

Evidence-Augmented Policy Optimization with Reward Co-Evolution for Long-Context Reasoning

Xin Guan^{1*}, Zijian Li³, Shen Huang², Pengjun Xie², Jingren Zhou², Jiuxin Cao^{1†}

¹ School of Cyber Science and Engineering, Southeast University

² Tongyi Lab, Alibaba Group

³ Hong Kong University of Science and Technology

{xin_guan, jx.cao}@seu.edu.cn, zijian.li@connect.ust.hk

{pangda, chengchen.xpj, jingren.zhou}@alibaba-inc.com

Abstract

While Reinforcement Learning (RL) has advanced LLM reasoning, applying it to long-context scenarios is hindered by sparsity of outcome rewards. This limitation fails to penalize ungrounded “lucky guesses,” leaving the critical process of needle-in-a-haystack evidence retrieval largely unsupervised. To address this, we propose **EAPO** (Evidence-Augmented Policy Optimization). We first establish the Evidence-Augmented Reasoning paradigm, validating via Tree-Structured Evidence Sampling that precise evidence extraction is the decisive bottleneck for long-context reasoning. Guided by this insight, EAPO introduces a specialized RL algorithm where a reward model computes a Group-Relative Evidence Reward, providing dense process supervision to explicitly improve evidence quality. To sustain accurate supervision throughout training, we further incorporate an Adaptive Reward-Policy Co-Evolution mechanism. This mechanism iteratively refines the reward model using outcome-consistent rollouts, sharpening its discriminative capability to ensure precise process guidance. Comprehensive evaluations across eight benchmarks demonstrate that EAPO significantly enhances long-context reasoning performance compared to SOTA baselines.

1 Introduction

The development of Large Language Models (LLMs) has profoundly revolutionized Natural Language Processing (NLP) (Xu et al., 2025). Models such as GPT (Agarwal et al., 2025) and DeepSeek-R1 (DeepSeek-AI et al., 2025) have propelled significant advancements in domains ranging from complex dialogue to summarization. However, as LLMs evolve into autonomous agents required to

* Work done during the author’s internship at Tongyi Lab.

† Corresponding author.

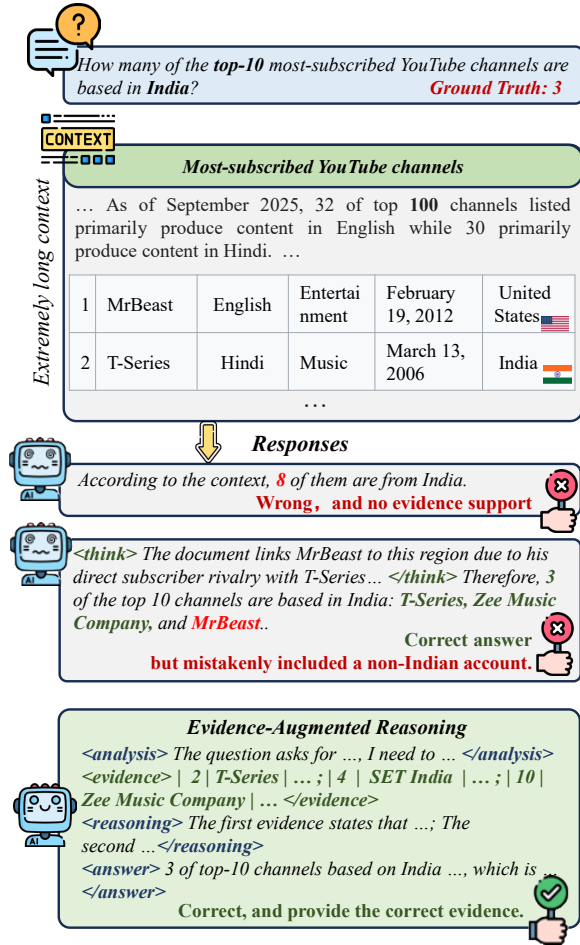


Figure 1: Challenges in reasoning over extremely long contexts. The examples illustrate critical failure modes, including answers with no evidence support, incomplete or error retrieval. Our work explicitly targets this bottleneck to ensure high-fidelity evidence extraction.

interact with dynamic environments, the demand for processing voluminous information has escalated. This imposes a critical challenge: as context length increases, model performance often suffers from severe degradation due to the “lost in the middle” phenomenon (Liu et al., 2024). In these high-noise scenarios, models frequently exhibit specific failure modes illustrated in Figure 1: they may gen-

erate answers with **no evidence support**, rely on **fragmentary retrieval**, or make **erroneous citations**, failing to discriminate “needle-in-a-haystack” facts from irrelevant distractors.

Reinforcement Learning (RL) (Zhang et al., 2025b) has emerged as a powerful tool for unlocking intrinsic reasoning capabilities. However, existing approaches in Long-Context Reasoning (LCR) predominantly rely on **outcome-based rewards** (Wan et al., 2025; Wang et al., 2026), guiding optimization solely based on the correctness of the final answer. This reliance exacerbates reward sparsity and may introduce noise: a model may correctly guess an answer via a flawed reasoning path (Shao et al., 2025). For instance, as shown in Figure 1, a model might correctly answer “3 channels” but cite an irrelevant entity (e.g., *MrBeast*) instead of the correct Indian channel. Such “shortcuts” satisfy the binary outcome reward but fail to generalize, leaving the underlying retrieval failure unaddressed.

To resolve these issues, we argue for a shift from result-oriented to evidence-augmented process-oriented supervision for LCR. To operationalize this, we first establish the Evidence-Augmented Reasoning (EAR) paradigm. By strictly enforcing a workflow where explicit evidence extraction precedes reasoning execution, EAR inherently resolves the problem of ungrounded responses, ensuring that no conclusion is derived without explicit citation. Building on this structured foundation, we conduct a preliminary study via Tree-Structured Evidence Sampling, which employs BFS to explore distinct evidence trajectories, evaluated via both semantic similarity and fine-grained LLM judgments. This analysis empirically validates our core hypothesis: securing high-quality evidence is the decisive prerequisite for unlocking superior performance in LCR.

Building on these insights, we propose **EAPO** (Evidence-Augmented Policy Optimization). EAPO is designed to internalize the benefits of precise evidence discovery directly into the model’s policy. It introduces a dense, group-relative evidence reward to provide fine-grained supervision at the process level, ensuring the model is rewarded not just for being right, but for being right for the high-quality evidence. However, relying on a static reward model carries risks, as the policy improves, a fixed evaluator may fail to distinguish subtle quality differences. To address this, we further incorporate an Adaptive Reward-Policy Co-Evolution mechanism. This

creates a self-reinforcing cycle where the reward model is iteratively refined using high-confidence, outcome-consistent rollouts generated by the policy itself. This continuous alignment sharpens the reward model’s discriminative capability, enabling it to assess complex evidence with increasing precision and dynamically synchronize its supervision signal with the evolving policy.

The core contributions of our paper are:

- We establish the EAR paradigm to enforce explicit evidence extraction prior to reasoning, and empirically identify this extraction step as the decisive bottleneck through Tree-Structured Evidence Sampling.
- We introduce the **EAPO** framework, shifting optimization from sparse outcome rewards to dense, evidence-augmented supervision. By leveraging a reward model to compute a group-relative evidence reward, EAPO guides the model to extract high-quality evidence, thereby enhancing reasoning capabilities.
- We design an Adaptive Reward-Policy Co-Evolution mechanism to establish a self-reinforcing loop. By iteratively refining the reward model with outcome-consistent rollouts, this mechanism sustains high-fidelity supervision.

2 Empirical Motivation: The Primacy of Evidence

Before detailing our reinforcement learning framework, we first establish empirical foundations of our approach. We introduce the Evidence-Augmented Reasoning (EAR) paradigm to structure the long-context workflow and utilize a Tree-Structured Evidence Sampling to validate a critical hypothesis that *evidence extraction is the decisive bottleneck in long-context reasoning*.

