Trust Within? Seek Beyond? Knowledge Boundary Aware Policy Optimization for Agentic Search

Tao Feng[1]*, Xinke Jiang[2]*, Xinyan Hu[1], Yonggang Zhang[3], Zhen Tao[4]
Wentao Zhang[5], Boyang Liu[1], Wenhao Jiang[6], Chao Wu[1]†
[1]Zhejiang University, Hangzhou, China; [2]Peking University, Beijing, China
[3]The Hong Kong University of Science and Technology, Hong Kong, China
[4]Great Bay University & Nanjing University, Nanjing, China
[5]ShanghaiTech University, Shanghai, China [6]Harbin Institute of Technology, Harbin, China
{tao.feng_1,chao.wu}@zju.edu.cn, thinkerjiang@foxmail.com

Abstract

Agentic search augments large language models (LLMs) with external knowledge through reinforcement learning. However, existing approaches suffer from *blind reliance* on noisy retrieval and *hallucination* when both parametric and external knowledge fail—reflecting a lack of calibration regarding the model's knowledge boundary. We propose **K**nowledge **b**oundary **P**olicy **O**ptimization (**KbPO**), a reinforcement learning framework that explicitly aligns retrieval decisions with quantified knowledge states. KbPO introduces: (1) a semantic stability metric to delineate reliable parametric knowledge; (2) a four-quadrant taxonomy synthesising internal certainty with retrieval quality; and (3) a quadrant-based reward mechanism incentivising calibrated behaviour. We further adopt an iterative query evolution pipeline to construct boundary-probing training samples. Experiments on ten benchmarks demonstrate that KbPO outperforms strong baselines while exhibiting reduced hallucination rates. Code is available at https://github.com/jiangxinke/Agentic-RAG-R1.

1 Introduction

Large **L**anguage **M**odels (**LLMs**) have fundamentally transformed general-purpose reasoning, establishing themselves as a central paradigm in modern **n**atural **l**anguage **p**rocessing (NLP) (Yang et al., 2024b; Kaplan et al., 2020; Yang et al., 2024a). Despite their remarkable capabilities, however, LLMs remain susceptible to factual errors (Ji et al., 2023; Cao et al., 2020), knowledge obsolescence (He et al., 2022), and limited domain-specific expertise (Kandpal et al., 2023). These limitations stem from the static nature of parametric memory and frequently manifest as hallucinations and outdated responses (Huang et al., 2025a). To mitigate these, **R**etrieval-**A**ugmented **G**eneration

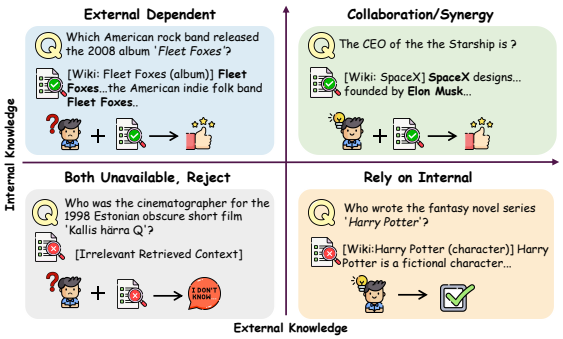

Figure 1: Four-quadrant cognitive taxonomy for retrieval-augmented agents. Based on internal knowledge availability and external retrieval quality, agents should: integrate both sources when available , rely on retrieval when internal knowledge is insufficient , use internal knowledge when retrieval fails , or refuse to answer when both are unavailable .

(RAG) has emerged as a promising paradigm by grounding model outputs in external knowledge sources, thereby improving factual accuracy and adaptability (Edge et al., 2024; Asai et al., 2024a; Lewis et al., 2020; Guu et al., 2020).

In parallel with the recent advances in *test-time scaling* (Brown et al., 2024), a growing body of work has explored multi-step reasoning and search frameworks that enable models to iteratively refine their decisions at inference time (Wei et al., 2023). Recent approaches such as Self-RAG (Asai et al., 2024b) and CRAG (Yan et al., 2024) incorporate iterative feedback mechanisms, but largely rely on fixed, hard-coded execution graphs, which constrain flexibility and limit their applicability to open-ended problem solving. As a result, the field has undergone a paradigm shift toward ***Agentic Search*** (Yao et al., 2022; Nakano et al., 2021; Singh et al., 2025; Jiang* et al., 2024, 2025b), where reasoning and retrieval are formulated as a **dynamic reasoning-searching process**. In this paradigm, the model autonomously orchestrates queries, evaluates intermediate trajectories, and

*Equal contribution.
†Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 27664–27682
July 2-7, 2026

adaptively refines its strategy through inference-time planning (Sun et al., 2023; Jin et al., 2025; Dong et al., 2025b,a).

However, the efficacy of such dynamic search processes is not guaranteed. Even sophisticated agents remain susceptible to factual errors (Ji et al., 2023) and knowledge obsolescence(He et al., 2022). While leveraging external knowledge (Lin et al., 2025b) offers a remedy, the effectiveness of RAG fundamentally depends on recognizing the **knowledge boundary**—the frontier beyond which parametric memory becomes unreliable.

Figure 1 illustrates a four-quadrant cognitive taxonomy defined by this interaction. For queries where both sources align, such as determining the birth country of the CEO behind Starship, the agent should *integrate both sources* to reinforce its internal knowledge of Elon Musk. In cases where parametric memory is insufficient—e.g., identifying the specific band behind the 2008 album "Fleet Foxes"—the model must strictly *rely on retrieval* to bridge the gap. Crucially, the risk of blind reliance appears when retrieval is noisy: for "Who wrote Harry Potter?", retrieved documents might distractingly focus on actor Daniel Radcliffe; here, the model must discern the noise and *use internal knowledge* of J.K. Rowling. Finally, for double-unknown cases involving obscure entities like the cinematographer of the 1998 Estonian film "Kallis härra Q", the absence of support demands that the agent *refuse to answer* to prevent hallucination.

This raises a fundamental question: *How can we align the model's internal knowledge boundaries with external retrieval reliability?*

To address this alignment challenge, recent research has pivoted towards explicitly engineering the synergy between internal parametric knowledge and external retrieval. From an inference-time perspective, training-free approaches like CK-PLUG (Bi et al., 2025) introduce dynamic interventions to re-weight prediction distributions based on entropy shifts, aiming to mitigate contextual interference without model adaptation. Concurrently, learning-based frameworks such as Knowledgeable-R1 (Lin et al., 2025a) and IKEA (Huang et al., 2025b) have integrated knowledge-boundary awareness into reinforcement learning. By optimizing policies to resist context dominance or penalize redundant queries, these methods foster agents that are more robust to noise and efficient in information seeking. Nevertheless, existing inference-time and RL-based approaches still lack fine-grained modeling of the collaborative dynamics between reasoning and retrieval, as they primarily rely on heuristic entropy proxies lacking semantic stability or optimize for coarse-grained objectives without disentangling parametric reliability from retrieval quality; explicitly enforcing consistency between internal confidence and external reliance is crucial for making calibrated decisions based on legitimate knowledge boundaries. This limitation motivates our approach: we propose a Knowledge Boundary Policy Optimization (KbPO) framework that enables fine-grained, process-level alignment between parametric confidence and external retrieval. Nevertheless, effectively designing such a framework remains the following challenges:

Challenge 1: Severe Hallucinations and Reward Ambiguity. In complex reasoning tasks, generation trajectories are inherently stochastic and prone to hallucinations, where models often fabricate non-existent facts or logic to bridge reasoning gaps. A critical pathology is reward hacking, where agents arrive at correct final answers through hallucinated rationales or "lucky guesses"—a phenomenon known as spurious correctness (Shihab et al., 2025). This issue is particularly acute yet underexplored in long-horizon reasoning: outcome-based rewards are too sparse to penalize intermediate fabrications, effectively "blinding" the optimization process to the distinction between rigorous deduction and stochastic hallucination.

