Escaping the Context Bottleneck: Active Context Curation for LLM Agents via Reinforcement Learning

Anonymous ACL submission

Abstract

Large Language Models (LLMs) struggle with long-horizon tasks due to the "Context Bottleneck" and the "Lost-in-the-Middle" phenomenon, where accumulated noise from verbose environments degrades reasoning over multi-turn interactions. To address this, we introduce a symbiotic framework that decouples context management from task execution. Our architecture pairs a lightweight, specialized policy model (ContextCurator) that actively curates the working memory for a powerful, frozen foundation model (TaskExecutor). Trained via reinforcement learning, the ContextCurator performs active information entropy reduction. It aggressively prunes environmental noise while meticulously preserving "reasoning anchors"—sparse data points critical for future deductions. On WebArena, our framework boosts Gemini-3.0-flash's success rate from 36.4% to 41.2% while reducing token consumption by 8.8% ($47.4K \rightarrow 43.3K$). On DeepSearch, it achieves 57.1% success (vs. 53.9%) with an $8\times$ token reduction. Remarkably, a 7B ContextCurator matches the context management performance of GPT-4o, providing a scalable, computationally efficient paradigm for autonomous long-horizon agents.

1 Introduction

The trajectory of artificial intelligence has fundamentally shifted from the development of static conversational models to the deployment of autonomous, goal-oriented Large Language Model (LLM) agents capable of sequential decision-making in complex environments (Wang et al., 2024; Liu et al., 2025; Shao et al., 2024; OpenAI, 2024). This advancement enables agents to tackle complex real-world challenges, from web browsing (Gur et al., 2024; Yao et al., 2022) to multi-hop information retrieval (Yang et al., 2024b; Xie et al., 2024). However, a critical bottleneck emerges in long-horizon tasks: while frontier models exhibit profound zero-shot reasoning capabilities, their performance degrades dramatically due to the inherent limitations of their *cognitive span* (Huang et al., 2025; Du et al., 2025).

The fundamental challenge to reliable long-horizon execution is the ability to maintain task-relevant working memory while systematically filtering environmental noise over extended trajectories. Real-world observation spaces are characterized by extreme informational entropy. In web tasks, DOM trees contain over 90% structural noise (e.g., advertisements, styling scripts, redundant navigational elements), while search paradigms flood the context with semantic noise and distractor paragraphs. When agents process these raw streams, the signal-to-noise ratio (SNR) deteriorates rapidly. Feeding this high-entropy data directly into a transformer inevitably triggers the *"Lost-in-the-Middle" phenomenon* and causes *cascading logical failures*, where early noise compounds into hallucinated reasoning in subsequent steps. Our experiments reveal that even frontier models like Gemini-3.0-flash achieve only 36.4% success on WebArena and 53.9% on DeepSearch (Tables 1 and 2). These results expose a critical limitation: brute-force expansion of context windows does not guarantee effective reasoning. The problem is not capacity, but signal-to-noise ratio.

Current mitigation strategies fall short of addressing the dynamic nature of these environments. *Passive memory systems* (Chhikara et al., 2025; Packer et al., 2024; Zhong et al., 2024; Ouyang et al., 2025) treat context management as a static retrieval problem, relying on semantic similarity to recall information. These methods suffer from retrieval bias; they frequently fail to retrieve implicit **reasoning anchors**—causally essential information that may be textually dissimilar to the current query. Conversely, *monolithic learned architectures* attempt to internalize memory management within the same task model (Zhou et al., 2025; Yu

et al., 2025). This creates a severe capacity split: small models lack the reasoning depth required for complex logical execution, while large proprietary models are too computationally expensive and opaque for direct on-policy fine-tuning.

To bridge this gap, we introduce **ActiveContext**, a framework that reframes memory management from a static storage utility into an **active, sequential decision-making process**. We adopt a complementary cognitive architecture that decouples the system into two specialized components: a powerful frozen foundation model (**TaskExecutor**) handles pure reasoning and task execution, while a lightweight specialized policy model (**ContextCurator**) actively curates its working memory.

Trained via reinforcement learning, the ContextCurator learns to act as a "Cognitive Architect." Driven by a distal reward signal based on the TaskExecutor's success, it develops a generalizable policy to perform active **information entropy reduction**. It aggressively prunes structural and semantic noise while meticulously preserving **reasoning anchors**—sparse data points critical for future deductions. This symbiotic design ensures that the TaskExecutor's attention heads are focused solely on high-fidelity, high-signal data.

Extensive evaluation demonstrates that our decoupled framework establishes a new Pareto frontier in agentic reasoning. On WebArena, ActiveContext boosts Gemini-3.0-flash from a 36.4% to a 41.2% success rate while reducing token consumption by 8.8% (47.4K $\rightarrow$ 43.3K). On DeepSearch, the success rate improves from 53.9% to 57.1% with nearly an $8\times$ token reduction (46.7K $\rightarrow$ 6.6K). Remarkably, our 7B ContextCurator matches the context management performance of GPT-4o. This result democratizes agentic pipelines, proving that active memory curation is a distinct cognitive skill that can be efficiently offloaded to smaller models.

Our work makes three key contributions:

- We propose a novel **symbiotic framework** that decouples context management from task execution. By pairing a specialized ContextCurator with a closed-source SOTA TaskExecutor, we transform memory from a passive storage utility into an active, decision-dependent control variable.
- We formulate an **on-policy RL pipeline** that trains lightweight models to actively manage context through dynamic environment interaction. Guided by distal rewards, the model learns a generalizable policy for **information entropy reduction**—aggressively pruning noise while meticulously preserving critical reasoning anchors.
- We demonstrate substantial empirical gains across complex long-horizon benchmarks. Our framework improves SOTA model performance while reducing token consumption, and proves that a specialized 7B model can match the context management capabilities of GPT-4o, providing a highly scalable and sustainable paradigm for autonomous agents.

2 Related Work

Memory for Agentic Systems. Context memory enables autonomous agents to maintain state and leverage history for complex reasoning (Zhang et al., 2024). Existing frameworks typically utilize passive storage (e.g., Mem0 (Chhikara et al., 2025), MemGPT (Packer et al., 2024), LangMem(LangChain, 2024)) via semantic retrieval, or procedural workflows (e.g., HiAgent (Hu et al., 2025), AWM (Wang et al., 2025)) to induce structured trajectories. However, passive systems suffer from retrieval bias, frequently missing causally essential "reasoning anchors" that lack textual similarity (Ming et al., 2025; Zhou et al., 2024), while procedural workflows struggle to generalize across high-entropy domains. Advanced cognitive graphs (e.g., Synapse (Zheng et al., 2024), MAGMA (Jiang et al., 2026)) mitigate this but introduce formidable computational latency. In contrast, our framework adopts a **symbiotic architecture** that decouples *context management* from *task execution*. Rather than relying on static storage or heavy graphs, our *ContextCurator* actively curates context to maximize real-time information density.

Reinforcement Learning for LLM Agents. RL is a pivotal paradigm for aligning agentic reasoning across diverse domains (e.g., web navigation, embodied environments) using algorithms like PPO and GRPO (Schulman et al., 2017; Shao et al., 2024; Yao et al., 2022; Wei et al., 2025; Jin et al., 2025b). Recognizing the limits of static memory, recent works (e.g., Mem1 (Zhou et al., 2025), MemAgent (Yu et al., 2025), MemoryR1 (Yan et al., 2025)) apply RL to optimize active memory operations as sequential decisions. However, these approaches train a *monolithic* model for both memory curation and task execution. This creates a severe capacity split: small open-source models lack the

deep reasoning required for complex deductions, while frontier models are opaque and computationally prohibitive for on-policy fine-tuning. Monolithic RL also risks inducing "lazy agent" shortcut behaviors (Zhang et al., 2025). *ActiveContext* deviates via a **collaborative symbiotic architecture**. We use RL to train a lightweight *ContextCurator* that strictly manages context for a powerful, frozen *TaskExecutor*, combining adaptive entropy reduction with frontier reasoning strength at a fraction of the cost.

3 Methodology

We propose **ActiveContext**, a symbiotic framework designed to resolve the "Context Bottleneck" in long-horizon tasks. Conventional approaches (e.g., Mem0 (Chhikara et al., 2025), MemoryBank (Zhong et al., 2024), ReasoningBank (Ouyang et al., 2025)) treat memory as a static retrieval problem. In high-entropy dynamic environments, these methods suffer from **rigidity** via brittle, hand-crafted workflows, and **retrieval bias**—semantic searches frequently fail to recall causally essential "reasoning anchors" lacking textual similarity to the current query. To transcend these limitations, ActiveContext reframes context management from a passive storage utility into an *active, sequential decision-making problem*.

