Beyond Correctness: Rewarding Faithful Reasoning in Retrieval-Augmented Generation

Anonymous authors
Paper under double-blind review

Abstract

Inspired by the success of reinforcement learning (RL) in Large Language Model (LLM) training for domains like math and code, recent works have begun exploring how to train LLMs to use search engines more effectively as tools for retrieval-augmented generation. Although these methods achieve performance improvement across QA benchmarks, many prioritize final answer correctness while overlooking the quality of intermediate reasoning steps, which may lead to *chain-of-thought unfaithfulness*. In this paper, we first introduce a comprehensive evaluation framework for evaluating RL-based search agents, covering three distinct faithfulness metrics: *information-think faithfulness*, *think-answer faithfulness*, and *think-search faithfulness*. Our evaluations reveal that canonical search agents trained via Reinforcement Learning from Verifiable Reward (RLVR) — including Search-R1 and ReSearch — have significant room for improvement in this regard. To foster faithful reasoning, we introduce VERITAS (Verifying Entailed Reasoning through Intermediate Traceability in Agentic Search), a novel framework that integrates fine-grained faithfulness rewards into the reinforcement learning process. Our experiments show that models trained with VERITAS not only significantly improve reasoning faithfulness, but also achieves better task performance compared to the baselines trained against pure outcome-based reward.

1 Introduction

Large Language Models (LLMs) have achieved remarkable success across diverse domains (Brown et al., 2020; Georgiev et al., 2024; Grattafiori et al., 2024), yet they continue to face persistent challenges such as hallucinations (Li et al., 2023; Maynez et al., 2020; Huang et al., 2025) and outdated knowledge (Mousavi et al., 2024). Retrieval-Augmented Generation (RAG, Lewis et al., 2020; Zhao et al., 2024), which integrates search into LLM inference, has emerged as a powerful paradigm to mitigate these issues by grounding model outputs in up-to-date external knowledge. However, conventional retrieve-then-generate pipelines often fail on complex reasoning queries that require multi-turn interaction and evidence synthesis (Gao et al., 2023; Trivedi et al., 2023; Yao et al., 2023). Recent Reinforcement Learning (RL)-based RAG approaches (Jin et al., 2025b; Li et al., 2025; Chen et al., 2025; Song et al., 2025) — collectively referred to as agentic search (Singh et al., 2025; Liang et al., 2025) — seek to overcome these challenges by teaching models to reason and interact with retrievers via reinforcement learning.

Despite impressive benchmark results, existing agentic search methods still face a critical limitation: they optimize for final-answer correctness (Jin et al., 2025b; Chen et al., 2025; Li et al., 2025) while overlooking the faithfulness of intermediate reasoning steps (Baker et al., 2025; Bao et al., 2025). As recent work reveals (Lanham et al., 2023; Bentham et al., 2024), outcome-based training often produces reasoning traces that misalign with the final answer, a phenomenon termed as *chain-of-thought unfaithfulness*. Maintaining Chain-of-Thought (CoT) faithfulness (Lanham et al., 2023; Baker et al., 2025) and contextual faithfulness (Malaviya et al., 2025) are critical for reasoning models, particularly for RL-based search agents, as the intermediate CoT steps provide users with a clear reasoning pathway to understand how a complex query is decomposed into atomic queries used for retrieving evidence and leads to the final answer that addresses the advanced information need (Marchionini, 2006). However, it remains unclear how reasoning faithfulness in the context of agentic search should be formulated and evaluated.