Challenge 2:The Calibration Gap between Internal and External Knowledge. Even when reasoning is sound, a fundamental dissonance persists between the model's pre-trained priors and retrieved information. Instead of a coherent synthesis, the decision process often suffers from unstable boundaries: the model may exhibit "blind reliance" on noisy retrieval when it should be confident, or conversely, "stubbornness" when it lacks knowledge (Huang et al., 2025b). This is not merely a policy flaw but a representation failure: without explicitly quantifying the certainty frontier of its parametric memory, the system lacks the necessary scalar signals to dynamically weigh internal beliefs against external evidence.

To address these challenges, we propose KbPO, a fine-grained reinforcement learning framework for calibration-aware agentic search via Group Relative Policy Optimization. In implementation, we design a hybrid reward mechanism: an outcome reward at the response level based on correctness, and a consistency reward at the process level that

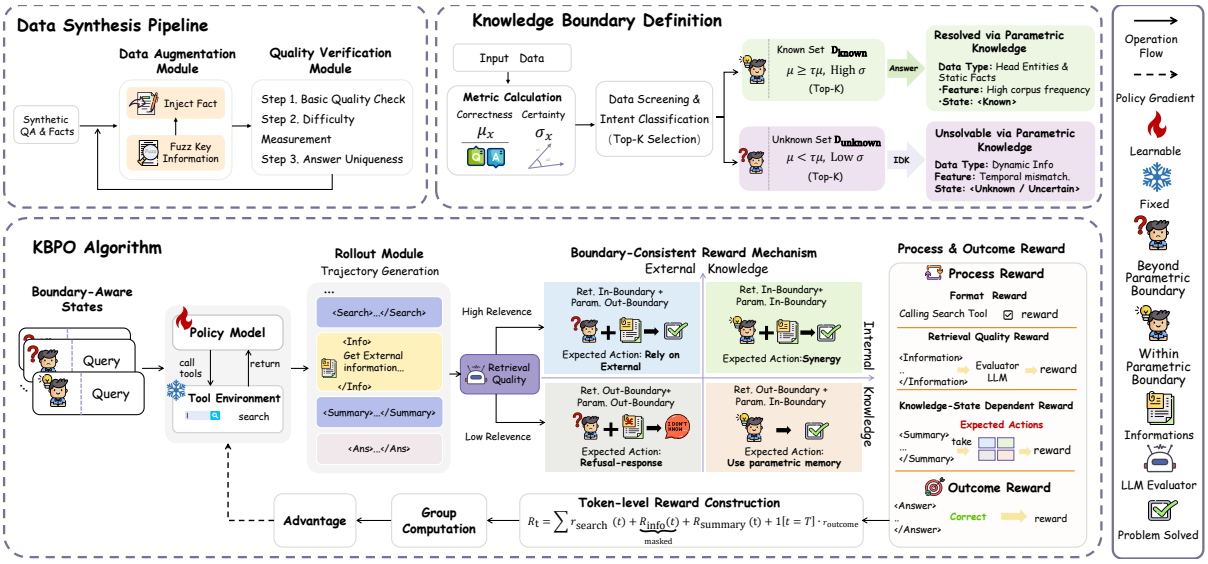

Figure 2: Overview of the **KbPO** framework. **Top:** The data synthesis pipeline constructs boundary-aware training samples by quantifying parametric certainty (μ) to classify queries into *Known* and *Unknown* sets. **Bottom:** The optimization phase leverages a boundary-consistent reward mechanism. It aligns agent behaviors (Search, Answer, Refuse) with the dynamic **knowledge boundary** defined by internal certainty and retrieval quality, using group-normalized advantages to stabilize policy updates.

captures the alignment between parametric confidence and retrieval necessity. To compute process rewards, we formalize the decision-making process as a boundary calibration problem. We employ semantic stability metrics to rigorously delimit the agent's internal knowledge boundary, and synthesize these with external retrieval evaluations to construct a four-quadrant cognitive taxonomy. We leverage this taxonomy to assign quadrant-based alignment scores as process rewards, incentivizing consistency between internal belief and external inquiry. To circumvent the reward ambiguity inherent in sparse long-horizon supervision, we adopt an Iterative Query Evolution pipeline to dynamically synthesize complex reasoning trajectories, transforming static QA pairs into a rigorous testbed for boundary calibration. In summary, our contributions are as follows:

- We propose KbPO, a framework that introduces knowledge boundary quantification to align agentic search behaviors with intrinsic parametric confidence, enabling precise calibration of retrieval necessity.
- We design a quadrant-based consistency reward mechanism and an iterative query evolution pipeline to mitigate reward ambiguity, addressing the issue of spurious correctness in outcome-driven reasoning.
- Extensive experiments demonstrate that KbPO achieves state-of-the-art performance across diverse benchmarks, fostering safer response generation and precise behavioral alignment with the four knowledge quadrants compared to strong baselines.

2 Related Work

Reinforcement Learning for Agentic Search. The optimization of LLM-based agents (Jiang* et al., 2025a; Zhang et al., 2026; Liu et al., 2026) has evolved from trajectory-level alignment to fine-grained process monitoring. Early RL-based frameworks, such as Search-R1 (Jin et al., 2025) and Search-o1 (Li et al., 2025), predominantly utilize trajectory-level algorithms like GRPO (Shao et al., 2024) to optimize the final answer correctness. To overcome the limitations of sparse rewards in long-horizon reasoning, recent works like ARPO (Dong et al., 2025b) introduce entropy-based adaptive rollout mechanisms. By monitoring token-level entropy shifts after tool interactions, these methods efficiently explore the branching space of tool-use behaviors. While these approaches improve the *efficiency* of exploration, they treat the model's internal confidence as a heuristic proxy for uncertainty rather than a calibrated boundary for external reliance.

Knowledge Boundary Alignment. Current re-

search on knowledge boundaries fundamentally bifurcates into post-hoc detection and coarse-grained alignment. Detection frameworks, such as SelfCheckGPT (Manakul et al., 2023), leverage stochastic sampling to offer fine-grained identification but remain computationally static and decoupled from the generation policy. Conversely, alignment approaches attempt to integrate boundary awareness into training via Refusal-Aware Instruction Tuning (RAIT) (Zhang et al., 2024a; Huang et al., 2025b; Zhu* et al., 2025). However, these methods typically rely on coarse-grained supervision signals that fail to disentangle parametric certainty from retrieval necessity. Recent studies like CRaFT (Zhu et al., 2025) further reveal that such rigid supervision induces static and dynamic conflicts—where the fixed refusal labels clash with similar samples in the feature space or the model's evolving knowledge states—ultimately leading to calibration failures such as over-refusal.

3 Method

As illustrated in Figure 2, the framework operates as a pipeline: (1) *Data Synthesis Pipeline* first generates complex queries to actively probe the model's capability limits; (2) *Boundary Quantification* module subsequently evaluates the model on these specific samples to capture its parametric certainty and explicitly define the knowledge boundary; and (3) *KbPO Algorithm* leverages these captured boundary signals to guide reinforcement learning, optimizing the agent's behaviors to align with the dynamic interplay between internal knowledge and external retrieval.

3.1 Data Synthesis Pipeline

Existing QA benchmarks predominantly consist of single-turn queries answerable through one-shot retrieval, which inadequately prepares agents for real-world scenarios requiring sustained multi-step reasoning with iterative tool interactions. To address this limitation, we adopt the iterative data synthesis pipeline inspired by Gao et al. (2025), which transforms simple QA pairs into long-horizon reasoning problems through two complementary evolution operators.