3.1 Problem Formulation

We formulate the interaction between an autonomous LLM agent and a complex, long-horizon environment (e.g., a web browser or search engine) as a Partially Observable Markov Decision Process (POMDP), defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{R})$. Here, $\mathcal{S}$ represents the latent, objective state of the environment (e.g., complete underlying DOM structures or full retrieval databases); $\mathcal{A}$ is the set of admissible executable actions (e.g., `click(element)`, `type(text)`, or search queries); and $\mathcal{O}$ denotes the high-dimensional observation space emitted at each discrete time step t as $o_t \in \mathcal{O}$. Finally, the environment transitions according to $\mathcal{T}(s_{t+1}|s_t, a_t)$ and yields a sparse, terminal reward $R(\tau) \in \{0, 1\}$ upon task completion.

The Standard Policy and the Context Bottleneck. Given a user instruction u, a standard autoregressive LLM agent attempts to learn a policy $\pi_\theta(a_t|h_t)$, conditioned on the complete, uncompressed interaction history: $h_t = \{u, o_0, a_0, o_1, a_1, \ldots, o_t\}$ The agent aims to maximize the expected return $J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)]$. However, in real-world tasks, the observation o_t is characterized by extreme informational entropy; it contains sparse, task-critical reasoning anchors k_t buried within massive environmental noise z_t, where $|z_t| \gg |k_t|$. For example, in web navigation, a raw Document Object Model (DOM) tree (o_t) typically consists of over 90% structural noise (z_t)—such as formatting tags, advertisements, and redundant navigational links—obscuring crucial, localized data points (k_t) like a specific product price or a hidden spatial coordinate.

As the temporal horizon t expands, the sequence length of h_t grows linearly. This raw concatenation triggers the **Context Bottleneck**: it inevitably distributes the model's probability mass across the accumulating noise $\sum z_t$, causing *Attention Dilution* (the "Lost-in-the-Middle" phenomenon) and inducing *Cascading Failures* as early noise compounds into hallucinated actions at step $t + 1$.

The Optimization Conflict of Monolithic Active Memory. To mitigate this unbounded context growth, recent RL-based architectures expand the agent's action space to include internal memory management operations: $\mathcal{A}' = \mathcal{A} \cup \mathcal{A}_{\text{mem}}$. The monolithic policy is reformulated to dynamically update a compressed state alongside executing environment actions: $\pi_\theta(a_t, m_t|\mathcal{M}_t, o_t)$.

While this actively reduces sequence length, it introduces a severe *optimization conflict*. Memory curation (which requires aggressive information entropy reduction) and task execution (which requires preserving deep logical dependencies) represent orthogonal cognitive objectives. Forcing a single parameter set θ to simultaneously optimize for both induces gradient interference, frequently diluting the reasoning trace and resulting in "lazy agent" behaviors. Furthermore, applying on-policy RL to align these dual objectives within massive, closed-source frontier models is computationally prohibitive.

Objective of this Work: Symbiotic Decoupling via RL. To resolve this optimization conflict, our objective is to map raw history into a heavily condensed, high-fidelity memory state ($|\mathcal{M}_t| \ll |h_t|$) without burdening the primary reasoning engine. We achieve this by decoupling the monolithic policy $\pi_\theta(a_t, m_t)$ into two distinct, specialized entities: an *active* context policy π_{ctx} (ContextCurator) to handle $\mathcal{A}_{\text{mem}}$, and a *frozen* execution policy π_{exec} (TaskExecutor) to handle $\mathcal{A}$.

The core technical challenge lies in cross-model

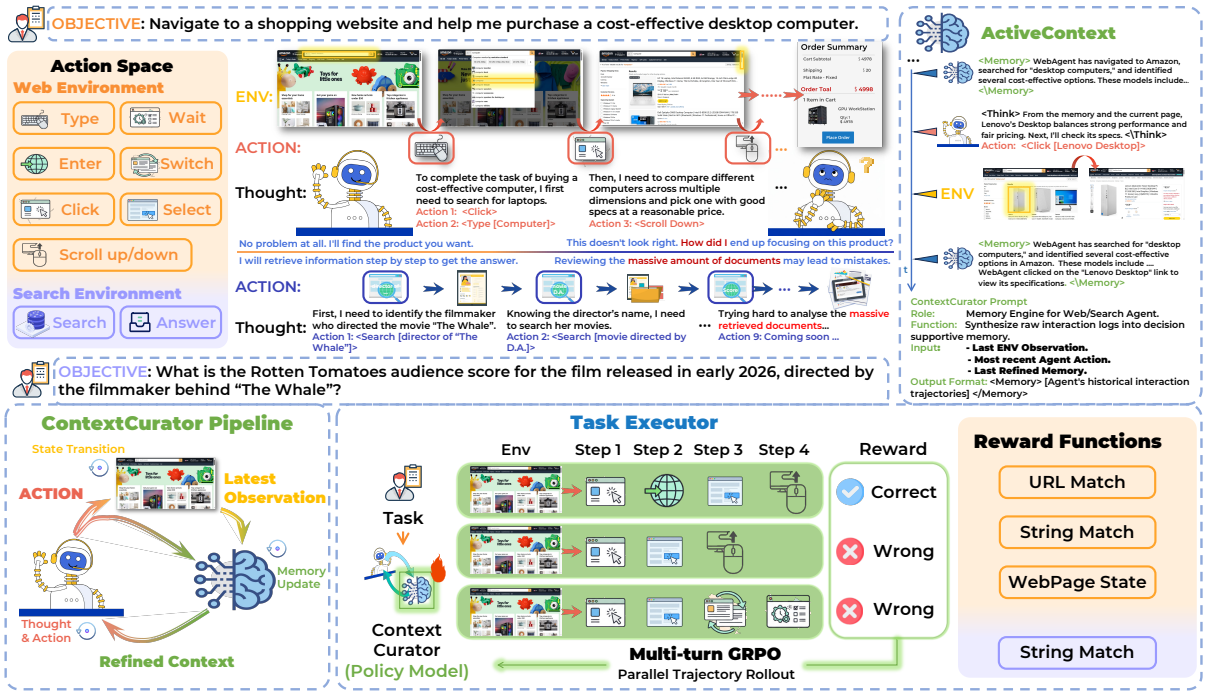

Figure 1: Overview of **ActiveContext**. **Top/Center:** Standard LLM agent interactions in web and search domains, illustrating the context bottleneck and "lost-in-the-middle" phenomenon during multi-turn reasoning. **Top-Right & Bottom-Left:** Our symbiotic architecture, where the *ContextCurator* actively refines verbose observations into a high-fidelity working memory for the *TaskExecutor*. **Bottom-Right:** The *Multi-turn GRPO* pipeline, which rolls out trajectories to collect distal rewards and optimize the context curation policy.

alignment. We resolve this by absorbing the frozen TaskExecutor into the environment's transition dynamics, effectively reducing the symbiotic setup into a tractable single-agent RL problem. We then employ a Multi-Turn Group Relative Policy Optimization (MT-GRPO) pipeline to train a lightweight, open-source ContextCurator. By optimizing against distal rewards generated by this augmented environment, the Curator learns to perform targeted information entropy reduction perfectly aligned with the reasoning prerequisites of the black-box TaskExecutor.

3.2 Symbiotic Cognitive Architecture

We achieve this objective by decoupling the cognitive architecture into two specialized LLM components, ensuring that reasoning is not obstructed by low-level information overload. At any time step t, the symbiotic interaction unfolds in two stages:

1. Active Context Curation (π_{ctx}). At each step t, the **ContextCurator** takes as input the current working memory $\mathcal{M}_t$, the latest observation o_t, and the previous action a_{t-1}, and produces an updated memory state. Unlike prior active memory frameworks, such as Mem1 (Zhou et al., 2025) and MemoryR1 (Yan et al., 2025), which formulate memory management as selecting from a small set of discrete operations (e.g., `navigate`, `scroll up`, and `rollback`), we instead model context curation as direct generation of the next memory state. Concretely, π_{ctx} autoregressively rewrites the current workspace into a refined textual memory m_t, which serves as the working memory for the next step: $\mathcal{M}_{t+1} = m_t \sim \pi_{\text{ctx}}(\cdot \mid \mathcal{M}_t, o_t, a_{t-1})$. This formulation allows the curator to compress redundant interactions, remove structural noise, and preserve information that is most relevant for subsequent reasoning and decision making.

2. Context-Conditioned Execution (π_{exec}). Given the curated memory m_t and the current observation o_t, the **TaskExecutor** generates the task action a_t. We instantiate the executor with a strong frozen foundation model, such as GPT-4o or Gemini-3.0-flash, which conditions on both the curated memory and the immediate environment signal:

$$a_t \sim \pi_{\text{exec}}(\cdot \mid m_t, o_t). \tag{1}$$

Here, m_t serves as a compact task workspace that retains relevant historical information, thereby directly affecting the quality of the generated action.

