Enhancing LLM-based Search Agents via Contribution Weighted Group Relative Policy Optimization

Junzhe Wang[1*†], **Zhiheng Xi**[1*], **Yajie Yang**[1], **Hao Luo**[1],
Shihan Dou[1], **Tao Gui**[1], **Qi Zhang**[1,2,3†]
[1] Fudan University [2] Shanghai Artificial Intelligence Laboratory
[3] Shanghai Key Laboratory of Intelligent Information Processing
jzwang24@m.fudan.edu.cn, qz@fudan.edu.cn

Abstract

Search agents extend Large Language Models (LLMs) beyond static parametric knowledge by enabling access to up-to-date and long-tail information unavailable during pretraining. While reinforcement learning has been widely adopted for training such agents, existing approaches face key limitations: process supervision often suffers from unstable value estimation, whereas outcome supervision struggles with credit assignment due to sparse, trajectory-level rewards. To bridge this gap, we propose Contribution-Weighted GRPO (CW-GRPO), a framework that integrates process supervision into group relative policy optimization. Instead of directly optimizing process rewards, CW-GRPO employs an LLM judge to assess the retrieval utility and reasoning correctness at each search round, producing per-round contribution weights. These weights are used to rescale outcome-based advantages along the trajectory, enabling fine-grained credit assignment without sacrificing optimization stability. Experiments on multiple knowledge-intensive benchmarks show that CW-GRPO outperforms standard GRPO by 5.0% on Qwen3-8B and 6.3% on Qwen3-1.7B, leading to more effective search behaviors. Additional analysis reveals that successful trajectories exhibit concentrated contributions in specific rounds, providing empirical insight into search agent tasks. Our code is available at https://github.com/zsxmwjz/CW-GRPO.

1 Introduction

Search agents empower large language models (LLMs) to move beyond fixed parametric knowledge by providing access to real-time information and specialized facts absent from pre-training (Li et al., 2025a; Zhang et al., 2025a). Through iterative retrieval of external evidence and integration into the reasoning process, search agents ground model predictions in verifiable sources and substantially enhance the factual reliability of LLM-based systems (Zhao et al., 2026; Li et al., 2025b; Zhou et al., 2024; Wang et al., 2025a). Consequently, they have become a key role in enabling LLMs to perform knowledge-intensive and fact-sensitive tasks (Glass et al., 2022; Sawarkar et al., 2024).

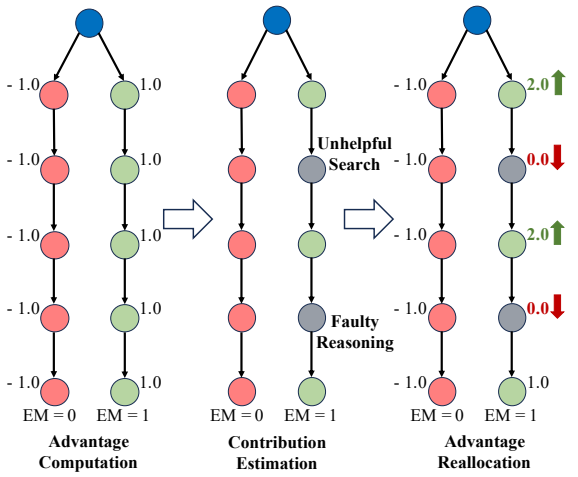

Figure 1: CW-GRPO assesses the quality of each search round within successful trajectories and to concentrate the learning advantage on high-quality search rounds, thereby enabling round-level credit assignment.

While reinforcement learning (RL) offers a significant paradigm for training such agents, current RL approaches suffer from fundamental limitations (Jin et al., 2025a). Depending on the level of supervision, they can be broadly categorized into process supervision and outcome supervision (Uesato et al., 2022). Process supervision (Lightman et al., 2023; Zhu et al., 2025; Wang et al., 2025b) assigns round-level rewards and is commonly optimized with actor–critic methods such as Proximal Policy Optimization (Schulman et al., 2017), but learning reliable critics over diverse intermediate states is unstable and frequently leads to biased advantages and brittle training (Liu et al., 2024; Kazemnejad et al., 2025). Outcome supervision, by contrast,

[*] Equal contribution.
[†] Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31704–31718

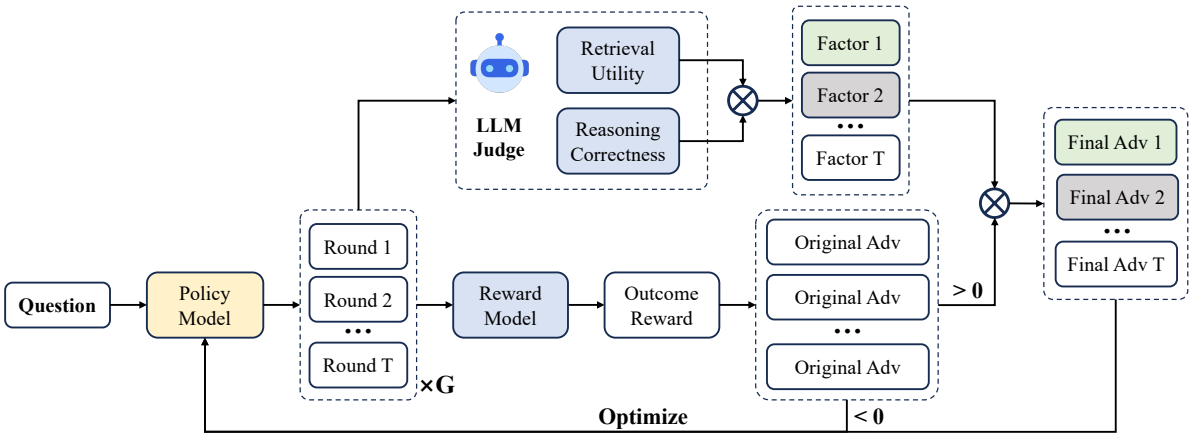

Figure 2: CW-GRPO employs an LLM judge to assess both the retrieval utility and the reasoning correctness at each individual search round. Based on these two signals, it computes a contribution factor for every search round, which is subsequently applied to the original advantage obtained from group-relative comparisons, thereby producing the final round advantage. As a result, training signals from highly contributive rounds are amplified, while those from low-contribution rounds are attenuated.

relies solely on final answer correctness, which makes it difficult to attribute task success to individual search rounds, leading to a **credit assignment problem** and obscuring the uneven contributions of intermediate decisions. Recent advances, such as Group Relative Policy Optimization (Shao et al., 2024), improve training stability and memory efficiency, making outcome-supervised optimization more practical; however, they (Jin et al., 2025b; Wang et al., 2025a) do not alter the sparsity of the reward signal, and the credit assignment issue remains unresolved.