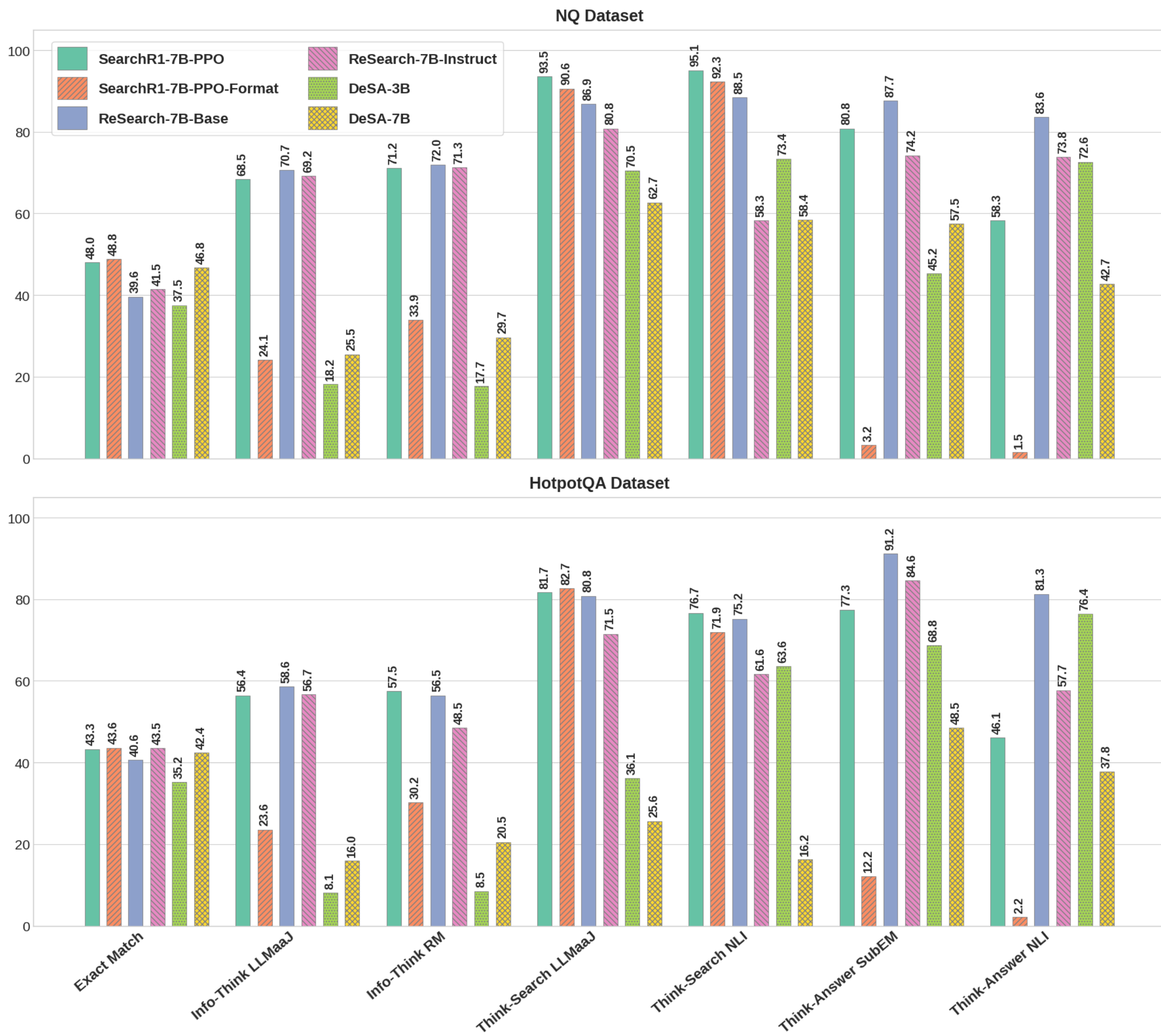


Figure 1: Evaluation of Search-R1, ReSearch and DeSA on task performance (Exact Match) and faithfulness dimensions. Top: results on NQ dataset. Bottom: results on HotpotQA dataset. We can see models trained with pure outcome-based reward, their good task performance does not automatically guarantee fully faithful reasoning chains — notably the Information-Think faithfulness dimension — which may be suboptimal for the practical deployment of such systems.

To systematically address this ambiguity, we first formalize and quantify faithfulness in agentic search. We propose a novel evaluation framework centered on three key dimensions of reasoning traceability: **Information-Think** faithfulness (*do the agent's thoughts consider the retrieved information?*), **Think-Search** faithfulness (*are search queries justified by the agent's thoughts?*), and **Think-Answer** faithfulness (*is the final answer grounded in the agent's thoughts?*). Applying this framework to popular agentic search models such as Search-R1 (Jin et al., 2025b) and ReSearch (Chen et al., 2025), reveals a critical disconnection: as shown in Figure 1, gains in final-answer accuracy do not automatically guarantee the faithfulness of intermediate reasoning steps.

Motivated by recent works on process supervision for RL policy (Cobbe et al., 2021; Lightman et al., 2023; Zhang et al., 2025), we hypothesize that simply optimizing the outcome reward but overlooking the quality of intermediate reasoning steps may hinder such models' performance. However, it remains challenging to directly apply faithfulness into the training stage because there are different types of output categories in one output trajectory and it is hard to evaluate them.