3.3 Policy Optimization via Multi-Turn GRPO

While system prompts can define the functional role of the ContextCurator, they cannot inherently teach a small LLM *what* information is causally salient for a given task. Relying solely on zero-shot prompting often leads to hallucinated pruning, where the model arbitrarily discards crucial reasoning anchors or retains irrelevant noise. To address this, we formulate context curation as a sequential decision-making problem and train the ContextCurator π_ϕ using an on-policy reinforcement learning approach based on Multi-Turn Group Relative Policy Optimization (MT-GRPO).

Reward Definition and The Causal Dependency Chain. Let $\tau = \{(o_t, m_t, a_t)\}_{t=0}^{L}$ be a sampled trajectory of length L. The environment provides a sparse binary reward $R(\tau) \in \{0, 1\}$ upon task completion. This reward is determined by task-specific outcome metrics, such as URL matches, string matches, or specific webpage state assertions. Since the TaskExecutor is strictly frozen, its policy π_{exec} can be absorbed into the environment's transition dynamics. Consequently, any variance in the terminal reward $R(\tau)$ is entirely attributable to the quality of the curated memory m_t. Given local context $c_t = (\mathcal{M}_t, o_t, a_{t-1})$, the causal dependency chain is defined as:

$$\underbrace{m_t \sim \pi_\phi(\cdot|c_t)}_{\text{Curator Decision}} \longrightarrow \underbrace{a_t \sim \pi_{\text{exec}}(\cdot|m_t, o_t)}_{\text{Executor Action}} \longrightarrow \underbrace{R(\tau)}_{\text{Outcome}} \tag{2}$$

Trajectory Sampling and Advantage Estimation. For a given instruction u, we sample a group of G independent trajectories $\{\tau_1, \tau_2, \ldots, \tau_G\}$ generated by the symbiotic system. To reduce gradient variance over long horizons, we compute the group relative advantage A_i for each trajectory τ_i by normalizing its reward against the group statistics:

$$A_i = \frac{R(\tau_i) - \text{mean}(\{R(\tau_1), \ldots, R(\tau_G)\})}{\text{std}(\{R(\tau_1), \ldots, R(\tau_G)\}) + \epsilon} \tag{3}$$

where ϵ is a small constant for numerical stability.

Optimization Process and Policy Update. The ContextCurator's policy π_ϕ is updated to maximize this expected advantage. During the backward pass, gradients are applied *only* to the tokens generated by the ContextCurator (m_t), dramatically reducing memory overhead compared to monolithic training. The multi-turn GRPO objective $\mathcal{J}_{\text{GRPO}}(\phi)$ is defined as the expectation over the sampled trajectories:

$$\mathcal{J}_{\text{GRPO}}(\phi) = \mathbb{E}\left[\frac{1}{L}\sum_{t=1}^{L}\left(\mathcal{O}_t^{\text{clip}}(\phi) - \beta \mathbb{D}_{\text{KL}}\left(\pi_\phi || \pi_{\text{ref}}\right)_t\right)\right] \tag{4}$$

where the per-step clipped surrogate $\mathcal{O}_t^{\text{clip}}(\phi)$ restricts excessive policy updates:

$$\mathcal{O}_t^{\text{clip}}(\phi) = \min\left(\rho_t(\phi)A_i, \text{clip}(\rho_t(\phi), 1-\epsilon, 1+\epsilon)A_i\right) \tag{5}$$

Here, $\rho_t(\phi) = \frac{\pi_\phi(m_t|c_t)}{\pi_{\phi_{\text{old}}}(m_t|c_t)}$ is the importance sampling ratio of the Curator's actions, and β controls the KL-divergence penalty against the reference model π_{ref}.

Through this optimization process, the MT-GRPO objective forces the ContextCurator to perform active information entropy reduction. If π_ϕ retains visual noise or drops a reasoning anchor, the TaskExecutor fails, yielding $R(\tau) = 0$. By optimizing against this distal reward, the Curator dynamically learns to adapt its context density, bridging the gap between raw environmental chaos and the strict reasoning prerequisites of the TaskExecutor.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate ActiveContext on two challenging long-horizon benchmarks: (1) *WebArena* (Zhou et al., 2024): A realistic execution environment simulating four domains (Shopping, Forums, Gitlab, CMS), evaluating the agent's ability to navigate raw DOM trees over extended trajectories characterized by high structural noise. (2) *DeepSearch* (Jin et al., 2025a): A complex RAG-based environment requiring multi-turn retrieval and reasoning. It specifically tests the agent's ability to synthesize answers from noisy search engine results across datasets like HotpotQA and 2WikiMultiHop. We report the Success Rate (SR) and the total context token consumption across both environments. The exact formulation for calculating context tokens is detailed in Appendix A.3.

Baselines & Models. We compare ActiveContext with standard baselines, including *No Memory*, which uses only the current observation, and *Full Context*, which concatenates the interaction history with verbose past observations replaced by placeholders to control context length. We

[1]Token counts are aggregated across the entire interaction trajectory. Because different methods require a varying number of steps to complete a task, the *Full Context* baseline may occasionally accumulate fewer total tokens than *No Memory*, as the latter typically requires longer exploration to succeed.

Table 1: Experimental results on the WebArena benchmark. Success rate (SR ↑) and context token consumption (Tokens (K) ↓)[1] are reported across 5 domain subsets. To prevent extreme context bloat, the *Full Context* baseline replaces historical webpage observations with placeholders. The bottom section extensively ablates different context curation methods using `gpt-4o-mini` as the fixed TaskExecutor. **The final average (Overall) is weighted by the number of instances in each subset.**

Methods	Shopping		Admin		Gitlab		Reddit		Map		Overall	
	SR	*Tokens*	*SR*	*Tokens*	*SR*	*Tokens*	*SR*	*Tokens*	*SR*	*Tokens*	*SR*	*Tokens*
TaskExecutor: Gemini-3.0-flash												
No Memory	13.0	50.7	22.2	57.2	36.7	75.1	26.1	42.8	6.7	27.2	20.0	51.2
Full Context	**28.3**	42.2	33.3	58.6	**66.7**	58.3	52.2	35.5	10.0	40.2	36.4	47.4
Synapse	19.6	44.6	**41.7**	71.5	43.3	70.0	**56.5**	44.4	26.7	36.6	35.2	53.6
Training-Free GRPO	17.4	55.4	36.1	70.6	36.7	76.9	**56.5**	47.8	30.0	50.1	32.7	60.6
ActiveContext (Ours)	26.1	39.0	41.7	50.5	**66.7**	53.8	47.8	43.1	**33.3**	30.9	**41.2**	43.3
TaskExecutor: Gemini-2.5-flash												
No Memory	10.9	45.2	11.1	75.0	36.7	71.1	26.1	37.7	6.7	36.4	17.0	53.8
Full Context	21.7	31.9	33.3	52.7	50.0	66.5	39.1	32.6	13.3	30.4	30.3	42.6
Synapse	13.0	27.2	**41.7**	54.0	40.0	70.2	47.8	33.3	13.3	33.5	29.1	42.9
Training-Free GRPO	21.7	51.7	36.1	62.3	33.3	80.6	**52.2**	43.7	3.3	64.4	27.9	60.5
ActiveContext (Ours)	**23.9**	28.0	33.3	51.6	**50.0**	53.1	47.8	34.3	**13.3**	28	**32.1**	38.6
TaskExecutor: gpt-4o												
No Memory	13.0	49.5	**33.3**	63.9	33.3	73.9	47.8	35.6	6.7	35.8	24.8	52.6
Full Context	13.0	38.9	25	46.4	53.3	54.1	43.5	27.5	13.3	37.5	27.3	41.5
Synapse	17.4	35.2	16.7	65.6	23.3	67.0	47.8	35.7	20.0	26.9	23.0	46.2
Training-Free GRPO	10.9	50.8	27.8	53.8	33.3	64.8	43.5	41.4	**23.3**	31.3	25.5	49.1
ActiveContext (Ours)	**19.6**	40.6	22.2	46.4	**56.7**	54.7	**52.2**	33.8	13.3	25.9	**30.3**	40.8
TaskExecutor: gpt-4o-mini (with Extended Ablations)												
Baselines & Retrieval-Based Memory												
No Memory	8.7	80.3	8.3	100.3	**26.7**	87.4	8.7	62.2	3.3	59.4	10.9	79.6
Full Context	8.7	46.3	**16.7**	51.0	23.3	68.2	13.0	43.0	3.3	36.9	12.7	49.1
Synapse	**10.9**	40.5	11.1	65.2	16.7	73.2	**39.1**	43.6	16.7	25.6	17.0	49.6
AWM	8.7	51.3	11.1	68.7	20.0	75.6	34.8	44.2	16.7	34.7	16.4	55.5
Mem0	4.3	44.7	**16.7**	71.2	20.0	76.8	26.1	50.3	16.7	33.6	15.2	55.1
Training-Free GRPO	**10.9**	60.0	13.9	71.1	**26.7**	78.7	**39.1**	47.5	**20.0**	42.1	**20.0**	60.8
Prompting-Based Memory Management (Zero-Shot Curators)												
Gemini-3.0-flash	**17.4**	37.8	**25.0**	54.6	**40.0**	48.1	30.4	26.1	**10.0**	32.0	**23.6**	40.7
Gemini-2.5-flash	10.9	38.1	25.0	55.2	26.7	48.9	30.4	27.0	**10.0**	32.6	19.4	41.2
gpt-4o	15.2	42.6	22.2	48.8	33.3	56.1	**39.1**	30.5	3.3	28.8	21.2	42.2
gpt-4o-mini	10.9	43.7	11.1	51.9	26.7	58.3	21.7	33.5	6.7	31.2	14.6	44.4
Qwen2.5-7B-Instruct	13.0	42.7	8.3	59.0	33.3	72.8	13.0	47.9	3.3	38.2	13.9	51.6
Active Memory Management (RL-Trained Curator)												
ContextCurator (Ours)	13.0	30.2	22.2	38.3	33.3	48.6	**39.1**	25.3	**10.0**	18.7	21.8	32.2

