BAPO: Boundary-Aware Policy Optimization for Reliable Agentic Search

Shiyu Liu[1,2,†,‡], **Yongjing Yin**[2,‡], **Jianhao Yan**[4], **Yunbo Tang**[3], **Qinggang Zhang**[5*]
Bei Li[2], **Xin Chen**[2], **Jingang Wang**[2], **Xunliang Cai**[2], **Jinsong Su**[1,3,6*]
[1]Institute of Artificial Intelligence, Xiamen University [2]Meituan Inc.
[3]School of Informatics, Xiamen University [4]Westlake University [5]Jilin University
[6]Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan (Xiamen University), Ministry of Culture and Tourism, China
shiyu.liu0709@foxmail.com; qinggangzhang@jlu.edu.cn; jssu@xmu.edu.cn

Abstract

RL-based agentic search enables LLMs to solve complex questions via dynamic planning and external search. While this approach significantly enhances accuracy with agent policies optimized via large-scale reinforcement learning, we identify a critical gap in reliability: these agents fail to recognize their reasoning boundaries and rarely admit "*I DON'T KNOW*" (*IDK*) even when evidence is insufficient or reasoning reaches its limit. The lack of reliability often leads to plausible but unreliable answers, introducing significant risks in many real-world scenarios. To this end, we propose **B**oundary-**A**ware **P**olicy **O**ptimization **(BAPO)**, a novel RL framework designed to cultivate reliable boundary awareness without compromising accuracy. BAPO introduces two key components: (i) a group-based boundary-aware reward that encourages an *IDK* response only when the reasoning reaches its limit, and (ii) an adaptive reward modulator that strategically suspends this reward during early exploration, preventing the model from exploiting *IDK* as a shortcut. Extensive experiments on four benchmarks demonstrate that BAPO substantially enhances the overall reliability of agentic search[1].

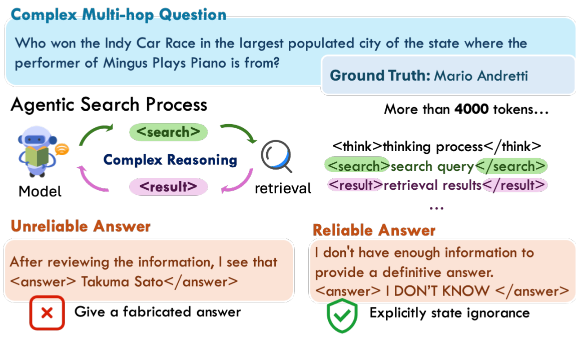

Figure 1: When the agentic search model produces wrong answers, its lengthy and complex reasoning makes it difficult for users to verify. To ensure reliability, the model should explicitly state when information is insufficient and that no answer is available.

1 Introduction

Recent advances leverage reinforcement learning (RL) to optimize Large Language Models (LLMs) as autonomous agents that actively plan and execute multi-turn searches (Jin et al., 2025a; Chen et al., 2025; Song et al., 2025). While the RL-based agentic search substantially improves accuracy on complex, knowledge-intensive questions, it introduces a critical reliability issue: these RL-based models almost never admit *"I DON'T KNOW"* (*IDK*) even when evidence is insufficient or reasoning reaches its limit (Zhang et al., 2025c; Joren et al., 2025). As presented in Figure 1, when an agentic search model cannot reach a concrete final answer, it should explicitly admit *IDK* rather than fabricating answers. This capability is especially valuable given the difficulty for users to verify the lengthy reasoning process, as honest acknowledgment of *IDK* helps users seek alternative information sources and prevents misinformation.

Previous research attributes the ability to acknowledge uncertainty to a model's awareness of its internal "boundary" (Li et al., 2025b; Zhang et al., 2025c). Specifically, EKBM (Zheng et al., 2025) verifies information against its parametric "knowledge boundary" to answer factual questions accurately, while CapBound (Zhang et al., 2025c) introduces a similar concept to determine whether a problem is solvable within its "capability boundary" for mathematical reasoning. However, the reasoning boundary for an agentic search model is inherently more complex and dynamic, as such a boundary is continuously shaped by the interplay between the model's internal reasoning capacity for search tool interaction

*Corresponding author.
‡Equal contribution.
†Work done during internship at Meituan Inc.
[1]Our code is available at https://github.com/Liushiyu-0709/BAPO-Reliable-Search.

Findings of the Association for Computational Linguistics: ACL 2026, pages 36408–36426
July 2-7, 2026

and the external retrieved information.

To investigate the impact of RL on an agentic search model's awareness of its reasoning boundaries, we conduct preliminary experiments measuring this awareness before and after RL training. A key finding is that while RL optimization improves accuracy on complex questions, it significantly degrades the model's awareness of its boundary. This occurs because standard RL rewards incentivize exhaustive exploration for correctness while simultaneously discouraging any acknowledgment of uncertainty (Jin et al., 2025a; Song et al., 2025). As a result, these models often lack awareness of their boundaries and fabricate plausible answers.

Enhancing the boundary awareness of agentic search models is crucial but challenging. (i) It is difficult to quantitatively measure a model's reasoning boundary, which makes it hard to construct a reliable reward signal for *IDK* responses. Unlike static knowledge or capability boundaries, which are tied solely to the model's inherent competencies, the boundary of an agentic search model is a dynamic, emergent property of the interaction between the agent's planning, the quality of its retrievals, and its iterative reasoning steps. Consequently, we lack a robust strategy to measure whether the agent has genuinely encountered its limit. (ii) It is challenging to integrate such a signal into the RL objective without creating unintended incentives. A naive reward that encourages *IDK* responses could be exploited as a low-effort shortcut, especially on difficult problems requiring deep exploration. If the reward function prioritizes *IDK* over incorrect answers, the model may learn to default to *IDK* prematurely, ultimately harming the reasoning accuracy and exploration depth.

To this end, we propose **B**oundary-**A**ware **P**olicy **O**ptimization **(BAPO)** for agentic search, designed to cultivate reliable self-awareness without compromising accuracy. Our research aims to address two fundamental research questions: (i) how to precisely model the learning signal for identifying the boundary of agentic search models, and (ii) how to integrate this signal into an RL framework to balance deep exploration and appropriate exploitation. Results across four challenging benchmarks demonstrate that BAPO consistently outperforms mainstream training-based and prompt-based methods in overall reliability. Remarkably, with only 5000 RL training samples, BAPO with Qwen2.5-7B-Instruct shows superior reliability compared to strong open-source agentic search models, without compromising average accuracy.

Our major contributions are listed as follows:

- We identify the key limitation of agentic search models and propose BAPO, a novel RL framework that dynamically rewards *IDK* responses to cultivate the model's self-awareness of its reasoning boundaries.
- BAPO introduces a group-based boundary-aware reward that encourages an *IDK* response only when the question is out of the model's boundary, and an adaptive reward modulator to prevent reward hacking.
- Extensive experiments on challenging benchmarks demonstrate that BAPO effectively enhances search reliability while preserving the model's ability to solve complex problems.

2 Background

2.1 Agentic Search

To enhance response quality in knowledge-intensive scenarios, Retrieval-Augmented Generation (RAG) supplements LLMs' parametric knowledge with external search (Lewis et al., 2020; Gao et al., 2023; Fan et al., 2024). Current prompting-based strategies often depend on predefined workflows for interleaving reasoning and search (Asai et al., 2024; Li et al., 2025c), resulting in limited effectiveness. While learning-based methods (SFT/DPO) offer an alternative (Lin et al., 2024; Wang et al., 2025; Fang et al., 2025; Li et al., 2025d), they remain constrained by complex data pipelines and suboptimal generalization.

Inspired by the success of reinforcement learning on mathematics and coding (Jaech et al., 2024; DeepSeek-AI et al., 2025), RL-based agentic search systems have emerged, such as Search-R1 (Jin et al., 2025a), ReSearch (Chen et al., 2025), R1-Searcher (Song et al., 2025), and Tool-Star (Dong et al., 2025). These methods employ format and outcome correctness rewards to enable autonomous query decomposition and flexible interaction with external search systems during reasoning, thereby achieving superior accuracy on complex queries. Nevertheless, current methods prioritize search accuracy while ignoring holistic reliability, which depends on both accuracy and precision. In contrast, our proposed BAPO leverages boundary-aware reward mechanisms to optimize overall reliability, thereby maximizing the practicality of agentic search models.

2.2 Formalization

To make it clear, we formally define Agentic Search as a sequential decision-making process following the ReAct paradigm (Yao et al., 2023). Given a query x, the agent aims to generate a trajectory τ that concludes with a final answer y. To structure this reasoning process, a system prompt is included to constrain the model's output format. For instance, the model is required to use tags <think>, <search>, and <answer> to indicate thoughts, search tool calls, and the predicted answers, respectively, with search results returned in <result>. The prompt template is detailed in Appendix C.1. As formalized below, the trajectory τ is a sequence of interleaved reasoning steps, actions, and observations:

$$\tau = \big((r_t, a_t, o_t)_{t=1}^{T-1}, r_T, y\big), \tag{1}$$

where each r_t denotes the reasoning state at step t that plans subsequent actions and r_T synthesizes accumulated information to conclude to the final answer y; a_t is the search action; and o_t is the returned search results.