To address the aforementioned challenges, we propose **C**ontribution-**W**eighted **GRPO (CW-GRPO)**, which reformulates process supervision as a problem of modulating outcome-derived advantages, rather than directly optimizing process rewards. Specifically, CW-GRPO retains the stability of the standard GRPO by computing outcome advantages through group-relative comparisons. We use an LLM judge to assess each search round based on retrieval utility and reasoning correctness. These metrics are synthesized into a contribution weight, which serves as a scaling factor to redistribute the outcome advantage across the trajectory. By avoiding direct comparison of process rewards, CW-GRPO incorporates process supervision into the GRPO framework, amplifying learning signals for informative rounds while suppressing those for redundant or unproductive ones, as illustrated in Figure 1.

Empirical results across multiple knowledge-intensive benchmarks demonstrate that CW-GRPO consistently outperforms both outcome-supervised and process-supervised baselines, enabling more effective search behaviors. Specifically, CW-GRPO achieves significant performance gains over standard GRPO, with relative improvements of 5.0% on Qwen3-8B and 6.3% on Qwen3-1.7B. The contributions of this work are summarized as follows:

- **A Reframing of Process Supervision**: We introduce a new perspective that treats process supervision for search agents as advantage reallocation guided by round contribution, rather than explicit process reward estimation.
- **Contribution-Weighted GRPO**: We propose a new optimization framework that brings process-level credit assignment into GRPO without requiring unstable value functions.
- **Empirical Characterization of Search Contribution**: We provide empirical evidence showing a structural characteristic in search agent tasks: the contribution to task success is highly concentrated in informative rounds rather than being uniformly distributed.

2 Related Work

Agentic Search. Large language models encode knowledge in static parameters, limiting access to up-to-date and long-tail knowledge. Retrieval-augmented generation (RAG) addresses this by enabling models to access external knowledge sources (Zhang et al., 2026; Zhao et al., 2026; Li et al., 2025b; Zhou et al., 2024). Pioneering

retrieval-augmented reasoning works such as IR-CoT (Trivedi et al., 2023) and FLARE (Jiang et al., 2023) use prompts to guide iterative reasoning and reduce model uncertainty, and Self-RAG (Asai et al., 2024) adopts supervised fine-tuning with reflection tokens for adaptive retrieval. More recently, agentic search enables LLMs to interact with real-world search engines for dynamic information access. WebGPT (Nakano et al., 2022) pioneers this direction using RLHF. Search-o1 (Li et al., 2025a) integrates agentic search into large reasoning models via prompting, while Search-R1 (Jin et al., 2025b), R1-Searcher (Song et al., 2025) and MM-Doc-R1 (Lin et al., 2026) adopt RL for end-to-end training with outcome-based rewards. However, outcome supervision assigns uniform credit across search rounds, failing to distinguish pivotal searches from redundant ones.

Outcome Supervision and Process Supervision. A key challenge in training search agents lies in credit assignment across multi-round search trajectories. *Outcome supervision*, widely adopted in RL methods like PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024), assigns rewards only based on the final answer, treating all immediate rounds equally and ignoring their unequal contributions. *Process supervision* addresses this via round-level feedback for finer-grained credit assignment. Early approaches to process supervision rely on learned Process Reward Models (PRMs), but often suffer from costly round-level annotation requirements and limited out-of-domain generalizability (Lightman et al., 2023; Huang et al., 2025). To avoid training explicit reward models, recent works increasingly leverage frozen LLMs as external evaluators to assess intermediate rounds. These methods use LLM-based judges to assess reasoning quality and then assign round-level rewards, enabling process supervision without training dedicated PRMs (Zhang et al., 2025b). These approaches typically optimize judge-produced signals directly as rewards, making them sensitive to evaluation noise and calibration errors. In contrast, CW-GRPO uses LLM-derived signals only to redistribute outcome-based advantages, avoiding direct optimization over noisy intermediate rewards and thereby enabling more stable training.

3 Contribution-Weighted GRPO

As motivated in Section 1, current search agents suffer from the credit assignment problem where sparse outcome rewards cannot distinguish between pivotal and redundant search rounds. To address this, we propose Contribution-Weighted Group Relative Policy Optimization (CW-GRPO). Unlike traditional process supervision that requires learning a fallible critic, CW-GRPO reframes process-level signals as dynamic scaling factors that redistribute trajectory-level advantages. The overall framework and its core mechanism of round-level credit assignment are presented in Figure 2.

3.1 Task Formulation

We consider a search agent policy π_θ tasked with answering a question q. The agent interacts with an environment (search engine) through a trajectory τ of T rounds:

$$\tau = \big((s^1, a^1), (s^2, a^2), \ldots, (s^T, a^T)\big), \tag{1}$$

where for each round $t < T$, the action a^t consists of a reasoning chain followed by an invocation of the search tool. The environment then returns retrieved documents to form the state s^{t+1}. The last action a^T is the final answer.

Training proceeds in groups. For each question, we sample a group of G trajectories $\{\tau_i\}_{i=1}^G$ under the current policy. Each trajectory τ_i receives a scalar outcome reward R_i based on the exact match (EM) of its final answer. As in GRPO, these outcome rewards are only used for relative comparison within the group.

3.2 Outcome-Level Advantage

Following the GRPO framework, we first establish a baseline for performance by comparing trajectories within the same group. This avoids the need for a value-function critic. For each trajectory τ_i, we compute a normalized outcome advantage:

$$A_i^O = \frac{R_i - \text{mean}\{R_i\}|_{i=1}^G}{\text{std}\{R_i\}|_{i=1}^G}. \tag{2}$$

While A_i^O effectively identifies which trajectories are better than others, it remains temporally coarse, assigning the same credit to every search round within a successful trajectory, regardless of whether a specific round was informative or distracting.

3.3 Round-Level Contribution Estimation

The effectiveness of CW-GRPO fundamentally depends on accurately identifying pivotal rounds

within a search trajectory. Accordingly, we decompose the quality of each round into two orthogonal binary signals: **retrieval utility** and **reasoning correctness**. Crucially, we adopt a *conjunctive* formulation, under which a round is considered contributive only if it satisfies both criteria, namely retrieving novel, task-relevant evidence and maintaining logically consistent reasoning. This design reflects an inherent structural property of multi-round search: effective task solving requires not only the acquisition of useful information but also its correct interpretation in context. Retrieval without sound reasoning may lead to incorrect interpretation or misuse of useful evidence; conversely, correct reasoning without informative retrieval fails to make meaningful progress toward solving the task. The conjunctive gating mechanism thus serves as a conservative filter, isolating rounds whose contributions are causally aligned with task success, thereby providing a low-noise and more reliable signal for advantage reallocation.

3.3.1 Signal Definition and Rubrics

To ensure the robustness of the LLM-derived signals, we employ a **rubric-based assessment** where an LLM judge is guided by a set of predefined criteria. For each search round $t < T_i$ in trajectory τ_i, the values of the two signals are determined as follows:

- **Retrieval utility** $u_i^t \in \{0, 1\}$: This signal assesses the information gained from external sources. A round is assigned $u_i^t = 1$ only if the retrieved documents contain novel, task-relevant evidence that was not present in the context of previous rounds. This prevents the agent from being rewarded for redundant or circular retrieval queries.