also compare against advanced memory methods, including Synapse (Zheng et al., 2024), AWM (Wang et al., 2025), Mem0 (Chhikara et al., 2025), and Training-Free GRPO (Cai et al., 2025). For on-policy RL, we initialize the ContextCurator from `Qwen-2.5-7B-Instruct` (Yang et al., 2024a). To evaluate cross-model generalization, we use frozen proprietary TaskExecutors, including `Gemini-3.0-flash`, `Gemini-2.5-flash` (Comanici et al., 2025), `GPT-4o`, and `GPT-4o-mini` (Hurst et al., 2024). In our formulation, these frozen models are treated as part of the environment dynamics during training.

4.2 Results on WebArena Benchmark

Table 1 presents the comparative evaluation on WebArena. The results empirically validate our primary hypothesis: *active context curation elevates the reasoning upper-bound while drastically reducing computational costs*.

Context Memory Quality Matters. Standard approaches face a severe dilemma: *No Memory* lacks historical environment feedback, while *Full Context* induces Attention Dilution due to context overload. ActiveContext breaks this trade-off, establishing a new Pareto frontier. With *Gemini-3.0-flash*, our framework achieves a Success Rate (SR) of **41.2%** (vs. 36.4% with full context) while si-

multaneously reducing token consumption by **8.8%** (43.3k vs. 47.4k). This confirms that the bottleneck in web agents is not the *quantity* of information, but the *Signal-to-Noise Ratio (SNR)*. By actively pruning redundant DOM elements (e.g., ads, sidebars), ActiveContext prevents cascading failures, allowing the backbone's attention mechanism to focus entirely on critical decision pivots.

Preserving Reasoning Anchors. In the **Map** domain, which requires precise retention of spatial coordinates and user arguments over long horizons, *No Memory* often fails near-completely (e.g., 6.7% SR for *Gemini-3.0-flash*). In contrast, ActiveContext achieves up to **33.3%** SR. This substantial gap confirms that ActiveContext successfully captures implicit "Reasoning Anchors"—data points like spatial coordinates that appear locally irrelevant but are globally causal—which heuristic summarization or full-context approaches routinely discard or bury.

ContextCurator Efficacy Across Model Scales. ActiveContext yields larger gains for smaller backbone models. For *GPT-4o-mini*, it improves SR from 12.7% to **21.8%** (+72% relative gain), compared with an 11% gain for *GPT-4o*. This suggests that frontier models are more robust to structural noise, whereas smaller models benefit more from a ContextCurator that constructs a compact, high-fidelity workspace.

Impact of MT-GRPO Optimization. Table 1 isolates the effect of our RL formulation by fixing `gpt-4o-mini` as the TaskExecutor. Compared with retrieval-based memories and prompted LLM curators, the RL-optimized ContextCurator achieves the best efficiency–performance trade-off. Relative to its base model `Qwen2.5-7B-Instruct`, RL fine-tuning improves SR from 13.9% to **21.8%** while reducing context tokens from 54.4k to 32.5k. It also outperforms a proprietary `gpt-4o-mini` curator in both SR (21.8% vs. 14.6%) and token usage (32.5k vs. 44.4k), and matches the much larger `gpt-4o` curator (21.2% SR). These results show that distal RL rewards can specialize a lightweight open-source model into an effective cognitive filter for context management.

4.3 Results on DeepSearch Benchmark

Table 2 presents results on DeepSearch. Unlike WebArena's visual structural noise, DeepSearch introduces severe *semantic noise*, where redundant RAG results degrade the model's attention mechanism and impair subsequent logic.

Information Entropy Reduction in Search. *ActiveContext* consistently outperforms the *Full Context* baselines. When utilizing *Gemini-2.5-flash* as the executor, it achieves a **24%** relative improvement in success rate (33.4% $\rightarrow$ 41.5%) while simultaneously reducing the average context length by a factor of $\sim 5\times$ (34.4k $\rightarrow$ 7.3k). By aggressively stripping away semantic noise, *ActiveContext* helps the executor maintain a pristine reasoning trace. This confirms that for complex multi-turn search tasks, actively filtering irrelevant paragraphs to maximize the Signal-to-Noise Ratio (SNR) in working memory is imperative for preserving logical coherence. We note that the trajectory-based *Synapse* method is also quite effective on these QA benchmarks; because many queries share underlying semantic structures, retrieving a successful historical trajectory as an exemplar can significantly boost performance. However, this comes at the cost of massive context bloat.

Domain Analysis: Handling Multi-Hop Complexity. The impact of ActiveContext is most pronounced in multi-hop reasoning tasks like *2WikiMultiHop* and *HotpotQA*. As shown in Table 2, for Gemini-3.0-flash executor, ActiveContext improves *2Wiki* performance from 62.0% to **68.0%**, and *HotpotQA* from 52.0% to **56.0%**. In these domains, distractor paragraphs often mislead the model into incorrect intermediate reasoning steps. ActiveContext effectively filters these distractors, preserving only the golden evidence. In contrast, simpler single-hop tasks like *NQ* (Natural Questions) show smaller gains (30.0% $\rightarrow$ 32.0%), suggesting that active curation is most critical when the reasoning chain is fragile and highly susceptible to noise accumulation.

RL-Driven Cognitive Decoupling. Table 2 provides strong evidence for our "Cognitive Decoupling" motivation in the search domain. The base `Qwen2.5-7B-Instruct` model (30.4% SR) underperforms the *Full Context* baseline (32.7% SR), indicating that generic instructions are insufficient for small LLMs to effectively distinguish signal from semantic noise in complex retrieved documents. However, the MT-GRPO optimization improves this performance to **34.9%**, proving that the RL objective successfully instills a robust, global context memory management capability. Remarkably, the specialized ContextCurator outperforms the generic `gpt-4o-mini` agent (33.0%) and achieves effective parity with the much larger `gpt-4o` (35.0%), all while consuming signif-

Table 2: Experimental results of *ActiveContext* on the DeepSearch benchmark. Success rate (SR ↑) and context token consumption (Tokens (K) ↓) are reported across 7 domain subsets for 4 different backbone LLMs. The bottom section extensively ablates different context curation methods using `gpt-4o-mini` as the fixed TaskExecutor.