As RL has demonstrated significant efficacy in enhancing LLM reasoning capabilities, Group Relative Policy Optimization (GRPO) (Shao et al., 2024) has been extensively employed in training agentic search models (Jin et al., 2025a; Chen et al., 2025; Dong et al., 2025). For a query x, a group of G trajectories $\{\tau_1, \tau_2, \ldots, \tau_G\}$ is sampled from the policy, and the objective function is defined as:

$$\mathcal{J}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{\tau_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \frac{1}{G}\sum_{i=1}^{G} \tag{2}$$
$$\left[\min\left(w_i(\theta)A_i, \text{clip}\left(w_i(\theta), 1-\epsilon, 1+\epsilon\right)A_i\right)\right],$$

where $w_i(\theta) = \frac{\pi_\theta(\tau_i|x)}{\pi_{\theta_{\text{old}}}(\tau_i|x)}$ and the KL term against a reference model is omitted. The advantage A_i is computed by normalizing the rewards within the group, $A_i = \frac{\mathcal{R}(\tau_i) - \text{mean}(\{\mathcal{R}(\tau_j)\}_{j=1}^{G})}{\text{std}(\{\mathcal{R}(\tau_j)\}_{j=1}^{G})}$.

Rule-based rewards have demonstrated robust empirical performance and are widely adopted in current research (Song et al., 2025; Chen et al., 2025). Typical reward design focuses on correctness and consists of two parts: (i) a format correctness reward that checks whether the output conforms to the prescribed structure across reasoning steps, tool calls, and answers; and (ii) an outcome correctness reward that measures the objective correctness of the final answer.

$$\mathcal{R}^{Correct} = \begin{cases} F1(\hat{y}, y), & \textit{correct format} \\ -1, & \textit{wrong format}, \end{cases} \tag{3}$$

where F1 means character-level F1 score between two strings, y is the prediction answer extracted from trajectory and $\hat{y}$ is the ground truth answer.

3 Preliminary Study

In this study, we first investigate the impact of RL training with correctness-based reward on the boundary awareness of agentic search models. Subsequently, we incorporate an additional reward term designed to encourage *IDK* responses and evaluate its effect. Our findings reveal an inherent trade-off between maximizing accuracy and maintaining effective boundary awareness.

3.1 Setup

Training. For training, we use a high-quality RL training dataset from Dong et al. (2025), comprising 5000 multi-hop QA samples from HotpotQA and 2WikiMultihopQA. Following the cold-start-free paradigm, we directlt apply GRPO to Qwen2.5-3B-Instruct (Yang et al., 2025b), bypassing preliminary in-domain SFT.

Evaluation. For evaluation, we use four multi-hop QA benchmarks: HotpotQA (Yang et al., 2018), MuSiQue (Trivedi et al., 2022), 2WikiMultiHopQA (Ho et al., 2020), and Bamboogle (Press et al., 2023). Specifically, HotpotQA, 2WikiMultiHopQA and MuSiQue are constructed from Wikipedia or Wikidata (Vrandečić and Krötzsch, 2014) using diverse multi-hop mining strategies, while Bamboogle is a manually curated dataset of 2-hop challenging questions. Following the setup in Dong et al. (2025), we construct a validation set of 80 multi-hop QA pairs, comprising 20 examples from the test set of each benchmark. In contrast, the results reported in Section 5 are evaluated on the complete test sets.

Metrics. To comprehensively evaluate model reliability, we follow Xu et al. (2024) to use three following metrics: accuracy (acc), precision (prec) and *IDK* rate (ρ_{IDK}). These metrics are defined as follows:

$$\text{acc} = \frac{N_c}{N}, \text{prec} = \frac{N_c}{N - N_r}, \rho_{IDK} = \frac{N_r}{N}, \tag{4}$$

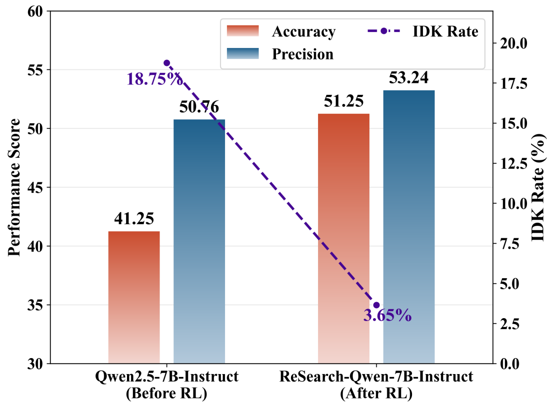

Figure 2: Evaluation results of accuracy, precision and *IDK* rate (ρ_{IDK}) of models before and after RL. The sharp drop in ρ_{IDK} coupled with the narrowing gap between accuracy and precision, indicates a diminished boundary awareness after RL.

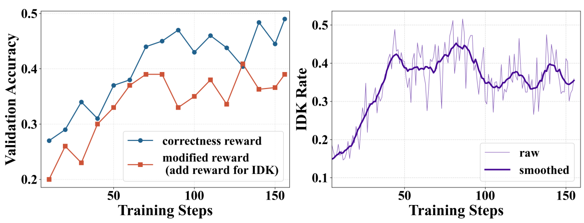

Figure 3: **Left:** Validation accuracy under different reward settings during RL training. **Right:** *IDK* rate (ρ_{IDK}) under the modified reward during RL training.

where $N = N_c + N_w + N_r$ is the total number of answers, with N_c, N_w, N_r denoting the number of correct answers, wrong answers and *IDK* answers, respectively. Following previous studies (Chen et al., 2025; Dong et al., 2025), we extract answers from the model output enclosed in \box{}, and then use GPT-4 as a judge to assess the correctness (Appendix C.2).

3.2 RL with Correctness Reward Impairs Boundary Awareness

To evaluate the impact of RL on boundary awareness, we compare Qwen2.5-7B-Instruct with its variant, ReSearch-Qwen-7B-Instruct (Chen et al., 2025), which is optimized via GRPO using a correctness-based reward. To elicit explicit *IDK* admissions, we follow Xu et al. (2024) to append the following reliable prompt to the original one:

Reliable Prompt

```
If you can't solve this question by
the reasoning process, you should
output \boxed{I DON'T KNOW}.
```

As illustrated in Figure 2, before RL training, Qwen2.5-7B-Instruct shows a notable precision advantage over its accuracy (50.76 vs. 41.25), accompanied by a substantial *IDK* rate of 18.75%, effectively filtering uncertain cases. However, after RL training (ReSearch-Qwen-7B-Instruct), while accuracy improves to 51.25, *IDK* rate drops sharply to 3.65%, and precision increases only marginally to 53.24. This indicates a weakened boundary awareness that suppresses *IDK* admission.

3.3 Encouraging *IDK* Responses Hinders Accuracy Advancement

Since correctness-based rewards alone weaken the model's awareness of when to acknowledge IDK, a natural approach is to incorporate appropriate positive rewards for *IDK* responses during RL training. Following the training setting in Section 3.1, we instruct the model with reliable prompt and modify the correctness reward function by assigning an additional reward of 0.5 to *IDK* responses within rollout groups that lacks correct answers. This design prioritizes *IDK* responses over incorrect answers when a correct answer is unavailable.

However, as shown in Figure 3, directly incentivizing *IDK* responses leads to reward hacking: the model learns to maximize rewards by defaulting to *IDK* rather than attempting to solve challenging problems. Consequently, this modified reward impedes accuracy gains compared to the vanilla reward. This finding underscores a fundamental challenge in training reliable models: balancing the model's exploration for correctness rewards against its exploitation for *IDK* rewards.

4 The Framework of BAPO

With observations and insights from the preliminary study, we propose Boundary-Aware Policy Optimization (BAPO), an RL algorithm built upon GRPO for training reliable agentic search models. As shown in Figure 4, BAPO uses boundary-aware reward to encourage *IDK* responses according to whether a problem is out of model's boundary. More importantly, it uses an adaptive reward modulator to balance exploration and exploitation, mitigating the reward hacking issue identified in Section 3.3.