- **Reasoning correctness** $v_i^t \in \{0, 1\}$: This signal estimates the internal consistency of the agent. A round is assigned $v_i^t = 1$ if the reasoning chain correctly interprets the current context and maintains a logical path toward the final answer. This ensures that the agent is not getting the right answer for the wrong reasons.

The conjunctive contribution signal is defined as the logical product $p_i^t = u_i^t \cdot v_i^t$. By enforcing a binary bottleneck, we apply a discrete gate to each search round: it either contributes a necessary building block to the solution or it does not.

3.3.2 Judge Calibration and Reliability

To ensure that the LLM judge aligns with expert intuition, we conducted a rigorous calibration process. We manually annotated a subset of trajectories (comprising 97 distinct search rounds) to serve as a gold standard.

The rubrics were iteratively refined to minimize ambiguity. In our final setup, the LLM judge achieved a **95% consensus rate with human expert annotations** on both retrieval utility and reasoning correctness. This high agreement provides a reliable basis for advantage redistribution, ensuring policy gradients are guided by high-fidelity process signals. Additional calibration details are deferred to Appendix B.

3.4 Adaptive Weighting Mechanism

We then convert these raw signals into a normalized **contribution weight** c_i^t. Importantly, we treat successful and failed trajectories differently, as a robustness-oriented design.

For **successful trajectories** ($R_i = 1$), rounds that introduce new task-relevant evidence and maintain logically consistent reasoning are typically associated with the final successful outcome. A temperature-controlled softmax is therefore adopted to emphasize high-contribution rounds:

$$c_i^t = \frac{\exp\left(\alpha p_i^t\right)}{\sum_{t'=1}^{T_i-1} \exp\left(\alpha p_i^{t'}\right)}, \quad t < T_i, \text{ if } R_i = 1, \tag{3}$$

where α is a hyperparameter controlling the "sharpness" of the redistribution. A higher α forces the model to learn primarily from the most pivotal rounds.

For **failed trajectories** ($R_i = 0$), round-level attribution is substantially more ambiguous. Many intermediate rounds exhibit reasonable behavior, such as attempting to retrieve missing evidence or issuing follow-up queries for verification. However, these rounds may still fail to obtain useful information because the required knowledge is not covered by the corpus or cannot be retrieved due to limitations of the retriever. In such cases, the failure cannot be readily attributed to identifiable errors in the agent's decisions at specific rounds. As a result, assigning differentiated weights based on round-level signals is unreliable and can introduce spurious supervision. A more detailed analysis of such failure patterns is provided in Appendix A. To mitigate this issue, we assign a uniform contribution across all rounds:

$$c_i^t = \frac{1}{T_i - 1}, \quad t < T_i, \text{ if } R_i = 0. \tag{4}$$

This design preserves the stability of outcome-based learning while avoiding noisy round-level credit assignment under ambiguous attribution. Failed trajectories are still utilized through outcome-level comparison, and credit redistribution is applied only when attribution is reliable. By construction, $\sum_{t=1}^{T_i-1} c_i^t = 1$.

3.5 Advantage Reallocation and Optimization

The core of CW-GRPO is the reallocated advantage A_i^t. For all search rounds $t < T_i$ of trajectory τ_i, we scale the outcome advantage by the normalized contribution. And for the final answer round $t = T_i$, we directly use the outcome advantage:

$$A_i^t = A_i^O \cdot c_i^t \cdot (T_i - 1), \text{ if } t < T_i; \quad A_i^{T_i} = A_i^O. \tag{5}$$

The scaling factor $(T_i - 1)$ ensures the magnitude of the total learning signal for the trajectory remains conserved, *i.e.,*

$$\frac{1}{(T_i - 1)} \sum_{t=1}^{T_i-1} A_i^t = A_i^O. \tag{6}$$

Finally, we optimize the policy using the clipped surrogate objective:

$$\mathcal{L}(\theta) = -\mathbb{E}\Big[\min\left(rA, \text{clip}(r, 1-\epsilon, 1+\epsilon)A\right)\Big], \tag{7}$$

where ϵ is a hyper-parameter for clipping, $r = \frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)} = \exp(\log \pi_\theta - \log \pi_{\theta_{old}})$ is the importance sampling ratio.

In summary, CW-GRPO achieves a seamless integration of process supervision into the GRPO framework. Our method achieves effective credit assignment across iterative search rounds by reallocating trajectory-level advantages based on each round's contribution, rather than estimating absolute intermediate rewards. This design preserves the training stability of group-relative optimization while ensuring the policy learns more effectively from the specific decisions that drive task success.

4 Setup

4.1 Datasets and Evaluation Metrics

We evaluate our method on two categories of knowledge-intensive tasks: (1) **General Question Answering**: Natural Questions (NQ) (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and PopQA (Mallen et al., 2023). (2) **Multi-Hop Question Answering**: HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), Musique (Trivedi et al., 2022), and Bamboogle (Press et al., 2023). We merge the training sets of NQ and HotpotQA as the training data. To mitigate the impact of decoding variance, we use Avg@4 Exact Match (EM) as the evaluation metric, averaging the EM scores over four sampled responses per task.

Hard-Case Evaluation Set. In standard settings, models often bypass external search by leveraging internal parametric knowledge, which obscures their true agentic search proficiency. In light of this, we adopt a more challenging protocol using **AgentGym-SearchQA-test**, a 400-sample test set from AgentGym-RL (Xi et al., 2025). These samples are specifically filtered from cases where the Qwen2.5-72B-Instruct (Team, 2024) model failed, representing the "hardest" distribution for current LLMs. This setup shifts the focus from simple parametric recall to complex information-seeking and reasoning capabilities, explicitly emphasizing whether agents retrieve and utilize external evidence rather than relying on parametric knowledge.

Strict Retrieval Constraint. To further emphasize the model's search agentic capabilities, we include a specific instruction in the system prompt: the model is prohibited from using any parametric knowledge beyond basic common sense.

4.2 Baselines

We compare our approach against two categories of search agents: (1) **Outcome-Supervised Methods**: Search-R1-PPO and Search-R1-GRPO (Jin et al., 2025b), which optimize the final answer via PPO and GRPO respectively. (2) **Process-Supervised Methods**: R3-RAG (Li et al., 2025c) and MT-PPO (Wei et al., 2025), which utilize explicit process rewards and PPO frameworks.

We also include results from state-of-the-art closed-source models (GPT-5.1 (Singh et al., 2025), Gemini-3-Pro-Preview (Google, 2025), Qwen3-Max (Team, 2025a)) as strong upper-bound references, alongside powerful open-source models (DeepSeek-V3.2 (DeepSeek-AI et al., 2025), Qwen3-32B (Team, 2025b)) for comparison.

Note on Comparison: Because of the uniquely challenging test set and the strict use of a no-parametric-knowledge prompt, the baseline performance reported in this paper differs from that reported in the original publications.