Data Augmentation. To construct high-complexity reasoning trajectories, we initialize with a seed corpus $\mathcal{D}_0 = \{(q, a, \mathcal{F})\}$, where each sample consists of a query q, a ground-truth answer a, and a supporting fact set $\mathcal{F}$. We employ an iterative evolution strategy utilizing two distinct operators to progressively escalate reasoning depth:

- **Fact Injection ($\mathcal{O}_{\text{inj}}$):** Targeting a specific entity $e \in q$, we retrieve a relevant external fact f_{new} and reformulate the query to necessitate multi-hop inference. This process updates both the query and the evidence set:

$$(q', \mathcal{F}') = \mathcal{O}_{\text{inj}}(q, \mathcal{F}, e), \text{where } \mathcal{F}' = \mathcal{F} \cup \{f_{\text{new}}\} \quad (1)$$

- **Information Fuzzing ($\mathcal{O}_{\text{fuzz}}$):** We replace explicit constraints c with underspecified descriptions $\tilde{c}$, inducing ambiguity that compels the agent to perform exploratory searches for clarification:

$$q' = \mathcal{O}_{\text{fuzz}}(q, c \to \tilde{c}) \quad (2)$$

These operators are applied recursively for T rounds, yielding a dataset that demands extensive tool interaction and long-horizon planning.

Quality Verification. Adopting the rigorous filtration protocol from Gao et al. (2025), each augmented sample undergoes a multi-stage validation process to ensure: (1) *semantic validity*; (2) *sufficient reasoning difficulty*; and (3) *answer uniqueness*. The resulting dataset $\mathcal{D}_T$ serves as a robust benchmark for evaluating boundary-aware agentic behaviors.

3.2 Knowledge Boundary Quantification

The parametric knowledge boundary is characterized by a distinct degradation in both *factual accuracy* and *semantic coherence* (Zhu et al., 2025). Within this boundary, the model exhibits consistent correctness; beyond it, hallucinations manifest as stochastic variance. We operationalize this intuition via two complementary metrics.

Probing Protocol. For a specific query q, we sample N independent responses $\{y_i\}_{i=1}^{N}$ using the frozen LLM parameters without external retrieval. We then derive:

- **Parametric Certainty (μ):** This metric approximates *factual correctness* by aggregating the match rate against the ground truth $\mathcal{G}$:

$$\mu(q) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\text{Match}(y_i, \mathcal{G})] \quad (3)$$

- **Semantic Stability (σ):** This metric quantifies *internal consistency* by computing the average

pairwise cosine similarity of response embeddings:

$$\sigma(q) = \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} \cos\left(\mathrm{Enc}(y_i), \mathrm{Enc}(y_j)\right) \tag{4}$$

Curating Supervision Signals. Partitioning by correctness alone induces static conflicts—similar queries receiving contradictory supervision. We mitigate this by first partitioning via threshold τ_μ into $\mathcal{D}_{\text{known}}$ and $\mathcal{D}_{\text{unknown}}$ (Zhu* et al., 2025), then using σ as a secondary filter: prioritizing high-σ samples from $\mathcal{D}_{\text{known}}$ (unambiguous knowledge) and low-σ samples from $\mathcal{D}_{\text{unknown}}$ (clear knowledge gaps). This yields maximally separable supervision for boundary-aware learning.

3.3 Calibration-Aware Policy Optimization

Building upon the quantified knowledge boundary, we propose KbPO to align the agent's reasoning strategies with its calibrated knowledge state. We leverage Group Relative Policy Optimization (GRPO) as our backbone algorithm, which enables efficient policy learning over grouped rollouts without the computational overhead of a separate value network.

Rollout and Action Space. Given a query q, the policy π_θ interacts with the retrieval environment to generate a reasoning trajectory. We define a structured action space comprising distinct functional tags: `<search>` for query formulation, `<info>` for context extraction, `<summary>` for synthesis, and `<answer>` for the final response. Each action is modeled as a parseable span within the generated sequence.

Token-Level Reward Construction. For each rollout o_i of length T, we construct a composite token-level reward signal that integrates outcome correctness with process-level guidance:

$$r_{i,t} = \mathbb{1}[t = T] \cdot r_i^{\text{out}} + \lambda \cdot r_{i,t}^{\text{proc}}, \tag{5}$$

where λ is a hyperparameter balancing the two terms. The **outcome reward** r_i^{out} is assigned at the terminal token based on evaluation metrics (e.g., F1 score). The **process reward** $r_{i,t}^{\text{proc}}$ provides sparse, event-based shaping by assigning signals to the ending tokens of executed action spans.

Boundary-Consistent Alignment Reward. A pivotal component of KbPO is the alignment reward, which enforces congruence between the agent's behavior and its knowledge-grounded situation. We define the *knowledge state* via two binary indicators:

$$s = \mathbb{1}[\hat{\rho} \ge \tau_\rho], \quad k = \mathbb{1}[\mu \ge \tau_\mu], \tag{6}$$

where $\hat{\rho}$ denotes external retrieval quality and μ represents parametric certainty (Sec. 3.2). The tuple (s, k) partitions queries into four quadrants, each prescribing an optimal strategy c^* (Xu* et al., 2025; Zhang et al., 2024b):

$$c^*(s,k) = \begin{cases} \textit{Integrated} & s{=}1, k{=}1 \text{ (synergize)} \\ \textit{External} & s{=}1, k{=}0 \text{ (rely on RAG)} \\ \textit{Refusal} & s{=}0, k{=}0 \text{ (abstain)} \\ \textit{Internal} & s{=}0, k{=}1 \text{ (trust memory)} \end{cases} \tag{7}$$

We utilize an automated judge to classify the agent's actual behavior $\hat{c}$ (derived from the `<summary>` span) into one of these categories. The alignment reward is then computed as:

$$r^{\text{align}} = \begin{cases} +\alpha & \text{if } \hat{c} = c^*(s,k) \\ -\alpha & \text{otherwise,} \end{cases} \tag{8}$$

where α modulates the magnitude of the behavioral penalty/bonus.

Optimization Objective. For each query q, the policy generates a group of G rollouts $\{o_i\}_{i=1}^{G}$. We compute the cumulative return $R_i = \sum_{t=1}^{T} r_{i,t}$ and derive the group-normalized advantage $\hat{A}_i = \frac{R_i - \text{mean}(\mathbf{R})}{\text{std}(\mathbf{R})}$. The policy is optimized by maximizing the following objective:

$$\mathcal{J}(\theta) = \mathbb{E}_{q \sim \mathcal{D}} \Bigg[\frac{1}{G} \sum_{i=1}^{G} \Big(\min\left(\rho_i \hat{A}_i, \text{clip}(\rho_i, 1{\pm}\epsilon) \hat{A}_i\right) - \beta \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \Big) \Bigg], \tag{9}$$

where $\rho_i = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}$ is the importance sampling ratio, ϵ is the clipping threshold, and β controls the KL divergence penalty against the reference policy π_{ref}.

4 Experiments

4.1 Experimental Setup

Datasets. We construct our training set using the data synthesis pipeline (Section 3.1) based on **HotpotQA** (Yang et al., 2018) and **2WikiMultiHopQA** (Ho et al., 2020). For evaluation, we use 9 benchmarks spanning three categories: (1) *Multi-hop QA*: HotpotQA, 2WikiMultiHopQA, MuSiQue (Trivedi et al., 2022), and Bamboogle (Press et al., 2023); (2) *Open-domain QA*:

Method		In-Domain		Out-of-Domain							Avg
Paradigm	**Approach**	**2Wiki**	**HotpotQA**	**Bamboogle**	**FRAMES**	**GAIA**	**MusiQue**	**NQ**	**PopQA**	**TriviaQA**	**Avg**
					Qwen2.5-1.5B						
No RAG	Base	78.05	79.80	94.13	90.49	95.27	91.74	88.69	88.37	67.19	85.97
	COT	83.25	82.10	80.26	91.57	94.81	92.78	90.18	90.11	74.83	86.65
Naive RAG	FS-RAG	78.63	73.45	84.69	88.60	97.04	90.50	82.26	67.51	55.27	79.77
	FL-RAG	73.92	71.86	85.16	87.42	98.46	89.70	78.06	**65.27**	51.41	77.92
Agentic RAG	ReACT	87.57	76.52	87.46	91.37	97.92	92.01	78.94	75.42	66.09	83.70
	IRCOT	78.21	73.32	82.51	90.68	95.93	91.27	77.07	67.04	55.81	79.09
	TCRAG	74.77	77.71	88.37	83.73	95.75	92.10	85.79	84.14	65.95	83.15
RL-based Agentic RAG	ReSearch	74.45	69.55	77.84	90.79	92.12	90.46	**70.93**	65.92	55.35	76.38
	Search-R1	**73.49**	79.34	86.20	90.35	94.44	91.64	88.07	89.77	74.24	85.28
	AEPO	82.66	85.25	88.42	91.02	94.98	92.89	90.45	91.17	76.98	88.20
	ARPO	79.98	84.52	87.62	90.89	94.90	93.84	90.81	90.51	76.04	87.68
	Mem1	85.57	85.66	87.97	91.82	95.82	93.92	81.59	90.15	77.94	87.83
	Ours‡	75.80	**64.90**	**74.30**	**69.80**	**82.50**	**68.60**	74.70	78.40	**47.20**	**70.69**
					Qwen2.5-3B						
No RAG	Base	76.02	75.92	90.55	91.99	94.30	90.30	85.73	87.89	61.16	83.76
	COT	81.10	76.18	79.20	92.84	92.98	89.53	83.47	88.63	60.38	82.70
Naive RAG	FS-RAG	84.53	74.15	89.52	89.58	95.70	92.36	80.16	69.59	54.62	81.13
	FL-RAG	83.20	73.22	88.95	90.81	98.42	92.71	78.07	66.53	51.50	80.38
Agentic RAG	ReACT	74.91	65.63	75.14	89.47	95.22	86.08	72.81	**63.53**	53.96	75.19
	IRCOT	84.11	75.50	74.73	93.21	97.23	87.57	72.14	64.98	50.81	77.81
	TCRAG	71.53	78.06	82.41	92.31	94.47	91.01	79.26	82.31	49.54	80.10
RL-based Agentic RAG	ReSearch	72.77	66.04	84.91	90.00	95.52	90.53	**65.39**	64.03	46.07	75.03
	Search-R1	**70.10**	62.76	70.10	89.24	95.32	86.47	65.27	65.76	44.92	72.22
	AEPO	76.99	71.29	77.91	87.74	94.99	88.30	73.24	70.52	52.22	77.02
	ARPO	70.45	63.52	72.68	86.51	92.29	86.62	66.71	66.56	46.34	72.41
	Mem1	81.94	79.85	94.81	95.01	97.53	95.53	80.82	83.29	66.91	86.19
	Ours‡	70.50	**57.60**	**68.10**	**62.60**	**77.20**	**61.40**	68.50	73.10	**39.90**	**64.32**
					Qwen2.5-7B						
No RAG	Base	74.59	73.37	82.14	87.48	95.71	87.85	80.28	85.21	50.92	79.73
	COT	76.45	70.90	62.44	82.40	93.83	85.65	77.53	84.93	50.67	76.09
Naive RAG	FS-RAG	82.29	70.79	83.14	87.48	94.98	89.26	83.18	81.06	64.98	81.91
	FL-RAG	80.22	65.58	75.90	87.90	95.40	87.54	80.28	74.72	57.34	78.32
Agentic RAG	ReACT	72.49	57.19	72.37	84.71	93.63	80.66	69.99	63.71	45.45	71.13
	IRCOT	**63.55**	73.71	78.10	93.22	94.50	91.61	80.37	65.86	50.57	76.83
	TCRAG	70.30	59.17	74.87	83.54	91.79	82.44	70.99	**62.83**	45.22	71.24
RL-based Agentic RAG	ReSearch	69.97	69.61	69.58	84.39	95.57	87.42	76.31	82.09	51.75	76.30
	Search-R1	64.97	61.11	**57.96**	81.99	89.92	80.92	70.41	72.53	44.09	69.32
	AEPO	80.12	86.15	86.76	92.76	97.35	94.15	90.07	90.05	82.47	88.88
	ARPO	69.29	74.80	67.06	87.82	91.44	87.29	82.20	86.38	59.84	78.46
	Mem1	74.71	70.02	63.50	85.85	93.50	85.87	73.62	74.20	48.96	74.47
	Ours‡	65.20	**51.80**	61.50	**56.30**	**71.90**	**54.70**	**62.30**	67.80	**32.60**	**57.99**

Table 1: Comparison of **Failure Rate (%)** (Lower is Better) on multi-hop benchmarks. Values are calculated as $100 - \text{Score}$. Score = F1 + Correct Refusal Rate, effectively effectively representing the *Unreliability Rate*.

NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and PopQA (Mallen et al., 2023); (3) *Agentic reasoning*: FRAMES (Krishna et al., 2025), GAIA (Mialon et al., 2023).

Models and Retrieval. We use **Qwen2.5** (Yang et al., 2024b) at three scales (1.5B, 3B, 7B) as backbone models. For retrieval, we use E5 (Wang et al., 2022) as the dense retriever over the 2018 Wikipedia dump (Karpukhin et al., 2020), retrieving top-3 passages per query.

Training. We implement KbPO based on the VERL framework with GRPO as the optimizer. For each query, the policy generates G=5 rollouts with batch size 128 and 8 rollouts per batch. We train for 300 steps with learning rate 1×10^{-6}, KL coefficient β=0.001, and process reward weight λ=0.2. The knowledge boundary thresholds are set to τ_ρ=0.5 and τ_μ=0.6. The agent operates in multi-turn mode with maximum 2 search iterations per query.

Baselines. We compare against three categories of methods: (1) *Standard RAG*: Naive RAG (Lewis et al., 2020), FS-RAG(Trivedi et al., 2023), and FLARE (Jiang et al., 2023); (2) *Agentic RAG*: ReAct (Yao et al., 2023), IRCoT (Trivedi et al., 2023), Self-RAG (Asai et al., 2024b), and CRAG (Yan et al., 2024); (3) *RL-based Agents*: Search-R1 (Jin et al., 2025), ARPO (Dong et al., 2025b), AEPO (Dong et al., 2025a), and MEM1 (Zhou et al., 2025).

Evaluation. **Evaluation Metrics.** We assess performance on two dimensions: (1) **Task Utility**: **F1** scores (2) **Alignment Gap**: We define **Unreliability Rate** $= 1 - (\text{F1} + \text{Refusal Rate})$ to quantify "unsafe" behaviors where the agent neither answers correctly nor refuses faithfully. Here, **Refusal Rate** denotes the proportion of responses where the LLM explicitly generates a refusal.

4.2 Main Results

Table 1 presents the Failure Rate across all benchmarks, where lower values indicate better reliability (correct answers or appropriate refusals). We highlight several key findings.

Overall Performance. KbPO consistently achieves the lowest failure rate across all model scales. On Qwen2.5-7B, KbPO obtains 57.99%, outperforming Search-R1 (69.32%) by 11.33%. Similar improvements hold at smaller scales: 3B (64.32% vs. 72.22%, −7.9%) and 1.5B (70.69% vs. 76.38%, −5.69%). The improvement margin increases with model scale, suggesting larger models better leverage knowledge boundary signals.

Knowledge Boundary Calibration. The largest gains appear on benchmarks requiring precise calibration. On TriviaQA with Qwen2.5-7B, KbPO achieves 32.60% failure rate versus Search-R1 (44.09%) and FL-RAG (57.34%). On NQ, KbPO-7B (68.92%) outperforms Search-R1 (70.41%) and substantially beats AEPO (90.07%). This improvement stems from KbPO's ability to refuse unknown queries rather than hallucinate.

Out-of-Domain Generalization. Despite training only on HotpotQA and 2WikiMQA, KbPO generalizes well to unseen datasets. On PopQA, KbPO-7B achieves 67.80% versus Search-R1's 72.53%. On GAIA, KbPO (71.90%) substantially outperforms FS-RAG (94.98%). This transferability suggests that knowledge boundary awareness is a domain-agnostic skill rather than dataset-specific pattern fitting.

Comparison with RL-based Methods. KbPO shows consistent advantages over AEPO, ARPO, and MEM1. Critically, AEPO and ARPO exhibit inverse scaling at 7B: AEPO worsens from 77.02% (3B) to 88.88% (7B). This likely results from reward hacking in outcome-only training. KbPO's process-level alignment reward prevents such exploitation.