4.1 Boundary-Aware Reward

We regard that a question exceeds the model's boundary if the model fails to get any correct

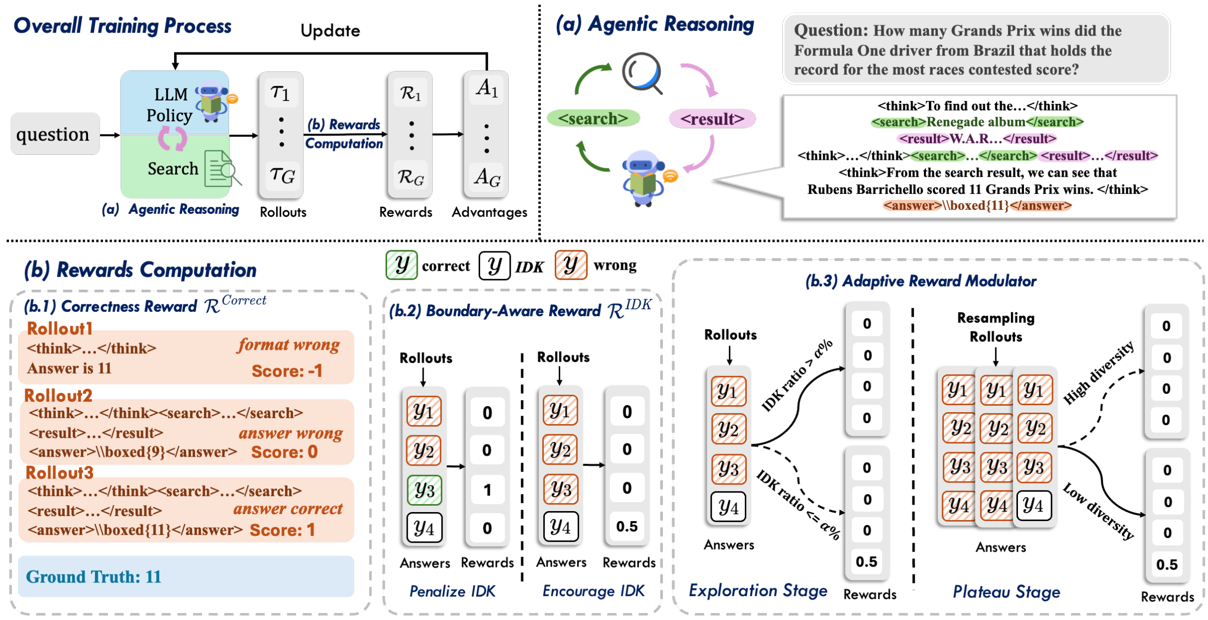

Figure 4: The overall framework of BAPO. Its training process begins with **(a) agentic reasoning**, where the model generates a group of rollouts for each question by interleaving thought processes with search environment interactions. These rollouts are then passed to the **(b) reward computation** module, which is comprised of **(b.1) correctness reward** $\mathcal{R}^{Correct}$ derived from the correctness of format and outcome, **(b.2) boundary-aware reward** $\mathcal{R}^{IDK}$ designed to incentivize *IDK* responses when no correct rollout exists within the group, and **(b.3) adaptive reward modulator** adaptively disabling $\mathcal{R}^{IDK}$ based on *IDK* ratio at the exploration stage and diversity of rollouts at the plateau stage.

answer among multiple rollouts. In this situation, we adjust the reward assignment for *IDK* responses. Formally, for a set of rollout outputs $\{\tau_i\}_{i=1}^G$ and their corresponding correctness rewards $\{\mathcal{R}_i^{Correct}\}_{i=1}^G$, the boundary-aware reward is calculated as:

$$\mathcal{R}^{IDK} = 0.5 \cdot \mathbb{I}(y_i = IDK), \quad \text{if } \forall \{\mathcal{R}^{Correct}(\tau_i)\}_{i=1}^G \leq 0. \tag{5}$$

The final reward $\mathcal{R}$ is the sum of $\mathcal{R}^{Correct}$ and $\mathcal{R}^{IDK}$:

$$\mathcal{R} = \mathcal{R}^{IDK} + \mathcal{R}^{Correct}. \tag{6}$$

4.2 Adaptive Reward Modulator

Our preliminary analysis suggests that purely incentivizing *IDK* responses leads to a degradation in exploration. To balance exploration and boundary awareness during RL training, we introduce an adaptive reward modulator for $\mathcal{R}^{IDK}$, which functions at two distinct levels:

Stage Level. In the early stages of training, the model needs to conduct extensive exploration to learn how to solve tasks. During this **exploration stage**, we avoid introducing additional rewards to interfere with the model's learning. $\mathcal{R}^{IDK}$ is deactivated by default and only activated when the proportion of *IDK* responses $\mathcal{R}^{IDK}$ falls below α. As training progresses, the task accuracy gradually stabilizes. At this **plateau stage**, we apply the reward $\mathcal{R}^{IDK}$ fully to make the model start paying more attention to its boundary awareness. Some difficult queries may not be easily answered correctly or admitted as unknown under the default group size, and thus we dynamically resample groups with no correct rollouts up to k times, until the model outputs *IDK* or provides a correct answer. The transition between the exploration and plateau stages is determined by the stagnation of the validation score, e.g., when the validation score plateaus for 5 consecutive steps.

Sample Level. The consistency of a model's generated responses can serve as a proxy for its confidence. To further preserve the model's exploration during the plateau stage, we adaptively modulate $\mathcal{R}^{IDK}$ based on the rollout diversity of each query. Specifically, for queries exhibiting high rollout diversity, indicating that the model is actively exploring the solution space, we deactivate $\mathcal{R}^{IDK}$ to prevent untimely convergence. In contrast, for samples with low rollout diversity, which signifies that the model has converged on a specific output, $\mathcal{R}^{IDK}$ is applied to refine its boundary awareness. Formally, a sample is considered to have high rollout diversity if its corresponding rollout answers $|\{y_{1\dots G}\}| \geq \frac{G}{2}$; otherwise, it is categorized as having low diversity.

5 Experiments

5.1 Experimental Setups

Baselines. We compare BAPO against (1) existing RL agentic search methods, and (2) our implementations of prompting and training-based methods. For the former, we select two representative approaches: Search-R1 (Jin et al., 2025a) and ReSearch (Chen et al., 2025). We use Qwen2.5-7B-Instruct as the backbone for a fair comparison. The prompt-based baselines include: (1) Naive RAG: A naive retrieval-based approach that directly concatenates retrieved context with the query. (2) IRCoT (Trivedi et al., 2023): An interleaving method where retrieval and CoT mutually guide each other. (3) TIR Prompt (Dong et al., 2025): a standard tool-integrated prompt, which instructs the model to use search tools during reasoning (Appendix C). (4) Reliable TIR Prompt: Building on TIR Prompt, it additionally encourages the model to acknowledge ignorance by responding with *IDK* (Section 3.2). For training-based baselines, we choose (1) Reliable Rejection Sampling Fine-Tuning (Reliable RFT), which generates multiple reasoning paths for each question under Reliable TIR Prompt, and incorporates both correct paths and reasonable *IDK* responses into the SFT dataset. Similar to BAPO, an *IDK* response is adopted when no correct path is available. (2) GRPO (Shao et al., 2024), which uses only the correctness reward, with other settings matching those of BAPO.

Reliability Metric. We follow Xu et al. (2024) to adopt a reliability metric to comprehensively balance accuracy and precision:

$$\text{reliability} = (1 - \rho_{IDK}) \cdot \text{prec} + \rho_{IDK} \cdot \text{acc}, \quad (7)$$

where ρ_{IDK} denotes the *IDK* rate. This metric functions as a dynamic interpolation between precision and accuracy based on the model's tendency to refuse. Specifically, when the agent rarely refuses to give an answer (low ρ_{IDK}), the metric prioritizes precision, enforcing strict correctness on generated answers. Conversely, as the agent tends to refuse more frequently (high ρ_{IDK}), the metric shifts focus toward standard accuracy . This mechanism effecively penalizes excessive "lazy" *IDK*.

Implementation Details. Regarding BAPO-specific hyper-parameters, we set the *IDK* ratio threshold α as 5%, and the resampling times k is set to 2. The retrieval environment is based on FlashRAG (Jin et al., 2025b). We use E5-base-v2 (Wang et al., 2024) as the retriever and Wikipedia data (Vrandečić and Krötzsch, 2014) from December 2018 as the knowledge base. During training and evaluation, we retrieve top-5 results for each query. We train the GRPO and BAPO models with a batch size of 64 for 2 epochs. The rollout size is set to 8, and the maximum tokens are set to 8192. The maximum number of tool invocations during training and inference is set to 3. Sensitivity analysis of the hyper-parameters are present in Appendix B.1.

5.2 Main Results

The performance of BAPO and other baselines conducted on Qwen2.5-7B-Instruct is presented in Table 1. Compared to all methods from both existing agentic search and our implemented baselines, BAPO achieves significant reliability improvements across all benchmarks (+15.8 scores in average). Notably, compared to existing agentic search models such as Search-R1 (Jin et al., 2025a) and ReSearch (Chen et al., 2025), which utilize large-scale training sets of 90k and 19k samples respectively, BAPO-trained model with **only 5k samples** achieves competitive accuracy and substantially higher reliability.

When compared to GRPO, BAPO achieves average improvements in reliability of 9.7%, with corresponding precision improvements of 11.8%, while incurring only marginal accuracy decreases of 2.2%. These results showcase BAPO's effectiveness and efficiency in training reliable agents.

Regarding methods with reliable techniques, such as Reliable RFT and Reliable TIR Prompt, we find that Reliable RFT tends to be over-conservative; although they achieve significant gains in precision, this come at the cost of a drastic decline in accuracy, suffering a 27-point accuracy drop compared to the TIR Prompt baseline. This ultimately undermines overall reliability. While Reliable TIR Prompt enhances precision without sacrificing accuracy, it inherently lacks the capacity to further improve problem-solving capability. On the contrary, BAPO achieves both boundary awareness and problem-solving capability.