Method	General QA			Multi-Hop QA				Overall
	NQ†	TriviaQA*	PopQA*	HotpotQA†	2wiki*	Musique*	Bamboogle*	
Proprietary Models								
GPT-5.1	31.0	60.0	38.5	36.0	43.5	13.8	39.0	34.44
Gemini-3-Pro-Preview	31.0	66.0	32.5	42.0	38.5	14.0	42.0	35.00
Qwen3-Max	31.5	62.5	30.5	39.5	49.0	18.5	42.5	36.56
Open-sourced Models								
DeepSeek-V3.2-Thinking	24.5	47.0	22.0	30.0	24.0	12.3	29.5	25.19
Qwen3-32B	20.5	53.0	22.0	28.0	33.5	8.5	30.0	25.50
Qwen3-1.7B	13.5	34.5	18.0	11.0	6.0	2.5	4.5	11.56
Search-R1-PPO	19.5	**50.5**	25.0	25.5	24.0	5.0	14.0	21.06
Search-R1-GRPO	18.0	45.5	25.0	**27.0**	22.5	6.8	16.5	21.00
R3-RAG	20.0	47.5	**25.5**	22.0	18.5	5.3	13.5	19.69
MT-PPO	**22.5**	49.0	25.0	24.0	**31.5**	6.0	13.5	22.19
CW-GRPO	22.0	47.5	25.0	24.0	28.5	**7.0**	**17.5**	**22.31**
Qwen3-8B	23.5	53.5	22.5	27.5	27.0	10.0	30.0	25.50
Search-R1-PPO	29.5	57.0	27.5	**33.5**	36.0	12.3	33.5	30.19
Search-R1-GRPO	33.5	57.0	26.0	**33.5**	32.5	12.5	31.5	29.88
R3-RAG	26.5	**57.5**	26.0	32.0	32.0	11.3	33.5	28.75
MT-PPO	25.5	56.5	26.0	32.0	**37.0**	12.5	31.5	29.19
CW-GRPO	**35.5**	**57.5**	**28.5**	**33.5**	33.0	**13.0**	**37.0**	**31.38**

Table 1: Avg@4 Exact Match on seven subsets of AgentGym-SearchQA-test. † / * indicate in-domain and out-of-domain datasets. CW-GRPO achieves the best performance among methods built on the same backbone.

4.3 Implementation Details

We use Qwen3-8B and Qwen3-1.7B (Team, 2025b) as base models due to their tool-calling capabilities. Training is conducted using the veRL (Sheng et al., 2024) framework. The maximum context length is set to 9192 tokens, with up to 10 interaction rounds. We employ SGLang (Zheng et al., 2024) as the inference engine for rollout, setting both temperature and top_p to 1.0.

Retrieval Infrastructure. We use the wiki-18-corpus dataset released by Search-R1 (Jin et al., 2025b), derived from the 2018 Wikipedia dump, as the knowledge source, and E5-base-v2 (Wang et al., 2022) as the retriever, with the number of retrieved documents k set to 3.

Training Hyperparameters. For all RL-based experiments, we use a KL-divergence regularization coefficient $\beta = 0.001$ and a clipping ratio $\epsilon = 0.2$. Specifically, for our proposed CW-GRPO, we apply GPT-oss-120B (Agarwal et al., 2025) as the judge model and set $\alpha = \infty$. This configuration ensures that for trajectories resulting in task success, the learning signals during search iterations are exclusively concentrated on those rounds where the conjunctive contribution signal is 1. Other specific parameters are detailed in Table 2. The hyper-parameter differences between GRPO-based and PPO-based methods stem from their distinct algorithmic characteristics (intra-group relative comparisons vs. independent trajectory optimization), and we ensure fairness by normalizing the total number of sampled trajectories per training step.

Parameter	GRPO / CW-GRPO	PPO-based (Actor)	PPO-based (Critic)
Learning Rate	1×10^{-6}	1×10^{-6}	1×10^{-5}
Warmup Ratio	0.285	0.285	0.015
Training Steps	200	200	200
Batch Size	32	128	128
Group Size	4	1	1
GAE γ/λ	-	1.0/1.0	1.0/1.0

Table 2: Detailed training hyperparameters of CW-GRPO and baselines.

5 Experimental Results

5.1 Main Results

Table 1 presents the main experimental results on hard knowledge-intensive QA benchmarks under no-parametric-knowledge constraints.

Overall Performance. Our experimental results demonstrate the robust superiority of CW-GRPO, which consistently achieves state-of-the-art perfor-

α	General QA			Multi-Hop QA				Overall
	NQ	TriviaQA	PopQA	HotpotQA	2wiki	Musique	Bamboogle	
0	33.5	57.0	26.0	33.5	32.5	12.5	31.5	29.88
1	27.0	58.0	28.0	32.5	35.0	12.8	31.5	29.69
3	33.0	58.5	31.0	29.5	36.0	13.0	31.0	30.63
5	27.0	58.5	27.0	35.0	39.0	12.0	33.5	30.50
∞	35.5	57.5	28.5	33.5	33.0	13.0	37.0	31.38

Table 3: Effect of the sharpness parameter α on search agent performance. Larger α value generally yields better overall accuracy, indicating that concentrating learning signals on high-contribution search rounds is beneficial.

mance across both model scales, outperforming all outcome- and process-supervised baselines built on the identical backbone. On Qwen3-8B, CW-GRPO achieves an overall score of 31.38, representing a 5.0% relative improvement over Search-R1-GRPO (29.88). On Qwen3-1.7B, CW-GRPO improves the overall performance from 21.00 to 22.31, corresponding to a 6.3% relative gain. These results indicate that CW-GRPO scales favorably across model sizes and is particularly effective in low-capacity regimes, where efficient credit assignment during search is critical.

Comparison with Open- and Closed-Source Models. With Qwen3-8B as the backbone, CW-GRPO surpasses all evaluated open-source models, including DeepSeek-V3.2-Thinking and Qwen3-32B, despite operating under a stricter evaluation setting. However, a clear performance gap remains between CW-GRPO and large closed-source models such as GPT-5.1, Gemini-3-Pro-Preview, and Qwen3-Max, suggesting that stronger base models and proprietary training data still provide advantages under this challenging hard-case distribution.

General QA vs. Multi-Hop QA. CW-GRPO demonstrates consistent improvements on both General QA and Multi-Hop QA tasks compared to outcome-supervised and process-supervised baselines. On Multi-Hop QA, CW-GRPO yields the most pronounced gains. On both Qwen3-1.7B and Qwen3-8B, it improves performance on 2WikiMultiHopQA, Musique, and Bamboogle simultaneously, indicating stronger long-horizon reasoning and evidence aggregation. On General QA, CW-GRPO remains competitive across datasets. While it achieves the best overall General QA performance on Qwen3-8B, its improvement on Qwen3-1.7B is smaller than that of MT-PPO. This suggests that for simpler single-hop questions, dense process supervision can still be effective for very small models, whereas CW-GRPO shows clearer advantages as reasoning depth increases.