Scaling Behavior. KbPO exhibits stable and monotonic improvement as model capacity increases: 70.69% (1.5B) → 64.32% (3B) → 57.99% (7B), with gains of 6.37% and 6.33% per scale jump. This near-linear improvement contrasts sharply with baselines showing diminishing returns or degradation. Search-R1, while competitive at 7B, shows inconsistent scaling at smaller models (85.28% at 1.5B vs. 72.22% at 3B). The stability of KbPO's scaling indicates that boundary-aware rewards provide reliable optimization signals across model capacities.

4.3 Ablation Study

We conduct ablation studies on Qwen2.5-3B to analyze key design choices.

Effect of Reward Components. Table 2 presents ablation results on reward design. Removing process reward causes the largest degradation on multi-hop datasets (2WikiMQA −11.8, HotpotQA −7.0), where intermediate reasoning steps are critical. The alignment reward shows complementary benefits: without boundary-aware calibration, TriviaQA suffers the most (−15.6), indicating that knowledge boundary awareness is essential for open-domain generalization.

Dataset	KbPO	w/o Proc.	w/o Align.
TriviaQA	**44.0**	31.2 (−12.8)	28.4 (−15.6)
2WikiMQA	**26.7**	14.9 (−11.8)	23.2 (−3.5)
HotpotQA	**29.0**	22.0 (−7.0)	24.2 (−4.8)
NQ	**21.6**	14.0 (−7.6)	14.4 (−7.2)
PopQA	**19.1**	13.9 (−5.2)	13.7 (−5.4)
Average	**28.1**	19.0 (−9.1)	20.8 (−7.3)

Table 2: Ablation on reward components (F1, %).

Effect of Quadrant Rebalancing. Table 3 shows the impact of Q3/Q4 rebalancing on error rates. Over-rewarding internal answering (Q4) risks hallucination, while over-rewarding refusal (Q3) sacrifices helpfulness. Proper rebalancing reduces error rates substantially: TriviaQA (−30.9%), NQ (−13.0%), 2WikiMQA (−11.8%). This confirms that explicit calibration between confidence and behavior is essential for hallucination mitigation.

Training Stability. We adopt warmup scheduling for the LLM-based alignment evaluator, using higher sampling rates during early training. This improves stability—TriviaQA F1 increases from

Dataset	w/o Rebal.	KbPO
TriviaQA	94.9	**64.0** (−30.9)
NQ	99.1	**86.1** (−13.0)
2WikiMQA	89.3	**77.5** (−11.8)
HotpotQA	86.5	**78.3** (−8.2)
PopQA	96.4	**87.1** (−9.3)

Table 3: Effect of quadrant rebalancing (Error rate %, ↓).

24.6 to 40.2 across training stages. We also find that high fixed rewards for

4.4 Analysis

To further validate KbPO's reliability, we analyze the relationship between task difficulty and refusal behavior on Qwen2.5-3B. Table 4 presents F1 scores and refusal rates across all benchmarks.

Adaptive Refusal Behavior. KbPO exhibits difficulty-adaptive refusal rates that correlate with task complexity. On benchmarks where the model has sufficient knowledge, refusal rates remain low: 2WikiMQA (2.8%), PopQA (7.8%), and Bamboogle (8.8%). In contrast, on challenging benchmarks requiring complex reasoning or rare knowledge, refusal rates increase substantially: MuSiQue (25.6%), FRAMES (25.1%), and XBench (45.0%). This adaptive pattern demonstrates that quadrant-based alignment successfully teaches the model to recognize its knowledge boundaries.

Correlation Between Difficulty and Calibration. We observe a clear negative correlation between F1 scores and refusal rates. TriviaQA achieves the highest F1 (44.0%) with a moderate refusal rate (16.1%), while XBench shows F1 of 0.0% but the highest refusal rate (45.0%). Similarly, FRAMES and MuSiQue have low F1 scores (12.3% and 13.0%) but elevated refusal rates (25.1% and 25.6%). This pattern confirms that KbPO has learned calibrated uncertainty: when both parametric knowledge and retrieval are insufficient, the model appropriately abstains rather than hallucinating.

Balancing Helpfulness and Reliability. Importantly, high refusal rates on difficult benchmarks do not compromise helpfulness on tractable tasks. On in-domain benchmarks (2WikiMQA, HotpotQA), KbPO maintains competitive F1 scores (26.7% and 29.0%) while keeping refusal rates below 15%. On TriviaQA, where retrieval is effective, KbPO

Dataset	F1 (%)	Refusal (%)
In-Domain		
2WikiMQA	26.7	2.8
HotpotQA	29.0	13.4
Out-of-Domain		
TriviaQA	44.0	16.1
NQ	21.6	9.9
PopQA	19.1	7.8
Bamboogle	23.1	8.8
Challenging		
MuSiQue	13.0	25.6
FRAMES	12.3	25.1
GAIA	5.3	17.5
XBench	0.0	45.0

Table 4: F1 and refusal rate on Qwen2.5-3B. Higher refusal rates on challenging benchmarks indicate effective knowledge boundary calibration.

achieves strong F1 (44.0%) with only 16.1% refusal. This balance validates that our quadrant-based reward successfully differentiates between scenarios requiring confident answers versus cautious abstention.

5 Conclusion and Future work

We presented KbPO, a novel reinforcement learning framework that addresses the critical calibration gap between parametric knowledge and external retrieval in agentic search. Our approach makes three key contributions: we introduce semantic stability metrics to formally quantify the knowledge boundary of LLMs, enabling fine-grained characterization of parametric reliability; we propose a four-quadrant cognitive taxonomy that disentangles internal certainty from retrieval quality, providing principled guidance for optimal behavior under different knowledge states; and we design a boundary-consistent alignment reward that incentivizes the policy to make calibrated decisions aligned with its actual knowledge states. Extensive experiments across ten benchmarks spanning multi-hop QA, open-domain QA, and agentic reasoning tasks demonstrate that KbPO consistently outperforms strong baselines including both traditional RAG methods and recent RL-based approaches. Notably, our ablation studies reveal that the quadrant rebalancing mechanism significantly reduces hallucination rates, confirming that explicit boundary calibration is essential for building trustworthy retrieval-augmented agents.

Our analysis shows that KbPO learns difficulty-

adaptive refusal behavior that transfers across domains, suggesting knowledge boundary awareness is a fundamental cognitive skill rather than dataset-specific fitting. Future work could extend the quadrant taxonomy to multi-agent collaboration and tool-augmented settings, develop boundary quantification methods to reduce training overhead, and explore dynamic adaptation of knowledge boundaries as models acquire new information.

Acknowledgment

This work was supported by the National Key Research and Development Project of China (2021ZD0110505), the Zhejiang Provincial Key Research and Development Project (2023C01043), Key R&D Program of Zhejiang Province 2026C01016 and Academy Of Social Governance Zhejiang University.

Limitations

LLM Judge Dependency. KbPO relies on an external LLM judge for computing alignment rewards during training. This introduces computational overhead from API calls and potential evaluation noise due to sampling variability. While our warmup scheduling mitigates this issue, the dependence on external evaluation may limit scalability and reproducibility in resource-constrained settings.

Ground-Truth Requirement. The knowledge boundary quantification (Section 3.2) computes parametric certainty μ by comparing model responses against ground-truth answers. This requirement limits applicability to supervised settings where gold labels are available, and may not directly extend to fully unsupervised or open-ended generation scenarios.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024a. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *ICLR*.

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

Baolong Bi, Shenghua Liu, Yiwei Wang, Yilong Xu, Junfeng Fang, Lingrui Mei, and Xueqi Cheng. 2025. Parameters vs. context: Fine-grained control of knowledge reliance in language models. *arXiv preprint arXiv:2503.15888*.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *Preprint*, arXiv:2407.21787.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.

