aij}, and case budget $C_{\max,i}$, define outcome-only cost-normalized success and its quartile contrast as

$$Y_{aij} = 100 s_{aij} \left(1 - \frac{C_{aij}}{C_{\max,i}}\right), \qquad \Delta_q = \overline{Y}_{A4,q} - \overline{Y}_{A1,q}, \quad q \in \{1,2,3,4\}.$$

Here $\overline{Y}_{a,q}$ is the weighted mean for arm a in history quartile q. Aggregation weights the two suites equally, task types equally within suite, and patients equally within task, averaging repeated cases within patient. Table 29 reports within-seed contrasts and the descriptive change $\Delta_4 - \Delta_1$ with paired five-seed t intervals.

Continuous analysis. Let $g(i)$ identify the source-suite-by-task-family stratum of case i and $\mathcal{I}_g$ its case-index set. Within each arm and policy seed, weighted least squares uses the same suite, task, and patient weights to fit

$$x_i = \log_2 H_i - \frac{1}{|\mathcal{I}_{g(i)}|} \sum_{r \in \mathcal{I}_{g(i)}} \log_2 H_r,$$
$$Y_{aij} = \alpha_{a,g(i),j} + \beta_{a,j} x_i + \varepsilon_{aij}.$$

Here r indexes cases in the stratum, x_i is centered log-base-two history length, $\alpha_{a,g(i),j}$ is the arm-, stratum-, and seed-specific intercept, $\beta_{a,j}$ is the seed-specific slope for arm a, and ε_{aij} is the residual. The across-seed mean slope β_a describes the CNS change per history-length doubling within task stratum. The primary arm-by-length contrast is $\theta_{\text{length}} = \beta_{A4} - \beta_{A1}$, computed within seed before averaging. A positive value indicates a widening VAPA–GRPO gap. Table 29 reports the individual slopes and their contrast.

Table 29. History-length contrasts in cost-normalized success, with paired five-seed 95% t intervals. Quartile differences are pp. Slopes are pp per history-length doubling.

Quartile contrast	Estimate [95% CI]	Slope	Estimate [95% CI]
Δ_1	$+8.47$ [7.53, 9.41]	β_{A1}	-3.71 [-4.39, -3.03]
Δ_2	$+9.88$ [8.76, 11.00]	β_{A2}	-3.28 [-4.11, -2.45]
Δ_3	$+10.42$ [9.59, 11.25]	β_{A3}	-3.02 [-3.86, -2.18]
Δ_4	$+11.73$ [10.66, 12.80]	β_{A4}	-2.46 [-3.27, -1.65]
$\Delta_4 - \Delta_1$	$+3.26$ [1.71, 4.81]	θ_{length}	$+1.25$ [0.72, 1.78]

Interpretation. Cost-normalized success declines with history length for every arm, while A4 retains the largest mean. History length is observational and can reflect task mix, disease burden, or evidence density. The analysis estimates associations across the observed record-length distribution.

E.4 Subgroup Breakdown

Strata. We repeat the A1 and A4 EHRSHOT comparison by age band at the prediction cutoff and recorded sex, using strata fixed before scoring and shared across arms and seeds. Estimates pool instances within stratum and weight patients equally, so they need not reproduce the task-macro means of Table 1. We omit race and ethnicity because several retained task-level categories have too few positives for stable macro estimates. Appendix E.3 separately analyzes pre-cutoff record volume.

Findings. The A4 advantage is positive in every reported stratum and every interval excludes zero, so the aggregate gain is not carried by one age or recorded-sex group. Baseline performance varies more across strata than the advantage does.

Table 30. EHRSHOT transfer by stratum (%). Entries pool instances within a stratum and weight patients equally. Δ is the paired A4−A1 difference with its five-seed 95% t interval.

		AUPRC			AUROC		
Stratum	Patients	A1	A4	Δ [95% CI]	A1	A4	Δ [95% CI]
A. AGE BAND AT CUTOFF							
Under 50	372	34.18	37.02	+2.84 [2.07, 3.61]	66.21	69.44	+3.23 [2.57, 3.89]
50 to 64	596	37.41	40.63	+3.22 [2.62, 3.82]	67.88	71.22	+3.34 [2.78, 3.90]
65 to 79	741	39.86	43.35	+3.49 [2.98, 4.00]	69.15	72.61	+3.46 [3.01, 3.91]
80 and over	471	41.02	44.61	+3.59 [2.91, 4.27]	69.92	73.38	+3.46 [2.89, 4.03]
B. RECORDED SEX							
Female	1,104	38.02	41.29	+3.27 [2.79, 3.75]	68.11	71.49	+3.38 [2.95, 3.81]
Male	1,076	38.94	42.38	+3.44 [2.91, 3.97]	68.74	72.21	+3.47 [3.01, 3.93]

F Case Studies

Selection rule. The 750 matched MedAgentBench episode-seed pairs share patient, task, cutoff, initial state, budget, and greedy decoding. Within each joint-success stratum, select the pair closest to its median CNS difference, breaking ties by episode identifier and seed. Selection ignores trace content and process rewards. Figure 12 shows the discordant strata. In Table 31, s^{R} and s^{N} are binary successes for replay and no fork, n is the stratum count, and d is replay-minus-no-fork CNS in percentage points. The table's marginals match Table 12.