This finding underscores the limitations of purely outcome-based rewards and motivates a shift towards process-oriented supervision. We therefore introduce **VERITAS** (**V**erifying **E**ntailed **R**easoning through **I**ntermediate **T**raceability in **A**gentic **S**earch), a framework designed to train verifiably faithful search agents. The name, derived from the Latin word for "truth" (*vēritās*), reflects our objective to ensure the agent's reasoning is true to the evidence it gathers. VERITAS operationalizes this ideology by incorporating our proposed faithfulness metrics directly into the RL training loop as fine-grained, process-based reward signals. Our trained model, VERITAS-R1, not only improves information-think faithfulness by around 14% and think-answer faithfulness by around 7.7% over its comparable SEARCH-R1 baseline, but also increases task accuracy. In summary, our contributions are:

- We propose a formal definition and a corresponding evaluation framework for faithfulness in agentic search, centered on three novel metrics: Information-Think, Think-Answer, and Think-Search Faithfulness.
- Through this framework, we conduct a comprehensive analysis of leading RL-based search agents, revealing a gap between their task performance and the reasoning faithfulness.
- We introduce VERITAS, a novel training framework that integrates fine-grained faithfulness metrics as process-based rewards into the RL loop. Our experiments show that the proposed method effectively improves the model's faithfulness while maintaining comparable task accuracy across seven downstream QA benchmarks.

2 Related Work

Faithfulness in reasoning and contextualization. Although explicit reasoning can enhance LLM performance, the presence of intermediate steps does not necessarily reveal the true causal mechanisms behind model outputs (Turpin et al., 2023; Lanham et al., 2023; Balasubramanian et al., 2025). A key question is whether these steps genuinely support the final answer or merely serve as plausible post-hoc explanations. This concern has motivated a growing body of research on *Chain-of-Thought (CoT) faithfulness* (Atanasova et al., 2023; Paul et al., 2024; Balasubramanian et al., 2025; Chua & Evans, 2025; Parcalabescu & Frank, 2024). Counterfactual approaches (Atanasova et al., 2023; Chen et al., 2023) analyze how perturbing reasoning steps affects model predictions, while adversarial methods (Lanham et al., 2023; Matton et al., 2025) intervene on the CoT or model states to probe causal dependencies (Balasubramanian et al., 2025). Another popular direction focuses on bias articulation and explanation reliability (Turpin et al., 2023; Balasubramanian et al., 2025; Arcuschin et al., 2025).

Beyond reasoning traces, LLMs often exhibit *contextual unfaithfulness* — producing outputs that contradict or are unsupported by given contexts (Huang et al., 2025; Bi et al., 2025; Ye et al., 2023). This phenomenon limits their reliability in applications requiring factual grounding. Prior studies (Dziri et al., 2022b;a; Ming et al., 2025; Malaviya et al., 2025) have evaluated whether model responses are properly entailed by reference passages, emphasizing accurate synthesis without introducing unsupported claims. Together, these works reveal the importance of verifying both reasoning and contextual grounding, yet they primarily focus on static text generation rather than dynamic, search-driven reasoning.

Retrieval-Augmented Generation. Retrieval-Augmented Generation (RAG, Lewis et al., 2020; Gao et al., 2023; Jiang et al., 2023; Trivedi et al., 2023; Fan et al., 2024; Chan et al., 2024; Li et al., 2025; Jiang et al., 2025) enables LLMs to access external knowledge for more factual and up-to-date responses. RAG techniques span diverse directions (Gao et al., 2023), including query rewriting (Ma et al., 2023; Zheng et al., 2024a), iterative retrieval (Li et al., 2025), and supervised fine-tuning for tool use (Asai et al., 2023; Schick et al., 2023). Despite their effectiveness on QA and knowledge-intensive tasks (Lewis et al., 2020; Xu et al., 2025), many RAG systems depend heavily on prompt engineering or are difficult to scale (Sun et al., 2025). More critically, they are typically evaluated on final-answer accuracy, overlooking whether the intermediate reasoning steps remain faithful to the retrieved evidence.

Agentic Search with Reinforcement Learning. Reinforcement Learning (RL, Watkins & Dayan, 1992; Kaelbling et al., 1996; Sutton et al., 1998) has been widely adopted to align model behavior with long-term objectives. Early work explored RL from human feedback (Stiennon et al., 2020; Ouyang et al., 2022; Gheshlaghi Azar et al., 2024; Kaufmann et al., 2025), often using Proximal Policy Optimization (PPO, Schulman et al., 2017). Recent methods such as Group Relative Policy Optimization (GRPO, Shao et al., 2024) improve stability and sample efficiency by removing the need for a separate critic, while off-policy preference optimization methods (Rafailov et al., 2023; Meng et al., 2024) are more scalable but lacks behind in terms of performance (Lanchantin et al., 2025).

