ECHO-GRAPH: SCALING LONG HORIZON RL WITH DEPENDENCY-AWARE CREDIT PROPAGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon agentic reinforcement learning requires not only extending the accessible interaction history, but also assigning effective learning signals across repeatedly reconstructed contexts. ECHO couples context management with credit assignment by selecting historical memories during reconstruction and reinforcing the turns retained in the final reconstruction trace. However, final-trace masking on successful trajectories can leave early prerequisite steps unreinforced as reconstruction depth grows, biasing learning toward later stages. Fully masking unsuccessful trajectories removes corrective feedback, risking rapid entropy collapse and reduced exploration. We propose **ECHO-Graph**, which propagates outcome credit backward through a DAG of turn dependencies induced by context reconstruction, reinforcing early prerequisite steps and repeatedly reused evidence in successful trajectories. For unsuccessful trajectories, where provenance does not reliably support fine-grained credit assignment, we apply attenuated dense penalties to provide corrective feedback and help sustain exploration. The method requires no additional rollouts, critics, or model evaluations. Experiments on BrowseComp-Plus and SWE-Bench Pro show gains over ECHO, especially on difficult instances, with stronger context-horizon scaling on the former.

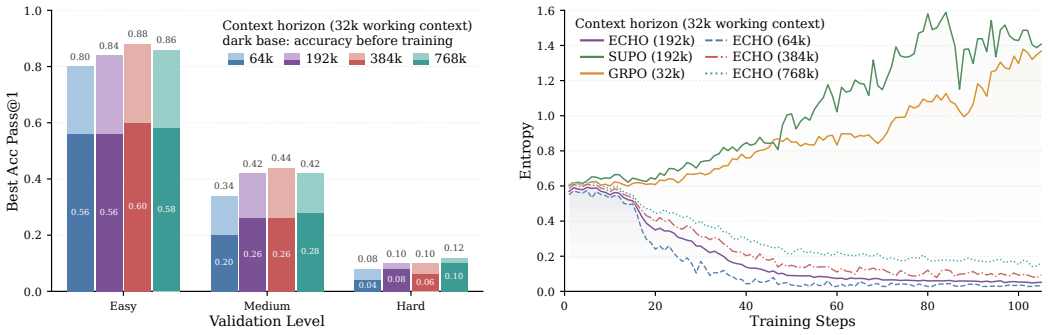
1 INTRODUCTION

Large language model (LLM) agents increasingly tackle complex tasks through extended interactions with tools and environments, interleaving reasoning, information gathering, and action over long trajectories (Yao et al., 2023; Zhou et al., 2024). Reinforcement learning (RL) provides a natural way to improve such agents from interaction outcomes (Luo et al., 2025), yet these outcomes are typically sparse and delayed, making it difficult to identify which intermediate decisions are responsible for eventual success or failure. Recent approaches therefore seek finer-grained credit through grouped step-level comparisons or turn-level progress estimation (Feng et al., 2025; Tao et al., 2026). At the same time, long-horizon interaction creates a second difficulty: finite context windows require agents to repeatedly compress, summarize, or reorganize their interaction history (Sun et al., 2025; Li et al., 2026). Consequently, information that ultimately supports the final answer may originate many reconstruction stages before it is eventually used.

ECHO (Xie et al., 2026) takes an important step toward connecting these two problems by coupling context reconstruction with credit assignment. It maintains source-indexed turn memories and reuses the provenance of the final reconstruction to route positive outcome credit back to selected historical turns, enabling traceable supervision without additional rollouts or model evaluations. However, this routing remains local to the final reconstruction: earlier turns may contribute through multiple intermediate selections and reconstructions yet receive no credit if they disappear from the final trace. Meanwhile, ECHO reinforces only a sparse subset of tokens on successful trajectories, leaving limited corrective feedback from failures. As shown in Figure 1, these limitations become more pronounced with longer horizons, where gains saturate on harder instances and policy entropy collapses rapidly.

To address these limitations, we introduce **ECHO-Graph**, which replaces sparse final-trace masking with graded dependency-aware positive credit and pairs it with attenuated dense negative feedback. This enables credit to propagate through reconstruction dependencies while maintaining a more balanced optimization signal.

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.



(a) Context-horizon scaling across difficulty levels.

(b) Actor entropy during training.

Figure 1: Motivation for ECHO-Graph. (a) ECHO performance across context horizons and difficulty levels. (b) Actor entropy over training steps for different methods and context horizons.

Our main contributions are:

- We characterize two limitations of ECHO through context-horizon scaling and training-dynamics analyses: limited gains from longer horizons and rapid policy entropy collapse.
- We propose ECHO-Graph, which combines dependency-aware positive credit propagation with attenuated dense negative feedback to reinforce reused evidence and learn from unsuccessful trajectories. The method requires no additional rollouts, critics, or model evaluations.
- Experiments on BrowseComp-Plus and SWE-Bench Pro demonstrate improved performance across difficulty levels, including strong gains on difficult instances. Further experiments on BrowseComp-Plus show stronger context-horizon scaling than ECHO.

2 Related Work

2.1 Context Management for Long-Horizon Agents

Context management differs in both its triggering mechanism and the representation of retained history. Rule-triggered methods reconstruct context at prescribed boundaries, using summary compression, as in ReSum and SUPO (Wu et al., 2025; Lu et al., 2025), or source-indexed memory selection, as in ECHO (Xie et al., 2026). Action-triggered methods let the policy initiate context editing or subtrajectory folding, as in MemAct and FoldGRPO (Zhang et al., 2025; Sun et al., 2025). These axes are distinct: rule-triggered reconstruction can still use learned selection, while policy-initiated folding can retain summaries. We build on selective memory to study how the dependencies created by repeated reconstruction can support long-horizon credit assignment.

2.2 Credit Assignment in Agentic Reinforcement Learning

Credit-assignment approaches can be organized by the source of their intermediate learning signals. Learned-critic approaches estimate returns parametrically, as in actor-critic PPO and SAO (Schulman et al., 2017; Hou et al., 2026). Resampling-based approaches evaluate intermediate decisions through additional continuations or local rerollouts, as in VinePPO and Memory-R2 (Kazemnejad et al., 2024; Yan et al., 2026). Rollout-statistical approaches reuse sampled experience through state grouping or transition graphs, including GiGPO, GraphGPO, and G2PO (Feng et al., 2025; Cheng et al., 2026; Wang et al., 2026). Likelihood-based approaches instead score answer predictability along a trajectory, as in IGPO and TRACE, requiring model evaluations without training a separate critic (Wang et al., 2025; Tao et al., 2026). Context-provenance methods route outcome credit through reconstruction records, with ECHO relying on the final reconstruction trace (Xie et al., 2026). Our work considers dependencies across successive reconstructions.