2.1 Evidence-Augmented Reasoning

We propose the **Evidence-Augmented Reasoning (EAR)** paradigm, as shown in Figure 1, which explicitly decouples the retrieval of information from its logical manipulation. EAR enforces a strict four-stage workflow: 1) **Task Analysis** <analysis>: Deconstructing the user’s query constraints. 2) **Evidence Extraction** <evidence>: Locating and quoting relevant text segments verbatim from the context. 3) **Reasoning Execution** <reasoning>: Deriving conclusions based exclusively on the extracted

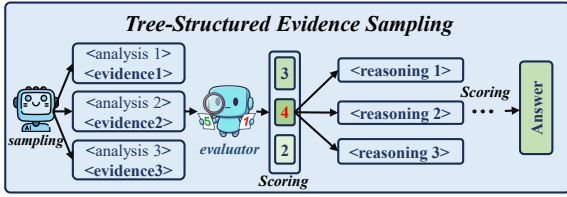


Figure 2: Overview of Tree-Structured Evidence Sampling. The process samples diverse evidence and reasoning paths, evaluates their scores, and derives the final answer.

evidence. 4) **Answer Formulation** `<answer>`: Synthesizing the final response. This structural decomposition, operationalized via special tokens, not only focuses attention but crucially exposes the intermediate evidence state for direct supervision.

2.2 Preliminary Experiment on Evidence

Tree-Structured Evidence Sampling. To quantify the critical role of evidence quality in determining final reasoning outcomes, we design a systematic exploration strategy modeled as a decision tree. As shown in Figure 2, we employ Breadth-First Search (BFS) to sample k distinct evidence trajectories for a given query and context, creating a diverse search space. To efficiently value these intermediate nodes without redundant full-context processing, we apply a Query-Centric Node Evaluation that scores candidates based strictly on their utility relative to the query. We investigate two specific valuation paradigms to identify the better evidence:

- **Semantic Similarity Assessment (SSA):** Using a reranker (Qwen3-8B-Reranker (Zhang et al., 2025c)) to measure the coarse-grained embedding alignment between the query and the extracted evidence.
- **LLM-as-a-Judge (LLM-E/R):** Leveraging an LLM to perform fine-grained evaluation on the utility of the extracted evidence (LLM-E) or the coherence of the reasoning (LLM-R) (Prompts in Appendix E).

Experiment Setting To quantify the factors constraining long-context performance, we conduct a controlled analysis on the Musique from LongBench benchmark (Bai et al., 2024) using Qwen3-30B-A3B-Instruct (Yang et al., 2025) as the base model. To establish an empirical performance ceiling, we construct an Evidence-Grounded (Oracle) baseline, where high-fidelity evidence is extracted

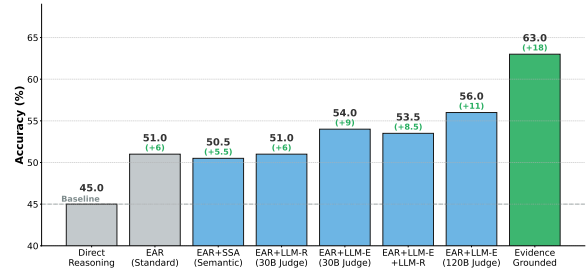


Figure 3: Performance analysis of Qwen3-30B-A3B-Instruct on Musique in LongBench. We compare Direct Reasoning, EAR variants with different evaluators, and an Oracle upper bound (*Evidence-Grounded*).

by a superior model (gemini-2.5-pro (DeepMind, 2025)) utilizing both the query and ground-truth answer, and then fed to the base model for inference. We employ Tree-Structured Sampling with a beam width of $k = 3$. As illustrated in Figure 3, our analysis yields the following insights:

Evidence Extraction is the Decisive Bottleneck.

Two observations isolate evidence retrieval as the primary bottleneck. First, providing high-fidelity evidence (Oracle) drastically elevates performance to 63.0%. Second, optimizing solely the reasoning step (EAR+LLM-R) yields no improvement over standard EAR. Moreover, adding a reasoning judge atop evidence evaluation (EAR+LLM-E+LLM-R) has a negligible effect, causing a slight performance drop compared to LLM-E alone. We attribute this to restricted reasoning diversity: once evidence is fixed, subsequent reasoning paths become highly homogeneous (>96% semantic consistency; see Appendix B), rendering LLM-R redundant. This implies that once high-quality evidence is secured, inherent reasoning capabilities suffice; finding the evidence is the true bottleneck.

Superior Process Supervision Drives Greater Performance Gains.

Our analysis reveals a direct correlation between the fidelity of process supervision and the magnitude of performance improvement. While SSA offers limited guidance (+5.5%), employing an LLM-based judge to provide process supervision significantly improves performance (+9.0%). Furthermore, elevating the supervision quality with a stronger gpt-oss-120B evaluator yields even greater gains (+11.0%).

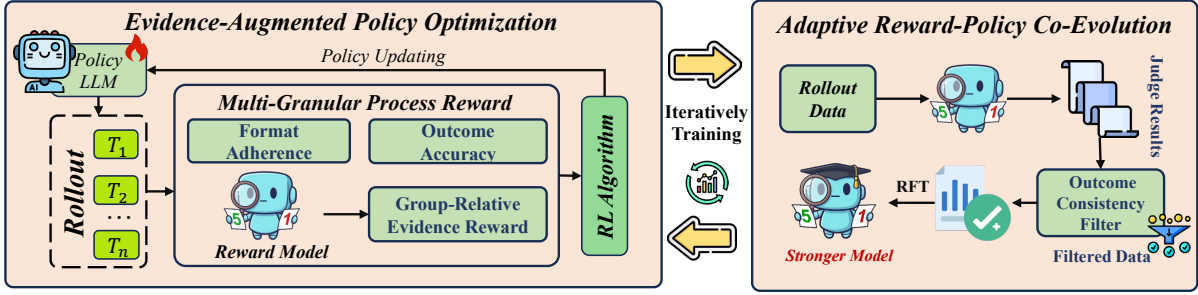


Figure 4: Overview of EAPO. **Left:** The Evidence-Augmented Policy Optimization phase, utilizing dense evidence scores to guide the high-quality evidence extraction process. **Right:** The Adaptive Reward-Policy Co-Evolution cycle, which iteratively refines the reward model to sharpen its discriminative capability for accurate supervision.

3 Evidence-Augmented Policy Optimization

Building on the empirical findings in Section 2.2, we propose Evidence-Augmented Policy Optimization (EAPO). Guided by the insight that evidence extraction is the decisive bottleneck, EAPO shifts supervision from sparse outcome signals to dense, process-oriented guidance. Furthermore, to leverage the critical correlation between evaluator strength and performance, we incorporate an Adaptive Reward-Policy Co-Evolution mechanism. This establishes a self-reinforcing loop where the reward model’s discriminative power is iteratively sharpened in tandem with the policy. The framework’s overview is depicted in Figure 4.

3.1 Preliminaries

Group Relative Policy Optimization (GRPO). We adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024) as the core reinforcement learning algorithm. Formally, for each query q sampled from the dataset distribution $P(Q)$, we generate a group of G outputs $\{o_1, \dots, o_G\}$ utilizing the old policy $\pi_{\theta_{old}}$. The optimization objective is to maximize the following surrogate loss function:

$$\mathcal{L}_{GRPO}(\theta) = \mathbb{E}_{q \sim P(Q), \{o_i\} \sim \pi_{\theta_{old}}} \left[\frac{1}{G} \sum_{i=1}^G \min \left(r_i(\theta) \hat{A}_i, \text{clip}(r_i(\theta), 1 - \epsilon_{low}, 1 + \epsilon_{high}) \hat{A}_i \right) \right] \quad (1)$$

where $r_i(\theta) = \frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}$ denotes the probability ratio between the current and old policies. The hyperparameters ϵ_{low} and ϵ_{high} constrain the policy update step size, set to 0.2 and 0.28 respectively. \hat{A}_i represents the advantage value for the i -th output,

which is derived from the standardized total reward: $\hat{A}_i = \frac{R_{total}^{(i)} - \text{mean}(R_{total})}{\text{std}(R_{total})}$. $R_{total}^{(i)}$ denotes the reward for the i -th sample. We define the specific composition of this fine-grained process reward R_{total} in the subsequent section.

