ZipRL: Adaptive Multi-Turn Context Compression with Hindsight Response Replay

Zhexin Hu[1,2*] **Li Wang**[1*] **Xiaohan Wang**[1,*†] **Jiajun Chai**[1] **Xiaojun Guo**[1]
Wei Lin[1] **Xin Zhou**[1] **Guojun Yin**[1†]

[1]Meituan
[2]Institute of Software, Chinese Academy of Sciences

Abstract

Adaptive context compression is vital for scaling Large Language Models (LLMs) to complex, multi-turn agent tasks. However, rule-based compression methods may discard task-critical nuances, while Reinforcement Learning (RL) approaches usually struggle to balance information retention and token efficiency under the sparse rewards inherent to long-horizon workflows. To bridge this gap, we propose **ZipRL**, a novel adaptive compression framework tailored for Reinforcement Learning from Verifiable Rewards (RLVR). ZipRL features a multi-granularity compression mechanism for active, non-uniform information reduction, coupled with Hindsight Response Replay (HRR), a technique designed to densify training signals during RLVR optimization. Theoretically, we prove ZipRL's superior task-relevant utility over uniform methods. Concretely, ZipRL utilizes coarse-to-fine prompts for macro-compression and incorporates HRR into GRPO via generalized advantage reshaping. Multiple models of varying versions and parameter scales validate the effectiveness of our approach. Benchmarks on five agent tasks show ZipRL outperforms state-of-the-art approaches by 27.9% and 34.7% across Qwen3-4B and Qwen3-8B models, while maintaining exceptional token efficiency and robustness under extreme 256-turn extrapolation stress tests. Our code is available at `https://github.com/huzhexin/ZipRL`.

1 Introduction

Large language models (LLMs) have emerged as powerful agents for solving long-horizon tasks that require extended reasoning, planning, and interaction with external environments (Ke et al., 2025; Wu et al., 2024, 2026). Effective context management is crucial for unlocking the full potential of these capabilities, as it strategically optimizes the use of the finite context window to ensure coherent, accurate, and efficient model responses (Shao et al., 2024; Song et al., 2023). For example, recent research has integrated LLMs with external environments, such as research agents, enabling LLMs to retrieve and process dynamic information for open-ended problem solving (Jin et al., 2025). However, in such multi-turn scenarios, particularly agentic search, the accumulation of retrieved documents and interaction history rapidly consumes the limited context window. This necessitates effective context compression strategies that preserve critical information while maintaining token efficiency, ensuring the agent operates within computational constraints without sacrificing performance.

Despite the critical role of context compression, existing approaches encounter two key challenges, as illustrated in Figure 1. First, current methods typically employ uniform processing granularity (Song et al., 2025; He et al., 2026), treating all retrieved segments with equal importance. This lack of relevance awareness is suboptimal from an information-theoretic perspective: as we demonstrate, uniform compression often fails to maximize the task-relevant information utility retained from the context for the target query. While irrelevant documents introduce noise, critical details in relevant documents may be lost if over-compressed. Second, optimizing compression policies via Reinforcement Learning (RL) is notoriously difficult due to sparse reward signals brought by long-horizon multi-turn interaction (Yu et al., 2026). In multi-turn tasks, supervision is often only available at the final outcome (e.g., success or failure), making it challenging to credit specific intermediate compression actions. Although Process Reward Models (PRMs) (Ma et al., 2023; Zhang et al., 2025b) attempt to mitigate this, their performance relies heavily on task-specific reward designs and may encourage greedy strategies.

To address these challenges, we introduce

*Equal contribution
†Corresponding authors

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 34371–34391
October 24-29, 2026 ©2026 Association for Computational Linguistics

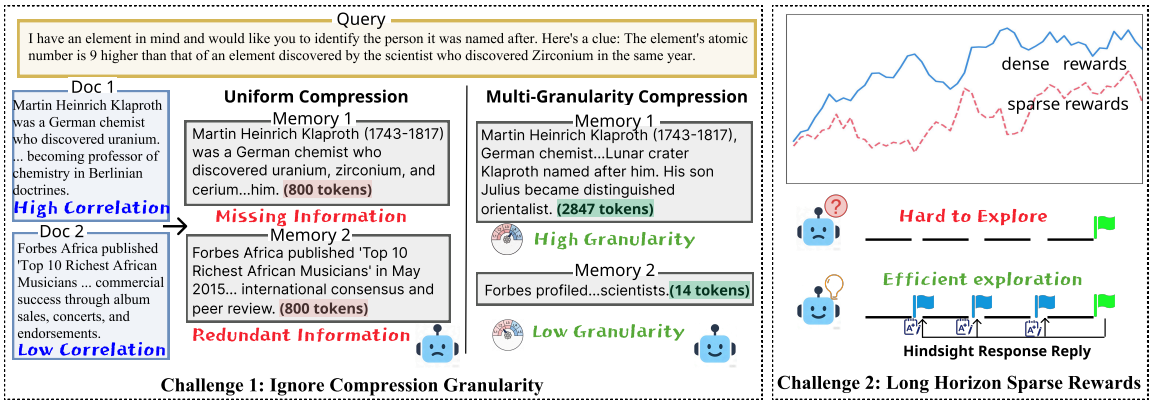

Figure 1: Challenges faced by current long-context compression and search agents, with ZipRL addressing these by incorporating the Multi-Granularity Mechanism and Hindsight Response Replay for improved compression performance.

ZipRL, an adaptive multi-turn context compression framework designed to balance information preservation and efficiency. Grounded in the principle that multi-granularity compression preserves more task-relevant information utility, ZipRL enables the agent to actively perceive the relevance of retrieved content and dynamically select macro-level compression strategies (coarse-to-fine) via in-context prompts. Furthermore, to tackle the sparse reward problem in RL training with context compression, we propose Hindsight Response Replay (HRR) inspired by the idea of Hindsight Experience Replay (HER) in conventional RL. Instead of relying on expensive external PRMs, HRR utilizes a context-oriented heuristic metric to score compression segments efficiently. These scores are then integrated into Group Relative Policy Optimization (GRPO) through advantage re-shaping, densifying the training signals and allowing the model to learn optimal compression policies from the final outcome reward. Experimental results on five Multi-hop QA and Web Browsing benchmarks across multiple models of varying versions and parameter scales validate the effectiveness of our approach. It outperforms the strongest baselines by 27.9% and 34.7% on average for 4B and 8B parameter models, respectively, while maintaining superior performance in long-context interactions (up to 256 turns) with high token efficiency. Our contributions are summarized as follows:

- We identify the limitations of uniform processing in long-context tasks and theoretically validate that multi-granularity compression better preserves task-relevant information utility under general utility functions satisfying concavity and monotone marginal utility. Based on this, we propose an active compression mechanism that adaptively determines compression levels based on document-query relevance.
- We propose ZipRL, a novel framework that integrates adaptive compression with robust RL optimization. We introduce Hindsight Response Replay (HRR), which leverages a heuristic-based metric to densify rewards via advantage re-shaping. We define a context-oriented metric to score compression segments, leveraging reward densification to optimize the average score towards a desired goal. HRR effectively mitigates the sparse reward issue in multi-turn context compression without the overhead of external reward models.
- Extensive experiments across five benchmarks in Web Browsing and Multi-hop QA tasks on multiple models of varying versions and parameter scales validate the effectiveness of our approach. Specifically, ZipRL outperforms the strongest same-scale specialized agents (e.g., ASearcher and AgentFold) by 27.9% and 34.7% average EM for Qwen3-4B and Qwen3-8B models respectively, during extreme extrapolation stress tests up to 256 turns (12.8× training horizon).

2 Related Work

2.1 Context Compression for Agents

Recent autonomous agents tackle long-horizon tasks via external retrieval (Jin et al., 2025; Nguyen et al., 2025; Chen et al., 2025), necessitating effective context management to fit limited context windows. Existing methods employ external storage (Zhong et al., 2024; Salama et al., 2025), internal memory updates (Zhou et al., 2026;

Ge et al., 2023; Huang et al., 2024), or multi-agent collaboration (Zhang et al., 2024). Static prompt compression (e.g., LLMLingua (Jiang et al., 2023), RECOMP (Xu et al., 2024)) reduces context pre-inference but lacks dynamic adaptability. While recent variable-rate compression methods (ACC-RAG (Guo et al., 2025), AttnComp (Luo et al., 2025), SARA (Jin et al., 2026)) improve single-turn token efficiency, they operate in *static* settings without multi-turn action adaptability or end-task RL feedback (see Appendix S for taxonomy). Crucially, most methods (Yu et al., 2026; Xu et al., 2026) overlook the non-uniform relevance across retrieved documents. To address this, our multi-granularity mechanism dynamically adapts compression levels based on information density, thereby retaining more task-critical information.

2.2 Sparse Rewards in RL

Sparse rewards in long-horizon interactions severely hinder RL efficiency (Riedmiller et al., 2018). Traditional RL mitigates this via Hierarchical RL (Pateria et al., 2021; Barto and Mahadevan, 2003), Curriculum Learning (CL) (Wang et al., 2021), or Hindsight Experience Replay (HER) (Andrychowicz et al., 2017). In the LLM context, approaches like CLPO (Zhang et al., 2025a) and Process Reward Models (PRMs) (Ma et al., 2023; Zhang et al., 2025b) attempt to densify training signals, but they often rely on expensive external models or task-specific designs. Recent RLVR advances instead tackle credit assignment directly: reshaping token-level policy gradients or distilling token-level credit (Lin et al., 2026b; Wei et al., 2026), correcting biased group-relative advantages (Yang et al., 2026), and improving rollout exploration or label efficiency (Lu et al., 2026; Wang et al., 2026; Lin et al., 2026a). In contrast, ZipRL draws inspiration from HER to integrate hindsight-based reward densification directly into multi-turn agent training, alleviating the sparse reward bottleneck without external PRM overhead.

3 Method

Background on Hindsight Experience Replay (HER) (Andrychowicz et al., 2017) and its relationship to our proposed HRR is provided in Appendix C. As illustrated in Figure 2, we present the proposed adaptive long-text search agent along with the ZipRL algorithm. The Cold Start phase is detailed in Appendix A, and the multi-granularity mechanism is described in Section 3.2. Section 3.3 discusses the evaluation of compression quality, while Section 3.4 introduces the hindsight response replay mechanism. Finally, the reinforcement learning training objective is formalized in Section 3.5.

3.1 Cold Start

Before RL training, we initialize the policy via a Cold Start SFT phase on 1,155 high-quality trajectories synthesized by GPT-4o (in line with recent data-synthesis pipelines for multi-hop search (Chen et al., 2025)), equipping the model with basic compression capabilities before RL optimization. Full details are in Appendix A.