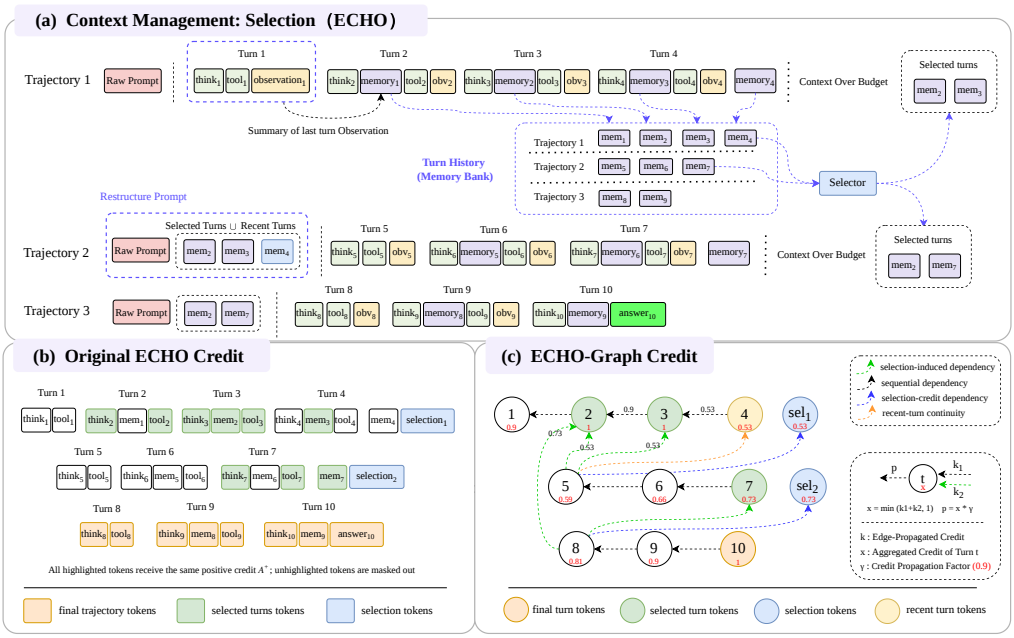


Figure 2: **ECHO-Graph credit assignment.** (a) Context reconstruction from selected memories and recent turns. (b) ECHO assigns positive credit through the final reconstruction trace. (c) ECHO-Graph propagates positive credit backward through dependencies across reconstruction stages.

3 PRELIMINARIES

3.1 CONTEXT RECONSTRUCTION

Long-horizon agents periodically reconstruct their working context as the interaction history approaches the context limit. This partitions a rollout into sub-trajectories,

$$\tau = (\sigma_1, \sigma_2, \ldots, \sigma_J), \tag{1}$$

where σ_j denotes the interactions within the j-th context segment.

A common strategy is rolling summarization, as adopted by SUPO (Lu et al., 2026). At reconstruction round j, previous interactions are compressed into a summary state,

$$z_j = \mathcal{C}(z_{j-1}, \sigma_{j-1}), \tag{2}$$

and the policy continues from the reconstructed context,

$$c_{j,t} = x \oplus \mathrm{Render}(z_j) \oplus H^{\mathrm{loc}}_{j,t}, \tag{3}$$

where x is the original task, $H^{\mathrm{loc}}_{j,t}$ contains the recent interactions in the working context, and $\oplus$ denotes concatenation.

ECHO (Xie et al., 2026) instead stores completed turns in a persistent source-indexed memory archive $\mathcal{E}_j = \{e_k\}$. As illustrated in Figure 2(a), the policy selects historical memories indexed by $\hat{I}_j$ and combines them with recent turns:

$$c_{j,t} = x \oplus \mathrm{Render}\left(\mathcal{E}_j[\hat{I}_j]\right) \oplus H^{\mathrm{loc}}_{j,t}. \tag{4}$$

This selective reconstruction preserves the source identities of reused historical turns, exposing explicit dependencies between each new segment and earlier turns.

3.2 Credit Assignment under Context Reconstruction

Let $\hat{A}_i$ denote the group-relative trajectory advantage of rollout τ_i. Standard GRPO and summarization-based methods such as SUPO assign this advantage densely to all trainable response tokens (Shao et al., 2024; Lu et al., 2026),

$$A_{i,t}^{\text{dense}} = \hat{A}_i. \tag{5}$$

Such dense assignment does not distinguish useful historical evidence from redundant interactions.

As illustrated in Figure 2(b), ECHO (Xie et al., 2026) assigns positive credit through the final reconstruction trace using a provenance-guided mask,

$$A_{i,t}^{\text{ECHO}} = M_{i,t}^{\text{ECHO}} \hat{A}_i, \qquad \hat{A}_i > 0, \tag{6}$$

where $M_{i,t}^{\text{ECHO}} \in \{0, 1\}$ is defined as

$$M_{i,t}^{\text{ECHO}} = \mathbb{I}\left[g_{i,t} = 1 \vee d_{i,t} \in I_i^{\text{src}} \vee f_{i,t} \in I_i^{\text{src}} \vee b_{i,t} = 1\right]. \tag{7}$$

Here, I_i^{src} denotes the source turns selected by the final reconstruction. The four conditions cover tokens in the final sub-trajectory, action/reasoning tokens from the selected source turns, their corresponding memory-finding tokens, and generated memory-selection tokens, respectively. Rollouts with non-positive advantage receive no ECHO credit update.

4 Methods

4.1 Motivation

ECHO (Xie et al., 2026) extends interaction horizons through repeated context reconstruction. However, Figure 1 shows limited gains from longer horizons on harder instances, alongside rapid policy entropy decline during training. These observations motivate examining how positive and negative learning signals interact under context reconstruction.

On successful trajectories, final-trace masking can leave earlier prerequisite steps unreinforced despite their contributions through intermediate reconstructions. Propagating positive credit across these dependencies would extend supervision to useful earlier decisions. However, if unsuccessful trajectories remain masked, this broader reinforcement is still driven exclusively by successes and may accelerate policy concentration and entropy decline.

Conversely, adding dense negative feedback while retaining sparse final-trace credit can create asymmetric supervision for early decisions. Useful prerequisite steps may be penalized when later errors cause a rollout to fail, yet receive no positive credit when they contribute to success but are omitted from the final trace. This asymmetry may suppress useful early behaviors and hinder long-horizon learning; higher entropy alone need not translate into productive exploration.

We therefore hypothesize that these two aspects of credit assignment should be designed jointly. Dependency-aware positive propagation provides task-directed reinforcement, while attenuated dense negative feedback supplies corrective signals and helps counter premature policy concentration. This motivates a design in which positive credit guides learning across reconstruction dependencies, with a tunable negative-feedback weight used to preserve exploration without overwhelming reinforcement of useful historical decisions.

4.2 Dependency-Aware Credit Assignment

For each rollout τ_i with positive advantage $\hat{A}_i > 0$, ECHO-Graph constructs a directed acyclic graph $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i)$ over interaction turns and memory-selection actions (Figure 2(c)). Each turn node shares its credit among its trainable reasoning, tool-call, and turn-memory tokens; selection-action nodes supervise their generated selection tokens. Environment observations remain masked.