Table 31. Outcome strata over 750 matched episode–seed pairs. Counts and percentages describe each stratum. Median and selected CNS differences are pp.

Stratum	$(s^{\mathrm{R}}, s^{\mathrm{N}})$	n (%)	Med. d	Sel. d
Replay-only success	$(1, 0)$	87 (11.6)	+75.00	+75.00
No-fork-only success	$(0, 1)$	49 (6.5)	−75.00	−75.00
Shared success	$(1, 1)$	233 (31.1)	+8.33	+8.33
Shared failure	$(0, 0)$	381 (50.8)	0.00	0.00

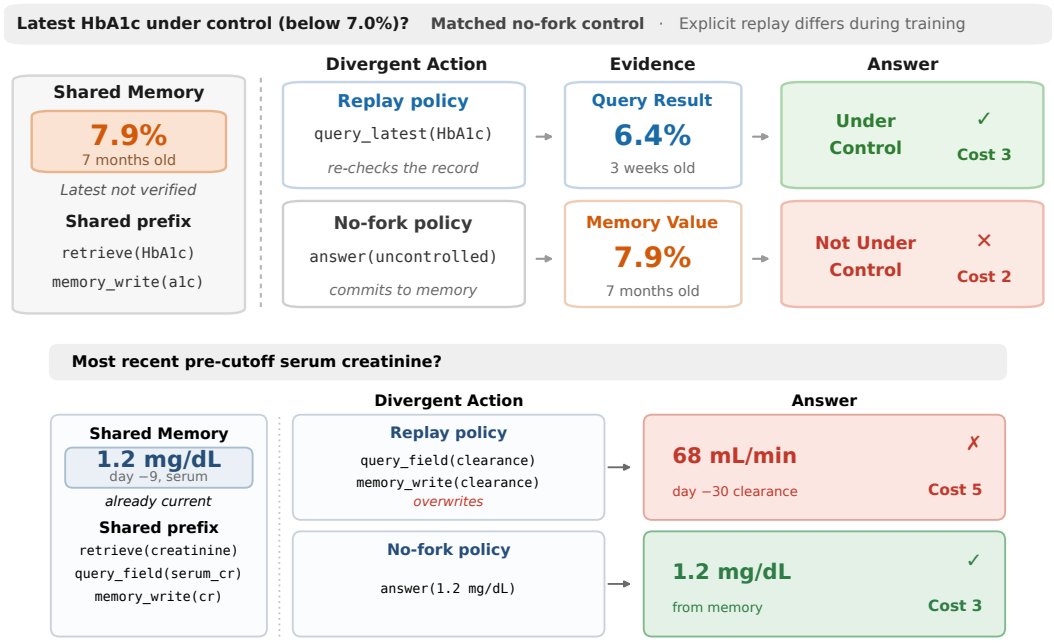


Figure 12. Matched deployment cases. Rechecking stale evidence succeeds in the top pair. In the bottom pair, further retrieval introduces an analyte and unit error.

G LIMITATIONS AND BROADER IMPACT

Scope of the evidence. Our conclusions concern critic-free credit assignment for bounded-memory EHR agents under the stated tasks, tool interface, budgets, and frozen-transfer protocol. The main factorial uses one backbone. The second backbone tests the overall VAPA–GRPO contrast rather than every component. The full-budget memory sweep compares A1 with A4, which tests whether the complete method's margin changes with capacity but does not establish a capacity trend for every baseline. The replay control matches sampled actor tokens, while Appendix B.5 separately reports GPU hours. It therefore supports a token-allocation comparison rather than a claim of equal wall-clock cost. Component ablations and audits should be read at their stated budgets. The turn-index outcome term and the baseline-specific pilot grids in Appendix B.3 are not independently varied at full budget. Published systems also require the documented interface adaptations, and target-trained clinical predictors remain unmatched references even when their label budget is reported. We therefore restrict the central empirical claim to the prespecified factorial and matched RL comparisons.

Observed limits. The gains do not imply uniform improvement on every task or population. Although A4 has the highest macro mean, it trails the other arms on ICU-transfer AUPRC in Table 21. That endpoint also has the lowest positive rate in Table 3, but the present evidence does not explain the reversal. We report it as a failure case rather than attributing it to class imbalance. Appendix E.4 covers age and recorded sex, while Appendix E.3 analyzes record length. Neither supports conclusions about race, ethnicity, or attributes without sufficient task-level samples. The learned-reward-model control evaluates one design rather than the full family. The fitted head in Appendix C.5 has no external validation and no multiplicity adjustment across its eight exploratory tasks, so it is evidence about the supervision gap rather than a deployable calibration procedure.

Broader impact and safeguards. Better action-level credit could help an agent preserve relevant evidence and avoid unnecessary retrievals under a fixed budget. The same mechanism could also reinforce misspecified checks, amplify gaps in the underlying records, or make an unreliable system appear more trustworthy. Deterministic verifiers improve auditability but do not establish clinical validity. This study uses deidentified structured releases, excludes post-cutoff evidence from agent inputs, keeps record-derived data on controlled compute, and reports subgroup and reward-channel audits. The evaluated agents answer benchmark questions. They do not diagnose, triage, or recommend treatment. Any clinical use would require institution-specific external validation, prospective safety and fairness evaluation, calibration, privacy and security review, monitoring for verifier failures and distribution shift, and clinician oversight. Benchmark improvements alone do not justify deployment.