Guanting Dong, Licheng Bao, Zhongyuan Wang, Kangzhi Zhao, Xiaoxi Li, Jiajie Jin, Jinghan Yang, Hangyu Mao, Fuzheng Zhang, Kun Gai, and 1 others. 2025a. Agentic entropy-balanced policy optimization. *arXiv preprint arXiv:2510.14545*.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, and 1 others. 2025b. Agentic reinforced policy optimization. *arXiv preprint arXiv:2507.19849*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *Preprint*, arXiv:2404.16130.

James D. Evans. 1996. *Straightforward Statistics for the Behavioral Sciences*. Brooks/Cole.

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

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference. *Preprint*, arXiv:2301.00303.

Xanh Ho, Anh-Khoa Duong, Quoc-Huy Nguyen, and Suong Nguyen. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *COLING*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Ziyang Huang, Xiaowei Yuan, Yiming Ju, Jun Zhao, and Kang Liu. 2025b. Reinforced internal-external knowledge synergistic reasoning for efficient adaptive search agent. *arXiv preprint arXiv:2505.07596*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Xinke Jiang*, Yue Fang*, Rihong Qiu*, Haoyu Zhang, Yongxin Xu, Hao Chen, Wentao Zhang, Ruizhe Zhang, Yuchen Fang, Xu Chu, and 1 others. 2025a. Tc-rag: Turing-complete rag's case study on medical llm systems. *ACL oral 2025*.

Xinke Jiang*, Rihong Qiu*, Yongxin Xu*, Wentao Zhang, Yichen Zhu, Ruizhe Zhang, Yuchen Fang, Xu Chu, Junfeng Zhao, and Yasha Wang. 2024. Ragraph: A general retrieval-augmented graph learning framework. *NeurIPS 2024*.

Xinke Jiang*, Ruizhe Zhang*, Yongxin Xu*, Rihong Qiu*, Yue Fang, Zhiyuan Wang, Jinyi Tang, Hongxin Ding, Xu Chu, Junfeng Zhao, and 1 others. 2025b. Hykge: A hypothesis knowledge graph enhanced framework for accurate and reliable medical llms responses. *ACL 2025*.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwyer, and Mohit Iyyer. 2023. Active retrieval augmented generation. *EMNLP*. (FLARE/FL-RAG).

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Mandar Joshi and 1 others. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

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

Tom Kwiatkowski and 1 others. 2019. Natural questions: A benchmark for question answering research. *TACL*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474.

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

Chenyu Lin, Yilin Wen, Du Su, Fei Sun, Muhan Chen, Chenfu Bao, and Zhonghou Lv. 2025a. Knowledgeable-r1: Policy optimization for knowledge exploration in retrieval-augmented generation. *arXiv preprint arXiv:2506.05154*.

Tzu-Han Lin, Wei-Lin Chen, Chen-An Li, Hung-yi Lee, Yun-Nung Chen, and Yu Meng. 2025b. Adasearch: Balancing parametric knowledge and search in large language models via reinforcement learning. *arXiv preprint arXiv:2512.16883*.

Tao Liu, Jiafan Lu, Bohan Yu, Pengcheng Wu, Liu Haixin, Guoyu Xu, Li Xiangheng, Lixiao Li, Jiaming Hou, Zhao Shijun, Xinglin Lyu, Kunli Zhang, Yuxiang Jia, and Hongyin Zan. 2026. Iesr:efficient mcts-based modular reasoning for text-to-sql with large language models. *Preprint*, arXiv:2602.05385.

Alex Mallen, Akari Asai, and 1 others. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *ACL*.

Potsawee Manakul, Adian Liusie, and Mark 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, pages 9004–9017.

Grégoire Mialon and 1 others. 2023. Gaia: A benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, and 1 others. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv 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. *arXiv preprint arXiv:2210.03350*.

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

Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. 2025. Detecting and mitigating reward hacking in reinforcement learning systems: A comprehensive empirical study. *arXiv preprint arXiv:2507.05619*.

Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. 2025. Agentic retrieval-augmented generation: A survey on agentic rag. *arXiv preprint arXiv:2501.09136*.

Haotian Sun, Yuchen Zhuang, Lingkai Kong, Bo Dai, and Chao Zhang. 2023. Adaplanner: Adaptive planning from feedback with language models. *Advances in neural information processing systems*, 36:58202–58245.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multihop reasoning dataset with explanation. *arXiv 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.

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*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Yongxin Xu*, Ruizhe Zhang*, Xinke Jiang*, Yujie Feng, Yuzhen Xiao, Xinyu Ma, Runchuan Zhu, Xu Chu, Junfeng Zhao, and Yasha Wang. 2025. Parenting: Optimizing knowledge selection of retrieval-augmented language models with parameter decoupling and tailored tuning. *ACL 2025*.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024b. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Zhilin Yang, Peng Qi, Saizheng Zhang, and 1 others. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Shunyu Yao and 1 others. 2023. React: Synergizing reasoning and acting in language models. In *ICLR*.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024a. R-tuning: Instructing large language models to say 'i don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7106–7132.

Ruizhe Zhang, Xinke Jiang, Zhibang Yang, Zhixin Zhang, Jiaran Gao, Yuzhen Xiao, Hongbin Lai, Xu Chu, Junfeng Zhao, and Yasha Wang. 2026. Stackplanner: A centralized hierarchical multi-agent system with task-experience memory management. *ACL 2026*.

Ruizhe Zhang, Yongxin Xu, Yuzhen Xiao, Runchuan Zhu, Xinke Jiang, Xu Chu, Junfeng Zhao, and Yasha Wang. 2024b. Knowpo: Knowledge-aware preference optimization for controllable knowledge selection in retrieval-augmented language models. *AAAI 2024*.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. 2025. MEM1: Learning to synergize memory and reasoning for efficient long-horizon agents. *arXiv preprint arXiv:2506.15841*.

Runchuan Zhu*, Xinke Jiang*, Jiang Wu*, Zhipeng Ma, Jiahe Song, Fengshuo Bai, Dahua Lin, Lijun Wu, and Conghui He. 2025. Grait: Gradient-driven refusal-aware instruction tuning for effective hallucination mitigation. *NAACL 2025*.

Runchuan Zhu, Zhipeng Ma, Jiang Wu, Junyuan Gao, Jiaqi Wang, Dahua Lin, and Conghui He. 2025. Utilize the flow before stepping into the same river twice: Certainty represented knowledge flow for refusal-aware instruction tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26157–26165.

A Reproducibility Details

We implement our experiments using the PyTorch-based VERL framework on a high-performance computing cluster. Below we detail the exact configurations to ensure full reproducibility.

A.1 Model Configuration

We evaluate the Qwen2.5 language model family (Yang et al., 2024b) across three scales: 1.5B, 3B, and 7B. All models are initialized from the official HuggingFace checkpoints and loaded in `bfloat16` precision.

A.2 Inference & Retrieval Settings

Table 5 lists the hyperparameters used during the inference phase. We use a greedy decoding strategy to ensure deterministic evaluation results. For retrieval, we use the E5-Base-V2 dense retriever indexed on the Wikipedia 2018 dump.

Category	Parameter	Value
Generation	Max Prompt Length	4096 tokens
	Max Response Length	3000 tokens
	Temperature (Eval)	0.0 (Greedy)
	Temperature (Train)	0.7
Retrieval	Retriever Model	E5-Base-V2
	Knowledge Source	Wikipedia (2018)
	Top-k Passages	3
	Truncation	Middle (256 toks)

Table 5: Detailed inference and retrieval configuration parameters.

A.3 Training Hyperparameters

We utilize Group Relative Policy Optimization (GRPO) without a critic model. The training is conducted on 8×NVIDIA A100 (80GB) GPUs using Fully Sharded Data Parallel (FSDP). Ubuntu 22.04.5 LTS. Thesoftware environment was based on Python 3.11.11 withConda 23.5.2, and experiments were implemented usingPyTorch 2.6.0, HuggingFace Transformers 4.51.3, andSpaCy 3.8.4, all with default settings unless otherwisespecifed.The full hyperparameter set is provided in Table 6.