Methods	**NQ**		**TriviaQA**		**PopQA**		**HotpotQA**		**2wiki**		**Musique**		**Bamboogle**		**Avg.**	
	SR	*Tokens*	*SR*	*Tokens*	*SR*	*Tokens*	*SR*	*Tokens*	*SR*	*Tokens*	*SR*	*Tokens*	*SR*	*Tokens*	*SR*	*Tokens*
TaskExecutor: Gemini-3.0-flash																
Full Context	30.0	62.4	76.0	20.4	62.0	36.4	52.0	53.5	62.0	38.9	33.0	83.6	62.0	31.6	53.9	46.7
Synapse	**86.0**	53.6	68.0	51.2	54.0	100.5	32.0	149.2	42.0	83.0	**44.0**	83.0	**72.0**	44.6	56.9	80.7
Training-Free GRPO	22.0	78.2	76.0	16.7	44.0	47.2	46.0	60.5	**68.0**	39.1	27.0	82.7	60.0	33.4	49.0	51.1
ActiveContext (Ours)	32.0	6.8	**78.0**	3.2	**64.0**	7.4	**56.0**	7.4	**68.0**	7.5	36.0	8.5	66.0	5.6	**57.1**	6.6
TaskExecutor: Gemini-2.5-flash																
Full Context	22.0	38.2	58.0	20.6	32.0	35.1	**32.0**	35.5	28.0	40.8	16.0	42.0	46.0	28.8	33.4	34.4
Synapse	**42.0**	81.4	35.0	86.1	**48.0**	76.4	26.0	135.3	32.0	92.4	**40.0**	77.8	**62.0**	52.0	40.7	85.9
Training-Free GRPO	10.0	81.2	60.0	38.4	26.0	75.9	28.0	76.5	24.0	78.3	17.0	119.0	42.0	63.5	29.6	76.1
ActiveContext (Ours)	27.9	8.1	**66.0**	3.7	40.9	7.3	31.8	8.4	**36.4**	9.6	25.6	8.3	61.9	5.6	**41.5**	7.3
TaskExecutor: gpt-4o																
Full Context	22.0	40.8	58.0	20.7	32.0	35.1	32.0	35.5	**28.0**	38.2	16.0	42.0	46.0	28.8	33.4	34.4
Synapse	**48.0**	38.9	36.0	42.0	**40.0**	47.8	18.0	52.1	**28.0**	45.9	**46.0**	38.9	**58.0**	27.4	**39.1**	41.9
Training-Free GRPO	26.0	35.3	58.0	19.6	36.0	31.0	34.0	35.5	**28.0**	38.9	15.0	42.8	34.0	31.9	33.0	33.6
ActiveContext (Ours)	28.0	7.2	**66.0**	3.6	26.0	7.7	**38.0**	7.4	26.0	9.3	15.0	9.1	44.0	6.4	34.7	7.2
TaskExecutor: gpt-4o-mini(with Extended Ablations)																
Baselines & Retrieval-Based Memory																
Full Context	20.0	35.5	50.0	19.7	34.0	33.8	30.0	37.2	40.0	33.8	15.0	42.1	40.0	30.0	32.7	33.2
Synapse	**52.0**	38.6	37.0	42.7	**38.0**	50.5	20.0	55.9	38.0	45.0	**42.0**	41.8	**60.0**	30.9	41.0	43.6
AWM	24.0	39.1	56.0	16.4	34.0	34.8	25.0	36.2	39.0	34.2	18.0	43.2	38.0	32.1	33.4	33.7
Mem0	23.0	46.7	**62.0**	27.6	32.0	47.9	**31.0**	52.1	**42.0**	46.7	15.0	59.7	40.0	45.4	34.7	46.6
Training-Free GRPO	20.0	40.1	64.0	18.3	32.0	36.2	30.0	39.1	40.0	36.1	16.0	45.4	34.0	35.0	33.7	35.7
Prompting-based Memory Management(Zero-Shot Curators)																
Gemini-3.0-flash	**34.0**	8.0	**66.0**	4.4	**40.0**	8.0	**38.0**	8.7	**42.0**	9.5	**19.0**	9.6	**56.0**	6.3	**42.1**	7.8
Gemini-2.5-flash	28.0	7.9	64.0	4.6	30.0	8.2	32.0	8.3	34.0	9.1	17.0	9.7	52.0	6.1	36.7	7.7
gpt-4o	30.0	11.2	58.0	6.9	34.0	12.0	30.0	11.4	32.0	11.2	17.0	13.4	44.0	9.2	35.0	10.8
gpt-4o-mini	28.0	11.2	56.0	7.3	32.0	10.8	28.0	10.9	30.0	10.0	17.0	12.6	40.0	8.9	33.0	10.2
Qwen2.5-7B-Instruct	22.0	10.7	52.0	6.5	28.0	9.9	28.0	10.0	28.6	10.5	10.0	12.9	44.0	8.7	30.4	9.9
Active Memory Management (RL-Trained Curator)																
ContextCurator (Ours)	28.0	7.0	62.0	5.0	38.0	7.6	34.0	7.3	38.0	7.4	12.0	9.3	32.0	7.3	34.9	7.3

icantly fewer tokens (7.3k vs 10.8k).

Case Study. As detailed in Appendix A.4, ActiveContext's performance on the WebArena benchmark practically demonstrates its ability to organically condense multi-turn interaction history into concise, actionable memory. This mechanism ensures the GPT-4o-mini model maintains focus on the distal task objective, avoids getting lost in verbose DOM observations, and efficiently arrives at the correct answer during complex multi-step navigation.

Collectively, these results across the Web and Search domains demonstrate that ActiveContext is not merely a compression heuristic, but a fundamental *reasoning enhancer*. By decoupling context management from execution and optimizing the former via reinforcement learning, we establish a highly scalable, efficient, and performant paradigm for autonomous long-horizon agents.

5 Conclusion

In this work, we study the *context bottleneck* that limits LLM agents in long-horizon environments. We introduce ActiveContext, a symbiotic framework that formulates context memory management as an active sequential decision-making process rather than a static storage paradigm. By decoupling the ContextCurator from the TaskExecutor and optimizing it with on-policy reinforcement learning, ActiveContext treats memory as a dynamic control variable for efficient reasoning. Experiments on WebArena and DeepSearch show that ActiveContext achieves a superior performance–efficiency trade-off. It improves the reasoning upper bound of strong LLM agents while substantially reducing token consumption, with up to an 85% reduction in retrieval-heavy settings such as Gemini-3.0-flash on DeepSearch. Notably, an RL-finetuned 7B curator can match the context-management performance of proprietary frontier models such as GPT-4o, suggesting that active context curation is a specialized capability distinct from general reasoning. These results suggest a scalable path toward sustainable autonomous agents: rather than relying on ever-larger context windows, agents can improve both efficiency and reasoning quality by actively managing their cognitive span.

Limitations

Due to computational resource constraints, this work only trains a 7B model as the ContextCurator. Although the proposed framework already demonstrates strong empirical performance, we believe that scaling ContextCurator to larger and more capable foundation models could further improve its context selection accuracy, reasoning robustness, and generalization ability. We leave the exploration of larger-scale ContextCurator models for future work.

References

Yuzheng Cai, Siqi Cai, Yuchen Shi, Zihan Xu, Lichao Chen, Yulei Qin, Xiaoyu Tan, Gang Li, Zongyi Li, Haojia Lin, Yong Mao, Ke Li, and Xing Sun. 2025. Training-free group relative policy optimization. *Preprint*, arXiv:2510.08191.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *ArXiv preprint*, abs/2504.19413.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *ArXiv preprint*, abs/2507.06261.

Yiming Du, Wenyu Huang, Danna Zheng, Zhaowei Wang, Sebastien Montella, Mirella Lapata, Kam-Fai Wong, and Jeff Z Pan. 2025. Rethinking memory in ai: Taxonomy, operations, topics, and future directions. *arXiv e-prints*, pages arXiv–2505.

Izzeddin Gur, Hiroki Furuta, Austin V. Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. 2024. A real-world webagent with planning, long context understanding, and program synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Mengkang Hu, Tianxing Chen, Qiguang Chen, Yao Mu, Wenqi Shao, and Ping Luo. 2025. HiAgent: Hierarchical working memory management for solving long-horizon agent tasks with large language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32779–32798, Vienna, Austria. Association for Computational Linguistics.

Jen-tse Huang, Kaiser Sun, Wenxuan Wang, and Mark Dredze. 2025. Language models do not have human-like working memory. *arXiv preprint arXiv:2505.10571*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Dongming Jiang, Yi Li, Guanpeng Li, and Bingzhe Li. 2026. Magma: A multi-graph based agentic memory architecture for ai agents. *arXiv preprint arXiv:2601.03236*.

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.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025b. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *Preprint*, arXiv:2503.09516.

LangChain. 2024. Langgraph: Building stateful, multi-actor applications with llms. `https://github.com/langchain-ai/langgraph`.

Bang Liu, Xinfeng Li, Jiayi Zhang, Jinlin Wang, Tanjin He, Sirui Hong, Hongzhang Liu, Shaokun Zhang, Kaitao Song, Kunlun Zhu, and 1 others. 2025. Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems. *ArXiv preprint*, abs/2504.01990.

Tianshi Ming, Xian Wu, Yingying Zhang, Zichuan Fu, and Dawei Cheng. 2025. Anchorcot: Anchors pave the way for multi-hop reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15522–15536.

OpenAI. 2024. Learning to reason with llms.

Siru Ouyang, Jun Yan, I-Hung Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long T. Le, Samira Daruki, Xiangru Tang, Vishy Tirumalashetty, George Lee, Mahsan Rofouei, Hangfei Lin, Jiawei Han, Chen-Yu Lee, and Tomas Pfister. 2025. Reasoningbank: Scaling agent self-evolving with reasoning memory. *Preprint*, arXiv:2509.25140.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2024. Memgpt: Towards llms as operating systems. *Preprint*, arXiv:2310.08560.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv: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. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, and 1 others. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. 2025. Agent workflow memory. In *Forty-second International Conference on Machine Learning*.