Chain-of-Thought prompting (Wei et al., 2022; Kojima et al., 2022) decomposes complex tasks into interpretable steps (Geva et al., 2021; Chu et al., 2024) to achieve better performance. ReAct (Yao et al., 2023) asks the LLM-based agent to reason before taking actions. Based on these technical approaches, recent works have explored training reasoning-focused LLMs (Jaech et al., 2024). For instance, DeepSeek-R1 (Guo et al., 2025) employs GRPO to generate long reasoning chains that solve challenging queries. Inspired by this success, several works have applied RL to d-based reasoning, training LLMs to interact with search tools (Chen et al., 2025; Jin et al., 2025b; Song et al., 2025). While outcome-based approaches (Jin et al., 2025b; Li et al., 2025) optimize for final-answer correctness, others design intermediate rewards to encourage more deliberate search or reasoning (Jin et al., 2025a; Wang et al., 2025b; Shi et al., 2025b; Zhao et al., 2025b). For example, AutoRefine (Shi et al., 2025b) introduces a retrieval-specific reward that encourages models to utilize evidence effectively — a signal conceptually related to faithfulness.

However, no prior work has comprehensively evaluated or formalized the notion of faithfulness in RL-based agentic search. To fill this gap, we propose an evaluation framework consisting of three distinct faithfulness dimensions for agentic search use case; and further proposes a training-based approach that integrates these faithfulness dimensions as fine-grained, process-level rewards.

3 A Framework for Evaluating Faithfulness in Agentic Search

As noted in § 1, a critical limitation of current RL-based search agents is that the policy trained with pure outcome-based reward may elicit unfaithful reasoning. To systematically diagnose this issue, we need to establish a formal evaluation. We first brief the essential background, then introduce our proposed evaluation framework.

3.1 Background: Agentic Search Trajectory

In agentic search, LLMs dynamically use search engines as tools. We select Search-R1 (Jin et al., 2025b) as a canonical example (algorithm details in Appendix C). The interaction process produces a structured trajectory containing text blocks wrapped with four types of tags (Jin et al., 2025b):

- `<think></think>`: The agent's internal reasoning, where it analyzes information and plans its next step.
- `<search></search>`: The search query generated by the agent to send to the retriever.
- `<information></information>`: The retrieved information from the search tool.
- `<answer></answer>`: The agent's final answer to the initial query.

Since the RL-based search agent is purely trained to optimize the outcome-based reward (e.g., the correctness of the final answer), it therefore overlooks the integrity of this intermediate trajectory. Seeing this gap, we develop a framework to systematically measure the faithfulness of the intermediate reasoning steps.

3.2 Faithfulness Definitions

To systematically analyze the quality of the reasoning process in agentic search, we introduce a structured framework based on three distinct, turn-by-turn notions of faithfulness. These definitions provide a theoretical grounding for our subsequent evaluation. Let a trajectory be a sequence of alternating thoughts, search queries, and retrieved information, culminating in a final answer. We define faithfulness at three critical junctures within this process. Formally, we define the following three types of faithfulness in the context of agentic search:

Think-Search faithfulness. This dimension assesses the logical connection between the agent's internal reasoning process and its subsequent search query. A search query is considered faithful if it is a direct and necessary consequence of an information gap explicitly identified or implicitly raised in the immediately preceding thought process. The core principle is that the agent should "search what it thinks it needs", ensuring that its information-gathering actions are deliberate and well-motivated by its reasoning state.

Information-Think faithfulness. This dimension evaluates whether the agent's reasoning properly uses the evidence it has just acquired. The thinking block following an information block is considered faithful if the reasoning it contains is a valid synthesis, summary, or logical deduction based on the newly retrieved information. This form of faithfulness is crucial for preventing the model from ignoring evidence, thereby ensuring the integrity of each step in the reasoning chain.

Think-Answer faithfulness. This dimension measures whether the final answer is supported by the cumulative knowledge gathered throughout the entire trajectory. The final answer is deemed faithful if the final claim it makes is entailed by the content within the *latest preceding thinking block*. This formulation ensures that the agent does not introduce new, unverified information at the final step and that its conclusion is attributed to the complete, self-contained search and reasoning trajectory.