Turn-memory dependency graph. Let b_j denote the first interaction turn after reconstruction j. We construct four types of dependencies:

- *Sequential dependency.* Consecutive turns within a sub-trajectory are connected as $v_t \to v_{t+1}$, preserving the local reasoning chain.
- *Recent-turn continuity.* The last automatically retained recent turn is connected to v_{b_j}, maintaining cross-segment continuity without treating every retained turn as an additional reuse branch.
- *Selection-induced dependency.* Each explicitly selected historical turn is connected to v_{b_j}. Repeated selections across reconstruction stages create distinct downstream dependencies, even when the selected memory sets overlap or are nested.
- *Selection-action dependency.* Each memory-selection action s_j is connected to v_{b_j}. Through backward propagation, this first turn collects discounted credit from subsequent reasoning and downstream reuse, serving as a credit proxy for the next sub-trajectory. The selection action receives this signal through the same propagation rule.

All edges follow interaction order, and each distinct source–destination edge is included once.

Credit propagation. The terminal interaction turn is initialized with $c_{i,T} = 1$. For all other nodes, credit is propagated in reverse topological order using sum aggregation with unit clipping:

$$c_{i,u} = \min\left(1, \sum_{v \in \mathrm{Ch}(u)} \gamma c_{i,v}\right), \tag{8}$$

where $\mathrm{Ch}(u)$ denotes the downstream neighbors of u and $\gamma \in (0, 1]$ controls discounting. Summation allows distinct reuse paths to contribute to a turn's credit, while clipping bounds its total credit by one. Thus, repeated reuse can strengthen supervision without amplifying a node's credit beyond the terminal credit. Selection actions follow the same rule, giving $c_{i,s_j} = \gamma c_{i,b_j}$.

Positive credit. For $\hat{A}_i > 0$, the token-level advantage is

$$A^{+}_{i,a} = \begin{cases} c_{i,t}\hat{A}_i, & a \text{ belongs to interaction turn } t, \\ c_{i,s_j}\hat{A}_i, & a \text{ belongs to selection action } s_j. \end{cases} \tag{9}$$

The graph reuses reconstruction provenance without extra rollouts, critics, or model evaluations.

Dense negative credit. For negative-advantage rollouts, reconstruction provenance alone does not reliably identify which intermediate decisions deserve negative credit. We therefore apply attenuated dense feedback to all trainable policy-generated tokens:

$$A^{\mathrm{neg}}_{i,a} = \lambda_{\mathrm{neg}}\hat{A}_i, \qquad \hat{A}_i < 0, \tag{10}$$

where $\lambda_{\mathrm{neg}} \in [0, 1]$ controls its strength. Zero-advantage rollouts receive no credit update. We analyze the positive-credit design and negative-credit strength in Section 4.4.

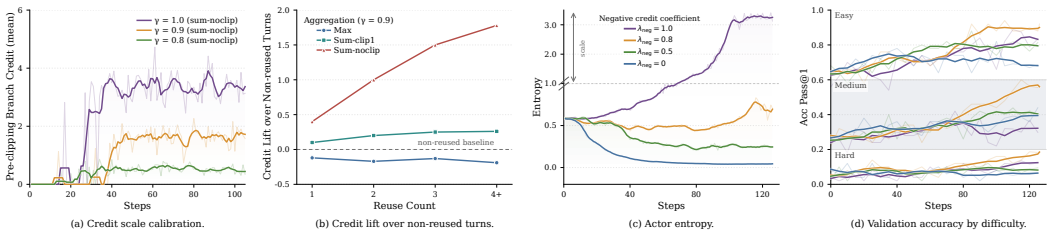
4.3 Propagation Properties and Complexity

For a purely sequential trajectory, Eq. (8) reduces to $c_t = \gamma^{T-t}$. More generally, let $d_G(u, T)$ denote the shortest directed path length from node u to the terminal turn T. Since $c_u \geq \gamma c_v$ for every downstream neighbor v, applying this inequality along a shortest path gives

$$\gamma^{d_G(u,T)} \leq c_u \leq 1. \tag{11}$$

Reconstruction dependencies can therefore provide shorter credit-propagation paths to early turns, mitigating attenuation over long interaction histories. Adding distinct reuse edges cannot decrease credit, while unit clipping bounds accumulation. At $\gamma = 1$, all nodes that can reach the terminal receive unit credit. These properties characterize the weighting rule; reconstruction provenance remains a proxy for contribution rather than a guarantee of causal relevance.

Given the graph, interaction order already provides a topological ordering. A reverse traversal processes each node once and each edge once, yielding $O(|V| + |E|)$ propagation time and $O(|V|)$ additional storage for node credits. This computation requires no additional model evaluations.



(a) Credit scale calibration. (b) Credit lift over non-reused turns. (c) Actor entropy. (d) Validation accuracy by difficulty.

Figure 3: **Credit design analysis.** (a) Mean branch credit under different propagation factors γ, using sum aggregation without clipping. (b) Credit lift over non-reused turns under different aggregation rules at $\gamma = 0.9$. (c,d) Actor entropy and validation Pass@1 under different λ_{neg}, with positive graph credit fixed to $\gamma = 0.9$ and Sum-clip1.

4.4 Credit Design Analysis

We conduct credit-design analyses on BrowseComp-Plus in two stages. We first use credit-level diagnostics to select the propagation factor and aggregation rule for positive graph credit. With this configuration fixed, we vary the strength of dense negative feedback and examine its effects on training dynamics and task performance.

Calibrating positive graph credit. Positive credit should reward historical reuse while limiting amplification from multiple dependency paths. Figure 3(a) shows that $\gamma = 0.8$ strongly attenuates branch credit, whereas $\gamma = 1.0$ produces substantial accumulation. The intermediate setting $\gamma = 0.9$ yields a moderate credit scale. At this setting, Figure 3(b) compares how aggregation rules reward reuse. Max provides little reuse-dependent lift, while Sum-noclip produces a steep increase with reuse count. Sum-clip1 preserves a positive reuse-dependent lift while limiting its growth. These credit-level diagnostics motivate using $\gamma = 0.9$ with sum aggregation and unit clipping.

Ablating dense negative credit. With this positive graph-credit configuration fixed, we vary $\lambda_{\text{neg}} \in \{0, 0.5, 0.8, 1.0\}$. Figure 3(c) reveals distinct entropy trajectories. Without negative credit, entropy rapidly approaches zero. A coefficient of 0.5 slows this decline but still leads to low entropy, whereas $\lambda_{\text{neg}} = 1.0$ produces a sharp increase. The intermediate setting $\lambda_{\text{neg}} = 0.8$ avoids both extremes, retaining moderate entropy throughout training.