3.2 Multi-Granular Process Reward Mechanism

EAPO constructs a dense, process reward R_{total} . This composite signal is composed of distinct components designed to guide the model through the structured Evidence-Augmented workflow:

Format Adherence (R_f). To ensure the output conforms to the EAR structure, we verify the presence of specific delimiters (e.g., <evidence>). If the format is correct, we assign $R_f = 1$; otherwise, $R_f = 0$. If the format is invalid, the evaluation terminates immediately, and the remaining reward components are not calculated.

Group-Relative Evidence Quality (R_e). This component serves as an intermediate process reward designed to assess the quality of the extracted evidence. Specifically, we extract all evidence segments generated within all rollouts and input this collective set simultaneously into the Reward Model θ_{RM} for assessment. The RM assigns an integer utility score $v \in \{1, \dots, 5\}$ (Prompt in the Appendix E). These scores are then normalized within the sampled group to yield $R_e \in [0, 1]$, enabling the policy to prioritize superior evidence through comparative feedback.

Outcome Accuracy (R_a). Following (Wan et al., 2025), we employ an LLM-based evaluator to assess answer correctness. This module assigns a binary reward by verifying if the generated answer is semantically consistent with the ground truth.

Total Reward. The composite reward is a weighted linear combination:

$$R_{total} = \alpha R_f + \beta R_e + \gamma R_a \quad (2)$$

where α, β, γ are hyperparameters balancing structural compliance, evidence quality, and final accuracy.

3.3 Adaptive Reward-Policy Co-Evolution

Our preliminary analysis in Section 2.2 revealed a critical phenomenon: the efficacy of policy model optimization is positively correlated with the quality of process supervision. To improve the capabilities of the process Reward Model (RM), EAPO incorporates an Adaptive Reward-Policy Co-Evolution mechanism, establishing a dynamic closed-loop where the RM iteratively improves alongside the Policy. This cycle proceeds in two key phases:

Exploratory Rollout and Scoring. During the optimization phase, the current policy π_θ generates diverse reasoning trajectories for each query. The current RM evaluates the evidence sets extracted within these rollouts, assigning integer quality scores to guide the immediate policy update.

Outcome-Consistent Rejection Fine-Tuning (RFT). To prevent supervision degradation, we employ a strategy inspired by Rejection Fine-Tuning (RFT). Rather than naively training on all correct paths, we apply a strict Outcome Consistency filter to curate a high-fidelity dataset. We retain only those trajectories where the RM’s judgment aligns with objective reality. Specifically, instances where high evidence scores correspond to correct answers and low scores to incorrect ones. These validated High-Confidence data are then used to fine-tune the Reward Model parameters θ_{RM} via a supervised objective:

$$\theta_{RM} \leftarrow \arg \min_{\theta} \mathbb{E}_{(q,o,y) \sim \mathcal{D}_{high}} [\mathcal{L}(R_\theta(q, o), y)] \quad (3)$$

where q denotes the input query, o represents the evidence set, and y indicates the target prediction derived from the consistency check. This process continuously sharpens the RM’s discriminative boundary, ensuring that the supervision signal remains accurate and precise as the policy explores increasingly complex reasoning patterns.

By synchronizing the evolution of the reward model with the policy, EAPO creates a self-

reinforcing cycle that guides the model toward autonomous, high-fidelity evidence discovery.

4 Experiments

Benchmarks We evaluate our model on a comprehensive suite of eight long-context reasoning benchmarks, categorized into three types: (1) SEAL-0 and (2) SEAL-Hard from SEAL (Pham et al., 2025), which represent challenging reasoning QA tasks¹; a collection of multi-hop QA datasets from LongBench-V1 (Bai et al., 2024), specifically (3) HotpotQA, (4) MuSiQue, and (5) 2WikiMulti-hopQA; and three diverse tasks from LongBench-V2 (Bai et al., 2025): (6) Single-Document QA (SDQ), (7) Multi-Document QA (MDQ), and (8) Long Structured-Data QA (LSA).

Baselines We compare our approach against state-of-the-art models divided into three distinct categories:

- *Compared LLMs:* High-performing closed-source models including GPT-4o (Hurst et al., 2024), Gemini-2.0-Flash (Team, 2025), Qwen3-Plus, and Claude-Sonnet-4 (Anthropic, 2025). Leading open weights models such as Qwen3-14B, Qwen3-30B-A3B-Instruct, Qwen3-30B-A3B-Thinking (Yang et al., 2025), and GPT-OSS-120B (Agarwal et al., 2025). **QwenLong-32B**, the SOTA model for specialized long-context reasoning trained via RL.
- *Control Baselines (GRPO):* To rigorously validate our algorithm, we include variants of our base models trained using *only* the outcome-based reward via standard GRPO. These serve as the primary control group to isolate the specific contribution of our evidence-based process rewards.

Evaluation & Training Details We employ gpt-4o-2024-11-20 as the judge model to evaluate answer accuracy; the exact evaluation prompt is provided in the Appendix E. To ensure consistent long-context assessment, we restrict all training and evaluation samples to a maximum context length of 128k tokens. For training, we utilize three base models to cover diverse scales, architectures, and reasoning patterns: Qwen3-14B (Dense), Qwen3-30B-A3B-Instruct (MoE), and Qwen3-30B-A3B-Thinking (Yang et al., 2025). The Reward Model

¹SEAL provides the Wikipedia URLs for the answers. We employ Jina to extract the page content as the context.

Models	SEAL		LongBench			LongBench-v2			AVG
	Seal-0	Seal-hard	Musique	Hotpotqa	2wiki	SDQ	MDQ	LSA	
<i>Comparison with other LLMs</i>									
GPT-4o	19.8	39.4	60.0	78.0	79.5	47.0	<u>50.0</u>	44.4	52.3
Gemini-2.0-Flash	24.3	36.6	52.5	75.0	84.5	43.1	44.1	48.2	51.0
Qwen3-Plus	24.3	32.3	56.5	71.5	80.0	50.3	44.1	55.6	51.8
Claude-Sonnet-4	<u>42.3</u>	56.3	60.5	81.5	81.0	50.3	48.0	<u>66.7</u>	<u>60.8</u>
GPT-OSS-120B	39.6	55.5	64.0	80.0	87.5	45.7	46.1	63.0	60.2
QwenLong-32B	40.5	<u>56.7</u>	<u>64.5</u>	77.5	84.5	49.0	52.9	37.0	57.8
<i>Comparison among Our Models</i>									
Qwen3-14B	35.1	49.6	56.0	79.5	83.5	40.4	42.2	38.9	53.2
+ GRPO	35.1	51.0	57.0	80.0	83.0	40.7	42.4	40.9	53.8 (+0.6)
+ EAPO	36.0	52.5	57.0	<u>81.0</u>	83.5	43.0	44.1	42.9	55.0 (+1.8)
Qwen3-30B-A3B-Instruct	23.4	30.3	45.0	73.0	75.0	39.7	43.1	55.6	48.1
+ GRPO	31.5	40.2	51.5	75.5	85.5	41.1	43.6	55.6	53.1 (+5.0)
+ EAPO	33.0	44.1	59.5	77.5	87.5	42.4	45.1	55.6	55.6 (+7.5)
Qwen3-30B-A3B-Thinking	33.3	55.5	60.0	79.0	84.5	47.0	45.1	59.3	58.0
+ GRPO	40.4	53.8	60.5	78.5	85.0	49.7	43.1	63.0	59.2 (+1.2)
+ EAPO	44.1	57.9	65.0	80.0	<u>86.5</u>	51.7	49.0	70.4	63.1 (+5.1)

Table 1: Main results on long-context reasoning benchmarks. The best result is highlighted in bold, and the second best result is shown underlined.

is initialized from the Qwen3-30B-A3B-Thinking. Regarding the reward coefficients, we set the format weight $\alpha = 0.1$ following Sheng et al. (2024), while $\beta = 0.3$ and $\gamma = 0.6$ are determined based on our subsequent hyperparameter analysis. The Reward Model is updated every 20 RL steps using the most recently collected high-quality rollouts. Detailed hyperparameters are listed in the Appendix A.