5.3 Generalization Across Model Scales

To verify BAPO's generalization across different model scales, we extend our evaluation to the 3B

Method	HotpotQA			MusiQue			2Wiki.			Bamboogle		
	Acc	Prec	Rel.	Acc	Prec	Rel.	Acc	Prec	Rel.	Acc	Prec	Rel.
Existing Agentic Search Methods												
Search-R1	49.0	49.0	49.0	22.5	22.5	22.5	39.0	39.0	39.0	52.0	52.0	52.0
ReSearch	**61.5**	61.5	61.5	**31.0**	31.0	31.0	54.2	54.2	54.2	54.4	54.4	54.4
Our Implementations												
Naive RAG	49.6	49.6	49.6	12.7	12.7	12.7	29.5	29.5	29.5	32.0	32.0	32.0
IRCoT	52.1	52.1	52.1	14.2	14.2	14.2	30.6	30.6	30.6	36.8	36.8	36.8
TIR Prompt	51.5	51.5	51.5	21.5	21.5	21.5	43.0	43.0	43.0	48.8	48.8	48.8
Reliable TIR Prompt	52.5	62.1	60.6	21.0	30.0	27.2	35.5	45.5	43.3	47.2	50.8	50.5
Reliable RFT	24.5	**68.0**	40.2	11.0	36.6	18.5	14.5	48.2	23.9	36.8	56.0	49.4
GRPO	60.0	60.0	60.0	29.5	29.5	29.5	59.5	59.5	59.5	**57.6**	57.6	57.6
BAPO (Ours)	58.0	66.6	**65.5**	29.5	**38.8**	**36.6**	57.0	**64.1**	**63.3**	**57.6**	**61.5**	**61.2**

Table 1: Performance on QA tasks with Qwen2.5-7B-Instruct as the backbones. Rel. indicates the reliability metric.

Method	Acc	Prec	Rel.
Qwen2.5-3B-Instruct			
TIR Prompt	29.2	29.2	29.2
Reliable TIR Prompt	25.4	31.8	30.2
GRPO	**45.1**	45.1	45.1
BAPO (Ours)	44.8	**52.9**	**51.3**
Qwen2.5-14B-Instruct			
TIR Prompt	49.7	49.7	49.7
Reliable TIR Prompt	47.6	56.7	55.0
GRPO	**56.6**	56.6	56.6
BAPO (Ours)	54.0	**65.7**	**63.3**

Table 2: Performance of BAPO and baselines across different model scales. Rel. indicates the reliability metric.

Acc	Prec	ρ_{IDK}	Reliability
BAPO			
44.8	52.8	16.8%	**51.3**
w/o Boundary-Aware Reward			
30.6	**62.4**	53.1%	44.8
w/o Sample Modulator			
43.3	52.0	20.4%	50.1
w/o Sample & Stage Modulator			
37.8	56.0	35.2%	49.0

Table 3: Ablation study. The metrics are averaged across four benchmarks on Qwen2.5-3B-Instruct.

and 14B versions of the Qwen2.5-Instruct series. As shown in Table 2, BAPO improves the reliability of its instruct model backbones with Reliable TIR Prompt by an average of 76.1% and 27.4% on the 3B and 14B scales, respectively. Furthermore, compared to the standard GRPO baseline, BAPO achieves substantial reliability advantages of 13.9% and 11.9%. These results are consistent with our main experiments, confirming that BAPO is generalizable for building reliable agentic search.

5.4 Ablation Study

Table 3 details the ablation study on Qwen2.5-3B-Instruct comparing BAPO with its variants. Through the experiments results, we can draw the following observations:

Fixed Reward Strategy Leads to Reward Hacking. When replacing our proposed reward mechanism with a fixed positive reward of 0.5 for *IDK* responses, which simulates the setting used in BARREL (Yang et al., 2025a), the model exhibits extreme over-conservatism. The *IDK* rate spikes to 53.1%; while this boosts precision, it drastically reduces accuracy and yields the lowest overall reliability score among all variants.

Adaptive Reward Modulator is Critical. Next, we ablate the two components of the Adaptive Reward Modulator to verify their effectiveness. We find that removing sample-level deactivation results in a mild degradation in reliability (-1.2). In contrast, removing both the stage-level and sample-level modulators causes the model to degenerate into a local optimum where it exploits *IDK* responses. Consequently, we observe an excessive surge in the *IDK* rate ($16.8\% \rightarrow 35.2\%$), accompanied by significant drops in accuracy ($44.8 \rightarrow 37.8$) and reliability ($51.3 \rightarrow 49.0$).

5.5 Reward Dynamics of Two Stage

In Figure 5, we visualize BAPO's training dynamics of the average correctness reward $R^{Correct}$ and the average boundary-aware reward R^{IDK}, along with the IDK ratio ρ_{IDK} across different stages on Qwen2.5-14B-Instruct. At the exploration stage (the first 60 steps), $R^{Correct}$ increases from 0.3 to

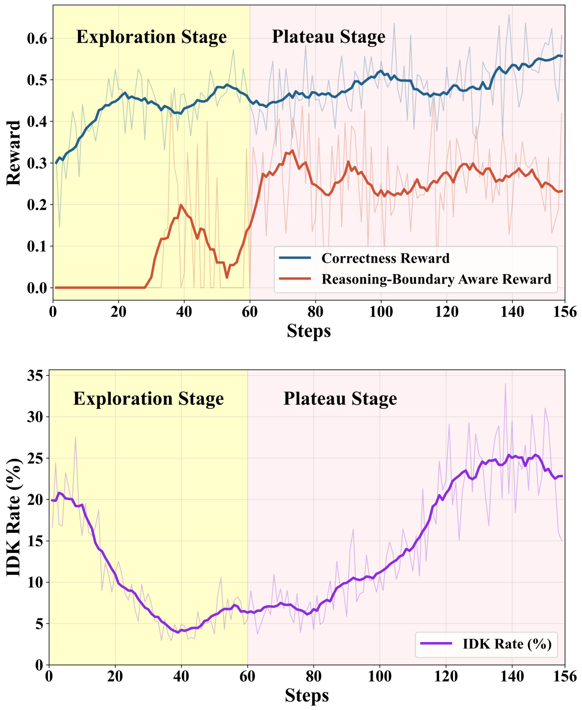

Figure 5: **Upper:** Dynamics of correctness reward $\mathcal{R}^{Correct}$ and boundary-aware reward $\mathcal{R}^{IDK}$. **Bottom:** Dynamics of *IDK* ratio ρ_{IDK} during BAPO training on Qwen2.5-14B-Instruct.

0.5 at the first 60 steps, with ρ_{IDK} drops from the initial 20% to 5%, indicating that the model prioritizes acquiring problem-solving skills over boundary awareness. In the plateau stage, the growth of $R^{Correct}$ slows down, whereas R^{IDK} rises quickly and stabilizes between 0.25 and 0.3. This shift drives a sustained increase in the *IDK* rate, eventually restoring ρ_{IDK} to over 25%.

These dynamics revel that our adaptive reward mechanism effectively varies the main optimization goals across training stages, thereby preventing $R^{Correct}$ and R^{IDK} from hampering each other. It achieves a superior balance between solving problems and refusing unknown questions.

5.6 Awareness of When to Refuse

To assess the rationality of *IDK* responses produced by BAPO, we use **rejection success rate**, defined as the error rate of the GRPO-trained model on the subset of problems where the BAPO-trained model refuses to give a final answer by responding with *IDK*:

$$\text{rejection success rate} = \frac{|S_{\text{refuse}}^{\text{BAPO}} \cap S_{\text{fail}}^{\text{GRPO}}|}{|S_{\text{refuse}}^{\text{BAPO}}|}, \quad (8)$$

where $S_{\text{refuse}}^{\text{BAPO}} = \{x \in \mathcal{D} \mid \mathcal{M}_{\text{BAPO}}(x) = \mathit{IDK}\}$ and $S_{\text{fail}}^{\text{GRPO}} = \{x \in \mathcal{D} \mid \mathcal{M}_{\text{GRPO}}(x) \neq \hat{y}\}$. To

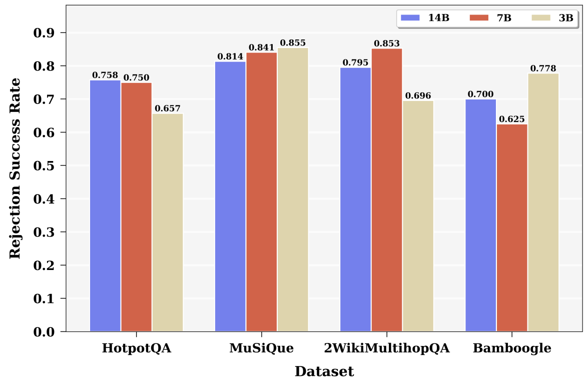

Figure 6: Rejection success rates calculated on Qwen2.5-Instruct series models.

mitigate randomness, we adopt the results of best-of-N (with N=5) as the criterion for determining failures and refusals.

Given that the GRPO-trained model acts as the upper bound of problem-solving capability under our settings, if a problem triggering a refusal is also unsolvable by GRPO, this suggests that the BAPO-trained model's refusal reflects a precise awareness of its reasoning boundaries. Figure 6 shows that the BAPO-trained models attain robust average rejection success rates of 74.7% (3B), 76.7% (7B), and 76.7% (14B). This validates BAPO's ability to enhance model self-awareness, ensuring the appropriate rejection of queries lying outside their reasoning scope. Besides that, it is worth noting that the rejection success rate does not grow with model size, primarily since the overall error rate of the GRPO baselines simultaneously decreases as the model size scales.