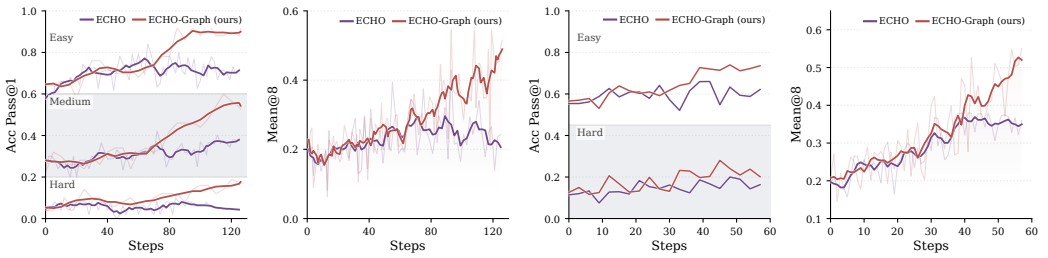
Task performance provides the basis for choosing among these settings. As shown in Figure 3(d), $\lambda_{\text{neg}} = 0.8$ achieves the strongest late-training validation performance across all three difficulty levels, with particularly clear gains on Medium and Hard instances. The weaker settings yield smaller gains, while the higher entropy at $\lambda_{\text{neg}} = 1.0$ does not translate into better accuracy. Together, these results support pairing selective positive graph credit with attenuated dense negative feedback. We therefore use $\lambda_{\text{neg}} = 0.8$ by default.

5 Experiment

5.1 Experimental Setup

Training uses eight B200 GPUs and eight sampled responses per prompt, with a fixed 32K working context across experiments. Within each benchmark, ECHO and ECHO-Graph share the same backbone, training data, rewards, rollout budget, and context-reconstruction mechanism, differing only in credit assignment. Both retain the three most recent turns verbatim at each reconstruction. Unless otherwise stated, ECHO-Graph uses the default credit-assignment configuration established in Section 4.4 on both benchmarks. Additional benchmark-specific settings are provided in Appendix A.1.

BrowseComp-Plus. Following Context-Folding (Sun et al., 2025), we use 680 queries for training and 150 for evaluation, with 50 evaluation queries each at Easy, Medium, and Hard difficulty. Difficulty is determined by accuracy over multiple rollouts. Harder instances involve more tool calls and longer interaction trajectories, making this split well suited to evaluating long-horizon credit



(a) BrowseComp-Plus accuracy. (b) BrowseComp-Plus training curve. (c) SWE-Bench Pro accuracy. (d) SWE-Bench Pro training curve.

Figure 4: Training dynamics of ECHO and ECHO-Graph on BrowseComp-Plus and SWE-Bench Pro, showing validation Pass@1 by difficulty and mean training score. ECHO-Graph exhibits stronger and more sustained learning, especially on challenging instances.

Table 1: Main results on curated difficulty subsets of BrowseComp-Plus and SWE-Bench Pro, measured by Pass@1. We additionally evaluate frontier models under the ReAct harness as reference baselines. *Base model* denotes the benchmark-specific backbone specified in the experimental setup. *Context Horizon* denotes the nominal cumulative context budget per episode. *Avg.* is the unweighted mean across the five benchmark–difficulty settings.

Method	**Context Horizon**	**BrowseComp-Plus**			**SWE-Bench Pro**		**Avg.**
		Easy	**Medium**	**Hard**	**Easy**	**Hard**	
ReAct Agent							
Claude-Opus-4.8	32K × 6	1.00	0.82	0.52	0.84	0.42	0.720
GPT-5.5	32K × 6	1.00	0.88	0.52	0.76	0.36	0.704
Deepseek-V4-Flash (284B-A13B)	32K × 6	0.88	0.66	0.36	0.74	0.40	0.608
GLM-5.3-Flash (320B-A18B)	32K × 6	0.80	0.80	0.40	0.70	0.30	0.600
Base model	32K	0.52	0.18	0.06	0.44	0.06	0.252
+ RL (GRPO)	32K	0.78	0.32	0.08	0.60	0.20	0.404
Summary Agent							
Base model	32K × 6	0.52	0.16	0.06	0.52	0.10	0.272
+ RL (GRPO)	32K × 6	0.82	0.38	0.08	0.64	0.22	0.428
Select Agent							
Base model	32K × 6	0.56	0.20	0.04	0.56	0.12	0.296
+ RL (ECHO)	32K × 6	0.84	0.42	0.10	0.66	0.20	0.444
+ RL (**ECHO-Graph**)	32K × 6	**0.94**	**0.60**	**0.18**	**0.74**	**0.28**	**0.548**

assignment. We initialize the policy from Qwen3-32B and use Qwen3-Embedding-8B for dense retrieval, with `search`, `open_page`, and `finish` tools. DeepSeek-V4-Flash serves as an LLM judge to assess final-answer correctness and assign rollout rewards.

SWE-Bench Pro. We use 581 repository-level software engineering instances for training and 100 for evaluation, with 50 evaluation instances each at Easy and Hard difficulty. The policy is initialized from Qwen3.5-35B-A3B and operates in isolated repository sandboxes using `execute_bash`, `str_replace_editor`, `think`, and `finish`. Rollouts receive binary rewards based on whether the required repository tests pass. We report Pass@1 using greedy decoding with one rollout per instance.

5.2 Main Results

For the main comparisons, ECHO and ECHO-Graph use up to five context reconstructions, corresponding to a nominal context horizon of $32\text{K} \times 6 = 192\text{K}$. For each trained method and benchmark,

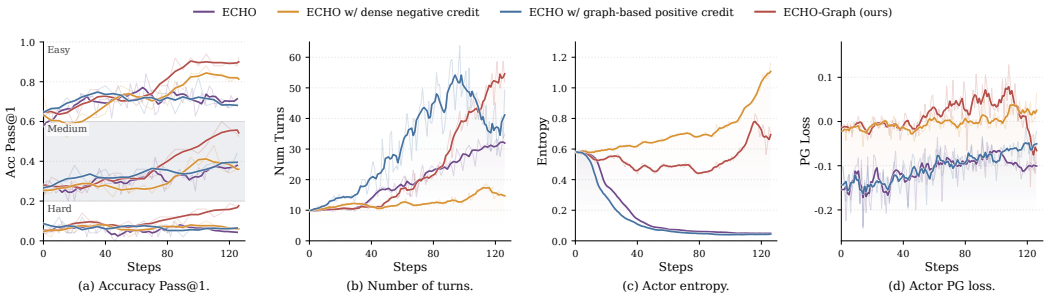


Figure 5: **Component ablations of ECHO-Graph on BrowseComp-Plus.** Graph credit uses $\gamma = 0.9$ with sum aggregation and unit clipping; dense negative credit uses $\lambda_{\text{neg}} = 0.8$.

we select the checkpoint with the highest mean validation Pass@1 across difficulty levels and report all difficulty-level results from that checkpoint.

Across the five benchmark–difficulty settings in Table 1, ECHO-Graph outperforms all evaluated baselines using the same benchmark-specific backbones, improving average Pass@1 over ECHO by 10.4 percentage points. It also consistently surpasses the GRPO-trained Summary Agent at matched context horizons, demonstrating gains beyond context extension alone. Its average Pass@1 is within six percentage points of DeepSeek-V4-Flash and GLM-5.3-Flash, despite using far fewer total parameters.