3.2 Multi-Granularity Mechanism

In real-world applications, documents and queries often exhibit varying degrees of relevance, which necessitates different compression granularities. To enable such adaptive compression behavior, we introduce a **Multi-Granularity Mechanism**. This mechanism leverages carefully designed prompts (see Appendix U) to guide the model in identifying the amount of query-relevant information within the retrieved text. It then selects an appropriate compression granularity, denoted by g, and allocates different weights according to textual relevance, ensuring that more valuable information is preserved. Specifically, we first analyze the text and select k compression granularities (see Appendix J.2 for details). For each granularity, we define a compression range $[\mathcal{L}_l^g, \mathcal{L}_h^g]$ based on the text length at that granularity. Documents with stronger relevance typically contain more important information; therefore, we apply a higher compression level and allocate a larger compression range to retain as much important information as possible. Conversely, we limit retention for weakly relevant documents to avoid noise. This adaptive strategy maximizes salient information retention while effectively filtering out irrelevancies.

3.3 Compression Quality Evaluation

To evaluate the quality of compressed context without relying on expensive human annotation or unstable external reward models, we design a multi-dimensional heuristic scoring function. A high-quality compressed summary should satisfy two complementary objectives: *structural conformity* (adhering to length and format constraints) and *semantic preservation* (retaining critical information). We construct the final score Q_{com} based on four

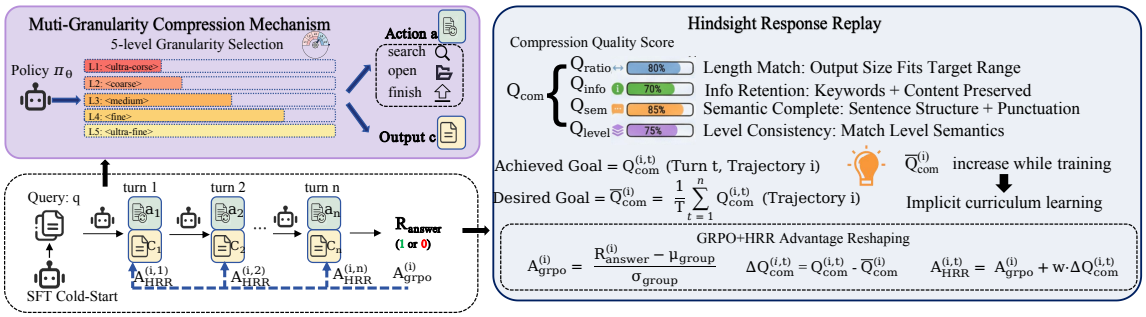

Figure 2: Overview of the ZipRL framework, including the Multi-Granularity Mechanism, Compression Quality Evaluation, and Hindsight Response Replay.

distinct dimensions. In this section, we outline the design rationale for each dimension; for detailed formulations, please refer to Appendix G.

Structure Objective: This objective enforces the length and style constraints specified by the target compression level g.

Compression Ratio Score Q_{ratio}. It measures whether the compressed length l_y falls within the target range $[\mathcal{L}_l^g, \mathcal{L}_h^g]$ defined for level g. Unlike a simple binary check, we implement a soft penalty mechanism (see Figure 6 in Appendix G). Short outputs incur a linear penalty for information loss, while long outputs receive a non-linear penalty for redundancy.

Level Strategy Consistency Score Q_{level}. Different compression levels require different granularities, not just different lengths. For instance, Level 1 (Ultra-coarse) should consist of few concise sentences, whereas Level 5 (Ultra-fine) allows for detailed narration. We define expected bounds for sentence count S_g and word count W_g for each level. The score R_{level} penalizes outputs that meet the character limit but violate the structural density.

Semantic Objective: Structural correctness is meaningless if the core meaning is lost or the text becomes incoherent. These dimensions evaluate the content quality.

Information Retention Score Q_{info}. To ensure the compressed text y retains key information from the original text x relevant to the query q, we employ a keyword-based overlap metric. We extract keyword sets K_q from the query and K_x from the original text via tokenization, stopword removal, and filtering for tokens with length ≥ 4 characters. Q_{info} is computed as a weighted combination of query-specific keyword coverage (S_{key}) and general content retention (S_{gen}), with $\lambda_1 = 0.7$ and $\lambda_2 = 0.3$. The denominator of S_{key} uses $|K_q \cap x|$ rather than $|K_q|$, restricting evaluation to keywords *actually present* in the source document to avoid penalizing irretrievable keywords. When $|K_q \cap x| = 0$, the system falls back to S_{gen} alone. Full formulations are provided in Appendix G.

Semantic Completeness Score Q_{sem}. Aggressive compression often leads to broken sentences or linguistic fragmentation. We evaluate semantic completeness by checking for valid sentence terminators (punctuation), reasonable sentence structures, and vocabulary coherence. This acts as a linguistic validity filter, ensuring the compressed text remains natural and readable.

Total Score: The final compression quality score is a weighted sum of these four complementary dimensions:

$$Q_{\text{com}} = \alpha_1 Q_{\text{ratio}} + \alpha_2 Q_{\text{level}} + \alpha_3 Q_{\text{info}} + \alpha_4 Q_{\text{sem}} \quad (1)$$

In our implementation, we prioritize information retention and length compliance, setting weights as $\alpha = \{0.3, 0.1, 0.4, 0.2\}$ respectively. This composite score guides the model to balance the trade-off between brevity (structure) and utility (semantic).

3.4 Hindsight Response Replay

In multi-turn task settings, it is often difficult to distinguish the contribution of each individual turn solely based on the final reward signal. Inspired by the success of Hindsight Experience Replay (HER) (Andrychowicz et al., 2017) in converting sparse binary feedback into dense learning signals, we propose **Hindsight Response Replay (HRR)** as an *advantage reshaping* technique tailored for multi-turn LLM agent training. While HER achieves reward densification by relabeling goals and replaying transitions under substitute objectives, HRR achieves an analogous effect through a different mechanism: it uses a turn-level compression quality score to redistribute the trajectory-level

advantage computed by GRPO, assigning denser credit to individual turns without requiring a replay buffer, goal-conditioned policy, or explicit goal substitution (see Table 6 for a detailed comparison). The key insight shared with HER is that non-optimal trajectories still contain valuable per-step information: even when the final task reward is zero, individual turns may exhibit high-quality compression behavior that should be reinforced. In GRPO, the advantage for each trajectory i is computed as:

$$A_{\text{GRPO}}^{(i)} = \frac{R_{\text{final}}^{(i)} - \mu_{\text{group}}}{\sigma_{\text{group}}}, \tag{2}$$

where $R_{\text{final}}^{(i)}$ denotes the final reward for trajectory i, and μ_{group} and σ_{group} represent the mean and standard deviation of the group rewards, respectively. Each turn j in trajectory i uses the same advantage $A_{\text{GRPO}}^{(i)}$, which makes it difficult for the agent to differentiate the contribution of each individual turn, potentially increasing the difficulty of training.

To address this limitation, HRR utilizes our compression quality score to reinterpret the per-turn contribution within each trajectory, alleviating the issue of sparse rewards by providing turn-level credit assignment. Concretely, we treat the achieved compression quality as a *dynamic reference point* (analogous to how HER treats the achieved state as a substitute goal), allowing us to "replay" the advantage calculation by measuring each turn's distance to this reference. Specifically, for each trajectory i consisting of T turns, we compute the average compression quality score as the trajectory-level reference:

$$\bar{Q}_{\text{com}}^{(i)} = \frac{1}{T} \sum_{j=1}^{T} Q_{\text{com}}^{(i,j)}. \tag{3}$$

For each turn in trajectory i, we treat the corresponding compression quality score as the state and compare it with the expected target $\bar{Q}_{\text{com}}^{(i)}$ to evaluate its quality. This allows us to reallocate the trajectory-level advantage from GRPO at the turn level. Formally, for the j-th turn in trajectory i, the final advantage is defined as:

$$A^{(i,j)} = A_{\text{GRPO}}^{(i)} + w \cdot \left(Q_{\text{com}}^{(i,j)} - \bar{Q}_{\text{com}}^{(i)} \right), \tag{4}$$

where w is an advantage reshaping coefficient that controls the strength of hindsight relabeling. When the j-th turn in trajectory i has a compression quality score greater than the trajectory average, the advantage value is increased. Conversely, if the score is smaller than the average, the advantage value is decreased. The magnitude of this adjustment is determined by the absolute difference in compression quality scores, thus providing a denser training signal for different turns in a trajectory and more accurately evaluating each turn's contribution. As training progresses, the model's compression quality improves (see Section 5), leading to a higher average compression quality score, thereby adaptively adjusting the expected training target. This approach alleviates the sparse reward problem and enables better credit assignment.

3.5 Training Objectives

Based on hindsight-relabeled dense advantages, ZipRL uses the following objective, which facilitates stable and efficient training through the $\mathcal{J}_{\text{ZipRL}}(\theta)$ objective:

$$\begin{aligned}
\mathcal{J}_{\text{ZipRL}}(\theta) = \mathbb{E}_{q,\{o^i\}} \Bigg[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o^i|} \sum_{j=1}^{T} \sum_{t=1}^{|o^{(i,j)}|} \\
\left(m_t^{(i,j)} - \beta\, \mathbb{D}_{\text{KL}}[\pi_\theta \| \pi_{\text{ref}}] \right) \Bigg], \\
m_t^{(i,j)} = \min\Big(\rho_t^{(i,j)} A^{(i,j)}, \\
\text{clip}\left(\rho_t^{(i,j)},\, 1-\epsilon,\, 1+\epsilon \right) A^{(i,j)} \Big), \\
\rho_t^{(i,j)} = \frac{\pi_\theta\left(o_t^{(i,j)} \mid q,\, o_{<t}^{(i,j)} \right)}{\pi_{\theta_{\text{old}}}\left(o_t^{(i,j)} \mid q,\, o_{<t}^{(i,j)} \right)}.
\end{aligned} \tag{5}$$

4 Theoretical Analysis

We provide a utility-theoretic justification for multi-granularity compression, showing that relevance-aware resource allocation can improve expected downstream utility over uniform compression under the same average resource budget.