Summary. Overall, these results demonstrate that CW-GRPO provides a robust and scalable improvement over existing outcome- and process-supervised search agents, particularly for multi-hop reasoning under strict retrieval constraints. The consistent gains across model sizes highlight the effectiveness of contribution-weighted optimization in guiding search behavior, even in extremely challenging evaluation settings.

5.2 Sharpness of Advantage Reallocation

Based on Qwen3-8B, we analyze the effect of the sharpness parameter α, which controls how outcome advantages are redistributed across search rounds based on their estimated contributions. Setting $\alpha = 0$ yields uniform weighting over the search rounds and reduces the method to standard GRPO, whereas $\alpha = \infty$ corresponds to CW-GRPO, where the advantage is fully concentrated on rounds with both retrieval utility and correct reasoning.

As shown in Table 3, increasing α generally leads to improved overall performance, with the best overall result achieved at $\alpha = \infty$. This trend suggests that, in successful search trajectories, the contribution to task success is highly uneven across rounds. Rather than accumulating gradually, effective progress is driven by high-quality search rounds that exhibit retrieval utility and correct reasoning. For search agent tasks, effective credit assignment requires acknowledging this highly concentrated contribution structure.

5.3 Supervision Decomposition

We ablate the two supervision signals in CW-GRPO by separately removing retrieval utility or

Method	General QA			Multi-Hop QA				Overall
	NQ	TriviaQA	PopQA	HotpotQA	2wiki	Musique	Bamboogle	
CW-GRPO	35.5	57.5	28.5	33.5	33.0	13.0	37.0	31.38
wo-retrieval	27.0	53.5	26.0	31.0	37.5	12.0	34.5	29.19
wo-reasoning	23.5	55.5	31.0	34.0	28.5	12.3	31.5	28.56

Table 4: Ablation on contribution supervision in CW-GRPO. We remove supervision on retrieval utility (wo-retrieval) or reasoning correctness (wo-reasoning), allowing a search round to contribute if only one condition is satisfied. Dropping either signal leads to inferior performance compared to full CW-GRPO, demonstrating that both are necessary for effective credit assignment in search agents.

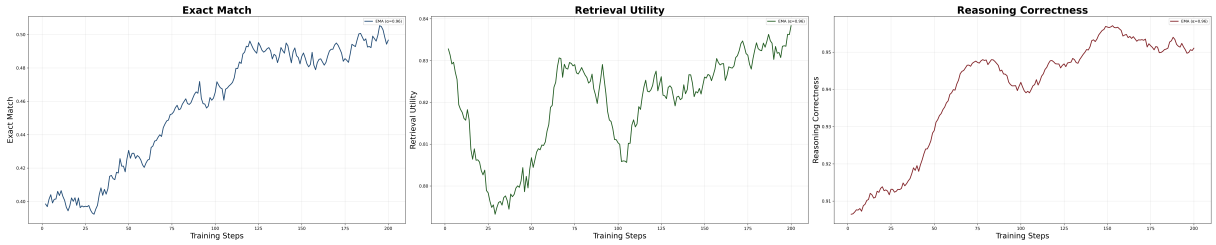

Figure 3: Training curves of CW-GRPO on Qwen3-8B. We report the exponential moving averages of Exact Match, retrieval utility, and reasoning correctness over training steps.

reasoning correctness supervision while keeping all other components unchanged.

According to Table 4, removing either signal consistently degrades performance across both single-hop and multi-hop QA tasks, suggesting that neither signal alone is sufficient for reliable credit assignment during search.

Without retrieval utility supervision, rounds with coherent but evidence-unhelpful reasoning can receive positive credit. Without reasoning correctness supervision, rounds can be assigned positive contribution weights for retrieving relevant documents even when the agent misinterprets existing evidence, weakening the link between evidence acquisition and correct reasoning. In either setting, contribution weights may be assigned to rounds that do not make a substantive contribution to task completion, introducing noise into advantage reallocation and harming overall performance.

These results highlight the necessity of joint supervision: effective search requires both acquiring novel information and correctly reasoning over it, as neither component alone is sufficient to achieve strong overall performance.

5.4 Training Dynamics

As shown in Figure 3, the optimization remains stable across training, with no signs of collapse. The Exact Match score steadily improves, accompanied by consistent upward trends in both retrieval utility and reasoning correctness, indicating progressive improvement in the agent's search behavior. These results show that CW-GRPO achieves stable optimization while improving both final accuracy and the quality of the search process.

We further analyze the effect of the conjunctive gate mechanism during training. Among all search rounds in successful trajectories, approximately 13.84% (2.2k out of 15.8k) are assigned zero contribution, indicating that the gate meaningfully reshapes the distribution of learning signals rather than acting as a near-identity transformation.

5.5 Case Study

Figure 4 compares the search trajectories of Qwen3-8B before and after CW-GRPO training on a factual question regarding John I. Jenkins's tenure start. The untrained model issues only a single query and fails to distinguish between the election year and the actual commencement of his term, leading to an incorrect answer based on this temporal confusion.

In contrast, the CW-GRPO-trained model adopts a qualitatively different strategy by issuing multiple parallel queries in the first round. This expanded search coverage allows the agent to recall a broader set of documents, including explicit evidence that Jenkins's term began in 2005. This case illustrates that CW-GRPO not only improves retrieval utility through parallelization but also strengthens the agent's resistance to distracting evidence. By

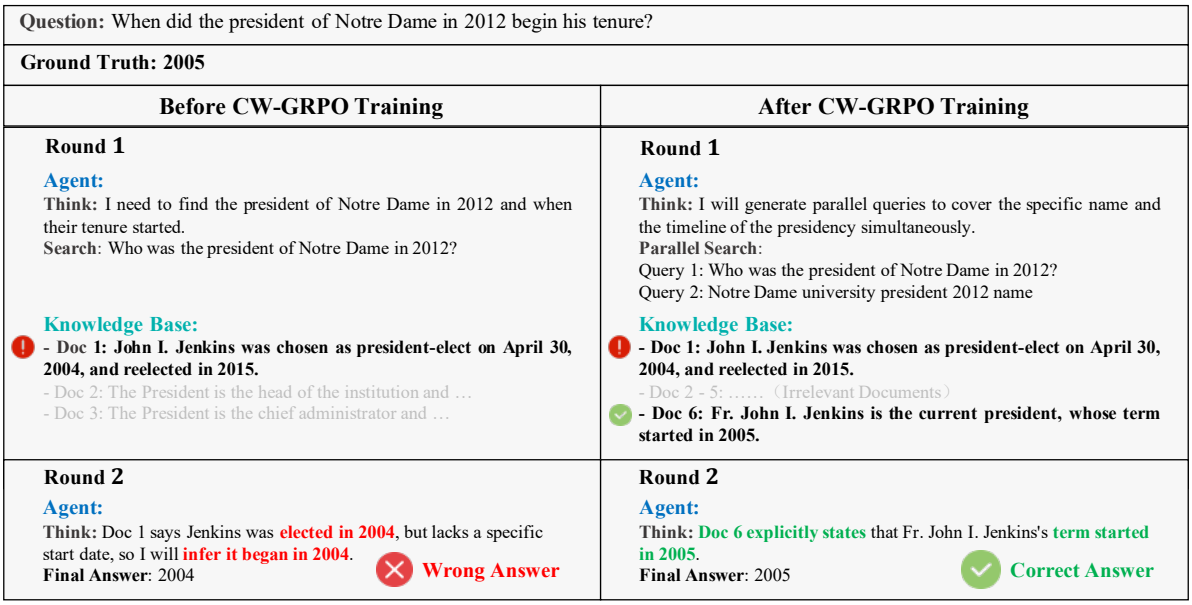

Figure 4: Case study of CW-GRPO on Qwen3-8B.

learning to prioritize direct answer-bearing evidence over superficially relevant but misleading signals, the model achieves more robust and grounded search-based decision making.