BrowseComp-Plus. Under the same context-management mechanism and horizon, ECHO-Graph improves over ECHO by 18 percentage points on Medium and nearly doubles Pass@1 on Hard. On Easy, it reaches 94% Pass@1, approaching GPT-5.5 and Claude-Opus-4.8. Figure 4(a,b) shows that these gains develop as training progresses: the validation accuracy gap widens particularly on Easy and Hard, while ECHO-Graph's training score continues to rise as ECHO's levels off and subsequently declines.

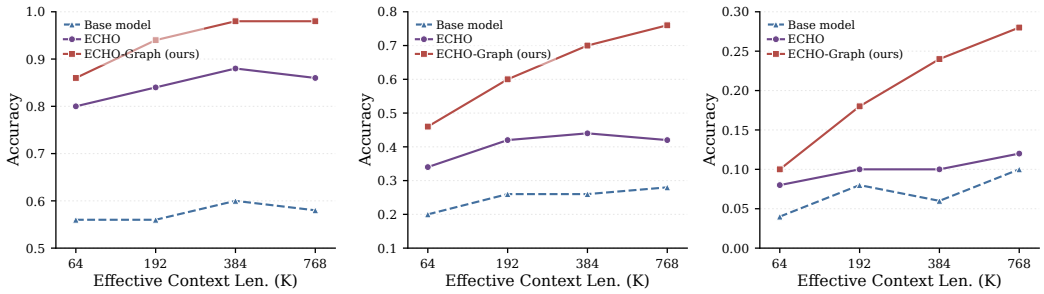
SWE-Bench Pro. The benefits extend to repository-level software engineering, where ECHO-Graph improves Hard Pass@1 over ECHO by 40% relative under matched context settings. On Easy, it matches DeepSeek-V4-Flash and surpasses GLM-5.3-Flash. Figure 4(c,d) shows similar early progress, followed by a clearer advantage for ECHO-Graph on Easy. Its training score continues to improve in the final stage, while ECHO's progress slows.

5.3 Ablation Study

Starting from ECHO, we evaluate graph-based positive credit and dense negative credit separately and jointly to assess their individual contributions and combined effect (Figure 5).

Positive graph credit alone. Positive graph credit produces longer rollouts but also faster early entropy decline, with limited accuracy gains. This is consistent with our hypothesis that reinforcing earlier dependencies can consolidate longer behavior sequences while accelerating concentration on patterns observed in successful trajectories. Broader positive supervision can therefore promote sustained interaction without necessarily producing more reliable problem-solving behavior.

Dense negative credit alone. Adding dense negative credit to ECHO increases entropy, but clear accuracy gains are largely confined to Easy instances, while Medium and Hard performance remains limited and rollouts stay short (Figure 5). These results support our hypothesis that the benefit of negative feedback depends on positive-credit coverage. With few reconstructions, final-trace masking can preserve much of a successful trajectory. With deeper reconstruction, however, useful prerequisite decisions may be penalized in failed rollouts yet receive no positive credit when they contribute to success. This asymmetric supervision helps explain why increased entropy yields limited gains on harder tasks: useful early behaviors remain exposed to negative updates without corresponding reinforcement from successful outcomes.



(a) Easy. (b) Medium. (c) Hard.

Figure 6: Context-horizon scaling on BrowseComp-Plus. Pass@1 of the base Select Agent, ECHO, and ECHO-Graph across effective context horizons from 64K to 768K, with a fixed 32K working context. ECHO and ECHO-Graph are trained separately for each horizon. (a) Easy, (b) Medium, and (c) Hard subsets.

Combining positive and negative credit. ECHO-Graph achieves the strongest late-training performance across all three difficulty levels, while maintaining non-vanishing entropy and developing longer trajectories. Its advantage over both single-component variants supports our hypothesis that positive-credit propagation and negative feedback should be designed jointly to reinforce useful historical dependencies while avoiding premature policy concentration.

5.4 Long-Horizon Scaling

We evaluate long-horizon scaling on BrowseComp-Plus by comparing the base Select Agent, ECHO, and ECHO-Graph. We vary the effective context horizon from 64K to 768K while fixing the working context at 32K. For each horizon, ECHO and ECHO-Graph are trained separately from Qwen3-32B and evaluated with the corresponding context budget.

As shown in Figure 6, increasing the context horizon provides only limited and often saturating gains for the base agent and ECHO. On Easy instances, ECHO improves moderately before plateauing, while on Medium and Hard instances the benefit of a longer horizon is much weaker. ECHO-Graph exhibits stronger scaling: performance continues to improve on Medium and Hard as the context horizon grows, while Easy performance saturates at a high level. In particular, on Hard instances, where ECHO shows little sensitivity to longer horizons, ECHO-Graph is able to convert additional context capacity into consistent performance gains.

This difference becomes increasingly important as longer horizons introduce more reconstruction stages and deeper historical dependencies. By propagating credit through these dependencies, ECHO-Graph can reinforce useful decisions beyond the final reconstruction trace and better exploit the additional history made available by context reconstruction. These results indicate that extending the context horizon alone is not sufficient; effective long-horizon scaling also requires a credit-assignment mechanism that can learn from increasingly deep reconstructed histories.

6 Conclusion

We presented ECHO-Graph, a dependency-aware credit assignment framework for long-horizon agentic reinforcement learning. By propagating positive credit across context-reconstruction dependencies and incorporating attenuated dense negative feedback, ECHO-Graph reinforces reused historical evidence while learning from unsuccessful trajectories. The method requires no additional rollouts, critics, or model evaluations. Experiments on BrowseComp-Plus and SWE-Bench Pro demonstrate consistent improvements over ECHO, while context-horizon scaling on BrowseComp-Plus shows greater benefits from longer interactions. These results highlight the importance of preserving learning signals across reconstruction boundaries, making credit assignment an essential complement to context management for long-horizon agents.

AI Assistance Disclosure

Generative AI tools were used to assist with identifying relevant literature, brainstorming, improving the readability and language of the manuscript, and providing suggestions on the organization and title of the paper. All AI-assisted outputs were reviewed and verified by the authors. The authors take full responsibility for the final content of the paper.

References

Xin Cheng, Shuo He, Lang Feng, HaiYang Xu, Ming Yan, Lei Feng, and Bo An. Beyond Trajectory-Level Attribution: Graph-Based Credit Assignment for Agentic Reinforcement Learning, 2026. URL https://arxiv.org/abs/2605.26684.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-Group Policy Optimization for LLM Agent Training, 2025. URL https://arxiv.org/abs/2505.10978.

Zhenyu Hou, Yujiang Li, Jie Tang, and Yuxiao Dong. Single-Rollout Asynchronous Optimization for Agentic Reinforcement Learning, 2026. URL https://arxiv.org/abs/2607.07508.

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, 2024. URL https://arxiv.org/abs/2410.01679.

Yujiang Li, Zhenyu Hou, Yi Jing, Jie Tang, and Yuxiao Dong. Compactionrl: Reinforcement learning with context compaction for long-horizon agents. *arXiv preprint arXiv:2607.05378*, 2026.