6 Conclusion

In this paper, we reveal a critical challenge that RL training with only correctness rewards undermines the model's reliability in providing truthful answers to users. To this end, we propose BAPO, a novel RL algorithm for training reliable agentic search models. Beyond correctness rewards, BAPO incorporates boundary-aware rewards to encourage appropriate *IDK* responses. To tackle the tradeoff between exploration and exploitation, we introduce an adaptive reward modulator to prevent the model from being over-encouraged to admit ignorance. Extensive experiments show that BAPO achieves superior reliability, enabling the model to maintain awareness of its reasoning boundaries without compromising its capability for deep exploration.

Limitations

Despite the promising results of BAPO in enhancing agentic search model's reliability, our work has several limitations. First, our evaluation primarily focuses on knowledge-intensive tasks. Consequently, the study leaves the generalizability of our method to other types of reasoning problems to be fully explored. Furthermore, constrained by computational resources, our experiments currently scale only up to models with 14B parameters. It remains to be seen how the proposed method performs on larger-scale LLMs. Finally, due to the high costs of commercial search APIs, we only consider a local RAG setup. While this ensures reproducibility, it does not fully replicate the noise, and dynamic nature of web search, which may present additional challenges for reliability.

Acknowledgements

The project was supported by National Key R&D Program of China (No. 2022ZD0160501), Natural Science Foundation of Fujian Province of China (No. 2024J011001), and the Public Technology Service Platform Project of Xiamen (No.3502Z20231043). We also thank the reviewers for their insightful comments.

References

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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Jiuhai Chen and Jonas Mueller. 2024. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 5186–5200. Association for Computational Linguistics.

Lida Chen, Zujie Liang, Xintao Wang, Jiaqing Liang, Yanghua Xiao, Feng Wei, Jinglei Chen, Zhenghong Hao, Bing Han, and Wei Wang. 2024. Teaching large language models to express knowledge boundary from their own signals. *CoRR*, abs/2406.10881.

Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. 2025. Research: Learning to reason with search for llms via reinforcement learning.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Guanting Dong, Yifei Chen, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Yutao Zhu, Hangyu Mao, Guorui Zhou, Zhicheng Dou, and Ji-Rong Wen. 2025. Tool-star: Empowering llm-brained multi-tool reasoner via reinforcement learning. *CoRR*, abs/2505.16410.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on RAG meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 6491–6501. ACM.

Tianqing Fang, Zhisong Zhang, Xiaoyang Wang, Rui Wang, Can Qin, Yuxuan Wan, Jun-Yu Ma, Ce Zhang, Jiaqi Chen, Xiyun Li, Hongming Zhang, Haitao Mi, and Dong Yu. 2025. Cognitive kernel-pro: A framework for deep research agents and agent foundation models training. *Preprint*, arXiv:2508.00414.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997.

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.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, and 80 others. 2024. Openai o1 system card. *CoRR*, abs/2412.16720.

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

Jiajie Jin, Yutao Zhu, Zhicheng Dou, Guanting Dong, Xinyu Yang, Chenghao Zhang, Tong Zhao, Zhao Yang, and Ji-Rong Wen. 2025b. Flashrag: A modular toolkit for efficient retrieval-augmented generation research. In *Companion Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025 - 2 May 2025*, pages 737–740. ACM.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. 2025. Sufficient context: A new lens on retrieval augmented generation systems. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *Preprint*, arXiv:2302.09664.

Abhishek Kumar, Robert Morabito, Sanzhar Umbet, Jad Kabbara, and Ali Emami. 2024. Confidence under the hood: An investigation into the confidence-probability alignment in large language models. *Preprint*, arXiv:2405.16282.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jiatong Li, Weida Wang, Qinggang Zhang, Junxian Li, Di Zhang, Changmeng Zheng, Shufei Zhang, Xiaoyong Wei, and Qing Li. 2025a. Mol-r1: Towards explicit long-cot reasoning in molecule discovery. *arXiv preprint arXiv:2508.08401*.

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, and Yang Deng. 2025b. Knowledge boundary of large language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 5131–5157. Association for Computational Linguistics.

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

Xinze Li, Sen Mei, Zhenghao Liu, Yukun Yan, Shuo Wang, Shi Yu, Zheni Zeng, Hao Chen, Ge Yu, Zhiyuan Liu, Maosong Sun, and Chenyan Xiong. 2025d. RAG-DDR: optimizing retrieval-augmented generation using differentiable data rewards. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Xuefeng Li, Haoyang Zou, and Pengfei Liu. 2025e. Torl: Scaling tool-integrated RL. *CoRR*, abs/2503.23383.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. RA-DIT: retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Shichao Ma, Zhiyuan Ma, Ming Yang, Xiaofan Li, Xing Wu, Jintao Du, Yu Cheng, Weiqiang Wang, Qiliang Liu, Zhengyang Zhou, and Yang Wang. 2026. Tspo: Breaking the double homogenization dilemma in multi-turn search policy optimization. *Preprint*, arXiv:2601.22776.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9004–9017. Association for Computational Linguistics.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. *Preprint*, arXiv:2210.03350.

Baochang Ren, Shuofei Qiao, Wenhao Yu, Huajun Chen, and Ningyu Zhang. 2025. Knowrl: Exploring knowledgeable reinforcement learning for factuality. *arXiv preprint arXiv:2506.19807*.

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

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. *CoRR*, abs/2503.05592.

Rui Sun, Jie Ding, Chenghua Gong, Tianjun Gu, Yihang Jiang, Juyuan Zhang, Liming Pan, and Linyuan Lü. 2026. Topodim: One-shot topology generation of diverse interaction modes for multi-agent systems. *Preprint*, arXiv:2601.10120.

Xin Sun, Jianan Xie, Zhongqi Chen, Qiang Liu, Shu Wu, Yuehe Chen, Bowen Song, Weiqiang Wang, Zilei Wang, and Liang Wang. 2025. Divide-then-align: Honest alignment based on the knowledge boundary of rag. *Preprint*, arXiv:2505.20871.

Shuchang Tao, Liuyi Yao, Hanxing Ding, Yuexiang Xie, Qi Cao, Fei Sun, Jinyang Gao, Huawei Shen, and Bolin Ding. 2024. When to trust llms: Aligning confidence with response quality. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 5984–5996. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*. Association for Computational Linguistics.

Wanxin Tian, Shijie Zhang, Kevin Zhang, Xiaowei Chi, Chunkai Fan, Junyu Lu, Yulin Luo, Qiang Zhou, Yiming Zhao, Ning Liu, Siyu Lin, Zhiyuan Qin, Xiaozhu Ju, Shanghang Zhang, and Jian Tang. 2025. Seea-r1: Tree-structured reinforcement fine-tuning for self-evolving embodied agents. *Preprint*, arXiv:2506.21669.

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. *Preprint*, arXiv:2212.10509.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Liang Wang, Haonan Chen, Nan Yang, Xiaolong Huang, Zhicheng Dou, and Furu Wei. 2025. Chain-of-retrieval augmented generation. *Preprint*, arXiv:2501.14342.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. Text embeddings by weakly-supervised contrastive pre-training. *Preprint*, arXiv:2212.03533.

Zhishang Xiang, Chuanjie Wu, Qinggang Zhang, Shengyuan Chen, Zijin Hong, Xiao Huang, and Jinsong Su. 2025. When to use graphs in rag: A comprehensive analysis for graph retrieval-augmented generation. *arXiv preprint arXiv:2506.05690*.

Hongshen Xu, Zichen Zhu, Da Ma, Situo Zhang, Shuai Fan, Lu Chen, and Kai Yu. 2024. Rejection improves reliability: Training llms to refuse unknown questions using rl from knowledge feedback. *ArXiv*, abs/2403.18349.

Junming Yang, Ning Xu, Biao Liu, Shiqi Qiao, and Xin Geng. 2026. Alignment through meta-weighted online sampling: Bridging the gap between data generation and preference optimization. *Preprint*, arXiv:2509.23371.

Junxiao Yang, Jinzhe Tu, Haoran Liu, Xiaoce Wang, Chujie Zheng, Zhexin Zhang, Shiyao Cui, Caishun Chen, Tiantian He, Hongning Wang, Yew-Soon Ong, and Minlie Huang. 2025a. Barrel: Boundary-aware reasoning for factual and reliable lrms. *Preprint*, arXiv:2505.13529.

Qwen: An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025b. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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. In *Conference on Empirical Methods in Natural Language Processing*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Hao Chen, Yilin Xiao, Chuang Zhou, Junnan Dong, and 1 others. 2025a. A survey of graph retrieval-augmented generation for customized large language models. *arXiv preprint arXiv:2501.13958*.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. 2025b. Faithfulrag: Fact-level conflict modeling for context-faithful retrieval-augmented generation. *arXiv preprint arXiv:2506.08938*.

Qingjie Zhang, Yujia Fu, Yang Wang, Liu Yan, Tao Wei, Ke Xu, Minlie Huang, and Han Qiu. 2025c. On the self-awareness of large reasoning models' capability boundaries. *Preprint*, arXiv:2509.24711.

Shijie Zhang, Guohao Sun, Kevin Zhang, Xiang Guo, and Rujun Guo. 2025d. Clpo: Curriculum learning meets policy optimization for llm reasoning. *Preprint*, arXiv:2509.25004.

Hang Zheng, Hongshen Xu, Yuncong Liu, Lu Chen, Pascale Fung, and Kai Yu. 2025. Enhancing llm reliability via explicit knowledge boundary modeling. *arXiv preprint arXiv:2503.02233*.