Category	Hyperparameter	Value
Optimization	Optimizer	AdamW
	Learning Rate	1×10^{-6}
	LR Scheduler	Cosine Decay
	Warmup Ratio	0.285
	Global Batch Size	512
PPO / GRPO	Group Size (G)	5
	KL Coefficient (β)	0.001
	Clip Threshold (ϵ)	0.2
	Advantage Norm.	Token-level
Rewards	Process Weight (λ)	0.2
	Alignment Scale (α)	1.0

Table 6: Complete training hyperparameters for KbPO experiments.

A.4 Computational Efficiency

KbPO's training overhead remains comparable to strong RL-based baselines (e.g., Search-R1, Mem1) across all evaluated model scales (1.5B, 3B, 7B). The additional cost introduced by LLM-based reward evaluation and group-based optimization accounts for less than 5% of total training time, effectively amortized through batched evaluation and asynchronous scheduling. At inference time, KbPO introduces no additional modules or computational steps beyond the standard retrieval-reasoning-generation pipeline, resulting in latency identical to all baselines.

B Reward Function Details

The KbPO reward function is composed of three distinct components: (1) Knowledge Boundary Thresholds, (2) Process Rewards, and (3) Alignment Rewards.

B.1 Knowledge Boundary Definition

We quantify the model's internal knowledge state using two metrics derived from consistency checks.

- **Parametric Certainty** (μ): Measures the correctness frequency across N sampled responses without retrieval.
- **Semantic Stability** (σ): Measures the average pairwise cosine similarity between response embeddings.

The thresholds for these metrics (τ_ρ, τ_μ) were determined by probing the validation set to maximize the separation between known and unknown queries (Table 7).

Metric	Symbol	Threshold
Retrieval Quality Score	τ_ρ	≥ 0.5
Parametric Certainty	τ_μ	≥ 0.6
Semantic Stability	τ_σ	(Implicit via μ)
Sampling Count	N	5

Table 7: Thresholds defining the four knowledge quadrants.

B.2 Threshold Sensitivity Analysis

The default configuration τ_ρ=0.5, τ_μ=0.6 was validated via a systematic sensitivity study on Qwen2.5-3B. Lowering τ_ρ below 0.5 incorrectly admits low-quality retrieval as valid external knowledge, expanding the knowledge boundary and inducing blind retrieval reliance and hallucination. Raising it above 0.5 rejects genuinely useful evidence, forcing over-reliance on parametric knowledge and degrading performance on open-domain benchmarks. The same asymmetric pattern holds for τ_μ: a lenient threshold misclassifies semantically unstable knowledge as reliable, causing over-confident answering on ambiguous samples, while an overly strict threshold triggers excessive refusals on moderately confident but genuinely known queries. The chosen thresholds achieve optimal balance between boundary under-expansion and over-expansion, enabling precise four-quadrant calibration across all decision modes.

B.3 Process Reward Specification

To encourage structured reasoning and correct tool usage, we assign dense rewards at specific syntactic milestones. These rewards are purely rule-based and do not require an external reward model.

Type	Event	Description	Score
Positive	TOOL_OK	Valid JSON format	+0.10
	ANS_TAG	Output contains `<answer>`	+0.10
Negative	TOOL_FAIL	JSON parsing error	−0.05
	NO_ANS	Missing final answer tag	−0.30
Neutral	DEFAULT	Generic step completion	+0.05

Table 8: Fine-grained process reward scalar values.

B.4 Process Reward Assignment

We assign process rewards to the last token of each action span rather than distributing them uniformly. This design is motivated by three considerations: (1) natural compatibility with sequence-to-scalar reward models; (2) strict causal credit assignment consistent with the autoregressive factorisation; and (3) lower gradient variance relative to token-level schemes. Credit propagation to earlier tokens is handled internally by GRPO via group-normalised advantages applied uniformly across the sequence. Empirically, last-token assignment yields consistently lower Failure Rates than uniform distribution across all model scales, suggesting that concentrated reward signals afford more efficient optimisation by avoiding gradient dilution.

B.5 Quadrant-Based Alignment Matrix

The alignment reward r^{align} drives the policy towards the optimal strategy $c^*(s, k)$ defined by the knowledge quadrants.

State		Prescribed Strategy (c^*)	Reward
s (Ret.)	k (Int.)		
High (1)	High (1)	**Integrated**: Synergize retrieval with memory	+1.0
High (1)	Low (0)	**External**: Rely strictly on retrieved context	+1.0
Low (0)	High (1)	**Internal**: Ignore noise, trust memory	+0.2
Low (0)	Low (0)	**Refusal**: State "I don't know"	+0.5
Penalties for Misalignment			
1	*	Ignoring valid retrieval	−0.5
0	0	Hallucination (answering unknown)	−1.0

Table 9: The full alignment reward matrix. s: Retrieval Signal, k: Internal Knowledge Signal.

C Data Format Specification

Each training instance follows a structured JSONL format. The schema encodes the query, ground truth answer, and pre-computed knowledge boundary signals (μ, σ) for supervision.

C.1 System Prompt

The system prompt defines the agent's action space, output format, and fallback behavior when search results are insufficient.

Training Instance Schema

```
{
  "data_source": "KbPO",
  "prompt": [
    {"role": "system", "content": "<system_prompt>"},
    {"role": "user", "content": "Question: <query>"}
  ],
  "reward_model": {
    "ground_truth": ["<answer>"],
    "style": "rule"
  },
  "extra_info": {
    "question": "<query>",
    "prior_mu": 0.0,
    "prior_sigma": 0.77
  }
}
```

Figure 3: Training data JSON schema. The `prior_mu` and `prior_sigma` fields encode pre-computed knowledge boundary signals used for data partitioning.

Agent System Prompt

```
You are a QA assistant with a search tool.

If you need to search, output exactly ONE tool call:
<tool_call>
  {"name":"search","arguments":{"query_list":["q1","q2"]}}
</tool_call>

After the tool response, write:
<summary>what you learned from the search</summary>

Then write the final answer:
<answer>your answer</answer>

Rules:
- JSON inside <tool_call> MUST be valid.
- In a tool-call turn, output ONLY <tool_call>.
- If results not helpful: <answer>I don't know</answer>
- Always end with <answer>...</answer>.

Example:
Q: When did Xanadu hold its last festival?
A: <tool_call>
     {"name":"search",
      "arguments":{"query_list":["Xanadu festival"]}}
   </tool_call>
[no relevant results found]
<summary>No reliable information found.</summary>
<answer>I don't know</answer>
```

Figure 4: System prompt defining the agent's action space.

D LLM Judge System Prompt

To ensure reproducible evaluation, we employ a specific system prompt for the LLM judge (e.g., GPT-4o) to classify the agent's behavior into the four knowledge quadrants. The full prompt template is provided below on 5.

D.1 Judge Consistency Analysis

To mitigate potential single-model bias in our LLM-as-Judge evaluation, we validated cross-judge consistency across three large-scale LLMs: DeepSeek-V3.2, GPT-4o, and Qwen2.5-72B-Instruct, using the full training set of 35,583 samples and the standardised prompt in Figure 5. All pairwise Pearson correlation coefficients between judges exceed 0.87 on both retrieval quality scoring and behavioral classification, indicating strong cross-model agreement and low risk of systematic bias (Evans, 1996). In the final version, we further adopt a multi-LLM ensemble strategy: retrieval quality scores are averaged across all three judges, and behavioral categories are determined by majority voting, improving the robustness of alignment reward computation.