Miao Lu, Weiwei Sun, Weihua Du, Zhan Ling, Xuesong Yao, Kang Liu, and Jiecao Chen. Scaling LLM Multi-turn RL with End-to-end Summarization-based Context Management, 2025. URL https://arxiv.org/abs/2510.06727.

Miao Lu, Weiwei Sun, Weihua Du, Zhan Ling, Xuesong Yao, Kang Liu, and Jiecao Chen. Beyond the context window: Scaling agentic RL via end-to-end optimized context compression. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21074–21125. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.966.

Xufang Luo, Yuge Zhang, Zhiyuan He, Zilong Wang, Siyun Zhao, Dongsheng Li, Luna K. Qiu, and Yuqing Yang. Agent lightning: Train any ai agents with reinforcement learning. *arXiv preprint arXiv:2508.03680*, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

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

Weiwei Sun, Miao Lu, Zhan Ling, Kang Liu, Xuesong Yao, Yiming Yang, and Jiecao Chen. Scaling Long-Horizon LLM Agent via Context-Folding, 2025. URL https://arxiv.org/abs/2510.11967.

Leitian Tao, Baolin Peng, Wenlin Yao, Tao Ge, Hao Cheng, Mike Hang Wang, Jianfeng Gao, and Sharon Li. TRACE: Turn-level Reward Assignment via Credit Estimation for Long-Horizon Agents, 2026. URL https://arxiv.org/abs/2607.13988.

Guoqing Wang, Sunhao Dai, Guangze Ye, Zeyu Gan, Wei Yao, Yong Deng, Xiaofeng Wu, and Zhenzhe Ying. Information Gain-based Policy Optimization: A Simple and Effective Approach for Multi-Turn LLM Agents, 2025. URL https://arxiv.org/abs/2510.14967.

Yunan Wang, Minghui Song, Zihan Zhang, Shaohan Huang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. Group-Graph Policy Optimization for Long-Horizon Agentic Reinforcement Learning, 2026. URL https://arxiv.org/abs/2606.22995.

Xixi Wu, Kuan Li, Yida Zhao, Liwen Zhang, Litu Ou, Huifeng Yin, Zhongwang Zhang, Xinmiao Yu, Dingchu Zhang, Yong Jiang, Pengjun Xie, Fei Huang, Minhao Cheng, Shuai Wang, Hong Cheng, and Jingren Zhou. ReSum: Unlocking Long-Horizon Search Intelligence via Context Summarization, 2025. URL https://arxiv.org/abs/2509.13313.

Zijun Xie, Binbin Zheng, Enlei Gong, Jihua Liu, Yuyang You, Lingfeng Liu, Jiayao Tang, Guanqun Zhao, Xiaoliang Fu, Aoqi Hu, and Zeyu Chen. ECHO: Prune To Act, Trace To Learn With Selective Turn Memory In Agentic RL, 2026. URL https://arxiv.org/abs/2606.31650.

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, 2026. URL https://arxiv.org/abs/2605.21768.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Yuxiang Zhang, Jiangming Shu, Ye Ma, Xueyuan Lin, Shangxi Wu, and Jitao Sang. Memory as Action: Autonomous Context Curation for Long-Horizon Agentic Tasks, 2025. URL https://arxiv.org/abs/2510.12635.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024.

A APPENDIX

A.1 TRAINING CONFIGURATION DETAILS

We provide additional settings that supplement the main experimental setup.

BrowseComp-Plus. Table 2 lists the configuration for the BrowseComp-Plus experiments.

Table 2: Additional training and tool settings for BrowseComp-Plus.

Parameter	Value
Search results	top-5 snippets with docids
Open-page excerpt limit	16K characters
Max parallel tool calls	5
Tool response limit	16,000 tokens
Prompt / response limit	4,096 / 32,768 tokens
Model context limit	40,960 tokens
Selected turns per reconstruction	at most 8
Training batch / mini-batch	32 / 8
Rollout mode	synchronous
Learning rate / warmup	10^{-6} / 0 steps
PPO clip range	$[0.20, 0.28]$
Entropy coefficient / KL loss	0 / disabled
Validation cadence	before training, then every 5 steps
Training epochs	5

SWE-Bench Pro. Table 3 lists additional training, sandbox, and tool settings for SWE-Bench Pro.

Table 3: Additional training, sandbox, and tool settings for SWE-Bench Pro.

Parameter	Value
Max parallel / total tool calls	1 / 100
Tool response limit	12,000 characters
Sandbox session timeout	150 minutes
Tool-step / reward timeout	300 / 1,200 seconds
Prompt / response limit	4,096 / 32,768 tokens
Model context limit	40,960 tokens
Selected turns per reconstruction	at most 8
Selection output limit	1,024 tokens
Training batch / mini-batch	32 / 8
Rollout mode	asynchronous
Optimizer / weight decay	AdamW / 0.01
Learning rate / warmup	10^{-6} / 0 steps
PPO epochs / clip ratio	1 / 0.20
Entropy coefficient / KL loss	0 / disabled
Training sampling	temperature 1.0, top-p 1.0
Validation cadence	before training, then every 3 steps
Training epochs	3

A.2 PROMPT TEMPLATES

We provide the prompt templates used in our experiments. Braces denote runtime-filled slots, and omitted protocol text is marked `[...]`. Every episode starts from two benchmark-specific messages: a system message that states the agent role and defines the available functions, and a user message that carries the task, the protocol to follow, and the tool-call format. Tool definitions therefore enter through the system side only — the dataset-level system prompt lists them in text, and the chat template additionally serializes the JSON tool schemas into the same system block (`inject_tool_schemas` is enabled) — while the user message carries usage examples only. The ECHO-specific prompts in the last paragraph are benchmark-independent: they are reused verbatim on both benchmarks, with the summary protocol appended to the end of whichever system message is in use and the selection prompt issued as a separate user message once the working context is full.

BrowseComp-Plus. All compared methods share the system and user prompts below.

System Prompt

You are a meticulous and strategic research agent. Your primary function is to conduct comprehensive, multi-step research to deliver a thorough, accurate, and well-supported report in response to the user's query. Your operation is guided by these core principles: *Rigor*, *Objectivity*, *Thoroughness*, *Transparency*. `[...]`

You have access to the following functions:

- `search`: supply a string `query` and optional `topk`; returns the top-`topk` results with docid, url, and document content (may be truncated).
- `open_page`: open a page by `docid` or `url` and return the complete content; the identifier must come from a prior search result.
- `finish`: return the final `answer` with a brief evidence-grounded `explanation` and an optional `confidence`.

Function calls must follow the specified format. Required parameters must be specified. You may provide optional reasoning before the function call, but not after. `[...]`

User Prompt

You need to answer the given question by interacting with a search engine, using the search and open tools provided. Please perform reasoning and use the tools step by step, in an interleaved manner.