Training Dataset We construct a composite training dataset totaling 4,664 samples, derived from two sources. First, adopting the long-context construction method of Zhu et al. (2025), we process the MuSiQue dataset by padding the context to a range of 32k–128k tokens. Second, we incorporate a Wikipedia-based Mixed QA subset, comprising both structured QA (reasoning over tables) and unstructured QA (free-text extraction). Detailed construction process in Appendix A. This hybrid composition ensures the model adapts to heterogeneous evidence formats.

4.1 Main Results

The overall performance comparisons are presented in Table 1. Our analysis yields three significant observations:

Robust Improvements Across Diverse LLMs. The proposed EAPO framework demonstrates con-

sistent superiority over base models across all tested architectures. For Qwen3-30B-Instruct, our method achieves a remarkable average improvement of +7.5%, boosting the score from 48.1% to 55.6%. Even on the highly capable Qwen3-30B-Thinking model, which already possesses advanced reasoning capabilities, EAPO delivers a substantial +5.1% gain, achieving a score of 63.1%. This confirms that optimizing the evidence extraction process is universally effective, functioning independently of model size (14B vs. 30B) or architecture type (Dense vs. MoE).

Efficacy of Evidence-Augmented Supervision. Comparing outcome-based GRPO with EAPO underscores the critical value of fine-grained guidance. While GRPO yields only marginal gains (e.g., +1.2% for 30B), EAPO’s dense evidence reward drives significant improvements—notably surpassing the GRPO baseline by over 4 points on SEAL-Hard (57.9% vs. 53.8%). This confirms that explicit evidence supervision effectively curbs parametric shortcuts, fostering robust and generalizable reasoning.

Competitiveness with Larger and Proprietary Models. Remarkably, our EAPO-30B-Thinking model (63.1%) outperforms significantly larger open-source models like GPT-OSS-120B (60.2%) and specialized long-context models

like QwenLong-32B (57.8%). Furthermore, it surpasses powerful proprietary models such as Gemini-2.0-Flash (51.0%), GPT-4o (52.3%), and Claude-Sonnet-4 (60.8%) on average, highlighting the efficiency of evidence-augmented optimization.

5 Analysis Experiment

Experiments in this section are conducted using the Qwen3-30B-A3B-Thinking model. We adopt this unified setting to rigorously assess the individual contributions of the EAPO framework components.

5.1 Ablation Study

To validate EAPO, we compare it against the GRPO outcome reward baseline and a static reward model variant (w/o RM Co-Evolution) in Figure 5 and 6.

Faster Convergence and Higher Ceiling. Figure 5 shows that EAPO achieves superior sample efficiency, rapidly outpacing GRPO. This confirms that dense process supervision provides clearer gradient signals than sparse outcome rewards. Crucially, while the static variant saturates after step 50, the full EAPO framework driven by Co-Evolution, sustains its upward trajectory, achieving a higher performance ceiling.

Evidence Supervision Enhances Retrieval. As shown in Figure 6, GRPO’s implicit optimization of evidence is inefficient. In contrast, EAPO maintains a substantial and consistent lead in evidence scores by explicitly targeting evidence utility. This validates that the dense reward mechanism acts as a critical accelerator, ensuring the model prioritizes high-fidelity retrieval from the outset. To rigorously validate that these model-based score improvements reflect true evidence quality, we conducted a targeted human verification study. As detailed in Appendix C, human annotators confirmed that 97.3% of the evidence extracted by EAPO strictly supports the final answer, outperforming the GRPO baseline (92.6%).

5.2 Reward Model Analysis

To quantify the gain of Reward Model, we evaluate the RM’s Best-of-N Selection Accuracy ($N = 6$) followed by Liu et al. (2025). Using a set of samples with grounded evidence constructed in Section 2.2, we define the Oracle optimum among 6 sampled trajectories as the one maximizing ROUGE-L Recall. Accuracy is measured by the frequency

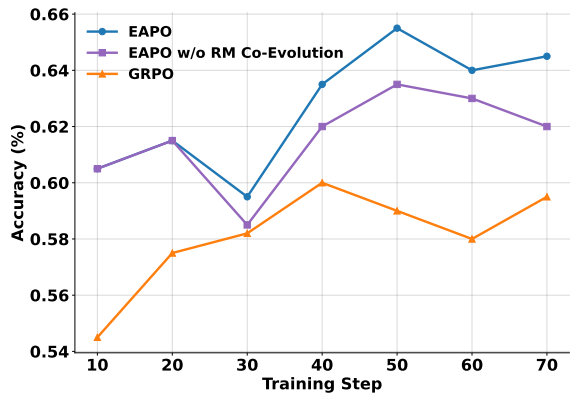


Figure 5: Task performance on the MuSiQue in Long-Bench. EAPO achieves superior convergence compared to the static reward baseline and the outcome-only GRPO.

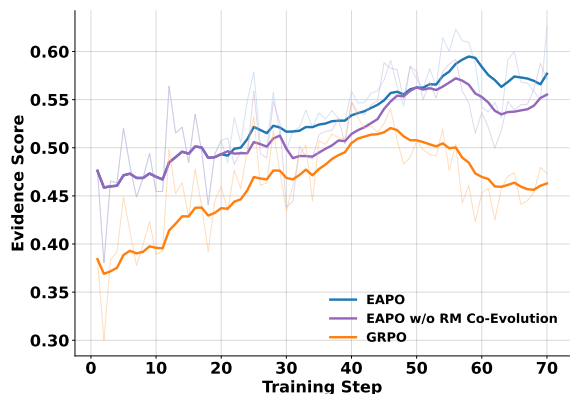


Figure 6: Evidence Quality scores during training. EAPO explicitly optimizes the retrieval process, leading to consistently higher-quality evidence extraction compared to the outcome-based GRPO baseline.

with which the RM’s top-ranked candidate matches this Oracle. As shown in Figure 7, the accuracy improves from 69.0% to **74.0%** over 60 steps, validating that co-evolution effectively aligns the RM with high-fidelity evidence quality.

5.3 Hyperparameter Experiment

We analyze the impact of β , which balances the evidence reward (R_e) against the outcome reward ($0.9 - \beta$). As shown in Figure 8, increasing β from 0.0 (outcome-only) to 0.3 yields consistent gains across all benchmarks, with Overall accuracy peaking at 63.1%. This confirms the value of dense evidence supervision. However, further increasing β to 0.5 causes a sharp decline back to 59.2%, suggesting that over-weighting intermediate steps distracts from the final answer. Thus, $\beta = 0.3$ offers the optimal trade-off between process and

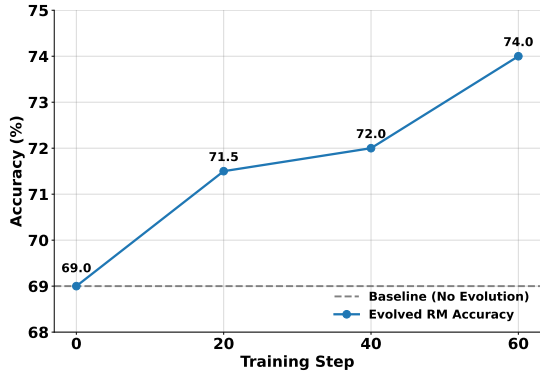


Figure 7: Evolution of RM Best-of-N Selection Accuracy ($N = 6$).