Zhepei Wei, Wenlin Yao, Yao Liu, Weizhi Zhang, Qin Lu, Liang Qiu, Changlong Yu, Puyang Xu, Chao Zhang, Bing Yin, Hyokun Yun, and Lihong Li. 2025. Webagent-r1: Training web agents via end-to-end multi-turn reinforcement learning. *Preprint*, arXiv:2505.16421.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. 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*.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z Pan, and 1 others. 2025. Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. *arXiv preprint arXiv:2508.19828*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024a. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2024b. Swe-agent: Agent-computer interfaces enable automated software engineering. 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*.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. WebShop: Towards scalable real-world web interaction with grounded language agents. In *NeurIPS*.

Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, and 1 others. 2025. Memagent: Reshaping long-context llm with multi-conv rl-based memory agent. *ArXiv preprint*, abs/2507.02259.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024. A survey on the memory mechanism of large language model based agents. *ACM Transactions on Information Systems*.

Zhiwei Zhang, Xiaomin Li, Yudi Lin, Hui Liu, Ramraj Chandradevan, Linlin Wu, Minhua Lin, Fali Wang, Xianfeng Tang, Qi He, and 1 others. 2025. Unlocking the power of multi-agent llm for reasoning: From lazy agents to deliberation. *arXiv preprint arXiv:2511.02303*.

Longtao Zheng, Rundong Wang, Xinrun Wang, and Bo An. 2024. Synapse: Trajectory-as-exemplar prompting with memory for computer control. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19724–19731. AAAI Press.

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. 2024. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. 2025. Mem1: Learning to synergize memory and reasoning for efficient long-horizon agents. *ArXiv preprint*, abs/2506.15841.

A Appendix

A.1 Use of Large Language Models

Large Language Models are used for grammar check and polishing in this paper.

A.2 Training Details.

Our policy model is trained on a single node equipped with 8 NVIDIA A100 GPUs (80GB), using full-parameter fine-tuning. For the MT-GRPO training, we utilize a constant learning rate of $1e-6$ with a batch size of 8. The rollout number G is set to 4 for WebArena-Lite and 8 for DeepSearch. In both tasks, the number of interaction rounds with environment is capped at 15. The KL divergence regularization coefficient β and the clip ratio ϵ are set to 0.001 and 0.2, respectively. The maximum context length and the number of new tokens are configured to 16,000 and 512. For efficient LLM rollouts during RL, we leverage vLLM with a tensor parallel size of 1 and a GPU memory utilization ratio of 0.6. Rollout sampling is performed with both the temperature and top-p set to 1.0.

A.3 Context Calculation

Effective context management is paramount for long-horizon tasks in web and search environments, where agents interact through multiple turns and observations can be highly dimensional. This subsection details how the context length is calculated in both the WebArena and DeepSearch tasks under three distinct memory configurations: *No Memory*, *Full Context*, and our proposed *ActiveContext*.

We define the context length for a single turn t as C_t. The total context length for a task spanning L turns is then $\sum_{t=1}^{L} C_t$.

A.3.1 No Memory

The *No Memory* configuration streamlines the agent's input at each turn by restricting the conversational context to only immediate necessities, completely discarding prior interaction history.

For web exploration tasks, the input at turn t is limited to the system prompt (S), the current web XTREE observation (O_t^{XTREE}), and the task objective (U). The context length for turn t under the *No Memory* configuration, C_t^{NoMem}, is formally expressed as:

$$C_t^{\text{NoMem}} = \text{len}(S) + \text{len}(O_t^{\text{XTREE}}) + \text{len}(U) \quad (6)$$

While this approach significantly reduces the per-turn context length compared to *Full Context*, it often leads to a loss of crucial long-term dependencies and sequential understanding necessary for complex, multi-step tasks.

Notably, for Search tasks, a *No Memory* configuration was not evaluated due to the prevalence of multi-hop search requirements, which are inherently impossible to solve without historical context. In contrast, Web tasks can still proceed under a *No Memory* setup, as the agent can often infer subsequent operations based purely on the current visual state of the web observation and the global objective.

A.3.2 Full Context

In the *Full Context* configuration, the agent's input at each turn t comprises the cumulative historical conversation. This includes the initial system prompt, user prompts (observations and objectives), and assistant prompts (reasoning and actions).

- **Web Tasks:** To prevent catastrophic context bloat from massive DOM trees, historical observations are replaced with a concise placeholder (P). The input at turn t includes the system prompt (S), the *current* observation (O_t^{AAT}), placeholders for all past observations ($\sum_{k=0}^{t-1} P$), the task objective (U), and all prior assistant prompts detailing reasoning (Re_k) and actions (A_k).

- **Search Tasks:** Because text retrieval is less structurally verbose than DOM trees, placeholders are not used. The input includes the system prompt (S), the entire history of retrieved information ($\sum_{k=0}^{t} R_k$), the final objective (U), and all previous assistant prompts ($\sum_{k=0}^{t-1}(Re_k, A_k)$).

The context length for turn t under the *Full Context* configuration, $C_t^{\text{FullContext}}$, is formally expressed as:

$$C_{t,\text{Web}}^{\text{FullContext}} = \text{len}(S) + \text{len}(O_t^{\text{AAT}}) + \sum_{k=0}^{t-1} \text{len}(P) + \text{len}(U) + \sum_{k=0}^{t-1} \text{len}(Re_k, A_k). \quad (7)$$

$$C_{t,\text{Search}}^{\text{FullContext}} = \text{len}(S) + \sum_{k=0}^{t} \text{len}(R_k) + \text{len}(U) + \sum_{k=0}^{t-1} \text{len}(Re_k, A_k). \quad (8)$$

A significant challenge with *Full Context* is that the input context length grows linearly with the number of interaction turns t. This cumulative growth inevitably triggers the "Context Bottleneck" described previously, significantly increasing computational costs and exacerbating the "Lost-in-the-Middle" phenomenon, thereby severely impeding reasoning capabilities.

A.3.3 ActiveContext

Under the *ActiveContext* framework, the TaskExecutor's input at each turn t is augmented with a refined working memory $\mathcal{M}_t$ dynamically generated by the ContextCurator. This module is specifically designed to condense salient information from the interaction history into a highly compact and relevant representation.

- **Web Tasks:** The TaskExecutor's input includes the system prompt (S), the current web XTREE observation (O_t^{XTREE}), the objective (U), and crucially, the refined memory ($\mathcal{M}_t^{\text{Web}}$). This memory represents a distilled summary of past interactions and reasoning anchors, rather than a raw, concatenated history.

- **Search Tasks:** The TaskExecutor's input includes the system prompt (S), the just-retrieved search results (R_t), the objective (U), and the refined memory ($\mathcal{M}_t^{\text{Search}}$) containing historical search content, key retrieved evidence, and past erroneous deductions.

The context length for turn t under the *ActiveContext* configuration, C_t^{Active}, is formally expressed as:

$$C_{t,\text{Web}}^{\text{Active}} = \text{len}(S) + \text{len}(O_t^{\text{XTREE}}) + \text{len}(U) + \text{len}(\mathcal{M}_t^{\text{Web}}). \quad (9)$$

$$C_{t,\text{Search}}^{\text{Active}} = \text{len}(S) + \text{len}(R_t) + \text{len}(U) + \text{len}(\mathcal{M}_t^{\text{Search}}). \quad (10)$$

This symbiotic approach strikes an optimal balance: it halts the linear growth of context length observed in the *Full Context* baseline while actively retaining essential historical anchors. By aggressively filtering noise, it mitigates the context bottleneck and maximizes the signal-to-noise ratio in the TaskExecutor's decision-making process. Although the per-turn input might occasionally be slightly longer than in the *No Memory* setup, *ActiveContext* achieves a substantially higher task success rate and requires fewer exploratory interaction turns per task. This ultimately results in a significantly shorter total context sequence required for successful task completion, demonstrating superior computational efficiency and reasoning effectiveness.

A.4 Case Study: Multi-turn Interaction with ActiveContext

To illustrate the effectiveness of our *ActiveContext* framework, we present a detailed case study of a multi-turn interaction from the WebArena benchmark. This example demonstrates how the `gpt-4o-mini` model, acting as the *TaskExecutor* and equipped with a RL-trained *ContextCurator*, navigates a complex web environment to achieve a specific objective, effectively managing context and refining its working memory.