6 Discussion: Generalization Across Tasks and Models

CW-GRPO is motivated by the general principle of redistributing outcome-derived advantages according to round-level contribution estimates. This principle is broadly applicable beyond the specific experimental settings considered in this work.

Task generalization. CW-GRPO can be applied to multi-round agentic tasks where the contributions of intermediate rounds can be reasonably estimated. In long-horizon decision-making environments(*e.g.,* Webshop (Yao et al., 2022)), individual actions should correspond to incremental progress toward latent subgoals. This enables a judge to assess state improvements at each round and facilitates round-level credit assignment.

The framework also extends to single-round but structurally decomposable reasoning tasks. In settings such as mathematical reasoning(*e.g.,* MATH (Hendrycks et al., 2021)), model outputs typically consist of multi-step derivations. When a trajectory fails, the error can often be localized to specific steps; these steps can then be assigned higher responsibility for failure, enabling a CW-GRPO variant based on step-level attribution.

However, CW-GRPO is less suitable for tasks where process contribution is ill-defined or inherently unstable. This includes highly subjective generation tasks (*e.g.,* open-ended creative writing) and perceptual generation tasks (*e.g.,* image synthesis), where holistic judgments dominate and assigning reliable credit to intermediate steps via low-variance signals is challenging.

Model generalization. CW-GRPO is model-agnostic and does not rely on a specific backbone architecture. Nevertheless, stronger base models tend to be more sensitive to the accuracy of round-level supervision, placing higher demands on the quality of the judge or rubric design, which may limit the practical gains of the method.

7 Conclusion

In this work, we propose CW-GRPO, which reallocates outcome advantages according to round-level contribution rather than explicit process rewards to tackle the credit assignment challenge in training LLM-based search agents. CW-GRPO retains GRPO's stability while enabling fine-grained credit assignment along search rounds. Our evaluations show that CW-GRPO consistently outperforms baselines, and task success is closely associated with the high-quality rounds, highlighting the importance of modeling contribution concentration in search trajectories. Overall, CW-GRPO offers a simple and effective way to incorporate process-level insights into outcome-supervised training for scalable agentic systems.

Limitations

Despite the advances demonstrated by CW-GRPO in alleviating credit assignment under outcome supervision, several limitations remain.

First, CW-GRPO reallocates advantages only for successful trajectories. For failed trajectories, it does not perform fine-grained credit assignment, as attributing task failure to specific search rounds remains challenging. Consequently, CW-GRPO lacks explicit modeling of which intermediate decisions contribute to failure, leaving failure cases underutilized for process-aware learning.

Second, CW-GRPO adopts a conjunctive gating mechanism for modeling round-level contributions, which improves robustness to judging noise but limits the expressiveness of contribution representation. In addition, the framework operates at the round level and does not naturally extend to token-level supervision, making finer-grained credit assignment beyond rounds difficult to incorporate.

Finally, CW-GRPO relies on an external LLM judge for contribution estimation. While we find the associated computational cost to be manageable in our setting (see Appendix C), this design still introduces additional inference cost and dependency on judge quality. Exploring lighter-weight supervision signals (*e.g.,* heuristic judges) may further improve scalability, especially in more resource-constrained or latency-sensitive scenarios.

Acknowledgments

We used ChatGPT and Gemini to assist with literature search and writing refinement, including correcting grammar errors and improving readability. However, we did not use the AI assistant for research innovation.

We wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by National Key R&D Program of China No.2025ZD0215702, National Natural Science Foundation of China (No. 62576106, 62521004, 62476061, 62376061).

References

OpenAI: Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, and 105 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2g: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715.

Google. 2025. Gemini 3 pro model card. Technical report, Google DeepMind. Accessed: 2026-01-06.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *Preprint*, arXiv:2011.01060.

Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. 2025. An empirical study of LLM-as-a-judge for LLM evaluation: Fine-tuned judge model is not a general substitute for GPT-4. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5880–5895, Vienna, Austria. Association for Computational Linguistics.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Bowen Jin, Jinsung Yoon, Priyanka Kargupta, Sercan O. Arik, and Jiawei Han. 2025a. An empirical study on reinforcement learning for reasoning-search interleaved llm agents. *Preprint*, arXiv:2505.15117.

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.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *Preprint*, arXiv:1705.03551.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. 2025. Vineppo: Refining credit assignment in rl training of llms. *Preprint*, arXiv:2410.01679.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025a. Search-o1: Agentic search-enhanced large reasoning models. *Preprint*, arXiv:2501.05366.

Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2025b. From matching to generation: A survey on generative information retrieval. *ACM Transactions on Information Systems*, 43(3):1–62.

Yuan Li, Qi Luo, Xiaonan Li, Bufan Li, Qinyuan Cheng, Bo Wang, Yining Zheng, Yuxin Wang, Zhangyue Yin, and Xipeng Qiu. 2025c. R3-rag: Learning step-by-step reasoning and retrieval for llms via reinforcement learning. *arXiv preprint arXiv:2505.23794*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *Preprint*, arXiv:2305.20050.

Jiahang Lin, Kai Hu, Binghai Wang, Yuhao Zhou, Zhiheng Xi, Honglin Guo, Shichun Liu, Junzhe Wang, Shihan Dou, Enyu Zhou, Hang Yan, Zhenhua Han, Tao Gui, Qi Zhang, and Xuanjing Huang. 2026. Mmdoc-r1: Training agents for long document visual question answering through multi-turn reinforcement learning. *Preprint*, arXiv:2604.13579.

Jiacai Liu, Chaojie Wang, Chris Yuhao Liu, Liang Zeng, Rui Yan, Yiwen Sun, Yang Liu, and Yahui Zhou. 2024. Improving multi-step reasoning abilities of large language models with direct advantage policy optimization. *Preprint*, arXiv:2412.18279.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *Preprint*, arXiv:2212.10511.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. Webgpt: Browser-assisted question-answering with human feedback. *Preprint*, arXiv:2112.09332.

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.

Kunal Sawarkar, Abhilasha Mangal, and Shivam Raj Solanki. 2024. Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers. In *2024 IEEE 7th international conference on multimedia information processing and retrieval (MIPR)*, pages 155–161. IEEE.