Theorem 4.1 (Utility Advantage of Relevance-Aware Allocation)**.** *Let $R \sim P$ be a relevance random variable supported on $\mathcal{R} \subseteq \mathbb{R}$. Let $\bar{\alpha} > 0$ be the average resource budget, and let $\mathcal{A} \subseteq \mathbb{R}_{++}$ be an open interval with $\bar{\alpha} \in \mathcal{A}$. Consider a utility function $\Phi : \mathcal{R} \times \mathcal{A} \to \mathbb{R}$ such that $\Phi(\cdot, \alpha)$ is measurable for every $\alpha \in \mathcal{A}$, and $\Phi(r, \cdot)$ is differentiable and strictly concave on $\mathcal{A}$ for every $r \in \mathcal{R}$. Let the uniform allocation be $\alpha_{\text{uni}}(r) \equiv \bar{\alpha}$, and let the relevance-aware adaptive allocation be $\alpha_{\text{ada}}(r) = f(r)$, where $f : \mathcal{R} \to \mathcal{A}$*

is measurable, non-decreasing, and satisfies

$$\mathbb{E}[f(R)] = \bar{\alpha}, \qquad \mathbb{P}(f(R) \neq \bar{\alpha}) > 0.$$

Define the allocated marginal utility

$$G_f(r) = \partial_\alpha \Phi(r, \alpha)\big|_{\alpha = f(r)}.$$

Assume that G_f is measurable and non-decreasing, and that $\Phi(R, f(R))$, $\Phi(R, \bar{\alpha})$, $G_f(R)$, and $G_f(R)f(R)$ are integrable. Then

$$\mathbb{E}[\Phi(R, f(R))] > \mathbb{E}[\Phi(R, \bar{\alpha})]\,.$$

Remark. The monotonicity of G_f means that the relevance-induced increase in marginal utility dominates the diminishing returns caused by larger allocations. Theorem 4.1 formalizes the benefit of multi-granularity compression: under a fixed average resource budget, allocating more resources to more relevant documents yields a strict utility gain when the allocation is positively aligned with the marginal utility of additional compression resources. The proof is deferred to Appendix E.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate the agent on Web Browsing and Multi-hop QA. For Web Browsing, we use BrowseComp (Wei et al., 2025), a dynamic web retrieval benchmark for long-context management. For Multi-hop QA, we use MusiQue (Trivedi et al., 2022), SQuAD (Rajpurkar et al., 2016), Frames (Krishna et al., 2025), and Bamboogle (Press et al., 2023), which require aggregating evidence across retrieval steps. MusiQue, Frames, and Bamboogle are multi-hop datasets, whereas SQuAD is primarily single-hop. Following Jin et al. (2025), we evaluate SQuAD in an *open-domain retrieval* setting without the source paragraph. ZipRL yields larger gains on harder multi-hop datasets, suggesting that adaptive compression is more beneficial for complex reasoning.

Baselines. We compare our method against the following baselines: **(1) ReAct** (Yao et al., 2022), which uses large-scale closed-source models with context concatenation, setting the performance upper bound; **(2) Summary-Only**, a training-free baseline using heuristic rules for context summarization; **(3) Search Agent Methods** (Search-R1 (Jin et al., 2025), WebSailor (Li et al., 2025a), NestBrowse (Li et al., 2026), AgentFold (Ye et al., 2025), ASearcher (Gao et al., 2025)), which utilize dynamic context condensing or specialized agents for long-context reasoning.

Retrieval System. To ensure fair comparison, all models within each benchmark share an identical retrieval backend. For Multi-hop QA datasets (MusiQue, SQuAD, Frames, Bamboogle), we deploy E5-base-v2 (mean pooling, max query length 256) over the Wikipedia 2018 corpus with FAISS Flat Index (FP16). For BC-Plus, we use Qwen3-Embedding-8B (last-token pooling, max query length 8192) over the dedicated BC+ corpus with pre-computed local vector indices. Top-k, corpus, and index configuration are strictly identical for all compared methods on each dataset.

Experimental Details. We use Qwen2.5-{3B, 7B, 14B} and Qwen3-{4B, 8B, 14B} (Yang et al., 2025) as base models. All models are trained on an 8×A100 (80GB) GPU cluster with at most 20 interaction turns, reward reshaping coefficient 0.2, learning rate $1\mathrm{e}{-6}$, and batch size 64. During evaluation, we use temperature 0.8 and top-p 1.0, and report EM and F1 (Li et al., 2025b). We adopt official baseline checkpoints, where ASearcher does not provide a 4B model and WebSailor provides 32B instead of 14B. Tool configurations, compression details, and additional experimental settings are provided in Appendix I and Appendix J.

5.2 Main Results

Performance Results. Table 1 reports results on Multi-hop QA and Web Browsing benchmarks. Among 8B-scale agents, ZipRL-8B achieves 30.3% average EM and 41.2% F1, outperforming the much larger Qwen3-235B-ReAct baseline with 29× fewer parameters and matching Gemini-3-Pro. Compared with similarly sized specialized agents, i.e., ASearcher-7B and AgentFold-8B, ZipRL-8B improves average EM by 7.8 points. This advantage is consistent across scales: ZipRL-4B and ZipRL-14B lead their parameter-matched groups, while ZipRL-32B further reaches 35.2% average EM, improving over ZipRL-14B by 4.1 points (Appendix K). ZipRL also substantially outperforms Summary-Only baselines across all datasets, indicating that its learned active compression policy preserves complex information dependencies beyond static heuristics. Moreover, on the same Qwen2.5 backbone series used by prior baselines, ZipRL-7B still yields a 5.5-point average EM gain over ASearcher-7B, further confirming that the im-

Table 1: Performance Comparison of Different Methods on Multi-hop QA and Web Browsing Tasks Across Five Datasets. BC-Plus denotes BrowseComp-plus. The **bold** and underline indicate the best and second-best results, respectively.

Model	MusiQue		SQuAD		Frames		Bamboogle		BC-plus		Average	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
ReAct-based Methods												
Qwen3-235B-ReAct	18.2	30.4	17.8	32.3	14.5	27.4	40.1	52.5	14.0	19.5	20.9	32.4
DeepSeek-V3.2-ReAct	9.6	17.3	10.2	22.9	10.1	18.4	32.0	40.6	8.8	13.3	14.1	22.5
Gemini-3-Pro-ReAct	**32.2**	**44.7**	**21.4**	**38.2**	**24.8**	**38.7**	**54.4**	**67.0**	22.8	27.8	**31.1**	**43.3**
GPT-4o-ReAct	11.8	28.9	8.6	24.5	15.8	32.6	39.2	55.3	**25.2**	**31.6**	20.1	34.6
LongCat-Flash-ReAct	24.4	37.5	18.6	35.3	23.5	38.0	48.8	62.0	24.4	28.5	27.9	40.3
Summary-based Methods												
Qwen3-235B-Summary	17.8	**33.2**	**13.0**	28.1	15.2	29.9	40.0	54.7	14.0	19.2	20.0	33.0
DeepSeek-V3.2-Summary	10.2	25.2	12.2	**30.6**	10.1	20.6	38.4	54.5	7.6	11.5	15.7	28.5
Gemini-3-Pro-Summary	14.6	27.4	10.4	25.5	**24.8**	**38.7**	35.2	49.4	23.2	27.4	21.6	33.7
GPT-4o-Summary	3.6	18.7	5.6	19.7	10.3	24.1	20.0	37.2	17.6	27.7	11.4	25.5
LongCat-Flash-Summary	**20.0**	**33.2**	12.8	29.5	18.5	32.3	**52.8**	**66.1**	**29.6**	**35.7**	**26.7**	**39.4**
Open-Source Large Model												
WebSailor-32B	22.0	34.6	17.4	37.3	14.0	29.5	30.4	45.1	17.6	25.6	20.3	34.4
Specialized Agent (Qwen2.5-3B)												
Search-R1-3B	18.2	22.4	8.8	17.6	3.2	8.0	12.0	21.0	0.8	2.8	8.6	14.4
WebSailor-3B	5.0	12.8	5.6	17.8	3.5	9.6	12.0	29.6	2.8	3.8	5.8	14.7
ZipRL-3B (ours)	**27.8**	**37.5**	**25.2**	**40.3**	**13.4**	**23.5**	**44.0**	**56.9**	**22.5**	**24.8**	**26.6**	**36.6**
Specialized Agent (Qwen3-4B)												
NestBrowse-4B	12.0	21.5	14.6	28.5	4.3	13.1	15.2	24.9	11.2	15.7	11.5	20.7
AgentFold-4B	16.0	26.5	19.4	33.8	14.3	24.7	44.0	57.9	15.6	19.8	21.9	32.5
ZipRL-4B (ours)	**31.4**	**40.4**	**23.0**	**39.0**	**14.6**	**26.2**	**47.2**	**60.9**	**24.0**	**26.9**	**28.0**	**38.7**
Specialized Agent (Qwen2.5-7B)												
Search-R1-7B	14.6	19.0	5.4	13.1	5.1	9.5	20.8	28.7	0.4	3.4	9.3	14.7
WebSailor-7B	7.6	14.9	13.0	26.4	7.2	14.1	36.8	49.1	6.4	8.3	14.2	22.6
ASearcher-7B	**32.6**	**42.9**	23.8	39.5	11.0	21.6	39.2	51.5	6.0	8.2	22.5	32.7
ZipRL-7B (ours)	31.2	41.4	**25.6**	**43.0**	**15.8**	**26.0**	**48.0**	**61.6**	**19.5**	**21.5**	**28.0**	**38.7**
Specialized Agent (Qwen3-8B)												
NestBrowse-8B	21.4	29.5	19.2	36.9	9.1	17.9	33.6	47.4	8.8	14.5	18.4	29.2
AgentFold-8B	21.6	30.1	20.8	34.9	12.3	21.2	42.4	54.1	15.6	19.3	22.5	31.9
ZipRL-8B (ours)	**36.8**	**46.7**	**25.8**	**42.4**	**18.1**	**29.3**	**49.8**	**64.7**	**20.8**	**23.0**	**30.3**	**41.2**
Specialized Agent (Qwen2.5-14B)												
Search-R1-14B	11.8	16.6	11.4	20.5	7.2	13.5	26.4	36.6	1.6	4.8	11.7	18.4
ASearcher-14B	**36.6**	**47.6**	**24.4**	**41.6**	16.1	27.7	48.8	60.0	13.2	17.9	27.8	39.0
ZipRL-14B (ours)	30.8	40.7	23.2	41.0	**19.2**	**31.3**	**51.2**	**64.4**	**23.6**	**26.1**	**29.6**	**40.7**
Specialized Agent (Qwen3-14B)												
NestBrowse-14B	24.4	35.2	15.2	30.3	13.8	25.7	42.4	56.9	7.6	11.7	20.7	32.0
AgentFold-14B	22.2	32.9	21.8	36.5	13.4	22.4	45.6	55.2	14.8	16.9	23.6	32.8
ZipRL-14B (ours)	**33.6**	**43.4**	**24.0**	**42.2**	**20.8**	**32.4**	**52.8**	**67.6**	**24.4**	**27.0**	**31.1**	**42.5**

provements stem from algorithmic design rather than base-model differences (Table 2).