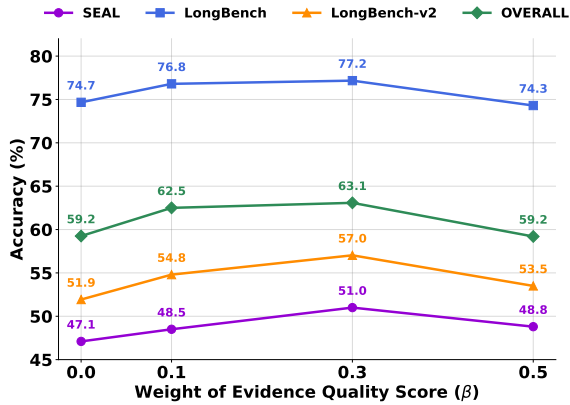


Figure 8: Comparison of different evidence quality reward weights.

outcome supervision.

5.4 Error Analysis

We conducted a fine-grained error analysis on LongBench-v2, centering our investigation on two critical failure modes: Evidence Error (failed or partial retrieval) and Reasoning Error (correct evidence but flawed logic). As visualized in Figure 9, EAPO demonstrates a simultaneous reduction in both highlighted categories: Evidence Errors decline from 17.7% to 13.5%, and Reasoning Errors from 20.7% to 15.4%. This trend validates our core hypothesis that dense process supervision not only minimizes retrieval failures but also generates a cascading positive effect, grounding subsequent reasoning in higher-fidelity information to mitigate downstream logic errors.

6 Related Work

6.1 Long-Context Reasoning

Long-context reasoning is a prominent area of focus in the development of LLMs, as the context

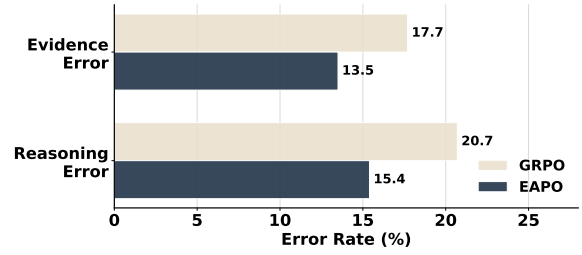


Figure 9: Error analysis on LongBench-v2. The comparison focuses on the reduction of Evidence and Reasoning errors achieved by EAPO relative to the GRPO baseline.

window determines the upper limit of a model’s single-pass reasoning capabilities. Mainstream approaches generally fall into four categories: 1) Retrieval-Augmented Generation (RAG) (Xu et al., 2024; Jin et al., 2025b), which relies on external retrievers that may become bottlenecks; 2) Agent-based Frameworks (Zhang et al., 2024; Yu et al., 2025), which decompose contexts via multiple agents but incur high computational latency; and 3) Architectural Modifications (Chen et al., 2024; Ding et al., 2024), which often struggle with complexity constraints. Distinct from these patches, our research focuses on 4) Reinforcement Learning (RL). Recent long-context RL methods primarily rely on sparse outcome signals or data enhancements. For instance, LongReward (Zhang et al., 2025a) optimizes automated outcome metrics via DPO, LoongRL (Wang et al., 2026) synthesizes complex reasoning data for standard GRPO, and QwenLong-11 (Wan et al., 2025) mitigates instability in outcome-based training. While effective, they inherently struggle with ungrounded reasoning. EAPO distinguishes itself by injecting dense, process-oriented supervision directly into the RL framework to explicitly tackle this challenge.

6.2 Reinforcement Learning and PRMs

RL has proven effective for enhancing LLM reasoning (DeepSeek-AI et al., 2025; Jin et al., 2025a). Beyond simple outcome-based rewards, Process-based Reward Models (PRMs) (Lightman et al., 2023; Khalifa et al., 2025) provide more granular, step-by-step guidance. However, recent PRM advancements remain heavily concentrated in structured domains like mathematics (e.g., Qwen2.5-Math (Zhang et al., 2025d), DeepSeekMath-v2 (Shao et al., 2025), LSRL (Ren, 2025)) and code generation (e.g., Posterior-GRPO (Fan et al., 2025)), which inherently allow for rigorous logical

verification. Because long-context reasoning lacks such strict linear verifiability, PRMs have been historically absent from this field. EAPO bridges this critical gap. Driven by our empirical finding that evidence extraction is the decisive bottleneck, we pioneer the PRM paradigm in long-context reasoning by isolating and explicitly rewarding this specific extraction process. Furthermore, we introduce an Adaptive Reward-Policy Co-Evolution mechanism to dynamically maintain PRM accuracy, effectively preventing reward degradation during training.

7 Conclusion

In this work, we introduced Evidence-Augmented Policy Optimization (EAPO) to overcome the limitations of sparse outcome rewards in long-context reasoning. Anchored in the Evidence-Augmented Reasoning paradigm, EAPO shifts the optimization focus from final results to process quality, utilizing a Reward Model to provide dense Group-Relative Evidence Rewards. Crucially, to sustain rigorous supervision, we integrated an Adaptive Reward-Policy Co-Evolution mechanism, which iteratively sharpens the reward model’s discriminative capability in tandem with the policy. Extensive evaluations demonstrate that EAPO significantly outperforms state-of-the-art baselines, paving the way for trustworthy, precision-centric autonomous agents.

Limitations

While EAPO demonstrates significant advancements in long-context reasoning, we acknowledge several limitations inherent to our approach. First, the efficacy of our method relies on the intrinsic capability of the base model to function as a discriminator. While we validated effectiveness on models in 30B, extremely small or weak models may struggle to provide the high-fidelity supervision signals required for the co-evolution loop to converge effectively. Second, our evaluation primarily focuses on multi-hop reasoning and extractive QA tasks. While these are critical benchmarks, the applicability of EAPO to other domains, such as open-ended creative writing or repository-level code generation, remains a subject for future exploration.

Ethics Statement

Long-context Reasoning is a significant area in natural language processing research. The test sets utilized for our evaluation are derived from publicly available and widely used datasets within the

community. For model training, we utilized data constructed entirely from public Wikipedia content. Similarly, the LLMs we employ are publicly accessible and have been developed and evaluated by the research community.

For the manual evaluation of evidence quality, we engaged three volunteer annotators from our research group with NLP backgrounds. Participation was strictly voluntary, and as internal academic collaboration, no monetary compensation was provided. The annotators were informed of the research objectives prior to the task. The annotation process was straightforward, relying exclusively on the single binary evaluation question detailed in Appendix C, thus requiring no additional instructional materials. As the task involved evaluating AI-generated text based on public datasets and collected no personal or sensitive information, formal institutional ethics review was deemed exempt.

Acknowledgement

We would like to thank the Tongyi Lab team at Alibaba Group for their valuable support and helpful discussions. This work is supported by National Natural Science Foundation of China under Grants No.62472092, No.62172089, No.62106045. Natural Science Foundation of Jiangsu Province under Grants No.BK20241751. Jiangsu Provincial Key Laboratory of Computer Networking Technology. Jiangsu Provincial Key Laboratory of Network and Information Security under Grants No.BM2003201, and Key Laboratory of Computer Network and Information Integration of Ministry of Education of China under Grants No.93K-9, Nanjing Purple Mountain Laboratories, Fintech and Big Data Laboratory of Southeast University.