Question: `{question}`

Follow this structured protocol to find the answer:

- **Phase 1 — Deconstruction & Strategy:** isolate key entities and constraints, brainstorm search vectors, and build a verification checklist.
- **Phase 2 — Iterative Research & Discovery:** use `search` for broad discovery and `open_page` as a mandatory follow-up on promising results; do not repeat identical queries; execute at least 5 and up to 50 tool calls; never simulate tool output.
- **Phase 3 — Synthesis & Analysis:** triangulate critical facts across at least two independent sources and maintain a running fact sheet.
- **Phase 4 — Verification & Final Report:** re-check every checklist item and return to Phase 2 if the evidence is weak. `[...]`

Tool calls use the Hermes-style format:

```
<tool_call>
{"name": "search", "arguments": {"query": "Query", "topk":
10}}
</tool_call>
```

Your final answer should contain an explanation with inline docid citations, an exact answer, and a confidence score. Use the `finish` tool to submit your answer.

SWE-Bench Pro. The repository-level task is supplied as the user message. The following system message is inserted by the agent loop, with the function schemas serialized separately by the model's chat template.

System Prompt

You are a software engineering agent working in `/app`. Inspect the repository, make the minimal source changes needed to solve the issue, and run relevant tests. Use `execute_bash` for terminal commands and `str_replace_editor` for file edits. Call `finish` only after the implementation has been verified; do not merely describe changes that still need to be made.

The injected function schemas expose:

- `execute_bash`: execute one shell `command` in the `/app` repository.
- `str_replace_editor`: view, create, edit, or undo edits to plain-text files under `/app`; supply a `command` and `path`, with operation-specific text or line arguments.
- `think`: record a private planning `content` note without modifying the repository.
- `finish`: submit the implementation and terminate the interaction, optionally with a concise `message`.

User Prompt

```
{instance-specific issue statement}
```

The issue statement may include a title, description, current and expected behavior, requirements, and new interfaces. No additional task protocol or tool-call example is appended to this user message. Tool calls instead follow the Qwen3 XML format supplied by the chat template, for example:

```
<tool_call>
<function=execute_bash>
<parameter=command>pwd</parameter>
</function>
</tool_call>
```

ECHO-Specific Prompts and Hints. The prompts below are shared by both benchmarks. The summary protocol is appended to the benchmark's system message for ECHO and ECHO-Graph so that every turn records a reusable finding, and the hint is re-sent whenever a turn omits the block.

ECHO Extra System Prompt

Tool Result Summary Protocol
After receiving a tool/function response, your next assistant message must include exactly one `<sum_last_turn>...</sum_last_turn>` block before any new tool/function call.
The block should be one concise factual sentence summarizing only the latest tool/function result.
After the block, continue in the same assistant message with the next tool/function call, or call `finish` if the final answer is ready. Do not stop after the summary.

Sum Last Turn Hint

Briefly record the latest tool result in `<sum_last_turn>...</sum_last_turn>`, then continue with the next action.

The selection prompt is sent as a single user message whose literal text begins with the `System:` label below; the turn list and runtime hints are appended to the same message.

Selection Prompt

System:
Your operational context is full. Select prior interaction turns that are necessary to continue solving the task.
Rules:

- Output exactly one `<selection>...</selection>` block.
- Do not call any function/tool in this turn.
- Do not answer the original task.
- Do not include `<think>`, tool calls, markdown fences, or text outside the selection tags.
- Select only turns that contain reusable evidence, constraints, failed attempts, or the next planned action.
- Prefer older turns that are not already covered by automatically retained recent turns.
- If no older turn is necessary, return an empty `<selection></selection>` block.

Valid turns:
`{turn_list}`

Output format:

```
<selection>
[zero or more lines, each formatted as turn_N: reason]
</selection>
```

Selection Hint

Valid selection indices: `turn_0` through `turn_{n_hist-1}`.
Select at most `{selection_max_turns}` historical turns; output an empty selection if the automatically retained recent turns are enough.
The latest turns `turn_{recent_start}` through `turn_{n_hist-1}` will be retained automatically; select any additional older turns that are still needed.

Current local context (the action and tool result that caused this boundary): `{local_context}`

In our BrowseComp-Plus experiments, `selection_max_turns` = 8 and `echo_recent_turns` = 3.

A.3 Case Study: Preserving Evidence Across Context Reconstruction

Figure 7 contrasts Selection with Summary on a BrowseComp-Plus question asking for the founders of a European watch company associated with a mountain's first ascent. Both agents initially explore Annapurna and encounter a document linking Hermann Buhl's 1953 ascent of Nanga Parbat to Hercules watches and their founders, Robert Pfaff and Robert Henzi. ECHO-Graph initially retains early search turns, but its second selection explicitly preserves `turn_7`, an earlier visit to document 3986. The corresponding action–finding pair retains the source identifier and the specific Buhl–Nanga Parbat–Hercules association in the reconstructed input, keeping an alternative to the current hypothesis available for reconsideration. The agent subsequently reopens this source, confirms the founders, and returns the correct answer after two selection events. SUPO's summaries, in contrast, weaken the same evidence into a generic association with high-altitude expeditions, omitting the specific link between Buhl, the first ascent, and the watch he wore. More importantly, the summaries label Annapurna as `[VERIFIED]` while stating that its first recorded attempt was in 1935, directly conflicting with the question's requirement of an attempt before 1899.

Thus, an unresolved hypothesis is carried forward as an established fact. By the third summary, the founder names have also disappeared, and Hercules is retained only as a discrepancy against the Annapurna hypothesis. Even after encountering the correct source again, SUPO continues along the Annapurna–Lip branch and ultimately answers Fred Lip incorrectly. This case illustrates the advantage of retaining source-linked evidence that can be revisited, rather than repeatedly rewriting



Figure 7: **Evidence selection versus rolling summarization. Selection** preserves a source-linked historical finding and is followed by a source revisit confirming the correct founders, whereas successive **summaries** omit key relations and retain a contradictory claim as `[VERIFIED]`, culminating in an incorrect answer. The figure shows selected milestones from evaluation trajectories of separately trained policies; the two columns have independent timelines. The reference answer chain summarizes document 3986 and the benchmark answer, rather than four independently verified retrieval steps.

a working hypothesis into a summary that loses discriminative details and reinforces an erroneous conclusion.

A.4 Step-aligned Evaluation on a Single Question

Table 4 follows ECHO and ECHO-Graph across evaluation checkpoints on the same question, whose reference answer is "Trudy Bronner, Mathematics." Each entry describes one recorded rollout, not an aggregate accuracy estimate. We preserve the answer strings and report ECHO-Graph's reconstruction counts to distinguish correct answers obtained with and without reconstruction. Equal checkpoint step indices alone do not establish matched training or inference budgets.

Trudy Bronner Question. In a blog post submitted online sometime after 2010 and before 2015, the writer pays tribute to what would have been their deceased parent's birthday. The writer of this particular blog post also mentions their other parent, living as of October 28, 2018. This other parent of the blog writer received an annual award sometime after this blog post was submitted online and before 2020. What is this other parent's first and last name as of October 28, 2018, and what area of study had this person (the other parent) earned a college degree sometime before 2019?