Same-Backbone Comparison (Qwen2.5). To eliminate confounds from base-model generation differences (e.g., ZipRL-8B uses Qwen3-8B whereas ASearcher-7B uses Qwen2.5-7B), we instantiate ZipRL on the identical Qwen2.5 series (3B/7B/14B) used by the baselines. As shown in Table 2, ZipRL consistently outperforms same-series baselines at all three scales, confirming that the gains stem from our algorithmic contributions rather than generational model differences (further discussion in Appendix L).

Comparison with Memory-Based Compression Methods. Beyond the specialized search agents reported in Table 1, we further compare ZipRL with two dedicated long-context memory compression methods, MemAgent (Yu et al., 2026) and MEM1 (Zhou et al., 2026), which rely on fixed-length memory overwrite or consolidation. To en-

Table 2: Strictly matched-scale comparison: ZipRL (Qwen2.5) vs. same-series baselines. ΔEM: absolute improvement over the best same-scale specialized baseline.

Scale	Model	Base	Avg EM	ΔEM
3B	Search-R1-3B	Qwen2.5-3B	8.6	—
	ZipRL-3B	Qwen2.5-3B	26.6	+18.0
7B	ASearcher-7B	Qwen2.5-7B	22.5	—
	ZipRL-7B	Qwen2.5-7B	28.0	+5.5
14B	ASearcher-14B	Qwen2.5-14B	27.8	—
	ZipRL-14B	Qwen2.5-14B	29.6	+1.8

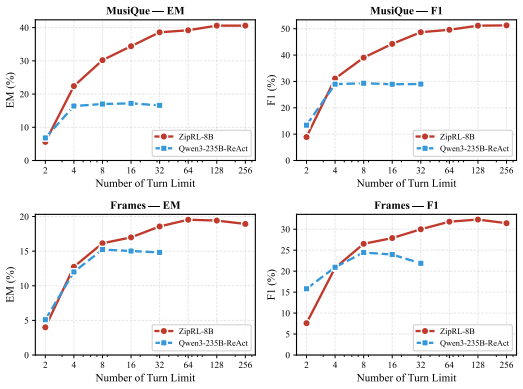

Figure 3: Performance variations of the MusiQue and Frames datasets under different turn constraints.

sure a fair comparison in our multi-turn retrieval setting, we adapt these baselines by injecting retrieved documents into their memory mechanisms as chunked observations. As summarized in Table 3, ZipRL consistently and substantially outperforms both baselines across all scales. Specifically, at the 8B scale, ZipRL achieves an average EM of 32.6%, outperforming MemAgent (20.1%) and MEM1 (21.5%) by **+12.5** and **+11.1** points, respectively (a relative improvement of **62.2%** and **51.6%**). At the 14B scale, ZipRL attains 32.8% EM, surpassing MemAgent (19.8%) and MEM1 (25.7%) by **+13.0** and **+7.1** points. These gains confirm that ZipRL's adaptive, per-document, RL-optimized compression is more effective than static or heuristic memory-based compression for long-horizon agentic search.

Long-Horizon Scalability. To assess robustness against context overload, we conduct extrapolation stress tests with extended interaction limits, as shown in Figure 3. We increase the maximum number of turns from 2 to 256 on MusiQue and Frames, where 256 turns serves as a stress-test boundary rather than a typical QA length. ZipRL-8B consistently outperforms the larger Qwen3-235B-ReAct baseline. While Qwen3-235B-ReAct saturates after 16 turns, ZipRL continues improving up to 256 turns, despite being trained with at most 20 turns. This 12.8× horizon extrapolation suggests that ZipRL learns relevance-aware retention rather than overfitting to short interaction patterns, enabling robust long-horizon deployment under accumulated context noise.

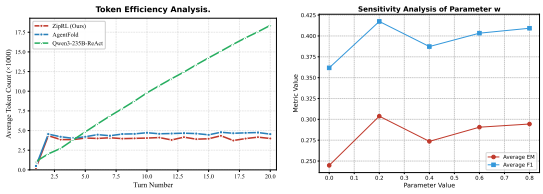

Figure 4: (Left) Average token count variation over training iterations. (Right) Sensitivity analysis of parameter w across five datasets with ZipRL.

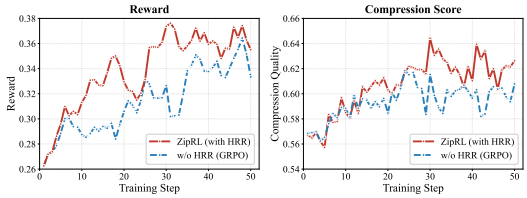

Figure 5: Training dynamics of ZipRL. (Left) Average reward and (Right) compression quality score.

Token Efficiency Analysis. As shown in Figure 4 (left), ZipRL achieves strong token efficiency. Compared with ReAct-based large models, ZipRL-8B substantially reduces token usage while preserving performance. Its adaptive multi-granularity compression further outperforms AgentFold by better balancing information retention and computational cost. A detailed analysis of premature termination is provided in Appendix P.

Fine-Grained Behavior Statistics. Fine-grained behavior statistics are provided in Appendix O (Table 11). ZipRL achieves a Finish Rate of 0.85 and adaptively increases retrieval intensity with question difficulty, while sustaining deep exploration on BC-Plus with 13.8 turns, avoiding under-retrieval and context-overload issues observed in baselines.

Training Dynamics and Multi-Run Stability. Figure 5 illustrates ZipRL's training dynamics: reward converges faster and higher than GRPO, and the compression quality score improves steadily from 0.57 to 0.64 with reduced variance. To validate this variance-reduction effect, we conduct 3

Table 3: Comparison with adapted memory-based methods on multi-turn retrieval QA (4 datasets, BC-Plus excluded).

Scale	Model	Bamboogle	MusiQue	SQuAD	Frames	Avg (EM/F1)
8B	MemAgent-8B	35.2/48.9	17.8/28.4	14.2/26.1	13.1/25.2	20.1/32.2
	MEM1-8B	38.4/52.4	18.8/28.0	15.2/28.3	13.6/25.8	21.5/33.6
	ZipRL-8B	**49.8/64.7**	**36.8/46.7**	**25.8/42.4**	**18.1/29.3**	**32.6/45.8**
14B	MemAgent-14B	36.8/50.4	16.4/27.2	13.4/25.9	12.6/24.7	19.8/32.1
	MEM1-14B	44.0/60.7	25.8/34.0	18.2/33.8	14.9/27.3	25.7/38.9
	ZipRL-14B	**52.8/67.6**	**33.6/43.4**	**24.0/42.2**	**20.8/32.4**	**32.8/46.4**

Table 4: Ablation Study of ZipRL on the Qwen3-8B Model Across Five Datasets. BC-Plus denotes BrowseComp-plus. The **bold** and underline indicate the best and second-best results, respectively.

Model Variant	MusiQue		SQuAD		Frames		Bamboogle		BC-Plus		Average	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
ZipRL (ours)	36.8	46.7	**25.8**	**42.4**	**18.1**	**29.3**	**49.8**	64.7	20.8	23.0	**30.3**	**41.2**
w/o RL	29.0	39.7	21.0	36.4	16.0	26.4	45.6	59.8	20.4	26.3	26.4	37.7
w/o Level 2&4	35.4	45.8	24.4	40.3	17.7	28.2	48.4	**64.9**	18.4	22.1	28.9	40.3
w/o Q_{ratio}	37.0	47.1	23.6	38.7	17.6	27.8	47.0	61.4	**21.6**	**26.5**	29.4	40.3
w/o Q_{info}	32.4	41.4	20.8	37.2	15.3	27.1	47.2	58.7	18.8	23.2	26.9	37.5
w/o Q_{sem}	32.4	42.5	23.0	38.5	16.9	28.3	46.4	62.0	20.0	24.1	27.7	39.1
w/o Q_{level}	**38.0**	**48.3**	25.0	42.0	16.9	28.8	46.4	59.9	20.8	23.9	29.4	40.6

Table 5: Cross-run reward statistics (mean $\pm$ std) over 3 independent runs.

Training Steps	GRPO ($w = 0$)	ZipRL (HRR)
Steps 21–25	0.5926 ± 0.0351	0.6171 ± 0.0111
Steps 26–30	0.6028 ± 0.0423	0.6263 ± 0.0113
Steps 31–35	0.6197 ± 0.0512	0.6379 ± 0.0124
Steps 36–40	0.6238 ± 0.0638	0.6439 ± 0.0099

independent training runs for both standard GRPO (w=0) and ZipRL with HRR. As shown in Table 5, HRR consistently achieves much lower cross-run standard deviation ($\sigma = 0.0099$–0.0124) than GRPO ($\sigma = 0.0351$–0.0638) across all training phases, confirming that turn-level credit assignment stabilizes multi-turn RL training.

6 Ablation Studies

To quantify the contribution of each component, we conduct ablation studies with Qwen3-8B as the base model, as shown in Table 4. First, removing the RL phase and using only the cold-start model leads to a substantial performance drop, highlighting the importance of reinforcement learning through environment interaction. Second, reducing the compression granularity by removing Levels 2 and 4 consistently lowers average EM and F1, indicating the benefit of maintaining five compression levels. Finally, we ablate each dimension of the Compression Quality Score, i.e., Q_{ratio}, Q_{level}, Q_{info}, and Q_{sem}. Removing any dimension degrades performance, with Q_{info} and Q_{sem} causing the largest drops. Results demonstrate that ZipRL benefits from fine-grained compression control and quality-aware reward design, particularly through information retention and semantic completeness.

7 Conclusion

Despite recent progress in LLM-based long-text search, existing models still lack proactive awareness of document relevance and suffer from sparse rewards in long-horizon tasks. We introduce ZipRL to address these challenges through a multi-granularity mechanism that adaptively assigns compression levels according to query relevance. We further design a rule-based evaluation scheme and incorporate compression scores into Hindsight Response Replay to mitigate reward sparsity during multi-turn training. Theoretically, we analyze the benefits of both multi-granularity compression and HRR. Extensive experiments across multiple models of varying versions and parameter scales on five Web Browsing and Multi-hop QA datasets validate the effectiveness of our approach and show that ZipRL consistently outperforms strong baselines.

Limitations

We note three limitations of ZipRL. (1) Language: Q_{info} relies on English stopwords, potentially degrading in multilingual or specialized domains (e.g., legal, code). (2) Robustness: Q_{com} ignores document credibility; EM drops heavily (85–99%) under fully adversarial retrieval (Appendix R). (3) Generalization: Relying on a single QA corpus for cold-start may limit adaptation to structured or code-heavy tasks.

Ethical Statements

Our experiments rely strictly on publicly available datasets, involving no sensitive user data. Large language models were used solely for language editing during manuscript preparation. All research outcomes and intellectual content remain the exclusive work of the authors.

References

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. 2017. Hindsight experience replay. *Advances in neural information processing systems*, 30.

Andrew G Barto and Sridhar Mahadevan. 2003. Recent advances in hierarchical reinforcement learning. *Discrete event dynamic systems*, 13(4):341–379.