References

- Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, et al. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *CoRR*, abs/2508.10925.
- Anthropic. 2025. [Introducing claude 4](#).
- Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. [Longbench: A bilingual, multi-task benchmark for long context understanding](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1:*

- Long Papers*), *ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3119–3137. Association for Computational Linguistics.
- Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025. [Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 3639–3664. Association for Computational Linguistics.
- Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2024. [Longlora: Efficient fine-tuning of long-context large language models](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Google DeepMind. 2025. [Gemini 2.5](#).
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, et al. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *CoRR*, abs/2501.12948.
- Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. [Longrope: Extending LLM context window beyond 2 million tokens](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.
- Lishui Fan, Yu Zhang, Mouxiang Chen, and Zhongxin Liu. 2025. [Posterior-grpo: Rewarding reasoning processes in code generation](#). *CoRR*, abs/2508.05170.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. 2025a. [Search-r1: Training llms to reason and leverage search engines with reinforcement learning](#). *CoRR*, abs/2503.09516.
- Jiajie Jin, Xiaoxi Li, Guanting Dong, Yuyao Zhang, Yutao Zhu, Yongkang Wu, Zhonghua Li, Ye Qi, and Zhicheng Dou. 2025b. [Hierarchical document refinement for long-context retrieval-augmented generation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 3502–3520. Association for Computational Linguistics.
- Gregory Kamradt. 2023. [Needle in a haystack - pressure testing llms](#).
- Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, and Lu Wang. 2025. [Process reward models that think](#). *CoRR*, abs/2504.16828.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. [Let’s verify step by step](#). *CoRR*, abs/2305.20050.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Trans. Assoc. Comput. Linguistics*, 12:157–173.
- Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025. [Inference-time scaling for generalist reward modeling](#). *CoRR*, abs/2504.02495.
- Thinh Pham, Nguyen Nguyen, Pratibha Zunjare, Weiyuan Chen, Yu-Min Tseng, and Tu Vu. 2025. [Sealqa: Raising the bar for reasoning in search-augmented language models](#). *CoRR*, abs/2506.01062.
- Hangliang Ren. 2025. [LSRL: process-supervised GRPO on latent recurrent states improves mathematical reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 12534–12545. Association for Computational Linguistics.
- Zhihong Shao, Yuxiang Luo, Chengda Lu, ZZ Ren, Jiewen Hu, Tian Ye, Zhibin Gou, Shirong Ma, and Xiaokang Zhang. 2025. [Deepseekmath-v2: Towards self-verifiable mathematical reasoning](#). *arXiv preprint arXiv:2511.22570*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *CoRR*, abs/2402.03300.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. [Hybridflow: A flexible and efficient rlhf framework](#). *arXiv preprint arXiv:2409.19256*.
- Gemini Team. 2025. [Gemini-2.0-flash](#).
- Fanqi Wan, Weizhou Shen, Shengyi Liao, Yingcheng Shi, Chenliang Li, Ziyi Yang, Ji Zhang, Fei Huang, Jingren Zhou, and Ming Yan. 2025. [Qwenlong-1l: Towards long-context large reasoning models with reinforcement learning](#). *CoRR*, abs/2505.17667.
- Siyuan Wang, Gaokai Zhang, Li Lyna Zhang, Ning Shang, Fan Yang, Dongyao Chen, and Mao Yang. 2026. [Loongrl: Reinforcement learning for advanced](#)

- reasoning over long contexts. In *The Fourteenth International Conference on Learning Representations, ICLR 2026*.
- Fengli Xu, Qian Yue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Si-jian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, and Yong Li. 2025. [Towards large reasoning models: A survey of reinforced reasoning with large language models](#). *CoRR*, abs/2501.09686.
- Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2024. [Retrieval meets long context large language models](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, et al. 2025. [Qwen3 technical report](#). *CoRR*, abs/2505.09388.
- Song Yu, Xiaofei Xu, Ke Deng, Li Li, and Lin Tian. 2025. [Tree of agents: Improving long-context capabilities of large language models through multi-perspective reasoning](#). *CoRR*, abs/2509.06436.
- Jiajie Zhang, Zhongni Hou, Xin Lv, Shulin Cao, Zhenyu Hou, Yilin Niu, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2025a. [Longreward: Improving long-context large language models with AI feedback](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 3718–3739. Association for Computational Linguistics.
- Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, Yu Fu, Xingtai Lv, Yuchen Zhang, Sihang Zeng, Shang Qu, et al. 2025b. [A survey of reinforcement learning for large reasoning models](#). *CoRR*, abs/2509.08827.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025c. [Qwen3 embedding: Advancing text embedding and reranking through foundation models](#). *CoRR*, abs/2506.05176.
- Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Ö. Arik. 2024. [Chain of agents: Large language models collaborating on long-context tasks](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025d. [The lessons of developing process reward models in mathematical reasoning](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 10495–10516. Association for Computational Linguistics.
- Dawei Zhu, Xiyu Wei, Guangxiang Zhao, Wenhao Wu, Haosheng Zou, Junfeng Ran, Xun Wang, Lin Sun, Xiangzheng Zhang, and Sujian Li. 2025. [Chain-of-thought matters: Improving long-context language models with reasoning path supervision](#). *CoRR*, abs/2502.20790.

A Experiment Details

Implementation Details We implement our reinforcement learning pipeline based on the VeRL framework (Sheng et al., 2024). For rollout generation, we set the sampling temperature to 1.0, with a maximum context window of 120k tokens for input and 8k tokens for output generation. The group size is configured as $G = 6$. Optimization is performed with a global batch size of 64 and a mini-batch size of 32, utilizing a constant learning rate of 2×10^{-6} . All experiments are executed on a computational cluster equipped with 16 NVIDIA H20 GPUs (90GB VRAM). The RL training time for 30B is approximately 72 GPU hours, and for 14B it is approximately 40 GPU hours. The SFT training for reward model is approximately 2 GPU hours.

Wikipedia-based Mixed QA Construction To detail the *Wikipedia-based Mixed QA* subset mentioned in the paper, we prioritize natural context continuity by sourcing Wikipedia articles exceeding 64k tokens. Through HTML parsing, we separate content into prose (unstructured) and tables (structured) to generate diverse reasoning tasks:

- **Structured QA (Single-Fragment):** Focuses on analytical reasoning over discrete table segments. These tasks require the model to perform filtering, ranking, and numerical calculation within a single structured evidence block.
- **Heterogeneous Mixed QA (Multi-Fragment):** Designed to bridge structured and unstructured modalities. This category includes *Text-and-Table QA*, which demands synthesizing information from free-text prose and tabular data, and *Multi-Table QA*, which requires linking logic across multiple interrelated tables.

B Analysis of Reasoning Diversity

In Section 2.2, we observed that optimizing solely the reasoning step (EAR+LLM-R) or adding a reasoning judge atop evidence evaluation (EAR+LLM-E+LLM-R) yields no performance improvement, and even causes a slight drop compared to evaluating evidence alone (LLM-E). To understand the underlying causes of this phenomenon, we empirically analyze the semantic diversity of the generated reasoning paths.

In our framework, evidence extraction strictly precedes reasoning execution. We hypothesize that both the content and the diversity of the downstream reasoning are heavily anchored to the extracted evidence. To verify this, we measured the semantic diversity of the reasoning content across the sampled paths for each method. We employed an LLM-as-a-Judge approach to evaluate whether the generated paths for a given query are semantically diverse or highly consistent (i.e., sharing identical core logic and conclusions).

Sampling Method	Semantically Diverse	Semantically Consistent
EAR + LLM-E	24.0%	76.0%
EAR + LLM-R	3.5%	96.5%
EAR + LLM-E + LLM-R	4.5%	95.5%

Table 2: Semantic diversity of reasoning outputs across different sampling configurations. When the reasoning evaluator (LLM-R) is applied without varying the evidence, the resulting reasoning paths become highly homogeneous.

The results, presented in Table 2, explain the limited efficacy of LLM-R. When the evidence content is fixed (as in EAR+LLM-R), the subsequent reasoning content becomes almost completely static, exhibiting 96.5% semantic consistency. Because fixing the evidence removes almost all variation in the downstream outputs, the exploration space for the reasoning step becomes severely restricted. Consequently, the reasoning judge is forced to evaluate candidate paths that share nearly identical reasoning content and lead to the exact same results, making the evaluation step highly redundant.