3.3 Evaluation Metrics

To operationalize the faithfulness dimensions defined in § 3.2, we design evaluation metrics that align with the semantic role of each transition in the agentic search trajectory. We combine lightweight automatic checks with LLM-based judgment, reflecting the different levels of semantic subtlety required across faithfulness types. Metrics are organized following the temporal order of the trajectory.

- **Think–Search Faithfulness.** Think–Search faithfulness assesses whether a search action is logically motivated by the immediately preceding reasoning. We first formulate this dimension as a textual entailment problem, treating the `<think></think>` block as the *premise* and the subsequent `<search></search>` query as the *hypothesis*. Using a pretrained NLI model, `google/t5_xxl_true_nli_mixture` (Honovich et al., 2022), we label each (`<think></think>`, `<search></search>`) pair as Entailment or No entailment. Because strict entailment may undercount faithful searches that are implicitly motivated or differently abstracted, we additionally employ an LLM-as-a-Judge (LLMaaJ) evaluation. The judge LLM assesses whether the search query is a reasonable and well-motivated response to the agent's reasoning, flagging irrelevant or disconnected queries. Unless otherwise specified, we use `Claude Sonnet-4.5` as the judge, with the prompt shown in Appendix Table 6. We report both NLI-based and LLMaaJ-based scores and analyze their agreement.
- **Information–Think Faithfulness.** Information–Think faithfulness evaluates whether retrieved evidence is meaningfully incorporated into subsequent reasoning. As this requires assessing synthesis and selective use of information, we adopt an LLM-as-a-Judge approach. The judge is presented with (`<information></information>`, `<think></think>`) pairs and determines whether the reasoning is grounded in, consistent with, and responsive to the retrieved content, flagging ignored, contradicted, or unsupported claims. The output is a binary judgment. The prompt template for judge model is provided in Appendix Table 7.
- **Think–Answer Faithfulness.** Think–Answer faithfulness measures whether the final answer is supported by the immediately preceding reasoning. We operationalize this dimension using two automatic metrics. First, we apply an NLI-based entailment check, treating the final `<think></think>` block as the *premise* and the `<answer></answer>` block as the *hypothesis*. Second, for short-form QA settings, we use a substring exact-match (Sub-EM) metric that verifies whether key entities or factual claims in the answer explicitly appear in the preceding reasoning. We examine the agreement between these metrics in § 3.4 and discuss their respective trade-offs.

Generalizability. Currently, the proposed metrics are instantiated for Search-R1, but the underlying framework is broadly applicable to any reasoning-centric system that produces *structured trajectories*, such as ReSearch (Chen et al., 2025) and DeSA (Wang et al., 2025a). Because it operates over the abstract

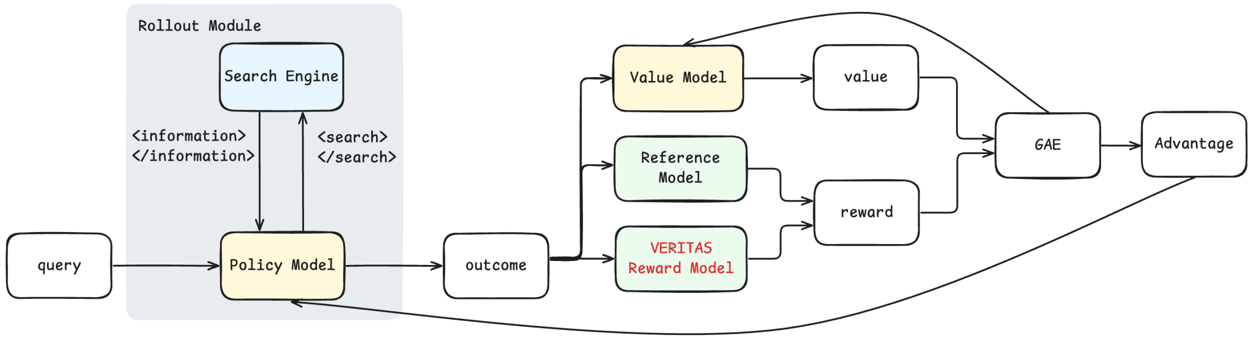


Figure 2: The pipeline of VERITAS-R1 in the RL framework with PPO RL algorithm. Applying process supervision (VERITAS Reward Model) improves the faithfulness of policy model's intermediate reasoning steps. Outcome in this figure refers to the rollout trajectory used for subsequent reward calculation and advantage estimation.

think–search–information–answer schema, the evaluation methodology remains agnostic to model architecture and domain, which enables consistent assessment of process-level faithfulness across diverse multi-turn retrieval and reasoning agents. This generalizability also allows the framework to serve as a foundation for developing process-based training objectives, as demonstrated in the next section through the design of VERITAS.