Luyao Zhuang, Shengyuan Chen, Yilin Xiao, Huachi Zhou, Yujing Zhang, Hao Chen, Qinggang Zhang, and Xiao Huang. 2025. Linearrag: Linear graph retrieval augmented generation on large-scale corpora. *arXiv preprint arXiv:2510.10114*.

A Implementation Details

A.1 Details of GRPO

We implement GRPO algorithm based on the verl[2] framework. The details of training parameters are listed in Table 4.

Hyper-Parameter	Value
Learning Rate	1e-6
LR Scheduler	Constent
Batch Size	64
Mini Batch Size	16
Training Epochs	2
Number of Rollout	8
Rollout Temperature	1.0
KL Loss Coefficient	0.001
Clip Ratio (ϵ)	0.1
Max Tokens	8192

Table 4: Training hyper-parameters setting of GRPO.

Consistent with previous agentic search models like Search-R1 and Tool-Star, we exclud all tool invocation results from the loss computation to avoid biasing the policy toward tool outputs. Only tokens from the text-based reasoning process and tool requests are considered during loss calculation. For fair comparison, all non-BAPO-specific parameters are set to match the standard GRPO configuration.

A.2 Details of Reliable RFT

As a variant of Rejection Sampling Fine-tuning (RFT), Reliable RFT aims to move beyond pure accuracy maximization and optimize the model to admit uncertainty in the fine-tuning process. We use the same training dataset as in the implementation of BAPO. For RFT data construction, we perform rollouts with 8 samples per question using a sampling temperature of 1.0. For each question, if no correct answer exists across the rollouts, we adopt the model's *IDK* response (if available); otherwise, we select the correct response(s) for training. We use LLaMA-Factory[3] training framework for conducting experiments. The training hyper-parameters are listed in Table 5.

Hyper-Parameter	Value
Learning Rate	7e-6
LR Scheduler	Cosine
Training Epochs	3
Warmup Ratio	0.1
Batch Size	8
Gradient Accumulation Steps	2

Table 5: Training hyper-parameters of Reliable RFT.

A.3 Details of Open-source Models

For a fair comparison, we directly use the released checkpoint of open-source models trained from Qwen2.5-7B-Instruct and follow the corresponding prompts in the original work to ensure the reproducibility of the results.

Search-R1 (Jin et al., 2025a) is a reinforcement learning framework that trains a model to autonomously invoke search engines during the reasoning process. Its reward is determined solely by the correctness of the final answer. The training dataset is a unified dataset merging Natural Questions (NQ) and HotpotQA, amounting to approximately 90k training samples.

ReSearch (Chen et al., 2025) is another framework designed to enable agentic search capabilities of LLMs. Its reward function incorporates both the prediction F1 score and format correctness. Models are trained using the MuSiQue training set, which comprises 19,938 samples.

A.4 Dataset Details

For training BAPO, we utilize the high-quality dataset introduced by (Dong et al., 2025)[4]. This training dataset is specifically constructed to be challenging for both Direct Reasoning and Tool-Integrated Reasoning. To target agentic search capabilities, we isolate the QA portion of this dataset. This yields 5,000 training samples derived from HotpotQA and 2WikiMultiHopQA.

A.5 Benchmark Details

For evaluation, we use four benchmarks of multi-hop QA tasks: HotpotQA (Yang et al., 2018), MuSiQue (Trivedi et al., 2022), 2WikiMultiHopQA (Ho et al., 2020), and Bamboogle (Press et al., 2023). Following the setup in Dong et al. (2025), the full test set consists of 200 test samples from HotpotQA, 200 from 2WikiMultiHopQA, 200 from MuSiQue, and 125 from Bamboogle.

[2]https://github.com/volcengine/verl

[3]https://github.com/hiyouga/LLaMA-Factory

[4]https://huggingface.co/datasets/dongguanting/Multi-Tool-RL-10K

Param.	Acc	Prec	ρ_{IDK}	Rel.
$\alpha = 0.0$	40.7	40.7	0.0%	40.7
$\alpha = 0.05$	**44.8**	52.8	16.8%	**51.3**
$\alpha = 0.2$	39.2	**53.9**	30.2%	49.2
$\alpha = 0.3$	41.4	53.8	25.2%	50.3
$k = 0$	43.4	50.9	15.4%	49.0
$k = 1$	44.5	50.3	13.2%	49.2
$k = 2$	44.8	**52.8**	16.8%	**51.3**
$k = 3$	**45.1**	52.7	16.3%	51.2

Table 6: Sensitivity analysis of BAPO's key hyperparameters: *IDK* ratio threshold α and resampling times k. Metrics are averaged across four benchmarks (HotpotQA, 2WikiMultihopQA, MuSiQue, Bamboogle) on Qwen2.5-3B-Instruct.

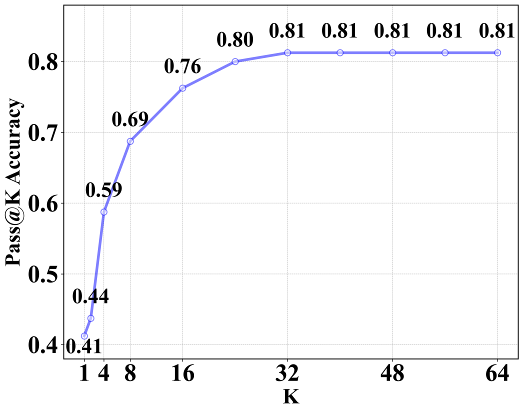

Figure 7: Accuracy of the Pass@K on Qwen2.5-7B-Instruct for different values of K. The accuracy improves with increasing K and stabilizes after $K = 32$.

B Additional Experiments

B.1 Hyper-Parameters Sensitivity Analysis

To investigate the impact of the values of BAPO's specific hyper-parameters *IDK* ratio α and resampling times k, we systematically evaluate model's performance under different hyper-parameter settings, while keeping all other hyper-parameters fixed to the default settings used in the main experiments. All experiments are conducted on Qwen2.5-3B-Instruct, and the results are reported in Table 6.

This hyper-parameters analysis reveals these key observations: (1) The value of α is closely tied to the model's tendency to acknowledge uncertainty. Since α directly controls the strength of the reward assigned to *IDK* responses, when $\alpha = 0$, the model is never encouraged to produce *IDK* responses in the early stage. As a result, the model has lost the ability to generate *IDK* responses in exploration stage and can no longer learn this behavior in plateau stage. In contrast, larger values of α promote more *IDK* responses during the exploration stage, enabling the model to cultivate boundary awareness. (2) The resampling times k allows us to dynamically enlarge the effective rollout size by repeatedly sampling within groups that contain no rewardable responses, thereby improving overall reliability. Increasing k from 1 to 2 yields clear gains, while further increasing k to 3 provides only marginal improvement, indicating that moderate resampling times is sufficient.

B.2 Why We Resample Rollouts

The resampling strategy is premised on the assumption that increasing the rollout size yields a more accurate estimation of the model's boundary. As illustrated in Figure 7, the model shows no further improvement in pass@K accuracy as K increases to 32. This indicates that failure stems from the problem lying beyond the reasoning boundary, rather than merely being an accidental failure to sample a correct solution. Consequently, our dynamic resampling design adaptively expands the group size to refine this boundary estimation. The results in Table 6 corroborate our findings, demonstrating that pass@24 to pass@32 (corresponding to resampling factors $k = 2$ and $k = 3$ with a rollout size of 8) are sufficient for accurate estimation.

B.3 Exact Match Performance

As a complement of the main experiment results based on LLM-as-a-Judge Equal (LE), we report the performance under the Exact Match (EM) metric in Table 7. Consistent with the LLM-based findings, models trained with BAPO achieve the highest reliability across the 3B, 7B, and 14B model scales. These results further demonstrate the effectiveness and efficiency of BAPO in developing reliable agents.

B.4 Case Study

Table 9 and Table 10 represent the reasoning process of Qwen2.5-14B-Instruct trained with GRPO and BAPO, respectively. When the retrieved results lack relevant information, the GRPO-trained model initially identifies the absence of necessary information and attempts to search for it. However, it subsequently hallucinates that the director is Ralph Pappier, resulting in an erroneous final outcome. In contrast, BAPO enables the model to recognize its