C Human Verification of Evidence Quality

We conducted an additional targeted manual annotation study to explicitly verify how often the extracted evidence truly supports the final answer under strict human review.

To maintain query consistency and ensure a fair comparison, we first identified the intersection subset of queries that were answered correctly by both the EAPO model and the GRPO baseline. From this exact intersection, we randomly sampled 100 queries and extracted their complete reasoning trajectories from both models.

We recruited three human annotators with NLP research backgrounds to perform a blind evaluation of these samples. They were tasked with answering a strict, boolean question: *"Does the information within the evidence block alone provide sufficient and logically sound support to derive the final answer?"*

The human verification results demonstrated that, on average, 97.3% of the EAPO samples were judged as having evidence that completely supports the final conclusion, compared to an average of 92.6% for the GRPO baseline. This direct assessment corroborates our automated metrics, clearly indicating that the EAPO training paradigm significantly enhances the model’s ability to accurately cite supporting evidence. It ensures that the generated reasoning is not just outcome-correct, but strictly grounded and human-verifiable.

D Needle in a Haystack Retrieval Experiment

We additionally evaluate retrieval performance using the "Needle in a Haystack" (NIAH) benchmark (Kamradt, 2023). As shown in Figure 10, both the **Base Model** and **GRPO** fall slightly short of perfection, failing to retrieve needles in a few specific instances. In contrast, **EAPO** achieves a flawless 100% retrieval rate. The "all-green" heatmap confirms that our dense evidence supervision further refines retrieval robustness, ensuring precise information localization across the entire context window.

E Detailed Prompt

We explicitly map each prompt to its corresponding module across our preliminary experiments (Section 2) and the EAPO framework (Section 3):

- **Prompt for EAR (Generation Prompt):** Utilized for generating reasoning paths in both the standard EAR baseline (Section 2) and the EAPO policy generation phase (Section 3).
- **Prompt for Evidence Evaluation:** Employed by the evidence judge (LLM-E) in the prelim-

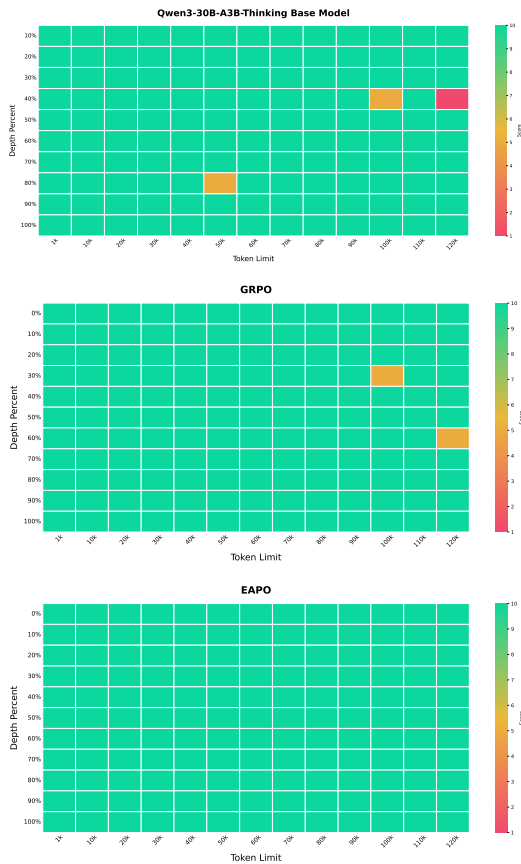


Figure 10: Needle in a Haystack retrieval performance across varying context lengths and depths.

inary study (Section 2.2) to score evidence utility, and also used by the Reward Model to calculate the Group-Relative Evidence Quality in EAPO (Section 3.2).

- **Prompt for Reasoning Evaluation:** Used specifically by the reasoning judge (LLM-R) in the preliminary study (Section 2.2) to assess the logic and coherence of the generated reasoning steps.
- **Prompt for Answer Evaluation (Outcome Evaluation):** Applied to compute the Outcome Accuracy Reward in EAPO (Section 3.2) and utilized for the final answer accuracy assessment in our main evaluation (Section 4).

Prompt for EAR

Please read the following content and answer the question below.

Content:

{content}

Question:

{question}

You must strictly adhere to the following four-part structure for your entire output. Do not include any introductory phrases or text outside of this format. Required Output Format:

<analysis>

[Articulate your analysis of the question, identify the specific information required, and outline your strategy for locating and extracting relevant evidence from the provided content. Do not answer the question directly.]

</analysis>

<evidence>

[Insert all verbatim, unmodified quotes from the content that directly support the final answer. Each piece of evidence should be on a new line and be as comprehensive as possible, using '...' to link related but separate text fragments if necessary.]

</evidence>

<reasoning>

[Provide a step-by-step logical explanation that connects the extracted evidence to the final answer. Clearly demonstrate how the evidence leads to your conclusion.]

</reasoning>

<answer>

[State the final, concise answer. This answer must be derived solely from the evidence provided above.]

</answer>

Prompt for Evidence Evaluation

You are a master strategist and information architect. Objective: Critically evaluate competing sets of 'Evidence'. Your mission is to act as a strategic filter, identifying the single most valuable set of evidence for constructing a high-quality, comprehensive, and accurate answer to the 'Original Question'.

Inputs:

* Original Question:

{question}

* Evidence Choices:

{choices}

Core Evaluation Principle: Comparative Analysis

Your evaluation must be relational, not absolute. Do not assess each choice in isolation. The score for any given choice must be determined by comparing its performance against the other available options across the criteria below. For a choice to earn a top score, it must be demonstrably superior *relative to its competitors*.

Evaluation Criteria: You must use these criteria as the dimensions for your comparison:

1. Relevance & Focus: How directly does the evidence address all facets of the 'Original Question'? Is it tightly focused, or diluted with irrelevant "noise"?

2. Depth & Breadth: Does the evidence provide comprehensive, multi-faceted coverage of the topic? Or is it superficial, one-sided, or incomplete?

3. Utility & Specificity: Is the evidence composed of concrete, factual, and directly usable information? Or is it vague, overly general, and requires heavy interpretation?

4. Coherence: Are the individual pieces of evidence within the set consistent with each other? Or do internal contradictions create ambiguity and undermine reliability?

Scoring Rubric (1-5 Scale): You must assign a score

to each choice based on the following strict rubric. Remember to apply these definitions through the lens of the Core Evaluation Principle.

* 1: Unusable: The evidence is fundamentally flawed. It is irrelevant, internally contradictory, or so vague that it provides no usable basis for an answer.

* 2: Poor: The evidence has major deficiencies. It may be only tangentially relevant, highly imbalanced, or lacking specific facts, making it a weak foundation, especially when compared to stronger options.

* 3: Adequate: The evidence is usable but clearly inferior to better options. It addresses the core question but has significant comparative gaps in depth, breadth, or specificity.

* 4: Good: The evidence is a strong contender. It is highly relevant and specific, but may be slightly less comprehensive or focused than the top choice. It provides a solid foundation for a high-quality answer.

* 5: Excellent: The evidence is perfectly tailored and clearly superior to all other options. It is the most comprehensive, balanced, specific, and coherent dataset available, with virtually no noise.

Output Requirements: Your entire output must be a single, valid JSON object structured exactly as follows, containing three keys: 'reason', 'scores', and 'best_choice'.

```
{
  "reason": "A string containing your detailed comparative analysis. Your analysis must: 1. Justify each score by explicitly comparing that choice to the others on the key 'Evaluation Criteria' (e.g., 'Choice A earns a 4 for Depth because it is more comprehensive than Choice B, but Choice C covers one additional aspect, making it superior'). 2. Use the language of the 'Scoring Rubric' to anchor your comparative judgments. 3. Conclude with a definitive statement explaining why the 'best_choice' is superior *relative to the alternatives*.",
  "scores": [An array of integers, where each integer is the 1-5 score corresponding to each choice in the provided order.],
  "best_choice": An integer representing the 0-based index of the single best evidence choice.
}
```

Prompt for Answer Evaluation

Your job is to look at a question, a gold target, and a predicted answer, and then assign a grade of either ["CORRECT", "INCORRECT", "NOT_ATTEMPTED"]. First, I will give examples of each grade, and then you will grade a new example. The following are examples of CORRECT predicted answers.