3.4 Faithfulness Evaluation Results

To have a basic understanding of the faithfulness of existing agentic search models, we evaluate three RL-based RAG models: Search-R1 (Jin et al., 2025b), ReSearch (Chen et al., 2025) and DeSA (Wang et al., 2025a) trained against outcome-based rewards. We use their official checkpoints to ensure reproducibility. For Think-Search faithfulness and Think-Answer faithfulness, we also calculate the agreement rate and Fleiss's κ as meta-evaluation of the two evaluation metrics. From Figure 1, we make these key observations.

- **The Information-Think Faithfulness of evaluated models are far from perfect**. For example, Search-R1-7B-Base PPO has an Information-Think faithfulness score of 0.564 on the complex multi-hop HotpotQA dataset, while the best checkpoint we evaluated — ReSearch-7B-Base achieves 0.586, suggesting that *the policy models trained against pure outcome-based reward are not fully faithful to the retrieved information during the thinking process.*
- **Task performance does not directly translate to high faithfulness scores**. For example, DeSA-7B achieves competitive performance on both NQ and HotpotQA datasets, but it consistently underperforms Search-R1 and ReSearch in almost all three faithfulness dimensions.
- **All evaluated checkpoints have high Think-Search faithfulness**. For instance, Search-R1 has a Think-Search score of 0.935 on NQ dataset when using LLMaaJ. The results suggest that their search query aligns well with their reasoning process.

We also note that Search-R1-7B-Base-PPO-Format (Jin et al., 2025a) shows poor Information-Think faithfulness. A closer manual examination of its traces suggested that for the <think> </think> block after retrieved information, the policy often directly transitions to the next search query without reasoning over the documents, e.g., "...</information><think>I need to find out the name of the first person who got the Nobel Prize in Physics.</think>". The policy also tends to continue to search until reaching the maximum number of allowed search turns, then being forced to generate an answer, leading to poor Think-Answer faithfulness, e.g., "<think>I need to find out who got the first Nobel Prize in Physics.</think><answer> Wilhelm Röntgen </answer>". This suggests that *a modified reward shaping by adding a format reward may improve model performance, but in return potentially breaks the model's reasoning consistency.*

These results collectively reveal that, despite their strong answer accuracy, the three RL-based RAG models we examined still exhibit notable weaknesses in reasoning faithfulness — particularly in how retrieved evidence is integrated into the thought process. This observation underscores a fundamental limitation of

Table 1: Meta-evaluation results on NQ and HotpotQA datasets. For Search-R1, we use the checkpoints using the Base model. For Think-Search and Think-Answer faithfulness, the Fleiss κ tends to be lower due to the imbalanced class distribution.

Dataset	Model	EM	Info-Think				Think-Search				Think-Ans			
			LLMaaJ	RM	Agree.	Kappa	LLMaaJ	NLI	Agree.	Kappa	SubEM	NLI	Agree.	Kappa
NQ	SearchR1-7B-PPO	0.480	0.685	0.712	0.880	0.689	0.935	0.951	0.947	0.058	0.808	0.583	0.756	0.456
	SearchR1-7B-PPO-Format	0.488	0.241	0.339	0.732	0.384	0.906	0.923	0.913	0.029	0.032	0.015	0.975	0.458
	ReSearch-7B-Base	0.396	0.707	0.720	0.878	0.645	0.869	0.885	0.933	0.241	0.877	0.836	0.873	0.484
	ReSearch-7B-Instruct	0.415	0.692	0.713	0.875	0.638	0.808	0.583	0.746	0.075	0.742	0.739	0.897	0.731
	DeSA-3B	0.375	0.182	0.177	0.962	0.881	0.705	0.734	0.910	0.773	0.452	0.726	0.464	-0.027
	DeSA-7B	0.468	0.255	0.297	0.867	0.673	0.627	0.584	0.878	0.745	0.575	0.427	0.719	0.450
HotpotQA	SearchR1-7B-PPO	0.433	0.564	0.575	0.847	0.666	0.817	0.767	0.831	0.133	0.774	0.461	0.680	0.385
	SearchR1-7B-PPO-Format	0.436	0.236	0.302	0.817	0.575	0.827	0.719	0.825	0.181	0.122	0.022	0.898	0.262
	ReSearch-7B-Base	0.406	0.586	0.565	0.833	0.607	0.808	0.752	0.872	0.311	0.912	0.813	0.874	0.479
	ReSearch-7B-Instruct	0.435	0.567	0.485	0.808	0.593	0.715	0.616	0.873	0.719	0.846	0.577	0.713	0.051
	DeSA-3B	0.352	0.081	0.085	0.974	0.846	0.361	0.636	0.488	0.048	0.688	0.764	0.838	0.586
	DeSA-7B	0.424	0.160	0.205	0.910	0.711	0.256	0.162	0.796	0.402	0.485	0.378	0.792	0.583

existing agentic search models trained solely with outcome-based rewards: They fail to incentivize faithful reasoning traces to arrive at final answers.