Reference reasoning chain. The reference answer is **Trudy Bronner, Mathematics**: document 73065 (2012) identifies Jim Bronner's birthday tribute and names his wife, **Trudy Bronner**, as the writer's mother; document 31881 (2018) identifies Trudy as the recipient of the **Todd Gutschow Lifetime Leadership Award**; and document 20701 states that she graduated with a college degree in **mathematics**. Thus the evidence chain is **Jim's tribute** $\rightarrow$ **Trudy** $\rightarrow$ **2018 award** $\rightarrow$ **Mathematics**. This is an author-reconstructed reference path, not a claim that every rollout selected all three documents or followed this order; in particular, document 20701, not document 73065, supplies the explicit degree evidence.

Table 4: Step-aligned evaluation outputs on the Trudy Bronner question. Results follow the benchmark evaluator. The last column reports ECHO-Graph's serialized trajectory segments (S) and completed context reconstructions (R); these are not parallel solution branches.

Step	ECHO		ECHO-Graph		S/R
	Result	Answer	Result	Answer	
0	wrong	The other parent's name is Andrew Foster, and their college degree was in education.	wrong	The other parent's name is Katheryne Staeger-Wilson, and their area of study was Education or Public Service.	3/1
5	wrong	The other parent's name is W. Joseph Campbell, and their college degree was in mass communication.	wrong	The information could not be determined with the available tools and data.	1/0
10	wrong	The other parent's name is Harry Harris, and their area of study was art. However, no award information aligns with the specified timeframe.	wrong	The other parent's name is Thomas A. Salek, and their college degree was in Political Communication.	1/0
15	wrong	Unable to determine the parent's name and college degree area based on available information.	wrong	The other parent's name is Andrew Foster, and their college degree was in Education from Gallaudet University.	1/0
20	wrong	The other parent's name is Michèle K. Spike, and their area of study was law and art history.	wrong	Adrienne Keene, American Studies	3/1
25	wrong	Maria Gravel, education	wrong	Khin Yew Chong, Engineering	3/1
30	wrong	The other parent's name is Katheryne Staeger-Wilson, and their college degree was in Education.	wrong	The other parent's name is Jessica Turner, and their college degree was in Business Administration (MBA).	1/0
35	correct	Trudy Bronner, Mathematics	wrong	Khin Yew Chong, Engineering	3/1
40	wrong	The other parent's name is Tom and their college degree was in Communications.	wrong	The other parent's name is Nicholas Gotta Have Sole, and their area of study is child advocacy.	3/1
45	wrong	The other parent's name and college degree field could not be determined from the available information.	wrong	Katheryne Staeger-Wilson, Education	3/1
50	wrong	Dmitri Kharzeev, Physics	wrong	The other parent's name is Stacy Robison, and their area of study was Public Health (MPH).	1/0

Continued on the next page

Step	ECHO		ECHO-Graph		S/R
	Result	Answer	Result	Answer	
55	wrong	The other parent's name is Khin Yew Chong, and their area of study is engineering.	wrong	Kenneth M. Hargreaves, Neurobiology	5/2
60	wrong	The other parent's name is Rose Pulliam, and their area of study was horticulture.	wrong	The other parent's name is Glenda Harris, and her field of study was not explicitly mentioned in the available documents. However, based on contextual clues from the blog post, it is possible she had a background in a field related to her role as a caregiver and homemaker, though no specific degree was noted.	1/0
65	wrong	The other parent's name is Cassandra Bird, and their area of study is Political Communication.	wrong	The other parent's name is Tom and their area of study was not explicitly mentioned in the provided documents.	3/1
70	correct	Trudy Bronner, mathematics	correct	The other parent's name is Trudy Bronner, and their college degree was in mathematics.	1/0
75	wrong	The other parent's name is Rhonda Rapp, and their college degree was in an unspecified field at St. Mary's University.	wrong	W. Joseph Campbell, mass communication	3/1
80	wrong	The other parent's name is Andrew Foster, and their area of study was education.	wrong	The other parent's name is Harry Harris, and their area of study was art.	5/2
85	correct	Trudy Bronner, mathematics	correct	Trudy Bronner, Mathematics	3/1
90	wrong	Meredith Neville-Shepard, Political Communication	wrong	The other parent's first and last name as of October 28, 2018, is Kevin Mattson, and their field of study is History.	5/2
95	wrong	The other parent's name is Donald J. Harris, and their area of study is economics.	correct	Trudy Bronner, Mathematics	1/0
100	wrong	The other parent's name is D'Neil and Michael Duffy, and their area of study is education (Montessori).	correct	Trudy Bronner, Mathematics	1/0
105	wrong	The other parent's name is Rhonda Rapp, and their field of study is unknown based on available data.	correct	Trudy Bronner, Mathematics	11/5
110	correct	Trudy Bronner, first name Trudy, last name Bronner. Area of study: Mathematics.	correct	Trudy Bronner, Mathematics	11/5
115	wrong	Jordi Alonso, health-related quality of life (HRQL) research	correct	Trudy Bronner, Mathematics	3/1
120	wrong	Unable to determine the parent's name and field of study based on available information.	correct	Trudy Bronner, Mathematics	1/0
125	wrong	The other parent's name is Tanya Foster, and their area of study is unspecified in the available documents.	wrong	no submitted answer (overlong)	11/5
126	wrong	no submitted answer (overlong)	correct	Trudy Bronner, Mathematics	5/2

Note. S counts all serialized trajectory segments, including separately stored selection-generation segments; R counts completed context reconstructions. "No submitted answer (overlong)" is the viewer's termination status rather than a generated answer. Individual successes and failures are retained, including ECHO-Graph's unsuccessful rollout at step 125.

A shared evidence pair, not an identical complete trajectory. All nine successful ECHO-Graph rollouts retrieve both document 31881, which identifies Trudy Bronner as the recipient of the 2018 Todd Gutschow Lifetime Leadership Award and the mother of David and Mike Bronner, and document 20701, which states that she graduated "with a degree in mathematics." Seven also retrieve document 73065, the 2012 birthday tribute connecting Jim Bronner to "his wife, my mom, Trudy Bronner." The exceptions, steps 85 and 105, reach the correct benchmark answer without retrieving that blog document. ECHO's four successful rollouts also retrieve the award–degree pair, and its step 35 success retrieves all three documents. The difference is therefore not access to an exclusive source, and the early ECHO success is evidence-backed rather than an unsupported lucky answer.

Later-checkpoint consistency across reconstruction depths. Neither method becomes consistently correct immediately after its first success. The clear distinction is at later checkpoints: over the eight evaluations at steps 95–126, ECHO-Graph succeeds seven times, versus once for ECHO. This reflects the stable consistency of ECHO-Graph's results in the later stages of training. ECHO-Graph's nine successes overall span zero, one, two, and five context reconstructions, whereas all four ECHO successes are unreconstructed, indicating that ECHO-Graph's selection and localization capability can withstand fluctuations arising from the search process and other sources of randomness, demonstrating strong result reproducibility and generalization.