Method	HotpotQA			MusiQue			2Wiki.			Bamboogle		
	Acc	Prec	Rel	Acc	Prec	Rel	Acc	Prec	Rel	Acc	Prec	Rel
Qwen2.5-7B-Instruct												
Naive RAG	31.9	31.9	31.9	6.2	6.2	6.2	25.8	25.8	25.8	20.8	20.8	20.8
IRCoT	30.3	30.3	30.3	6.9	6.9	6.9	21.6	21.6	21.6	24.8	24.8	24.8
Search-R1	39.5	39.5	39.5	13.5	13.5	13.5	34.0	34.0	34.0	40.8	40.8	40.8
ReSearch	**46.5**	46.5	46.5	**15.5**	15.5	15.5	37.5	37.5	37.5	47.2	47.2	47.2
TIR Prompt	19.5	19.5	19.5	4.0	4.0	4.0	11.0	11.0	11.0	26.4	26.4	26.4
Reliable TIR Prompt	24.5	29.6	28.3	7.0	9.5	8.8	16.0	20.1	19.3	25.6	27.5	27.4
Reliable RFT	16.5	45.8	27.0	5.0	16.6	8.5	5.5	19.6	9.4	28.0	42.6	37.6
GRPO	44.0	44.0	44.0	14.0	14.0	14.0	**43.5**	43.5	43.5	**48.0**	48.0	48.0
BAPO (Ours)	42.0	**52.0**	**50.1**	14.0	**20.0**	**18.1**	42.0	**50.2**	**48.6**	46.4	**49.5**	**49.3**
Qwen2.5-3B-Instruct												
TIR Prompt	7.5	7.5	7.5	2.0	2.0	2.0	4.5	4.5	4.5	11.2	11.2	11.2
Reliable TIR Prompt	16.5	22.1	20.7	3.5	5.2	4.6	16.0	20.2	19.3	16.8	17.7	17.6
GRPO	**38.5**	38.5	38.5	**11.0**	11.0	11.0	37.0	37.0	37.0	**39.2**	39.2	39.2
BAPO (Ours)	37.0	**44.8**	**43.5**	**11.0**	**15.9**	**14.4**	36.0	**40.0**	**38.9**	**39.2**	**42.2**	**42.0**
Qwen2.5-14B-Instruct												
TIR Prompt	37.0	37.0	37.0	9.0	9.0	9.0	32.5	32.5	32.5	45.6	45.6	45.6
Reliable TIR Prompt	36.0	43.6	42.3	12.5	16.5	15.5	34.0	42.5	40.8	43.2	45.7	45.5
GRPO	**46.0**	46.0	46.0	14.5	14.5	14.5	**44.5**	44.5	44.5	**53.6**	53.6	53.6
BAPO(Ours)	45.0	**53.8**	**52.3**	**16.0**	**22.6**	**20.6**	41.5	**53.2**	**50.6**	**53.6**	**58.2**	**57.8**

Table 7: Exact Match (EM) Performance of multi-hop question answering tasks.

Config	Acc	Prec	ρ_{IDK}	Rel.
Internal Confidence				
$x = 0.2$	36.7	45.8	20.0%	43.9
$x = 0.3$	34.1	48.7	30.0%	44.3
$x = 0.5$	27.5	54.8	50.0%	41.2
Expression Certainty				
$x = 0.2$	40.4	40.4	0.0%	40.4
$x = 0.3$	28.6	39.6	27.6%	36.5
$x = 0.5$	26.0	43.1	38.2%	36.6
Self-Reflection				
-	17.5	56.2	68.8%	29.5

Table 8: Performance of uncertainty based methods on Qwen2.5-Instruct-7B, with different values of x .

boundary. When no relevant information about the film "Winds of the Pampas" is found, the model trained with BAPO appropriately responds with *"I DON'T KNOW"* (*IDK*).

B.5 Detailed Analysis of Uncertainty Estimation Methods

Although uncertainty estimation methods have proven effective in standard QA or math reasoning tasks (Chen and Mueller, 2024; Kuhn et al., 2023; Zhang et al., 2025c), their effectiveness remains unexplored in agentic search scenario. In this section, we systematically evaluate three distinct uncertainty estimation methods: 1) **Internal Confidence** based on token probabilities, 2) **Expression Certainty** measuring confidence via the model's verbalized expressions, and 3) **Self-Reflection** utilizing the model's self-verification capabilities. The implementation details are as follows:

- **Internal Confidence.** Model confidence is conventionally quantified via the probability distribution of its output (Kuhn et al., 2023; Kumar et al., 2024). In our approach, we adopt the average log-probability as the proxy for internal confidence.

- **Expression Certainty.** To assess uncertainty based on model's expression, we follow Zhang et al. (2025c) to calculate the proportion of confident versus uncertain expressions within the reasoning trajectory. Specifically, we utilize a predefined lexicon of confident and uncertain phrases (as illustrated in Figure 8), which are manually curated and subsequently verified by an LLM. We quantify the degree of certainty as the ratio of confident expressions to uncertain expressions.

- **Self-Reflection.** Leveraging the inherent capacity of LLMs for self-reflection (Manakul et al., 2023; Madaan et al., 2023), we use a prompt-based self-reflection approach to instruct the model to retrospectively audit its own reasoning trajectory to identify potential

Confident Expression

yes, perfect, good, great, right, correctly, True, indeed, no mistake, don't think I made any mistakes, don't see any mistakes, can't see any mistakes, didn't make any mistakes, no doubt, excellent, no issue, stick with my answer, acceptable, correct, go with, feel confident, can be confident, I'm confident, I am confident, so confident that, confidently, pretty confident, more confident, definitely, reasonable, that's solid, fine, perfect answer, exactly, seems consistent, consistent solution, I think that's the answer, I think that's it, I think that is it, method is solid, its solid, seems solid, approach is solid, calculation is solid, the way I did it is solid, this is solid, any errors, any error, both solutions satisfy, they satisfy, that satisfy, satisfied, this is the only solution, it seems like this is the only solution, suggests that this is the only solution, valid, convinced, Therefore, must be, I believe, I think this is the answer, I think that is the answer, no problem, the answer should be, seems okay, satisfies, nice, plausible, Lets trust, easy, same as before

Uncertain Expression

a mistake, a fundamental mistake, my mistake, mistakenly, miscalculated, missed, wrong, messed up, incorrect, incorrectly, not correct, messy, complicated, convoluted, no!, abandon, misinterpreted, misinterpret, overthinking, not good, error, incorrect, failed, confused, unclear, misunderstanding, misunderstood, problematic, same issue, the issue is, difficulty, difficult, need to stop, will stop, should stop, too confident, not confident, must have an error, I made an error, must be an error, must be missing, must be wrong, must be mistake, I see the error, I found the error, I made a calculation error, I had a calculation error, must have made an error, mixed up, doesn't make sense, misunderstood, misunderstand, that's bad, even worse, that's worse, flawed, contradicts, contradiction, no solution, no solutions, invalid, not satisfy, cannot satisfy, can't satisfy, can't be right, not satisfying, not satisfied, not a valid, give up, I apologize, not true, didn't work, sorry, I'm stuck, Given complexity, failure, time limited, not nice, Need more precise, So not, maybe not, Not, no easy, not easy, not 100% sure, can't confirm, might, the best I can do, unsure, guess, tentatively, hard to say, not entirely sure, not certain, not absolutely certain, approximately, don't know, not entirely confident, not confident, not solid, more or less, wait, not sure, suspect, the answer might be, Probably not, give approximate, likely, confusion, confusing, not exact, uncertainty, Given approximate, Given complexity, might be possible, is that true, is that correct, is that possible, is that right, is that sufficient, hmm

Figure 8: Confident and uncertain expressions.

logical pitfalls. The specific prompt utilized for this verification is detailed in Section C.3.

For Internal Confidence and Expression Certainty, we apply a percentile-based thresholding method: responses ranked within the bottom $x \in [0, 1]$ of confidence scores across the dataset are classified as *IDK*. For Self-Reflection, we rely on the model's judgment, where any response flagged as uncertain by the LLM is directly labeled as *IDK*.

As shown in Table 8, although employing these methods improves precision, it comes at the cost of a substantial drop in accuracy, resulting in negligible gains or even degradation in reliability. This underscores a critical challenge in agentic search scenario: the validity of reasoning is inextricably linked to the quality of retrieved content. This external dependency induces a decoupling between the actual reachability of the answer and the model's internal or explicit uncertainty signals, thereby limiting the effectiveness of standard estimation paradigms. It is important to note that uncertainty estimation methods do not cultivate model's inherent capability to explicitly admit *IDK*, which often leads to a significant discrepancy between the reasoning process and the final answer.

C Prompt Template

C.1 TIR System Prompt

Table 11 presents the Tool-Integrated Reasoning Prompt for formulating the reasoning process and the search tool call format.

C.2 Prompt for LLM-as-a-Judge Evaluation

Table 12 presents the evaluation prompt for LLM-based scoring. It guides the judge model to determine the equivalence of the predicted and labeled answers.

C.3 Prompt for Self-Reflection

Table 13 details the Self-Reflection Prompt, which tasks the model with identifying logical inconsistencies within its reasoning trajectory.

D Detailed Related Work

D.1 Agentic Search

To enhance response quality in knowledge-intensive scenarios, Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Gao et al., 2023; Zhang et al., 2025b) supplements LLMs' parametric knowledge with external search (Fan et al., 2024; Zhang et al., 2025a; Xiang et al., 2025; Zhuang et al., 2025). Current prompting-based strategies often depend on predefined workflows for interleaving reasoning and search (Asai et al., 2024; Li et al., 2025c), resulting in limited effectiveness. While learning-based methods (SFT/DPO) offer an alternative (Lin et al., 2024; Wang et al., 2025; Fang et al., 2025; Li et al., 2025d,a), they remain constrained by complex data pipelines and suboptimal generalization.

Inspired by the success of reinforcement learning on mathematics and coding (Jaech et al., 2024; DeepSeek-AI et al., 2025), RL-based agentic search systems have emerged, such as Search-R1 (Jin et al., 2025a), ReSearch (Chen et al., 2025), R1-Searcher (Song et al., 2025), and Tool-Star (Dong et al., 2025). These methods employ format and outcome correctness rewards to enable autonomous query decomposition and flexible interaction with external search systems during reasoning, thereby achieving superior accuracy on complex queries. Nevertheless, current methods prioritize search accuracy while ignoring holistic reliability, which depends on both accuracy and precision. In contrast, our proposed BAPO leverages boundary-aware reward mechanisms to optimize overall reliability, thereby maximizing the practicality of agentic search models.