A.4.1 Task Objective

The agent's objective for this task is: *"Tell me the count of comments that have received more downvotes than upvotes for the user who made the latest post on the Showerthoughts forum."*

A.4.2 Interaction Log Analysis

The following sequence details the agent's steps, its reasoning, and the dynamically refined memory provided by the *ContextCurator*. We highlight two critical, consecutive turns (Turn 7 and Turn 8) in the interaction to demonstrate the information entropy reduction.

Turn 7: Navigating to the Comments subsection

1. ContextCurator (Memory Update):
The WebAgent identified 'MarvelsGrantMan136' as the user who made the latest post under the 'Postmill' tab. This post had no comments. The agent then navigated to 'MarvelsGrantMan136's profile.
Initial traversal of the user's profile and previous posts in the 'Postmill' tab did not yield relevant results regarding comments or vote counts, except for one old comment with a single downvote. To view all comments, the WebAgent clicked on the 'Comments' link on the user's profile page.
The next page loaded, displaying a list of comments made by 'MarvelsGrantMan136'.
The current page structure, as observed:

- Main content area for ‘MarvelsGrantMan136’
- Navigation links: ‘Overview’, ‘Submissions’, ‘Comments’ (link ID: [3018])
- An article subsection, potentially containing comment details or links to posts that have comments.

- -

2. Latest Web Observation:

```
Tab 0 (current): MarvelsGrantMan136
  [2008] RootWebArea ‘MarvelsGrantMan136’ focused: True
    [2287] HeaderAsNonLandmark
      [2288] link ‘Home’
        [2334] StaticText ‘Postmill’
    [2335] link ‘Forums’
    [2336] link ‘Wiki’
    [2344] searchbox ‘Search query’
    [2345] link ‘Notifications (0)’
    [2346] link ‘Submit’
    [2019] button ‘MarvelsGrantMan136’ hasPopup: menu expanded: False
    [2021] main
      [3016] link ‘Overview’
      [3017] link ‘Submissions’
      [3018] link ‘Comments’
      [2022] article
        [2025] HeaderAsNonLandmark
          [2354] heading ‘How Machine Learning Can Revolutionize the Book Industry’
            [3019] link ‘How Machine Learning Can Revolutionize the Book Industry’
          [3021] StaticText ‘Submitted by ’
          [2028] link ‘MarvelsGrantMan136’ expanded: False
          [3023] time ‘January 13, 2026 at 12:55:18 PM EST’
            [3609] StaticText ‘16 hours ago’
          [3024] StaticText ‘ in ’
          [3025] link ‘books’
        [3612] link ‘No comments’
        [3614] link ‘Edit’
        [4077] button ‘Delete’
        [2358] button ‘Retract upvote’
        [2361] button ‘Downvote’
      [2029] article
        [2303] checkbox ‘Hide –’ checked: true
        [2032] HeaderAsNonLandmark
          [2033] heading ‘MarvelsGrantMan136 wrote 16 hours ago’
            [2035] link ‘MarvelsGrantMan136’ expanded: False
            [3039] time ‘January 13, 2026 at 12:51:42 PM EST’
        [2376] StaticText ‘Reply to ’
        [2377] link ‘AI experts disown Musk-backed campaign citing their research’
        [2378] StaticText ‘ by ’
        [2039] link ‘Don_Gato1’ expanded: False
        [2380] StaticText
        [3628] link ‘Reply’
        [3629] link ‘Permalink’
        [4082] button ‘Delete’
        [3631] link ‘Edit’
        [2384] button ‘Retract upvote’
        [2387] button ‘Downvote’
      [2040] article
        [2043] HeaderAsNonLandmark
          [2391] heading ‘How Machine Learning Can Transform the Corresponding Field How Machine Learning Can Transform the Corresponding Field’
            [3060] link ‘How Machine Learning Can Transform the Corresponding Field How Machine Learning Can Transform the Corresponding Field’
          [3062] StaticText ‘Submitted by ’
          [2046] link ‘MarvelsGrantMan136’ expanded: False
          [3064] time ‘January 13, 2026 at 9:09:04 AM EST’
            [3646] StaticText ‘20 hours ago’
          [3065] StaticText ‘ in ’
          [3066] link ‘books’
        [3649] link ‘No comments’
        [3651] link ‘Edit’
        [4088] button ‘Delete’
        [2395] button ‘Upvote’
        [2398] button ‘Downvote’
      [2047] article
        [2304] checkbox ‘Hide –’ checked: true
        [2050] HeaderAsNonLandmark
          [2051] heading ‘MarvelsGrantMan136 wrote 20 hours ago’
            [2053] link ‘MarvelsGrantMan136’ expanded: False
            [3080] time ‘January 13, 2026 at 9:05:14 AM EST’
        [2413] StaticText ‘Reply to ’
        [2414] link ‘AI experts disown Musk-backed campaign citing their research’
        [2415] StaticText ‘ by ’
        [2057] link ‘Don_Gato1’ expanded: False
        [2417] StaticText
        [3665] link ‘Reply’
        [3666] link ‘Permalink’
        [4093] button ‘Delete’
        [3668] link ‘Edit’
        [2421] button ‘Retract upvote’
        [2424] button ‘Downvote’
      [2058] article
        [2305] checkbox ‘Hide –’ checked: true
        [2061] HeaderAsNonLandmark
          [2062] heading ‘MarvelsGrantMan136 wrote 23 hours ago’
            [2064] link ‘MarvelsGrantMan136’ expanded: False
            [3102] time ‘January 13, 2026 at 5:59:50 AM EST’
        [2439] StaticText ‘Reply to ’
        [2440] link ‘Driving in Pittsburgh summed up by one traffic sign.’
        [2068] link ‘Dare_County’ expanded: False
        [2443] StaticText ‘Yeah, pittsburgh traffic, you know...Yeah, pittsburgh traffic, you know...’
        [3681] link ‘Reply’
        [3682] link ‘Permalink’
        [4099] button ‘Delete’
        [3684] link ‘Edit’
        [2447] button ‘Retract upvote’
        [3115] StaticText ‘1’
        [2450] button ‘Downvote’
    [2277] complementary
      [2322] HeaderAsNonLandmark
        [2902] heading ‘MarvelsGrantMan136’
          [3574] link ‘MarvelsGrantMan136’
        [3577] StaticText ‘Registered ’
        [3578] time ‘June 24, 2020 at 11:02:10 AM EDT’
          [4061] StaticText ‘6 years ago’
        [3581] StaticText ‘Not whitelisted’
      [2905] StaticText ‘t2_5adwlxvnAwesome Prompt ArtistFreelance Web DeveloperAwesome Prompt ArtistAwesome Prompt ArtistAwesome Prompt ArtistFreelance Web DeveloperI am a robotI am a robotPro Python Developer withFreelance Web Developer Freelance Web Developer Freelance Web Developer 20 years of Experience’
      [2906] link ‘Edit biography’
      [2325] heading ‘Toolbox’
      [3586] link ‘Hidden forums’
      [3587] link ‘Trash’
      [2327] heading ‘MarvelsGrantMan136 is a moderator on:’
      [3588] link ‘PlantsForCatParents’
      [3589] link ‘cmu_lti’
      [3590] link ‘KaraokeKaraoke’
      [3591] link ‘Karaoke’
      [3592] link ‘sci_fi’
      [3593] link ‘cmu_lticmu_lti’

URL: http://3.151.53.169:9999/user/MarvelsGrantMan136
```

- -

3. TaskExecutor's Reasoning and Action (`gpt-4o-mini`):
Let's think step-by-step. According to the memory management agent, I have navigated to the comments subsection for the user 'MarvelsGrantMan136'. The previous action was to click on the 'Comments' link, which should have led to a list of comments made by this user. However, the observation does not provide specific details about the comments or their vote counts.
Since I need to find out how many comments have received more downvotes than upvotes, I should look for the relevant information regarding the comments made by 'MarvelsGrantMan136'. The next logical step is to ensure that I am viewing the comments subsection properly and to check if there are any comments listed that indicate their vote counts.
In summary, the next action I will perform is "'click [3017]"' to access the comments made by 'MarvelsGrantMan136'.

Turn 8: Task Completion

1. ContextCurator (Memory Update):
The WebAgent initially found that 'MarvelsGrantMan136' made the latest post under the 'Postmill' tab, which had no comments. The agent then traversed the user's profile to view posts and comments. After viewing the profile, the 'Comments' link was clicked to see all comments made by 'MarvelsGrantMan136'.
The page then loaded a list of comments from 'MarvelsGrantMan136'. Subsequently, the WebAgent clicked on the 'Comments' link (ID [3017]) to ensure proper navigation within the comments subsection.