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.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

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. *Preprint*, arXiv:2503.05592.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Qwen Team. 2025a. Qwen3-max: Just scale it.

Qwen Team. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Preprint*, arXiv:2108.00573.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

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

Hongru Wang, Cheng Qian, Wanjun Zhong, Xiusi Chen, Jiahao Qiu, Shijue Huang, Bowen Jin, Mengdi Wang, Kam-Fai Wong, and Heng Ji. 2025a. Acting less is reasoning more! teaching model to act efficiently. *Preprint*, arXiv:2504.14870.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Ziliang Wang, Xuhui Zheng, Kang An, Cijun Ouyang, Jialu Cai, Yuhang Wang, and Yichao Wu. 2025b. Stepsearch: Igniting llms search ability via step-wise proximal policy optimization. *Preprint*, arXiv:2505.15107.

Quan Wei, Siliang Zeng, Chenliang Li, William Brown, Oana Frunza, Wei Deng, Anderson Schneider, Yuriy Nevmyvaka, Yang Katie Zhao, Alfredo Garcia, and Mingyi Hong. 2025. Reinforcing multi-turn reasoning in llm agents via turn-level reward design. *Preprint*, arXiv:2505.11821.

Zhiheng Xi, Jixuan Huang, Chenyang Liao, Baodai Huang, Honglin Guo, Jiaqi Liu, Rui Zheng, Junjie Ye, Jiazheng Zhang, Wenxiang Chen, Wei He, Yiwen Ding, Guanyu Li, Zehui Chen, Zhengyin Du, Xuesong Yao, Yufei Xu, Jiecao Chen, Tao Gui, and 4 others. 2025. Agentgym-rl: Training llm agents for long-horizon decision making through multi-turn reinforcement learning. *Preprint*, arXiv:2509.08755.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Preprint*, arXiv:1809.09600.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, volume 35, pages 20744–20757. Curran Associates, Inc.

Ming Zhang, Kexin Tan, Yueyuan Huang, Yujiong Shen, Chunchun Ma, Li Ju, Xinran Zhang, Yuhui Wang, Wenqing Jing, Jingyi Deng, Huayu Sha, Binze Hu, Jingqi Tong, Changhao Jiang, Yage Geng, Yuankai Ying, Yue Zhang, Zhangyue Yin, Zhiheng Xi, and 4 others. 2026. Opennovelty: An llm-powered agentic system for verifiable scholarly novelty assessment. *Preprint*, arXiv:2601.01576.

Wenlin Zhang, Xiaopeng Li, Yingyi Zhang, Pengyue Jia, Yichao Wang, Huifeng Guo, Yong Liu, and Xiangyu Zhao. 2025a. Deep research: A survey of autonomous research agents. *arXiv preprint arXiv:2508.12752*.

Yaocheng Zhang, Haohuan Huang, Zijun Song, Yuanheng Zhu, Qichao Zhang, Zijie Zhao, and Dongbin Zhao. 2025b. Criticsearch: Fine-grained credit assignment for search agents via a retrospective critic. *Preprint*, arXiv:2511.12159.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. 2026. Retrieval-augmented generation for ai-generated content: A survey. *Data Science and Engineering*, pages 1–29.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. 2024. Sglang: Efficient execution of structured language model programs. In *Advances in Neural Information Processing Systems*, volume 37, pages 62557–62583. Curran Associates, Inc.

Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Zhicheng Dou, Tsung-Yi Ho, and Philip S Yu. 2024. Trustworthiness in retrieval-augmented generation systems: A survey. *arXiv preprint arXiv:2409.10102*.

Dawei Zhu, Xiyu Wei, Guangxiang Zhao, Wenhao Wu, Haosheng Zou, Junfeng Ran, Xun Wang, Lin Sun, Xiangzheng Zhang, and Sujian Li. 2025. Chain-of-thought matters: improving long-context language models with reasoning path supervision. *arXiv preprint arXiv:2502.20790*.

A Analysis of Failure Trajectories and Credit Assignment Ambiguity

In this section, we discuss why round-level credit assignment is unreliable in failed trajectories, complementing the design choice discussed in Section 3.4. The key challenge is that failed trajectories typically lack clearly identifiable low-quality rounds that can be reliably attributed to the final failure. Such ambiguity leads to a high risk of introducing noise when selectively reweighting search rounds. Concretely:

- **Redundant/verification queries**: These rounds look redundant but may serve an implicit verification role (re-checking ambiguous evidence).

Heuristically marking such rounds as low-quality would risk removing legitimate verification behavior.

- **Cascading Errors from Misinterpreted Evidence**: Later retrieval rounds may be misled by earlier misinterpretations of superficially relevant yet non-supportive documents, and therefore should not be regarded as the primary source of failure.
- **Attempts that failed to obtain key information**: When a search round recalls no useful documents, it usually shows no obvious error — the failure is caused by insufficient corpus coverage or inadequate retriever recall, not a clearly "bad" decision the agent made.
- **Premature overconfidence**: This usually occurs at the answer round, not during intermediate search rounds, and thus falls outside the scope of round-level search credit assignment.

A failure pattern that yields a reliably detectable low-quality signal is **similar-entity confusion**. Because most failure patterns are inherently ambiguous at round granularity, treating failed trajectories differently (uniform contribution) avoids reinforcing spurious signals and preserves GRPO's stable outcome-based baseline.

B LLM Judge Calibration Details

To improve transparency, we describe the calibration protocol of our LLM judge and summarize the associated dataset statistics below.

The calibration set contains 97 search rounds sampled from successful GRPO trajectories with diverse trajectory lengths. Each search round is annotated by a human annotator with respect to two criteria: retrieval utility and reasoning correctness. In addition, the annotator provides a confidence score indicating the certainty of the judgment for each labeled round.

To analyze how judgment quality varies with trajectory length, Table 5 reports the distribution of search rounds across different trajectory lengths, distinguishing whether each round simultaneously satisfies both retrieval utility and reasoning correctness. This allows us to examine how the quality of search rounds evolves as trajectories become longer.

To further understand the uncertainty structure of the calibration data, Table 6 breaks down the annotation confidence conditioned on whether a search round satisfies both quality criteria.

Number of Search Rounds	Satisfies Both Criteria	Does not Satisfy
1	5	0
2	17	3
3	18	12
$\geq$4	18	24

Table 5: Quality distribution of search rounds across different trajectory lengths.

Confidence	High	Medium	Low
Satisfies both criteria	55	3	0
Does not satisfy	34	5	0

Table 6: Confidence distribution across quality labels.

Overall, among the 89 high-confidence rounds, the LLM judge agrees with the human annotator on 87 rounds; among the remaining 8 medium-confidence cases, agreement occurs in 6 cases, yielding an overall agreement rate of 95.8%. This demonstrates strong alignment between the LLM judge and human judgments, particularly in high-confidence annotations.