Hao Chen, Zhexin Hu, Jiajun Chai, Haocheng Yang, Hang He, Xiaohan Wang, Wei Lin, Luhang Wang, Guojun Yin, and 1 others. 2025. Toolforge: A data synthesis pipeline for multi-hop search without real-world apis. *arXiv preprint arXiv:2512.16149*.

Jiaxuan Gao, Wei Fu, Minyang Xie, Shusheng Xu, Chuyi He, Zhiyu Mei, Banghua Zhu, and Yi Wu. 2025. Beyond ten turns: Unlocking long-horizon agentic search with large-scale asynchronous rl. *arXiv preprint arXiv:2508.07976*.

Tao Ge, Jing Hu, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. 2023. In-context autoencoder for context compression in a large language model. *arXiv preprint arXiv:2307.06945*.

Shuyu Guo, Shuo Zhang, and Zhaochun Ren. 2025. Enhancing rag efficiency with adaptive context compression. *arXiv preprint arXiv:2507.22931*.

Guanzhong He, Zhen Yang, Jinxin Liu, Bin Xu, Lei Hou, and Juanzi Li. 2026. Webseer: Training deeper search agents through reinforcement learning with self-reflection. In *International Conference on Learning Representations*, volume 2026, pages 106335–106372.

Chensen Huang, Guibo Zhu, Xuepeng Wang, Yifei Luo, Guojing Ge, Haoran Chen, Dong Yi, and Jinqiao Wang. 2024. Recurrent context compression: Efficiently expanding the context window of llm. *arXiv preprint arXiv:2406.06110*.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. Llmlingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 13358–13376.

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

Yiqiao Jin, Kartik Sharma, Vineeth Rakesh, Yingtong Dou, Menghai Pan, Mahashweta Das, and Srijan Kumar. 2026. Sara: Selective and adaptive retrieval-augmented generation with context compression. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14508–14528.

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese, and 1 others. 2025. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems. *arXiv preprint arXiv:2504.09037*.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2025. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4745–4759.

Baixuan Li, Jialong Wu, Wenbiao Yin, Kuan Li, Zhongwang Zhang, Huifeng Yin, Zhengwei Tao, Liwen Zhang, Pengjun Xie, Jingren Zhou, and 1 others. 2026. Nested browser-use learning for agentic information seeking. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22911–22923.

Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, and 1 others. 2025a. Websailor: Navigating super-human reasoning for web agent. *arXiv preprint arXiv:2507.02592*.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025b. Search-o1: Agentic search-enhanced large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5420–5438.

Zihan Lin, Xiaohan Wang, Jie Cao, Jiajun Chai, Li Wang, Xiaodong Lu, Wei Lin, Ran He, and Guojun Yin. 2026a. Resrl: Boosting llm reasoning via negative sample projection residual reinforcement learning. *arXiv preprint arXiv:2605.00380*.

Zihan Lin, Xiaohan Wang, Jie Cao, Jiajun Chai, Guojun Yin, Wei Lin, and Ran He. 2026b. Rest: Reshaping token-level policy gradients for tool-use large language models. In *International Conference on Learning Representations*, volume 2026, pages 83901–83945.

Xiaodong Lu, Xiaohan Wang, Jiajun Chai, Guojun Yin, Wei Lin, Zhijun Chen, Yu Luo, Fuzhen Zhuang, Yikun Ban, and Deqing Wang. 2026. Contextual rollout bandits for reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2602.08499*.

Lvzhou Luo, Yixuan Cao, and Ping Luo. 2025. Attncomp: attention-guided adaptive context compression for retrieval-augmented generation. *arXiv preprint arXiv:2509.17486*.

Qianli Ma, Haotian Zhou, Tingkai Liu, Jianbo Yuan, Pengfei Liu, Yang You, and Hongxia Yang. 2023. Let's reward step by step: Step-level reward model as the navigators for reasoning. *arXiv preprint arXiv:2310.10080*.

Xuan-Phi Nguyen, Shrey Pandit, Revanth Gangi Reddy, Austin Xu, Silvio Savarese, Caiming Xiong, and Shafiq Joty. 2025. Sfr-deepresearch: Towards effective reinforcement learning for autonomously reasoning single agents. *arXiv preprint arXiv:2509.06283*.

Shubham Pateria, Budhitama Subagdja, Ah-hwee Tan, and Chai Quek. 2021. Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 54(5):1–35.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 2383–2392.

Martin Riedmiller, Roland Hafner, Thomas Lampe, Michael Neunert, Jonas Degrave, Tom Wiele, Vlad Mnih, Nicolas Heess, and Jost Tobias Springenberg. 2018. Learning by playing solving sparse reward tasks from scratch. In *International conference on machine learning*, pages 4344–4353. PMLR.

Rana Salama, Jason Cai, Michelle Yuan, Anna Currey, Monica Sunkara, Yi Zhang, and Yassine Benajiba. 2025. Meminsight: Autonomous memory augmentation for llm agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33124–33140.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2023. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2998–3009.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Li Wang, Xiaodong Lu, Xiaohan Wang, Yikun Ban, Jiajun Chai, Wei Lin, Tianhao Peng, and Guojun Yin. 2026. When self-belief misleads: Active label acquisition for reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2605.25864*.

Xin Wang, Yudong Chen, and Wenwu Zhu. 2021. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4555–4576.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*.

Zhenlin Wei, Pu Jian, Yingzhuo Deng, Xiaohan Wang, Jiajun Chai, Zhexin Hu, Wei Lin, Shanbin Zhang, and Guojun Yin. 2026. Amr-sd: Asymmetric meta-reflective self-distillation for token-level credit assignment. *arXiv preprint arXiv:2605.18529*.

Jinyang Wu, Mingkuan Feng, Shuai Zhang, Feihu Che, Zengqi Wen, Chonghua Liao, and Jianhua Tao. 2024. Beyond examples: High-level automated reasoning paradigm in in-context learning via mcts. *arXiv preprint arXiv:2411.18478*.

Jinyang Wu, Guocheng Zhai, Ruihan Jin, Jiahao Yuan, Yuhao Shen, Shuai Zhang, Zhengqi Wen, and Jianhua Tao. 2026. Atlas: Orchestrating heterogeneous models and tools for multi-domain complex reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 17503–17535.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024. Recomp: Improving retrieval-augmented lms with context compression and selective augmentation. In *International Conference on Learning Representations*, volume 2024, pages 43478–43502.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2026. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Fengkai Yang, Zherui Chen, Xiaohan Wang, Xiaodong Lu, Jiajun Chai, Guojun Yin, Wei Lin, Shuai Ma, Fuzhen Zhuang, Deqing Wang, and 1 others. 2026. Your group-relative advantage is biased. *arXiv preprint arXiv:2601.08521*.

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

Rui Ye, Zhongwang Zhang, Kuan Li, Huifeng Yin, Zhengwei Tao, Yida Zhao, Liangcai Su, Liwen Zhang, Zile Qiao, Xinyu Wang, and 1 others. 2025. Agentfold: Long-horizon web agents with proactive context management. *arXiv preprint arXiv:2510.24699*.

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. 2026. Memagent: Reshaping long-context llm with multi-conv rl-based memory agent. In *International Conference on Learning Representations*, volume 2026, pages 39458–39486.

Shijie Zhang, Zheng Xiao, Shiyu Liu, Guohao Sun, Kevin Zhang, Xiang Guo, Rujun Guo, Shaoyu Liu, Wangxiao Zhao, and Guanjun Jiang. 2025a. Clpo: Curriculum learning meets policy optimization for llm reasoning. *arXiv preprint arXiv:2509.25004*.

Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Ö Arık. 2024. Chain of agents: Large language models collaborating on long-context tasks. *Advances in Neural Information Processing Systems*, 37:132208–132237.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025b. The lessons of developing process reward models in mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10495–10516.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 19724–19731.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Bryan Kian Hsiang Low, and Paul Liang. 2026. Mem1: Learning to synergize memory and reasoning for efficient long-horizon agents. In *International Conference on Learning Representations*, volume 2026, pages 58413–58438.

A Cold Start Details

Before RL training, we implement a Cold Start phase to initialize the policy. We constructed a demonstration dataset, $\mathcal{D}_{\text{demo}}$, by sampling 3,000 queries from ASearcher-LRM-35k (Gao et al., 2025) and synthesizing trajectories via GPT-4o. To ensure high quality, we applied two filters: (1) **Correctness**: retaining only correct trajectories, and (2) **Format**: discarding outputs with structural violations. This yielded 1,155 valid trajectories, which were then decomposed into transition-level samples. The model is optimized using the standard Supervised Fine-Tuning (SFT) objective:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(q,o)\sim\mathcal{D}_{\text{demo}}}\left[\sum_{t=1}^{|o|}\log \pi_\theta(o_t \mid q, o_{<t})\right]$$

where π_θ denotes the policy network, and $o = (o_1, o_2, \ldots, o_{|o|})$ is the output sequence. Each o_t represents the observation or generated context token at step t (including retrieved documents, reasoning traces, and tool outputs accumulated during the interaction). This phase equips the model with essential capabilities, adequately preparing it for subsequent RL optimization.

B Diagnosing the Performance Plateau at 256 Turns

To understand what limits performance as interaction turns increase beyond $\sim$128, we conduct a targeted diagnostic analysis on MusiQue along three axes. **(1) Dataset ceiling:** An oracle experiment (providing all gold-standard evidence documents directly to the model without retrieval) achieves 51.3% EM on MusiQue, substantially above ZipRL-8B's 36.8% at 256 turns, so the dataset ceiling is not the primary bottleneck. **(2) Retriever saturation:** We track the *novel-entity rate*: the fraction of retrieved documents at each turn that introduce at least one named entity not seen in any prior retrieved document. This rate drops from 0.81 at turn 10 to 0.29 at turn 128 and reaches 0.17 at turn 256, indicating that the retriever returns largely redundant results in later turns. **(3) Compression capacity:** The context compression ratio

(compressed tokens / raw retrieved tokens) remains stable at ≈ 0.41–0.44 throughout all turn budgets, ruling out compression capacity as a limiting factor. Together, these results indicate that **retriever saturation** (i.e., the exhaustion of novel information in the retrieval corpus) is the dominant bottleneck at ultra-long horizons.

C Preliminary: Hindsight Experience Replay

Hindsight Experience Replay (HER) (Andrychowicz et al., 2017) is a powerful technique designed to improve sample efficiency in sparse-reward RL tasks. In many real-world scenarios, agents receive meaningful feedback only upon achieving specific goals, making traditional RL inefficient due to the scarcity of reward signals. HER addresses this challenge by transforming unsuccessful episodes into useful training data through the concept of substitute goals. Formally, a trajectory under HER is represented as:

$$\tau = (s, a, r, s', g, g') \tag{6}$$

where s denotes the current state, a the executed action, r the reward with respect to the original goal g, s' the next state, and g' a substitute goal. The reward is recomputed with respect to g', denoted as r', allowing the agent to learn from outcomes that were originally considered failures. This relabeling converts sparse, binary rewards into denser signals, accelerating learning.