- -

2. Latest Web Observation:

```
Tab 0 (current): Submissions
  [4254] RootWebArea 'Submissions' focused: True
    [4458] HeaderAsNonLandmark
      [4459] link 'Home'
        [4487] StaticText 'Postmill'
    [4488] link 'Forums'
    [4489] link 'Wiki'
    [4497] searchbox 'Search query'
    [4498] link 'Notifications (0)'
    [4499] link 'Submit'
    [4265] button 'MarvelsGrantMan136' hasPopup: menu
expanded: False
    [4267] main
      [4946] link 'Overview'
      [4947] link 'Submissions'
      [4948] link 'Comments'
      [4268] article
        [4271] HeaderAsNonLandmark
          [4507] heading 'The Last Of Us' Moves
Production to Vancouver For Season 2
            [4949] link 'The Last Of Us' Moves
Production to Vancouver For Season 2
          [4509] link 'deadline.com'
          [4952] StaticText 'Submitted by '
          [4274] link 'MarvelsGrantMan136' expanded:
False
          [5536] StaticText 't3_127rvdp'
          [4956] time 'March 31, 2023 at 1:02:17 PM
EDT'
            [5540] StaticText '3 years ago'
          [4957] StaticText ' in '
          [4958] link 'television'
        [5543] link 'No comments'
        [5545] link 'Edit'
        [6200] button 'Delete'
        [4513] button 'Upvote'
        [4516] button 'Downvote'
      [4275] article
        [4278] HeaderAsNonLandmark
          [4520] heading "'Secret Invasion' Revealed:
Inside Samuel L. Jackson's Eye-Opening New Marvel
Series"
            [4971] link "'Secret Invasion' Revealed:
Inside Samuel L. Jackson's Eye-Opening New Marvel
Series"
          [4522] link 'vanityfair.com'
          [4974] StaticText 'Submitted by '
          [4281] link 'MarvelsGrantMan136' expanded:
False
          [5562] StaticText 't3_127jaxx'
          [4978] time 'March 31, 2023 at 8:12:27 AM
EDT'
            [5566] StaticText '3 years ago'
          [4979] StaticText ' in '
          [4980] link 'television'
        [5569] link 'No comments'
        [5571] link 'Edit'
        [6205] button 'Delete'
        [4526] button 'Upvote'
        [4529] button 'Downvote'
      [4282] article
        [4285] HeaderAsNonLandmark
          [4533] heading 'Star Trek: Starfleet Academy
Series, From Alex Kurtzman and Nancy Drew Creator,
Ordered at Paramount+'
            [4993] link 'Star Trek: Starfleet Academy
Series, From Alex Kurtzman and Nancy Drew Creator,
Ordered at Paramount+'
          [4535] link 'tvline.com'
          [4996] StaticText 'Submitted by '
          [4288] link 'MarvelsGrantMan136' expanded:
False
          [5588] StaticText 't3_126sut3'
          [5000] time 'March 30, 2023 at 1:00:27 PM
EDT'
            [5592] StaticText '3 years ago'
          [5001] StaticText ' in '
          [5002] link 'television'
        [5595] link 'No comments'
        [5597] link 'Edit'
        [6210] button 'Delete'
        [4539] button 'Upvote'
        [4542] button 'Downvote'
      [4289] article
        [4292] HeaderAsNonLandmark
          [4546] heading "'Fargo' Season 5: FX Unveils
First Look at Jon Hamm, Premieres in Late 2023"
            [5015] link "'Fargo' Season 5: FX Unveils
First Look at Jon Hamm, Premieres in Late 2023"
          [4548] link 'tvinsider.com'
          [5018] StaticText 'Submitted by '
          [4295] link 'MarvelsGrantMan136' expanded:
False
          [5614] StaticText 't3_12686kh'
          [5022] time 'March 29, 2023 at 10:05:47 PM
EDT'
            [5618] StaticText '3 years ago'
          [5023] StaticText ' in '
          [5024] link 'television'
```

```
        [5621] link 'No comments'
        [5623] link 'Edit'
        [6215] button 'Delete'
        [4552] button 'Upvote'
        [5029] StaticText '0'
        [4555] button 'Downvote'
      [4296] article
        [4299] HeaderAsNonLandmark
          [4559] heading 'Disney Lays Off Ike Perlmutter, Chairman of Marvel Entertainment'
            [5037] link 'Disney Lays Off Ike Perlmutter, Chairman of Marvel Entertainment'
          [4561] link 'nytimes.com'
          [5040] StaticText 'Submitted by '
          [4302] link 'MarvelsGrantMan136' expanded: False
          [5640] StaticText 't3_125s0rj'
          [5044] time 'March 29, 2023 at 11:49:11 AM EDT'
            [5644] StaticText '3 years ago'
          [5045] StaticText ' in '
          [5046] link 'television'
        [5647] link 'No comments'
        [5649] link 'Edit'
        [6220] button 'Delete'
        [4565] button 'Upvote'
        [5051] StaticText '3'
        [4568] button 'Downvote'
      [4303] article
        [4306] HeaderAsNonLandmark
          [4572] heading 'House Of The Dragon' To Get Sorter Season 2 (8 Episodes) As HBO Series Eyes Season 3 Greenlight
            [5059] link 'House Of The Dragon' To Get Shorter Season 2 (8 Episodes) As HBO Series Eyes Season 3 Greenlight
          [4574] link 'deadline.com'
          [5062] StaticText 'Submitted by '
          [4309] link 'MarvelsGrantMan136' expanded: False
          [5666] StaticText 't3_1257zaj'
          [5066] time 'March 28, 2023 at 8:42:17 PM EDT'
            [5670] StaticText '3 years ago'
          [5067] StaticText ' in '
          [5068] link 'television'
        [5673] link 'No comments'
        [5675] link 'Edit'
        [6225] button 'Delete'
        [4578] button 'Upvote'
        [5073] StaticText '1'
        [4581] button 'Downvote'
      [4310] article
        [4313] HeaderAsNonLandmark
          [4585] heading"'Star Trek: Strange New Worlds' (Season 3) and 'Lower Decks' (Season 5) Renewed at Paramount+; 'Strange New Worlds' Season 2 Premieres June 15, 'Lower Decks' Season 4 Premieres Summer 2023"
            [5081] link "'Star Trek: Strange New Worlds' (Season 3) and 'Lower Decks' (Season 5) Renewed at Paramount+; 'Strange New Worlds' Season 2 Premieres June 15, 'Lower Decks' Season 4 Premieres Summer 2023"
        [4591] button 'Upvote'
    [4448] complementary
      [4475] HeaderAsNonLandmark
        [4832] heading 'MarvelsGrantMan136'
          [5500] link 'MarvelsGrantMan136'
        [5503] StaticText 'Registered '
        [5504] time 'June 24, 2020 at 11:02:10 AM EDT'
          [6184] StaticText '6 years ago'
        [5507] StaticText 'Not whitelisted'
      [4835] StaticText 't2_5adwlxvnAwesome Prompt ArtistFreelance Web DeveloperAwesome Prompt ArtistAwesome Prompt ArtistAwesome Prompt ArtistFreelance Web DeveloperI am a robotI am a robotPro Python Developer withFreelance Web Developer Freelance Web Developer Freelance Web Developer 20 years of Experience'
      [4836] link 'Edit biography'
      [4478] heading 'Toolbox'
      [5512] link 'Hidden forums'
      [5513] link 'Trash'
      [4480] heading 'MarvelsGrantMan136 is a moderator on:'
      [5514] link 'PlantsForCatParents'
      [5515] link 'cmu_lti'
      [5516] link 'KaraokeKaraoke'
      [5517] link 'Karaoke'
      [5518] link 'sci_fi'
      [5519] link 'cmu_lticmu_lti'

URL: http://3.151.53.169:9999/user/MarvelsGrantMan136/submissions
```

---

3. TaskExecutor's Reasoning and Action (`gpt-4o-mini`):

Let's think step-by-step. According to the memory management agent, I have navigated to the comments subsection for the user 'MarvelsGrantMan136'. The comments listed show that there are a total of 4 comments made by this user, and all of them either have no votes or upvotes, but none have received more downvotes than upvotes. Therefore, the count of comments that have received more downvotes than upvotes is 0. In summary, the next action I will perform is "'stop [0]"'

---

4. Environment Feedback:

The agent correctly deduces that the count is 0 and issues the 'stop' action, indicating successful task completion and a correct final answer (reward 1).

A.4.3 Conclusion of Case Study

This case study vividly demonstrates the *Active-Context* architecture's ability to effectively condense a verbose multi-turn interaction history into concise, actionable memory. This mechanism allows the `gpt-4o-mini` model to maintain focus on the complex distal objective, avoid Attention Dilution from massive DOM trees, and ultimately arrive at the correct answer efficiently. The dynamic refinement of working memory by the *ContextCurator* is crucial for enabling robust performance in long-horizon tasks, contributing significantly to both token efficiency and reasoning accuracy.