C Computational Cost of LLM Judge

We provide a brief analysis of the computational cost introduced by the LLM judge during training. The judge model GPT-oss-120B has 116.8B parameters and adopts a Mixture-of-Experts (MoE) architecture, where only 5.1B parameters are activated per token. According to official deployment guidelines, the model can be served on a single NVIDIA H100 GPU, making it feasible to integrate into the training pipeline. In practice, incorporating the LLM judge leads to a moderate increase in training cost. Specifically, CW-GRPO training takes 11.7 hours on 5 GPUs (including one GPU dedicated to the judge), compared to 8.8 hours on 4 GPUs for standard GRPO, corresponding to approximately a 33% increase in wall-clock time. Overall, while the LLM judge introduces additional computational cost, we find it manageable in our experimental setup.

The Prompt for RL Training

Your task is to retrieve knowledge from the knowledge base to answer the given question.

You must follow this conversational protocol exactly:
- Reason in <think>...</think>: list steps and immediate plan for the next retrieval(s).
- Issue a single retrieval with the search tool: provide ONE complete natural-language query (a full question) suitable for a dense retriever (FAISS).
- After each retrieval, you MUST wait for external <tool_response>...</tool_response> before continuing. Do not proceed or speculate until information is received.
- When ready to conclude, reply in <answer>...</answer>. Keep your answer as concise as possible (such as <answer>Beijing</answer>).
- Every step must begin with <think> before issuing <tool_call> or <answer>.

Search guidance and KB metadata:
- Queries must be natural-language questions (no keywords, no fragments). e.g. "Who is the author X?" not "author X".
- The KB was last updated in 2018.
- Documents with identical titles should be treated as parts of the same entry and may be combined when explicitly supported by their content. Otherwise treat docs independently.

Granularity and commonsense:
- When the question requires multiple reasoning hops, decompose it into smaller sub-questions and search sequentially for the information needed at each hop.
- You may use common-sense (general knowledge a broadly-educated adult would know) to guide reasoning, formulate searches, or fill obvious gaps — but mark its use in <think> (e.g., "using commonsense: ...").

Entity disambiguation and query rephrasing:
- If the question is ambiguous, infer the likely entity type (person/place/org/work) and include that type as a constraint in your query.
- If <tool_response> lacks the target, retry with rephrasing: use synonyms or near-synonyms, reverse the relation (e.g., "Who founded X" ↔ "Who is the founder of X"), or adjust the information granularity (broaden or narrow the scope). Do not use the same query twice.

Behavioral rules:
- Do not invent facts. Only assert in <answer> what follows from retrieved documents or clearly-marked commonsense used.
- Keep all text outside the tags minimal and task-focused.

Question: {{question}}

Figure 5: The prompt for RL training.

D Prompts

We present the detailed prompts that guide the behavior of our search agent and LLM judge in this section.

Figure 5 shows the input prompt used during model training, defining how the search agent structures its reasoning, interacts with the retrieval tool across rounds, and generates the final answer.

Figure 6 illustrates the prompt used by the LLM judge, which specifies the rubrics for assessing each search round in terms of retrieval utility and reasoning correctness, producing the binary contribution signals used in CW-GRPO.

The Prompt for LLM Judge

You are an assistant that evaluates the validity of the agent's reasoning in a question-answering task.

You will be given a partial trace consisting of alternating messages between the agent and the information source. Your task is to judge **only the latest action** in the trace, based on the agent's latest <think> section and all previously retrieved information.

Evaluation Criteria

Retrieval Reward
For the latest action, give a retrieval_reward of **1** if and only if:

1. **Relevance:**
The retrieved information is genuinely relevant to the main question and is likely to be helpful in answering it.

2. **Novelty:**
The retrieved information should offer new, useful content for answering the question that was not already obtained in previous rounds. If the same information (or its substance) was retrieved in previous rounds, do **not** assign a retrieval_reward, even if it is relevant.

Thinking Reward
For the latest action, give a thinking_reward of **1** if and only if:

1. **Reasoning Support:**
The reasoning in the latest <think> section is logically grounded in the previously retrieved documents. The agent's claims, assumptions, and deductions must be supported by the information already obtained.

2. **Action Usefulness:**
- If the action is a **search**, the proposed retrieval query must aim to obtain missing information that is necessary or beneficial for producing a correct answer.
- If the action is an **answer**, the reasoning must align with and be properly supported by the retrieved information.
- When assigning the thinking reward, you must only focus on the content and quality of the agent's latest <think> section; do NOT consider the relevance of the retrieved documents or whether the final answer matches the ground truth.

If **either** the reasoning is unsupported **or** the action is not helpful toward answering the question, assign a score of **0**.

Additional Notes

- Your judgment must rely **only** on the conversation history and retrieved documents; do not use outside knowledge.
- The knowledge base (KB) used for retrieval was last updated in 2018.
- If multiple documents share the same title, treat them as parts of the same entry. But if the documents have similar but not the same title, treat them as separate and independent.
- The KB may contain documents with names similar to the target entity but that are actually irrelevant.
If the agent incorrectly treats such similar-but-unrelated documents as relevant and bases reasoning on them, you **must** assign 0.
- Be objective and consistent.

Analysis Steps
Before you assign the rewards, you should first analyze the latest action in detail. Please follow these steps:

1. Extract the factual claims, reasoning logic, search intent, and assumptions made in the latest action's <think> section.
2. For each factual claim, check whether it is supported by previously retrieved passages or is a matter of common sense.
3. Analyze whether the reasoning logic is rigorous, and whether the search intent aligns with the reasoning and constitutes information still needed to answer the question. **Attention**:
- If the agent attempts to retrieve information that was previously searched for but not successfully obtained—by rephrasing, using synonyms, or otherwise varying the query—and if that information would be helpful for answering the question, you should consider this retrieval attempt useful.
- If the agent makes an assumption in place of unavailable information, and the assumption is logically justified, do not penalize the agent for this.
4. Extract information from the retrieved documents that may be relevant to the main question. For statements similar to the question, analyze carefully whether they are truly relevant or only superficially similar but unrelated. If no relevant information is present, skip step 5 and assign a retrieval_reward of 0.
5. For each relevant information, check whether it was already retrieved in previous rounds and, if so, in which round. Finally, give the retrieval_reward based on whether genuinely new relevant information was retrieved in this turn.

Please conduct your analysis in the order above and justify your scoring.

Format

Input format: a partial multi-round conversation between the agent and the information source. Example:
```
Question: the question to answer
Agent: <think>...</think><tool_call>the first search tool call</tool_call>
Information: the information retrieved by the first search tool call
Agent: <think>...</think><tool_call>the second search tool call</tool_call>
Information: the information retrieved by the second search tool call
...
Agent (the last action): <think>...</think><tool_call|answer>...</tool_call|answer>
(Information: the information retrieved by the last search tool call)
```
Your should **only** evaluate the **last** action.

Return format: a JSON object, wrapped in ```json and ```. Example:
```json
{"analysis": "your detailed analysis of the latest action", "thinking_reward": 0/1, "retrieval_reward": 0/1}
```

Figure 6: The prompt for LLM Judge to specify the rubric for assessing each search round in terms of retrieval utility and reasoning correctness.