Relationship Between HER and Our Approach. While HER inspires our method philosophically (both convert sparse binary feedback into denser learning signals), the two differ substantially in mechanism (see Table 6 below). HER operates on goal-conditioned MDPs by relabeling the desired goal post-hoc and replaying stored transitions under the new goal, requiring an explicit replay buffer and a goal-conditioned reward function. In contrast, our proposed Hindsight Response Replay (HRR) operates within the GRPO framework by reshaping the trajectory-level advantage at turn granularity using a heuristic compression quality score. HRR requires neither a replay buffer nor goal substitution; instead, it treats the achieved compression quality as a reference point for credit assignment, redistributing the sparse final reward across turns proportional to each turn's compression quality relative to the trajectory mean. The shared philosophical principle is that *information from non-optimal outcomes should not be discarded*: HER extracts value by asking "what goal *would* this trajectory have achieved?", while HRR extracts value by asking "which turns contributed more to the compression objective within this trajectory?"

D Detailed Comparison: HER vs. HRR

Table 6 summarizes the key distinctions between Hindsight Experience Replay (HER) and our proposed Hindsight Response Replay (HRR).

Table 6: Comparison between Hindsight Experience Replay (HER) and Hindsight Response Replay (HRR).

Dimension	HER	HRR (Ours)
Signal source	Relabeled goal reward	Compression quality score Q_{com}
Core mechanism	Goal substitution + transition replay	Advantage reshaping
Replay buffer	Required	Not required
Goal conditioning	Explicit (goal-conditioned policy)	Implicit (trajectory-mean baseline)
Reward densification	Via counterfactual goal relabeling	Via turn-level quality comparison
Action space	Continuous control / discrete actions	Token-level generation
Shared philosophy	Extract learning signal from non-optimal experiences	

E Proof of Theorem 4.1

Theorem E.1 (Utility Advantage of Relevance-Aware Allocation)**.** *Let $R \sim P$ be a relevance random variable supported on $\mathcal{R} \subseteq \mathbb{R}$. Let $\bar{\alpha} > 0$ be the average resource budget, and let $\mathcal{A} \subseteq \mathbb{R}_{++}$ be an open interval with $\bar{\alpha} \in \mathcal{A}$. Consider a utility function $\Phi : \mathcal{R} \times \mathcal{A} \to \mathbb{R}$ such that $\Phi(\cdot, \alpha)$ is measurable for every $\alpha \in \mathcal{A}$, and $\Phi(r, \cdot)$ is differentiable and strictly concave on $\mathcal{A}$ for every $r \in \mathcal{R}$. Let the uniform allocation be $\alpha_{\text{uni}}(r) \equiv \bar{\alpha}$, and let the relevance-aware adaptive allocation be $\alpha_{\text{ada}}(r) = f(r)$, where $f : \mathcal{R} \to \mathcal{A}$ is measurable, non-decreasing, and satisfies*

$$\mathbb{E}[f(R)] = \bar{\alpha}, \qquad \mathbb{P}(f(R) \neq \bar{\alpha}) > 0.$$

Define the allocated marginal utility

$$G_f(r) = \partial_\alpha \Phi(r, \alpha)\big|_{\alpha = f(r)}.$$

Assume that G_f is measurable and non-decreasing, and that $\Phi(R, f(R))$, $\Phi(R, \bar{\alpha})$, $G_f(R)$, and $G_f(R)f(R)$ are integrable. Then

$$\mathbb{E}[\Phi(R, f(R))] > \mathbb{E}[\Phi(R, \bar{\alpha})]\,.$$

Proof. For each fixed relevance value r, define the utility difference between adaptive and uniform allocation as

$$\Delta(r) = \Phi(r, f(r)) - \Phi(r, \bar{\alpha}).$$

Since $\alpha \mapsto \Phi(r, \alpha)$ is differentiable and strictly concave on $\mathcal{A}$, the first-order concavity inequality gives

$$\Phi(r, \bar{\alpha}) \leq \Phi(r, f(r)) + G_f(r)\big(\bar{\alpha} - f(r)\big),$$

with strict inequality whenever $f(r) \neq \bar{\alpha}$. Therefore,

$$\Delta(r) \geq G_f(r)\big(f(r) - \bar{\alpha}\big).$$

To handle strictness rigorously, define the slack term

$$S(r) = \Delta(r) - G_f(r)\big(f(r) - \bar{\alpha}\big).$$

By strict concavity, $S(r) \geq 0$, and $S(r) > 0$ whenever $f(r) \neq \bar{\alpha}$. Since the adaptive allocation is non-trivial,

$$\mathbb{P}\big(f(R) \neq \bar{\alpha}\big) > 0.$$

Thus, $S(R)$ is non-negative and strictly positive on an event with positive probability. Under the stated integrability assumptions,

$$\mathbb{E}[S(R)] > 0.$$

Taking expectation over R gives

$$\begin{aligned}\mathbb{E}[\Delta(R)] &= \mathbb{E}\big[G_f(R)\big(f(R) - \bar{\alpha}\big)\big] + \mathbb{E}[S(R)] \\ &> \mathbb{E}\big[G_f(R)\big(f(R) - \bar{\alpha}\big)\big].\end{aligned}$$

It remains to show that the last expectation is non-negative. Let

$$A = f(R), \qquad M = G_f(R).$$

Using the budget constraint $\mathbb{E}[A] = \bar{\alpha}$, we have

$$\begin{aligned}\mathbb{E}[M(A - \bar{\alpha})] &= \mathbb{E}[MA] - \bar{\alpha}\mathbb{E}[M] \\ &= \mathbb{E}[MA] - \mathbb{E}[A]\mathbb{E}[M] \\ &= \mathrm{Cov}(M, A).\end{aligned}$$

Because both f and G_f are non-decreasing functions of r, this covariance is non-negative. Let R' be an independent copy of R. Then

$$\begin{aligned}2\,\mathrm{Cov}(M, A) = \mathbb{E}\big[&\big(G_f(R) - G_f(R')\big) \\ &\cdot\big(f(R) - f(R')\big)\big].\end{aligned}$$

Since both G_f and f are non-decreasing, we have

$$\big(G_f(R) - G_f(R')\big)\big(f(R) - f(R')\big) \geq 0$$

almost surely. Therefore,

$$\mathrm{Cov}(M, A) \geq 0.$$

Consequently,

$$\mathbb{E}\big[G_f(R)\big(f(R) - \bar{\alpha}\big)\big] \geq 0.$$

Combining the above inequalities yields

$$\mathbb{E}[\Delta(R)] > 0.$$

Equivalently,

$$\mathbb{E}\big[\Phi(R, f(R))\big] > \mathbb{E}\big[\Phi(R, \bar{\alpha})\big].$$

This completes the proof. □

F Human Evaluation of Compression Quality

To validate that Q_{com} correlates with human-perceived compression quality, we conducted a targeted human evaluation study. We randomly sampled 100 compression outputs from ZipRL-8B inference trajectories on the MusiQue and Frames test sets (50 per dataset, 20 per compression level). Three annotators with graduate-level NLP backgrounds rated each compressed output on two dimensions: **Fidelity** (1–5) and **Coherence** (1–5). Inter-annotator agreement was measured by Cohen's κ, yielding $\kappa = 0.73$ for Fidelity and $\kappa = 0.71$ for Coherence, indicating substantial agreement.

Table 7: Spearman correlation between Q_{com} sub-scores and human ratings ($n = 100$ samples).

Metric	**Fidelity** ρ	**Coherence** ρ
Q_{info}	0.64	0.29
Q_{sem}	0.27	0.73
Q_{ratio}	0.43	0.46
Q_{level}	0.39	0.54
Q_{com} (composite)	**0.68**	**0.69**

The weighted composite Q_{com} achieves the highest correlation on both dimensions ($\rho = 0.68$ / 0.69), outperforming all individual sub-scores and justifying the multi-dimensional design.

G Details of Compression Quality Scoring

G.1 Compression Ratio Score Q_{ratio}

Let l_y be the character length of the generated compressed text y, and $[\mathcal{L}_l^g, \mathcal{L}_h^g]$ be the target length interval for compression level g. We employ a

piecewise function:

$$Q_{\text{ratio}} = \begin{cases} 1.0, & \text{if } \mathcal{L}_l^g \le l_y \le \mathcal{L}_h^g \\ 0.8 \cdot \frac{l_y}{\mathcal{L}_l^g}, & \text{if } l_y < \mathcal{L}_l^g \\ 1.0 - 0.3 \cdot \frac{l_y - \mathcal{L}_h^g}{0.5\mathcal{L}_h^g}, & \text{if } \mathcal{L}_h^g < l_y \le 1.5\mathcal{L}_h^g \\ 0.3 \cdot \frac{1.5\mathcal{L}_h^g}{l_y}, & \text{if } l_y > 1.5\mathcal{L}_h^g \end{cases} \tag{7}$$

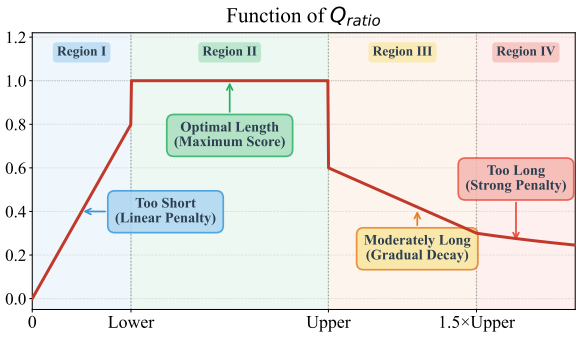

Figure 6: Compression ratio score versus text length.

G.2 Information Retention Score Q_{info}

Let K_q and K_x denote the sets of keywords extracted from the query q and the original text x, respectively. The retention score is:

$$Q_{\text{info}} = 0.7S_{\text{key}} + 0.3S_{\text{gen}}, \tag{8}$$

where:

$$S_{\text{key}} = \frac{|K_q \cap y|}{|K_q \cap x| + \epsilon}, \tag{9}$$

$$S_{\text{gen}} = 0.6 \cdot \frac{|y \cap K_x|}{|K_x| + \epsilon} + 0.4 \cdot \min\left(1.0, 2.0 \cdot \frac{|y|}{|x|}\right). \tag{10}$$

G.3 Semantic Completeness Score Q_{sem}

We evaluate linguistic validity using a multiplicative penalty model:

$$Q_{\text{sem}} = 1.0 \cdot p_{\text{str}} \cdot p_{\text{len}} \cdot p_{\text{punc}}, \tag{11}$$

with the penalties defined as:

- $p_{\text{str}} = 0.7$ if the text contains only a single sentence (unless level $g = 1$), and 0.3 if no valid sentence structure is detected; otherwise 1.0.
- $p_{\text{len}} = 0.5$ if word count < 20, and 0.9 if word count > 1200; otherwise 1.0.
- $p_{\text{punc}} = 0.8$ if no sentence terminators are present; otherwise 1.0.