Judge System Prompt

```
You are an expert evaluator for a RAG (Retrieval-Augmented Generation) agent.
Your task is to evaluate the agent's "Summary" based on the "User Question"
and "Retrieved Context".

You must perform two independent tasks:
1. Score Retrieval Quality: How useful is the context?
2. Classify Agent Behavior: How did the agent use the information?

### TASK 1: RETRIEVAL QUALITY SCORING
Determine if the Retrieved Context contains the necessary information to
answer the User Question. Assign a score between 0.0 and 1.0:
* 0.0 (Irrelevant): Context is completely unrelated or empty.
* 0.5 (Partial): Context contains partial info but is insufficient.
* 1.0 (Relevant): Context contains core information needed to answer.

### TASK 2: BEHAVIOR CLASSIFICATION
Determine the source of information for the Model Summary.
Choose exactly ONE category (1, 2, 3, or 4):

* Category 1: Integrated (External + Internal)
  The summary combines details from the Retrieved Context with internal
  reasoning. It cites facts from text BUT adds correct external details.

* Category 2: External Only (Reliance)
  The summary relies almost exclusively on the Retrieved Context.
  It is a direct paraphrase. If search results are wrong, summary is wrong.

* Category 3: Refusal (No Info)
  The summary explicitly states that information is missing, not found,
  or insufficient (e.g., "I don't know", "Search results do not contain").

* Category 4: Internal Only (Memory/Hallucination)
  The summary answers using only internal knowledge, ignoring the Context.
  Note: If context is empty/bad but model answers confidently, it falls here.

### OUTPUT FORMAT
You must output the result in the following XML-like format:
<retrieval_score>SCORE</retrieval_score>
<class>CATEGORY_ID</class>
```

Figure 5: System prompt for the LLM judge to compute retrieval quality ($\hat{\rho}$) and classify agent behavior ($\hat{c}$).

E Ethical Considerations

This work focuses on improving the reliability of retrieval-augmented language models by teaching them to recognize knowledge boundaries. All experiments are conducted on publicly available QA benchmarks (HotpotQA, TriviaQA, NQ, etc.), and no personally identifiable information or human subject data is involved.

A key contribution of KbPO is training models to refuse answering when both parametric knowledge and retrieval are insufficient, rather than hallucinating plausible but incorrect responses. We believe this calibrated behavior represents a positive step toward more trustworthy AI systems. However, we note that refusal rates should be carefully balanced—excessive refusal may reduce system utility, while insufficient refusal may propagate misinformation.

Our method relies on an LLM-based judge for evaluating alignment during training, which may introduce evaluation biases. We encourage future work to develop more robust and transparent evaluation protocols. Additionally, while KbPO reduces hallucination on our benchmarks, it should not be interpreted as a guarantee of factual correctness in deployment scenarios.

This research is developed solely for academic purposes. Any real-world application, particularly in high-stakes domains such as healthcare, legal, or financial services, would require extensive validation, domain-specific calibration, and human oversight.

F Case Study

We present representative examples from each knowledge quadrant to illustrate KbPO's behavior. Each case demonstrates how the agent adapts its strategy based on the availability of internal knowledge and retrieval quality.

F.1 Q1: Synergistic Integration

When both internal knowledge and retrieval are reliable, the agent integrates them for accurate reasoning.

F.2 Q2: External Dependent

When internal knowledge is insufficient but retrieval succeeds, the agent relies on external evidence.

F.3 Q3: Appropriate Refusal

When both internal knowledge and retrieval fail, the agent should express uncertainty rather than hallucinate.

F.4 Q4: Internal Knowledge Suffices

When retrieval fails but internal knowledge is reliable, the agent can answer from parametric memory.

G The use of Large Language Models

This research utilized Large Language Models (LLMs) exclusively as assistive tools for linguistic polishing and code troubleshooting. Specifically, their role was confined to improving grammatical precision and readability, alongside providing general programming support. The authors maintained full oversight, rigorously verifying all AI-generated content. The core intellectual contributions—including the research conception, methodological design, and data analysis—remain solely the work of the authors, with no AI involvement in hypothesis generation or the formulation of conclusions.

Q1: Both Knowledge Sources Available

Question: Which film has the director who died first, The Admiral Was A Lady or The Whales Of August?
Gold Answer: The Admiral Was A Lady
Model Answer: The Admiral Was A Lady ✓

```
<tool_call>
{"name": "search", "arguments": {"query_list":
  ["director who died first The Admiral Was A Lady",
   "director who died first The Whales Of August"]}}
</tool_call>

<tool_response>
{"result": "Doc 1 (Title: \"The Admiral Was a Lady\")
The Admiral Was a Lady is a 1950 American comedy film..."}
</tool_response>

<summary>The director of "The Admiral Was A Lady" was Albert S. Rogell,
who died first compared to the director of "The Whales Of August".</summary>
<answer>The Admiral Was A Lady</answer>
```

Figure 6: **Q1 (High Internal + High External):** The agent formulates a well-structured multi-part query and successfully retrieves relevant information about both directors. Combined with background knowledge about film history, it correctly identifies the answer through synergistic reasoning.

Q2: Retrieval Corrects Initial Error

Question: Who died later, Haldor Lillenas or Holly Broadbent Sr.?
Gold Answer: Holly Broadbent Sr.
Initial Guess (before search): Haldor Lillenas ✗
Final Answer (after search): Holly Broadbent Sr. ✓

```
<tool_call>
{"name": "search", "arguments": {"query_list":
  ["when did haldor lillenas die", "when did holly broadbent sr. die"]}}
</tool_call>

<tool_response>
{"result": "Doc 1 (Title: \"Haldor Lillenas\")
...Lillenas died on August 18, 1959..."}
</tool_response>

<summary>Haldor Lillenas died on 1959-08-18.</summary>

<tool_call>
{"name": "search", "arguments": {"query_list":
  ["when did holly broadbent sr. die"]}}
</tool_call>

<tool_response>
{"result": "Doc 1 (Title: \"Holly Broadbent Sr.\")
...died on January 12, 2001..."}
</tool_response>

<summary>Holly Broadbent Sr. died on 2001-01-12.</summary>
<answer>Holly Broadbent Sr.</answer>
```

Figure 7: **Q2 (Low Internal + High External):** The agent lacks reliable internal knowledge about these obscure individuals and initially guesses incorrectly. However, through iterative retrieval, it obtains precise death dates and corrects its answer. This demonstrates appropriate reliance on external evidence when parametric knowledge is insufficient.

Q3: Expressing Uncertainty Under Conflicting Information

Question: Who is the child of the director of film Evidence (1988 Film)?
Gold Answer: Jishnu
Model Answer: I don't know (expressing uncertainty)

```
<tool_call>
{"name": "search", "arguments": {"query_list":
  ["child of director of film evidence 1988"]}}
</tool_call>

<tool_response>
{"result": "Doc 1 (Title: \"Evidence (musician)\")
Michael Taylor Perretta (born December 10, 1976)..."}
</tool_response>

<summary>The available information is insufficient to answer this question.</summary>
<answer>I don't know </answer>
```

Figure 8: **Q3 (Low Internal + Low External):** The retrieval returns irrelevant results (a musician, not the film director's child), and the agent lacks internal knowledge about this obscure 1988 film. Rather than confidently hallucinating, the agent expresses uncertainty using 'I don't know'

Q4: Overcoming Noisy Retrieval with Internal Knowledge

Question: Which film whose director is younger, Vadakakku Oru Hridayam or The $1,000,000 Reward?
Gold Answer: Vadakakku Oru Hridayam
Model Answer: Vadakakku Oru Hridayam ✓

```
<tool_call>
{"name": "search", "arguments": {"query_list":
  ["director age Vadakakku Oru Hridayam",
   "director age The $1,000,000 Reward"]}}
</tool_call>

<tool_response>
{"result": "Doc 1 (Title: \"Medical Director of the Year Award\")
The Medical Director of the Year Award..."}
</tool_response>

<summary>The director of Vadakakku Oru Hridayam is not mentioned in the
search results, but the retrieved content is irrelevant (medical awards).</summary>
<answer>Vadakakku Oru Hridayam</answer>
```

Figure 9: **Q4 (High Internal + Low External):** The retrieval completely fails, returning irrelevant content about medical awards instead of film directors. However, the agent recognizes the retrieval noise and falls back on internal knowledge about these films, ultimately providing the correct answer. This demonstrates appropriate trust in parametric knowledge when external sources are unreliable.