D.2 LLM Reliability

One established approach to enhancing LLM reliability involves uncertainty-based methods, which

Question: What is the place of birth of the director of film Winds Of The Pampas?
Answer: Italian

<think> To answer this question, I need to find out who the director of the film "Winds of the Pampas" is and then find their place of birth. I will start by searching for the director of the film "Winds of the Pampas". </think>

<search> director of film Winds Of The Pampas </search>

<result> "The Wind Journeys" The Wind Journeys The Wind Journeys () is a 2009 Colombian-German-Argentine-Dutch drama film written and directed by Ciro Guerra... ...
"Ralph Pappier" Ralph Pappier Ralph Pappier (16 January 1914 in Shanghai – 29 August 1998 in Buenos Aires) was an Argentine production designer, set decorator and film director... ...
"Savage Pampas (1945 film)" Savage Pampas (1945 film) Savage Pampas (Spanish:Pampa bárbara) is a 1945 Argentine historical film directed by Lucas Demare ... ... </result>

<think> From the search result, I can see that the film "Winds of the Pampas" is not directly mentioned. However, I found that Ralph Pappier is an Argentine film director who has directed several films. I will check if "Winds of the Pampas" is one of his works. If not, I will look for the director of "Winds of the Pampas" again. </think>

<search> Winds of the Pampas director </search>

<result> "Ralph Pappier" Ralph Pappier Ralph Pappier (16 January 1914 in Shanghai – 29 August 1998 in Buenos Aires) was an Argentine production designer, set decorator and film director ... ...
"Pampa Ilusión" Pampa Ilusión Pampa Ilusión (Illusion of Pampa) is a Chilean telenovela produced by TVN. It was written by Víctor Carrasco, Larissa Contreras, María José Galleguillos and Alexis Moreno. Directed by Vicente Sabatini ... ...
"Law of the Pampas" Law of the Pampas Law of the Pampas is a 1939 American Western film directed by Nate Watt and written by Harrison Jacobs ... ...
</result>

<think> From the search result, I can see that Ralph Pappier is the director of "Winds of the Pampas". Now, I need to find his place of birth. </think>

<search> Ralph Pappier place of birth </search>

<result> "Ralph Pappier" Ralph Pappier Ralph Pappier (16 January 1914 in Shanghai – 29 August 1998 in Buenos Aires) was an Argentine production designer, set decorator and film director... ...
"Ralph Pappier" film ""Circus cavalcade"". The Argentine Academy of Cinematography Arts and Sciences gave him awards for Best Scenography for ""En el viejo Buenos Aires"" (1942) and ""Su mejor alumno"" (1944)... ... </result>

<think> From the search result, I can see that Ralph Pappier was born in Shanghai, China. </think>

<answer> The final answer is \boxed{Shanghai, China} </answer>

Table 9: A case study on Qwen2.5-14B-Instruct trained with GRPO. Text in red indicates that the model is hallucinating and providing a conclusion without relevant information in retrieval results.

Question: What is the place of birth of the director of film Winds Of The Pampas? **Answer:** Italian
<think> To answer this question, I need to find out the director of the film "Winds Of The Pampas" and then find out their place of birth. I will start by searching for information on the film "Winds Of The Pampas" </think>
<search> Winds Of The Pampas film </search>
<result> "Savage Pampas (1945 film)" Savage Pampas (1945 film) Savage Pampas (Spanish:Pampa bárbara) is a 1945 Argentine historical film directed by Lucas Demare and Hugo Fregonese and starring Francisco Petrone, Luisa Vehil and Domingo Sapelli... ... "Law of the Pampas" Law of the Pampas Law of the Pampas is a 1939 American Western film directed by Nate Watt and written by Harrison Jacobs. The film stars William Boyd, Russell Hayden, Sidney Toler, Steffi Duna, Sidney Blackmer, Pedro de Cordoba and William Duncan... ... "Savage Pampas (1966 film)" Savage Pampas (1966 film) Savage Pampas is a 1966 western film directed by Hugo Fregonese and starring Robert Taylor, Ron Randell and Marc Lawrence... ... </result>
<think> From the search result, I couldn't find any information about the film "Winds Of The Pampas". However, I found information about other films set in the Pampas region, such as "Savage Pampas" (1945 and 1966) and "Pampa Ilusión". Since there is no information about "Winds Of The Pampas", I will have to conclude that the film may not be well-known or may not have a lot of information available online. Therefore, I don't have enough information to answer the question. </think>
<answer>\boxed{I DON'T KNOW} </answer>

Table 10: A case study on Qwen2.5-14B-Instruct trained with BAPO.

quantify prediction confidence through token probabilities and consistency across diverse outputs (Manakul et al., 2023; Chen and Mueller, 2024). Alternatively, calibration strategies such as prompting (Tian et al., 2023) and fine-tuning (Tao et al., 2024) are employed to align model confidence with accuracy. Recently, Zhang et al. (2025c) find that reasoning models exhibit self-awareness regarding their capability boundaries and proposed test-time monitoring strategies to terminate unproductive reasoning for unsolvable questions. Distinctly, our work focuses on activating the model's intrinsic boundary awareness through RL, which is orthogonal to these uncertainty-based methods.

Sharing this focus on model's internal capacity to refuse queries out of its answerable boundary, Xu et al. (2024), Chen et al. (2024) and Sun et al. (2025) develop specialized reward models by constructing preference pairs of refusal and standard responses for subsequent RL training. KnowRL (Ren et al., 2025) incorporates a fact-level verification reward into RL to improve factual accuracy, yet it remains dependent on manually curated factual databases. BARREL (Yang et al., 2025a) fine-tunes models on distilled reasoning traces that align with expected patterns, followed by RL training with a static medium-level reward for uncertain responses. However, our empirical findings suggest that such static reward leads to excessively high rejection rates, thereby impairs accuracy. Unlike prior work, our approach targets the agentic search scenario, which is characterized by a sophisticated synergy between reasoning and search tool interaction.

D.3 Policy Optimization

DeepSeek-R1 (DeepSeek-AI et al., 2025) shows that RL with formatting and result-only rewards can steer LLMs toward complex chain-of-thought reasoning, boosting performance on challenging tasks. Subsequently, Search-R1 (Jin et al., 2025a) and TORL (Li et al., 2025e) explored extending this R1-style approach to tool use. However, to accommodate diverse training requirements, various RL variants have been proposed to adaptively adjust training objectives and task difficulty (Zhang et al., 2025d; Tian et al., 2025; Yang et al., 2026; Ma et al., 2026; Sun et al., 2026). Unlike prior work, our proposed BAPO aims to balance two conflicting rewards by introducing a stage-based adaptive reward modulator, thereby achieving an equilibrium between the model's boundary awareness and problem-solving capabilities during training.

E The Use of Large Language Models

In preparing this paper, we made limited use of Large Language Models (LLMs). Specifically, LLMs were employed for two purposes: (i) to aid in polishing the writing by improving grammar, readability, and clarity without altering the scientific content, and (ii) to assist in retrieval and discovery tasks, such as identifying and organizing related work. No LLMs were used for generating novel research ideas, designing experiments, or analyzing results. All conceptual and technical contributions presented in this paper are the sole work of the authors.

TIR System Prompt

You are a helpful assistant that can solve the given question step by step with the help of the wikipedia search tool.
Given a question, you need to first think about the reasoning process in the mind and then provide the answer. During thinking, you can invoke the wikipedia search tool to search for fact information about specific topics if needed.
The reasoning process and answer are enclosed within `<think> </think>` and `<answer> </answer>` tags respectively, and the search query and result are enclosed within `<search> </search>` and `<result> </result>` tags respectively.
For example, `<think>` This is the reasoning process. `</think> <search>` search query here `</search> <result>` search result here `</result> <think>` This is the reasoning process. `</think> <answer>` The final answer is `[ \boxed{answer here} ] </answer>`.
In the last part of the answer, the final exact answer is enclosed within `\boxed{}`.

Table 11: Tool-Integrated-Reasoning System Prompt

LLM-as-a-Judge Evaluation

You are an evaluation assistant. Please determine if the model output is equivalent to the labeled answer.
Question: {question}
Labeled Answer: {labeled answer}
Model Output: {pred answer}
Did the model give an answer equivalent to the labeled answer? Please respond with "Correct" if they are equivalent, or "Incorrect" if they are not equivalent. Do not include any other text.

Table 12: LLM-as-a-Judge Evaluation Prompt

Self-Reflection Prompt

Below is the reasoning trajectory of an agentic search model. `<search>...</search>` indicates search queries executed via tools, `<result>...</result>` shows the returned search results, and `<answer>...</answer>` contains the model's final answer.
Analyze the following reasoning trajectory for logic errors or inconsistencies.
Reasoning Trajectory: {reasoning trajectory}
Please determine if the response exhibits any of the following issues:
1. Logical contradictions or self-contradictions
2. Obvious flaws in the reasoning process
3. Discrepancy between the answer and the reasoning process

Please answer only "Yes" or "No" without explanation. If any of the above problems exist, answer "Yes"; if there are no obvious issues, answer "No".

Table 13: Self-Reflection Prompt