Table 8: Distribution of compression level selections by ZipRL-8B across 624 test trajectories.

Compression Level	Count	Percentage
Level 1 (Ultra-coarse)	65	10.42%
Level 2 (Coarse)	10	1.60%
Level 3 (Medium)	339	54.33%
Level 4 (Fine)	51	8.17%
Level 5 (Ultra-fine)	159	25.48%
Total	**624**	**100.00%**

G.4 Level Strategy Consistency Score Q_{level}

Given the target bounds $[\mathcal{W}_{min}^g, \mathcal{W}_{max}^g]$ and maximum sentence count S_{max}^g:

$$Q_{\text{level}} = 0.7S_{\text{word}} + 0.3S_{\text{sent}}, \tag{12}$$

where:

$$S_{\text{word}} = \begin{cases} 1.0, & \text{if } \mathcal{W}_{min}^g \le W_y \le \mathcal{W}_{max}^g \\ W_y / W_{min}^g, & \text{if } W_y < \mathcal{W}_{min}^g \\ 0.7, & \text{if } \mathcal{W}_{max}^g < W_y \le 1.5\mathcal{W}_{max}^g \\ 0.3, & \text{otherwise} \end{cases} \tag{13}$$

$$S_{\text{sent}} = \min\left(1.0, \max\left(0.3, \frac{S_{max}^g}{S_y}\right)\right). \tag{14}$$

H Analysis of Learned Compression Strategy

To understand the adaptive compression behavior learned by ZipRL, we analyzed the compression level selection patterns across 624 multi-turn interaction trajectories in our held-out test set.

The predominance of Level 3 (54.33%) shows the model learned that moderate compression suffices for most documents while conserving computational resources. The distribution aligns with Theorem 4.1, validating that adaptive allocation outperforms uniform compression.

I Tool Configuration

To enable the agent to perform automated information retrieval tasks, we designed three tools.

Search Tool: Accepts textual queries and returns the most relevant document fragments, document IDs, and URLs.

Open-Page Tool: Accepts either a URL or a document ID and returns the complete webpage content, enabling detailed retrieval following search results.

Finish Tool: Invoked when the agent has obtained a definitive answer or can no longer proceed.

Dataset	Description	Samples	Usage
ASearcher-LRM	Long-Horizon Search	4,500	Train
BC-Plus	Web Browsing	830	Test
MusiQue	Multi-hop QA	2,459	Test
SQuAD	Multi-hop QA	2,694	Test
Frames	Multi-hop QA	824	Test
Bamboogle	Multi-hop QA	125	Test

Table 9: Statistics of the training and test datasets used by all methods.

By coordinating these three tools, the agent executes the full "Search, Browse, Analyze, and Output" pipeline for automated information retrieval. ZipRL enforces a "compress-on-every-turn" mechanism: regardless of the specific tool invoked, the agent outputs a compressed reasoning state or observation summary before deciding on the next action.

J Experimental Details

J.1 Datasets for Experiments

Table 9 summarizes the statistics of all training and test datasets used in our experiments.

J.2 Selection of Compression Granularity

To establish a principled framework for compression layer division, we analyzed 1,949 samples of successfully resolved real-world dialogues. Utilizing GPT-4, we generated compressed summaries for each dialogue turn and performed K-means clustering on the summary lengths of these samples, exploring cluster counts (K) ranging from 2 to 8. As illustrated in Table 10, the Davies-Bouldin Index reaches its minimum value of 0.4988 at $K = 5$, indicating that a five-cluster configuration yields the optimal separation quality.

Table 10: Davies-Bouldin Index Comparison Across Different K Values.

K	2	3	4	$5^\star$	6	7	8
DB Index	0.616	0.570	0.512	**0.499**	0.506	0.521	0.513
$\Delta\%$	—	-7.5	-10.2	-2.5	$+1.4$	$+3.0$	-1.5

Based on these clustering results, we formally defined five compression levels:

- *Level 1 (100–500 chars):* Ultra-coarse compression, retaining only core conclusions.
- *Level 2 (500–1000 chars):* Coarse compression, retaining primary questions and key conclusions.
- *Level 3 (1000–2000 chars):* Medium compression, retaining decision points and important context.
- *Level 4 (2000–3000 chars):* Fine compression, retaining the complete logical chain and important details.
- *Level 5 (3000–6000 chars):* Ultra-fine compression, preserving nearly all critical information.

K Scaling to 32B Parameters

To empirically prove that our adaptive multi-granularity compression method does not encounter a performance ceiling at the 14B scale, we extended our framework to a 32B parameter base model (ZipRL-32B). Evaluated across the benchmarks, ZipRL-32B achieves an average EM of 35.2%, representing a +4.1% absolute EM improvement over the ZipRL-14B model, demonstrating that ZipRL's performance continues to scale effectively with larger base models.

L Same-Base-Model Comparison (Qwen2.5)

The strictly matched-scale comparison on the identical Qwen2.5 series (3B/7B/14B) used by the baselines is presented in the main text (Table 2 in Section 5). The average EM gains are $+18.0$, $+5.5$, and $+1.8$ at the 3B, 7B, and 14B scales, respectively. The gains shrink at larger scales because stronger base models possess better implicit context management capabilities, and ASearcher-14B is trained with roughly $6\times$ more RL data than ZipRL; nevertheless, on the Qwen3 series the gains remain substantial (e.g., $+7.5$ EM over AgentFold-14B). These controlled results confirm that ZipRL's performance gains stem from algorithmic contributions rather than generational model differences.

M Comparison with Memory-Based Methods

MemAgent (Yu et al., 2026) and MEM1 (Zhou et al., 2026) are designed for processing pre-given long documents via fixed-length memory overwrite or consolidation. To enable fair comparison in our multi-turn retrieval setting, we adapt both methods by merging retrieved documents and splitting them into 500-word chunks as tool observations, injected into their respective memory mechanisms. The full comparison, now presented in the main

text (Table 3 in Section 5), shows that ZipRL consistently outperforms both adapted baselines across all scales and datasets.

N Hyperparameter and Sensitivity Analysis

N.1 Hyperparameter Analysis

To analyze the sensitivity of the hyperparameter w in Equation (4), we evaluate ZipRL with the Qwen3-8B base model at different values of w, as shown in Figure 4 (right). When $w = 0$, the model reverts to standard GRPO, lacking turn-level dense advantage, and performs the worst, validating the effectiveness of HRR. For w values between 0.2 and 0.8, ZipRL shows consistent strong performance, with the best results at $w = 0.2$, which we adopt as the default.

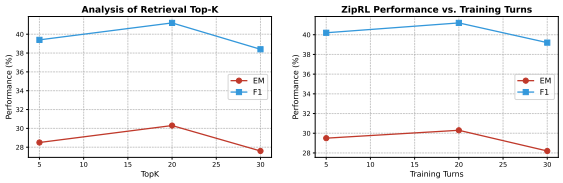

Figure 7: Impact of retrieval top-k (left) and training turns (right) on average EM and F1 scores across five datasets.

N.2 Retrieval Top-K Analysis

To assess the impact of the number of retrieved documents on ZipRL's performance, we evaluate the model at top-k values of 5, 20, and 30. As shown in Figure 7 (left), the best performance occurs when top-k is set to 20, which we adopt as the default parameter.

N.3 Training Turns Analysis

To investigate the impact of the maximum number of interaction turns on ZipRL's performance, we evaluate the model at maximum training turns of 5, 20, and 30. As shown in Figure 7 (right), the model achieves the best performance when the maximum number of turns is set to 20, which we select as the default parameter.

O Agent Behavior Analysis

We highlight four key observations. **(1) Adaptive difficulty-aware search depth:** ZipRL dynamically adjusts retrieval intensity based on question difficulty: it issues only 2.5 searches on Bamboogle but scales to 8.5 on Frames (3.4× increase), an emergent behavior arising from outcome-driven RL training. **(2) Excessive retrieval without convergence (Search-R1):** Search-R1-7B issues ∼9.6 searches on Bamboogle (3.8× more than ZipRL) yet achieves significantly lower accuracy, with a Finish Rate of only 0.20, indicating frequent context budget exhaustion without conclusive answers. **(3) Under-retrieval is not efficiency:** ReAct and NestBrowse models issue very few searches (0.87–1.26 per query), but this reflects under-retrieval rather than genuine efficiency; these models underperform substantially on multi-hop datasets requiring evidence chaining. **(4) Reliable termination:** ZipRL achieves a Finish Rate of 0.85, substantially outperforming AgentFold (0.58) and Search-R1-7B (0.20), demonstrating that ZipRL reliably commits to answers after sufficient evidence collection.

Table 11: Fine-grained agent behavior statistics (averaged across MusiQue, SQuAD, Frames, Bamboogle).

Model	Avg Turns	Avg Search/q	Avg Open/q	Finish Rate
ZipRL (Ours)	7.07	5.83	0.19	0.85
AgentFold	6.23	4.94	0.54	0.58
ASearcher	4.02	3.02	0.00	0.98
NestBrowse-8B	3.51	1.35	0.00	0.83
Search-R1-7B	8.90	8.09	0.07	0.20

Table 12: Average actual interaction turns and performance on the BC-Plus deep research benchmark.

Model	Avg Turns	Average EM	Average F1
ZipRL-8B (Ours)	13.8	0.3233	0.4082
GPT-4o-ReAct	6.0	0.2520	0.3161
DeepSeek-v3.2-ReAct	3.1	0.0880	0.1329
Qwen3-235B-ReAct	2.1	0.2067	0.2355
NestBrowse-8B	1.3	0.0880	0.1453

To demonstrate deep interaction capabilities on the BrowseComp-Plus (BC-Plus) benchmark, which requires complex deep research, ZipRL maintains effective exploration for an average of 13.8 turns, while baselines like Qwen3-235B-ReAct and Search-R1 terminate prematurely (2.1 to 6.0 turns) due to under-retrieval or context overload (Table 12).

The training dynamics (reward evolution and compression quality curves) and multi-run stability analysis are presented in the main text (Table 5 in Section 5): HRR consistently achieves much lower cross-run standard deviation ($\sigma = 0.0099$–0.0124) than standard GRPO ($\sigma = 0.0351$–0.0638) across 3 independent runs.