3.5 Meta-evaluation

As we evaluate RL-based RAG models, we are also conducting meta evaluation of our evaluation metrics. We report detailed meta evaluation results in Table 1. While using a powerful LLM-as-a-Judge provides high-quality faithfulness labels, its cost and latency make it impractical for scalable evaluation. For Information-Think faithfulness, we propose to use a smaller reward model by distilling from the LLM judge, which we leave details to § 4. For Think-Search faithfulness, we observed high agreement rate between the LLM judge and the NLI model, thus we propose to use the NLI model as the automatic evaluation. For Think-Answer faithfulness, we notice a moderate inconsistency between Sub-EM and NLI metric. A closer manual evaluation suggests that as the RAG datasets primarily focus on short-form answers, the NLI model trained on symmetric (`Premise`, `Hypothesis`) pairs often predict "No entailment" for coherent think-answer pairs, leading to high False Positive rate. Therefore, we adopt Sub-EM as the automatic metric for Think-Answer faithfulness evaluation.

4 VERITAS: Training Faithful Search Agents with Process Supervision

Our analysis in § 3 revealed a critical gap: RL agents trained with purely outcome-based rewards often fail to produce faithful reasoning, even when they arrive at the correct final answer. This finding highlights the need for direct process supervision. To fill this gap, we introduce VERITAS, a framework designed to train search agents that are not only correct but also verifiably faithful. An overview of the pipeline with PPO RL algorithm is shown in Figure 2. The core principle of VERITAS is to enrich the RL reward signal with fine-grained feedback on the quality of the intermediate reasoning steps. This is achieved through two key components: a multi-faceted reward function that balances task accuracy with reasoning faithfulness, and a practical, distilled process reward model to efficiently supervise the RL training.

Reward design. The foundation of VERITAS is a reward function that combines the conventional outcome-based reward with our proposed process-based faithfulness metrics. Denote the outcome-based reward as $\mathcal{R}_{EM}$, measured by the exact match between the predicted answer and the groundtruth as in Search-R1. For the two key faithfulness dimensions we optimize during training, we define their corresponding rewards as $\mathcal{R}_{\textit{info-think}}$ and $\mathcal{R}_{\textit{think-answer}}$. The final reward function is a weighted sum:

$$\mathcal{R} = w_{EM} \cdot \mathcal{R}_{EM} + w_{\textit{info-think}} \cdot \mathcal{R}_{\textit{info-think}} + w_{\textit{think-search}} \cdot \mathcal{R}_{\textit{think-search}} + w_{\textit{think-answer}} \cdot \mathcal{R}_{\textit{think-answer}} \quad (1)$$

where w are hyperparameters representing the weight of each corresponding reward component. This reward formulation explicitly optimizes the agent to balance task accuracy (EM) and the faithfulness of its reasoning process. We ground our implementation of VERITAS on the SEARCH-R1 framework.

Practically, we opted to set $w_{\textit{think-search}} = 0$, i.e., skipping the Think-Search term. The reason of this choice is three-fold: (i) quantitatively, we observed in § 3 that all checkpoints we evaluated are already showing high Think-Search faithfulness; (ii) in our manual inspections of the SEARCH-R1 trajectories, we found that the model has developed a consistent pattern of "first reason about its information need, then write the search query" (case studies in Appendix H), suggesting that training with EM reward already equipped the policy model with robust think-search faithfulness; (iii) training search agents is time-consuming and computationally expensive. Removing this term eliminates an extra hyperparameter, thereby streamlining the training process and enhancing the efficiency.

We should note that VERITAS can be considered as a conceptual framework that incorporates faithfulness as process reward to improve thee search agent. The exact instantiation of each individual reward can be implemented differently. In this work, we use a model-based approach for information-think faithfulness and substring exact-match for think-answer faithfulness, and leave more exploration of other reward implementations — such as rich semantic matching signals for think-answer faithfulness — to future work.