Question: What are the names of Barack Obama's children?

Gold target: Malia Obama and Sasha Obama

Predicted answer 1: sasha and malia obama

Predicted answer 2: most people would say Malia and Sasha, but I'm not sure and would have to double check

Predicted answer 3: Barack Obama has two daughters. Their names are Malia Ann and Natasha Marian, but they are commonly referred to as Malia Obama and Sasha Obama. Malia was born on July 4, 1998, and Sasha was born on June 10, 2001.

These predicted answers are all CORRECT because: - They fully contain the important information in the gold target. - They do not contain any information that contradicts the gold target. - Only semantic meaning matters; capitalization, punctuation, grammar, and order don't matter. - Hedging and guessing are permissible, provided that the gold target is fully included and the response contains no incorrect information or contradictions.

The following are examples of INCORRECT predicted answers.

Question: What are the names of Barack Obama's children?

Gold target: Malia and Sasha

Predicted answer 1: Malia.

Predicted answer 2: Malia, Sasha, and Susan.

Predicted answer 3: Barack Obama does not have any children.

Predicted answer 4: I think it's either Malia and Sasha. Or it could be Malia and Jackie. Or it could be Joey and Malia.

Predicted answer 5: It's possible you may mean Betsy and Olivia. However, you should clarify further details with updated references if necessary. Is that the correct answer?

Predicted answer 6: It may be the case that Obama's child is named James. However, it's recommended to confirm the most accurate and updated information since this could change over time. This model may not always reflect the most current information.

These predicted answers are all INCORRECT because: - A factual statement in the answer contradicts the gold target. Incorrect statements that have some hedging (e.g., "it is possible that", "although i'm not sure, i think") are also considered incorrect.

The following are examples of NOT_ATTEMPTED predicted answers.

Question: What are the names of Barack Obama's children?

Gold target: Malia and Sasha

Predicted answer 1: I don't know.

Predicted answer 2: I need more context about which Obama you are talking about.

Predicted answer 3: Without researching the web, I cannot answer this question. However, I can tell you that Barack Obama has two children.

Predicted answer 4: Barack Obama has two children. I know that one of them is Malia, but I'm not sure about the other one.

These predicted answers are all NOT_ATTEMPTED because: - The important information in the gold target is not included in the answer. - No statements in the answer contradict the gold target.

Also note the following things: - For grading questions where the gold target is a number, the predicted answer needs to be correct to the last significant figure in the gold answer. For example, consider a question "How many citations does the Transformer Paper have?" with gold target "120k".

- Predicted answers "120k", "124k", and "115k" are all CORRECT.

- Predicted answers "100k" and "113k" are INCORRECT.

- Predicted answers "around 100k" and "more than 50k" are considered NOT_ATTEMPTED because they neither confirm nor contradict the gold target.

- The gold target may contain more information than the question. In such cases, the predicted answer

only needs to contain the information that is in the question.

- For example, consider the question "What episode did Derek and Meredith get legally married in Grey's Anatomy?" with gold target "Season 7, Episode 20: White Wedding". Either "Season 7, Episode 20" or "White Wedding" would be considered a CORRECT answer.

- Do not punish predicted answers if they omit information that would be clearly inferred from the question. - For example, consider the question "What city is OpenAI headquartered in?" and the gold target "San Francisco, California". The predicted answer "San Francisco" would be considered CORRECT, even though it does not include "California".

- Consider the question "What award did A pre-trainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity win at NAACL '24?", the gold target is "Outstanding Paper Award". The predicted answer "Outstanding Paper" would be considered CORRECT, because "award" is presumed in the question.

- For the question "What is the height of Jason Wei in meters?", the gold target is "1.73 m". The predicted answer "1.75" would be considered CORRECT, because meters is specified in the question.

- For the question "What is the name of Barack Obama's wife?", the gold target is "Michelle Obama". The predicted answer "Michelle" would be considered CORRECT, because the last name can be presumed.

- Do not punish for typos in people's name if it's clearly the same name. - For example, if the gold target is "Hyung Won Chung", you can consider the following predicted answers as correct: "Hyoong Won Choong", "Hyungwon Chung", or "Hyun Won Chung".

Here is a new example. Simply reply with either CORRECT, INCORRECT, NOT ATTEMPTED. Don't apologize or correct yourself if there was a mistake; we are just trying to grade the answer.

Question: question

Gold target: correct_answer

Predicted answer: response

Grade the predicted answer of this new question as one of:

A: CORRECT

B: INCORRECT

C: NOT_ATTEMPTED

Just return the letters "A", "B", or "C", with no text around it.

Prompt for Reasoning Evaluation

You are a Strategic Logic Evaluator and Reasoning Analyst.

Objective: Evaluate a set of candidate outputs representing the **NEXT STEP** in a reasoning chain. You must determine which candidate represents the most logical, accurate, and valuable progression towards answering the 'Original Question', given the 'Historical Reasoning Context'.

Inputs: **Original Question:** question

Historical Reasoning Context (The Unfinished Process): approved_reasoning_process

Candidate Next Steps: choices

Core Evaluation Principles: Assess each choice

based on its logical validity as a continuation of the provided context.

1. **Logical Validity & Advancement:** If the 'Historical Reasoning Context' is sound, the candidate should logically follow it, deepening the analysis or moving to the next deduction.

Correction/Pivot: If the 'Historical Reasoning Context' contains flaws, contradictions, or weak assumptions, a candidate that identifies and corrects these issues is superior to one that blindly follows a wrong path. **Flexibility:** The candidate is NOT required to strictly adhere to the conclusions of the history, but it must strictly adhere to the facts/evidence cited within it (no external hallucinations).

2. **Evidence Grounding:** While the logical path can change, facts cannot be invented. The candidate must still rely on the factual evidence provided within the context (or implicitly known general knowledge if applicable to the domain), without hallucinating new specific data.

3. **Efficiency:** Does this step actually move the needle? Avoid circular reasoning or restating what has already been established in the history.

Scoring Rubric (1-4 Scale):

1: Critical Failure (Hallucination/Illogical Jump): The output invents facts not present in the context. The output makes a logical leap that makes no sense, neither following the history nor offering a valid correction. The output completely ignores the 'Original Question'.

2: Poor (Redundant/Weak): The output merely restates the 'Historical Reasoning Context' without adding new value. The output blindly follows a clearly erroneous path established in the history (fails to critical thinking). The logic is circular or trivial.

3: Good (Valid Progression): A solid, logical next step. It accepts the context and moves forward reasonably. If the history is sound, this choice extends it correctly. No hallucinations; safe and compliant.

4: Excellent (Insightful Advancement or Critical Correction): The output demonstrates superior reasoning. It might provide a necessary pivot if the previous logic was heading in the wrong direction. It synthesizes previous points to reach a breakthrough or a highly distinct intermediate conclusion. It maximizes the probability of correctly answering the 'Original Question'.

Output Requirements: You must output a single, valid JSON object containing the following keys. Do not include markdown formatting or any text outside the JSON object.

```
{
  "reason": "A detailed analysis string. You must:
  1) Evaluate the state of the 'Historical Reasoning Context' (is it on the right track?).
  2) Analyze each choice's logical contribution (is it a continuation or a pivot?).
  3) Justify why the 'best choice' is the most valuable next step towards the solution.",
  "scores": [An array of integers, where each integer is the 1-4 score corresponding to each choice in the provided order.],
  "best_choice": An integer (0-based index of the best choice)
}
```