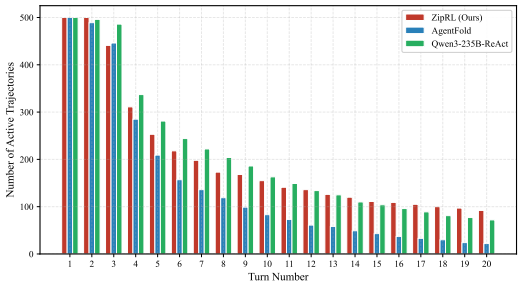

Figure 8: Distribution of active trajectories across different interaction turns on 500 difficulty queries from the MusiQue dataset. Sharp drops in baseline models between turns 5–10 indicate premature termination, while ZipRL maintains consistent reasoning depth up to 20 turns.

P Trajectory Distribution Analysis

As shown in Figure 8, the baseline models exhibit sharp trajectory interruptions between turns 5 and 10, but the failure mechanisms differ. **Context Overload (Qwen3-235B-ReAct):** Despite its vast parameter scale, Qwen3-235B-ReAct achieves an EM of only 18.2% under a full history retention strategy. **Cognitive Dead-ends (AgentFold):** AgentFold's fixed compression strategy lacks dynamic perception of task relevance, and aggressive compression often prunes "bridge entities" essential for multi-hop reasoning. In contrast, ZipRL demonstrates significant reasoning maintenance in deep interaction intervals (10–20 turns).

Q Pseudocode of ZipRL

Algorithm 1 presents the complete training procedure of ZipRL, covering both the cold-start SFT phase and the HRR-based RL optimization phase.

R Robustness Under Noisy Retrieval

We evaluate whether ZipRL adaptively adjusts its compression strategy when retrieval quality degrades. Using ZipRL-8B, we inject noise into the retrieved document pool at five levels: Clean (0%), Low (25%), Mid (50%), High (75%), and Adversarial (100% adversarial documents).

Robustness to Random Noise (0–75%). Performance degrades gracefully under random noise. Across four datasets, the EM drop from Clean to High (75%) noise is modest: -9.8% on Bamboogle, -4.2% on MusiQue, -11.2% on SQuAD, and -15.3% on Frames. ZipRL retains approximately 85–90% of clean performance at 75% noise.

Adversarial Retrieval Boundary. Under fully adversarial conditions (100% adversarial documents), performance degrades substantially across all datasets (e.g., MusiQue: $0.330 \rightarrow 0.002$), consistent with the shared assumption of retrieval trustworthiness in RAG-based systems. Notably, ZipRL's average retrieval turns increase from $\sim$8.6 to $\sim$35.9 (Table 13), suggesting emergent contradiction-resolution behavior. Extending the compression scoring framework with retrieval credibility estimation remains future work.

Table 13: ZipRL-8B performance and average retrieval turns under varying noise levels (EM / Avg Turns).

Noise Level	Bamboogle EM / Turns	MusiQue EM / Turns	SQuAD EM / Turns	Frames EM / Turns
Clean (0%)	0.488 / 8.6	0.330 / 15.4	0.232 / 6.4	0.163 / 20.7
Low (25%)	0.440 / 8.1	0.324 / 15.5	0.256 / 6.5	0.160 / 21.0
Mid (50%)	0.440 / 8.4	0.312 / 15.3	0.218 / 8.4	0.160 / 22.4
High (75%)	0.440 / 9.8	0.316 / 17.0	0.206 / 9.2	0.138 / 23.6
Adversarial	0.072 / 35.9	0.002 / 36.7	0.012 / 31.6	0.018 / 37.7

S Comparison with Adaptive RAG Compression Methods

Table 14: Taxonomy of context compression paradigms. ✓: supported; ✗: not supported.

Method	Multi-turn	Per-doc Adaptive	RL-optimized	Joint Retrieval
LLMLingua / RECOMP	✗	✗	✗	✗
ACC-RAG / SARA	✗	✓	✗	✗
AttnComp	✗	✓	✗	✗
MemAgent / MEM1	✓	✗	✗	✗
ZipRL (ours)	✓	✓	✓	✓

Recent adaptive RAG compression methods such as ACC-RAG (Guo et al., 2025), AttnComp (Luo et al., 2025), and SARA (Jin et al., 2026) pursue variable-rate context compression in single-turn settings, with no mechanism to issue new retrieval actions or receive optimization signal from downstream task outcomes via RL. ZipRL addresses a fundamentally different problem: *online, multi-turn agentic search*, where compression and agentic behavior are co-optimized through sparse end-task reward.

T Robustness of Compression Quality Score Against Gaming Behaviors

1. Mutual Structural Constraints. While generating redundant query keywords might marginally increase Q_{info}, this behavior is strictly penalized

Algorithm 1 ZipRL: Adaptive Multi-Turn Context Compression with Hindsight Response Replay

Input: Training Query Set $\mathcal{Q}$, SFT Model π_{sft}, Reference Model π_{ref}, Group Size G, Reshaping Weight w.
Output: Optimized Policy π_θ.
Stage 1: Supervised Fine-Tuning (SFT)
Initialize $\theta \leftarrow \text{Train}(\pi_{\text{init}}, \mathcal{D}_{\text{demo}})$
Stage 2: Reinforcement Learning with HRR
while not converged **do**
  Sample a batch of queries $q \sim \mathcal{Q}$
  for each query q **do**
    Sample G trajectories $\mathcal{O} = \{o^{(1)}, \ldots, o^{(G)}\}$ using $\pi_\theta(\cdot|q)$
    Obtain sparse outcome rewards $\mathcal{R}_{\text{final}} = \{R^{(1)}_{\text{final}}, \ldots, R^{(G)}_{\text{final}}\}$
  end for
  for each trajectory $o^{(i)}$ in $\mathcal{O}$ **do**
    for each turn $j = 1$ to T **do**
      Calculate **Turn Quality**: $Q^{(i,j)}_{\text{com}} \leftarrow \text{Score}(o^{(i,j)})$
    end for
    Calculate **Trajectory Baseline**: $\bar{Q}^{(i)}_{\text{com}} \leftarrow \frac{1}{T}\sum_j Q^{(i,j)}_{\text{com}}$
  end for
  Compute group stats: $\mu_R \leftarrow \text{mean}(\mathcal{R}_{\text{final}}), \sigma_R \leftarrow \text{std}(\mathcal{R}_{\text{final}})$
  for each trajectory i and turn j **do**
    $A^{(i)}_{\text{GRPO}} \leftarrow (R^{(i)}_{\text{final}} - \mu_R)/(\sigma_R + \varepsilon)$
    $A^{(i,j)}_{\text{HRR}} \leftarrow A^{(i)}_{\text{GRPO}} + w \cdot (Q^{(i,j)}_{\text{com}} - \bar{Q}^{(i)}_{\text{com}})$
  end for
  Update θ by maximizing $\mathcal{L}(\theta)$
end while

by Q_{ratio} (length overflow) and Q_{sem} (destroyed grammatical coherence). Since the final Q_{com} is a weighted sum, the penalty in structural and semantic dimensions heavily outweighs the illicit gain in Q_{info}.

2. Mathematical Bounding. The keyword coverage term $S_{\text{key}} = |K_q \cap y|/(|K_q \cap x| + \epsilon)$ naturally caps at 1.0, preventing unbounded reward through keyword repetition.

3. Anchoring by Final Task Reward. HRR functions as an advantage reshaping technique rather than replacing the environment reward. The final advantage remains fundamentally anchored by the GRPO advantage, which is strictly determined by the exact match (EM/F1) of the final answer. This is empirically confirmed by the monotonically increasing Pearson correlation between Q_{com} and task reward ($r : 0.25 \rightarrow 0.51$).

U Prompts

We used the prompts shown in Figure 9 and Figure 10.

Design of Prompt for ZipRL-Agent (Part I)

System Prompt:

You are an intelligent Agent assistant with tool invocation and context compression capabilities.
Your task is to understand user needs, invoke appropriate tools, and intelligently compress context based on conversational relevance.
[Workflow]
1. Understand Context
- Analyze current input (user question or compressed summary)
- Identify core intent and key information
2. Reasoning and Decision-Making
- **Tool Selection**: Determine if tool invocation is needed and select appropriate tool
- **Relevance Assessment**: Evaluate response relevance to original question (0%-100%)
- **Compression Level**: Determine compression level based on relevance (Levels 1-5)
3. Output Results
- Output reasoning process
- Provide compressed summary (if applicable)
- Give final response or invoke tool
[Available Tools]

---- BEGIN FUNCTION #1: search ----
Description: Performs web search with 'query' and optional 'topk' parameters
Parameters:
(1) query (string, required): Search query string
(2) topk (integer, optional): Number of top results (default: 10)
---- END FUNCTION #1 ----

---- BEGIN FUNCTION #2: open_page ----
Description: Open page by docid or URL from search results
Parameters:
(1) docid (string, optional): Document ID from search results
(2) url (string, optional): URL from search results
---- END FUNCTION #2 ----

---- BEGIN FUNCTION #3: finish ----
Description: Return final answer with explanation
Parameters:
(1) answer (string, required): Concise final answer
(2) explanation (string, required): Evidence-grounded explanation with citations [docid]
(3) confidence (string, optional): Confidence score (0%-100%)
---- END FUNCTION #3 ----

<IMPORTANT>
Function Call Format:
<function=function_name>
 <parameter=param_name>value</parameter>
</function>
</IMPORTANT>

Figure 9: Prompt Design for ZipRL, Part 1.

Design of Prompt for ZipRL-Agent (Part II)

[Compression Level Strategy]

Level	Relevance	Char Count	Scenarios
Level 1	<20%	200-500	Topic completely diverged
Level 2	20%-40%	500-1000	Topic extended to related direction
Level 3	40%-60%	1000-2000	Topic extended/deepened
Level 4	60%-80%	2000-3000	Topic closely related
Level 5	>80%	3000-6000	Topic highly relevant

Compression Strategies

Level 1 - Extremely Coarse: Core theme + final conclusion only
Level 2 - Coarse: Core theme + main questions + key conclusions
Level 3 - Medium: Theme + key decision points + important context
Level 4 - Fine: Complete logic chain + important details
Level 5 - Extremely Fine: All important information + language refinement
Naturally integrate `docid` and `url` citations into the summary if the tool output provides relevant content.

[Output Format Example] **All four components below are required! Each must be wrapped with corresponding tag symbols**

<think>
Reasoning process
</think>

<summary_level>
Number
</summary_level>

<summary>
Compressed summary
</summary>

<function=function_name>
<parameter=param_name>value</parameter>
</function>

User Prompt:
[Question]
{original_question}

[Previous_Summary]
{previous_summary if exists}

[Tool result]
{tool_result if exists}

Please perform reasoning and output according to the workflow in the system prompt.

Figure 10: Prompt Design for ZipRL, Part 2.