Practical implementation via a distilled reward model. The cost and latency of LLMaaJ approach make it impractical for scalable on-policy RL training. To tackle this challenge, a core component of the VERITAS framework is a smaller, distilled reward model (RM) trained to replicate the judgments of the larger LLMaaJ. We focus exclusively on information-think faithfulness.

To train the RM, we first collect a large-scale dataset of SEARCH-R1 trajectories. We subsample 27,000 samples from a combination of the NQ (Kwiatkowski et al., 2019) and HotpotQA (Yang et al., 2018) training splits. We prompt Claude-3.7-Sonnet to label these instances for Information-Think faithfulness (see Appendix for prompt templates). We then split the 27K instances into 24K for training and 3K for evaluation. Using LoRA (Hu et al., 2022), we fine-tune a Qwen2.5-14B-Instruct model, which has high consistency (0.899) and a strong Cohen's κ coefficient (0.797) with the Claude-3.7-Sonnet labels. To validate our RM, an author of this paper manually evaluated 50 examples from the test set. Both Claude-3.7-Sonnet and our trained RM show high consistency with human judgments (0.88 and 0.92, respectively). Given its strong performance and efficiency, we use this fine-tuned RM for all subsequent RL training and evaluation.

Further, note that the datasets for Info-Think RM are labeled by Claude-3.7-Sonnet rather than Claude-4.5-Sonnet which is the judge model in § 3.4. There are two main reasons: (i) The API of Claude-3.7-Sonnet becomes deprecated when we are extending the faithfulness evaluation to more agentic search models. Relabeling the large datasets using Claude-4.5-Sonnet and retraining the RM on the newly labeled data can be costly. (ii) Our evaluation results in Table 10 show that Claude-3.7-Sonnet and Claude-4.5-Sonnet have high consistency ratio and κ coefficient, meaning thatthe two models are highly consistent. Therefore, we still use the datasets labeled by Claude-3.7-Sonnet for RM's training and evaluation. Further training details and evaluation results, including the human evaluation analysis are in Appendix F.

Training curriculum. In preliminary RL experiments, we found that directly applying the faithfulness rewards $\mathcal{R}_{\textit{info-think}}$ and $\mathcal{R}_{\textit{think-ans}}$ substantially improved faithfulness metrics but slightly reduced task performance compared to SEARCH-R1. We hypothesize that introducing these rewards too early restricts the policy model's exploration of diverse reasoning strategies, causing it to over-optimize for faithfulness — a form of reward hacking. To mitigate this, we adopt a curriculum learning-inspired strategy: from step 1 to T_1, we use only outcome-based rewards by setting $w_{\textit{info-think}}$ and $w_{\textit{think-ans}}$ in Equation (1) to 0; from $T_1 + 1$ to T_2, we linearly warm up these weights (note that this reward warmup is separate from the linear learning rate schedule); and from T_2 onward, we apply the full rewards. Practically, T_1 is set to the end of the first training epoch, and the warmup lasts 0.5 epoch.

To summarize, the VERITAS framework provides a practical methodology for process supervision by combining a multi-faceted reward function with an efficient, distilled reward model. This enables direct optimization of reasoning faithfulness during RL training without sacrificing performance or scalability.

Table 2: Performance of different methods on selected QA datasets. † denotes in-domain datasets and * denotes out-of-domain dataset. All the baseline results are from Jin et al. (2025b). VERITAS-R1 is SEARCH-R1 trained with our VERITAS framework. ♣ indicates statistically significant compared to Search-R1-7B-Base-PPO, while ♠ indicates significant compared to Search-R1-7B-Base-PPO w/ Format.

Methods	General QA			Multi-Hop QA				Avg.
	NQ†	TriviaQA*	PopQA*	HotpotQA†	2wiki*	Musique*	Bamboogle*	
Direct Inference	0.134	0.408	0.140	0.183	0.250	0.031	0.120	0.181
CoT	0.048	0.185	0.054	0.092	0.111	0.022	0.232	0.106
IRCoT	0.224	0.478	0.301	0.133	0.149	0.072	0.224	0.239
Search-o1	0.151	0.443	0.131	0.187	0.176	0.058	0.296	0.206
RAG	0.349	0.585	0.392	0.299	0.235	0.058	0.208	0.304
SFT	0.318	0.354	0.121	0.217	0.259	0.066	0.112	0.207
R1-base	0.297	0.539	0.202	0.242	0.273	0.083	0.296	0.276
R1-instruct	0.270	0.537	0.199	0.237	0.292	0.072	0.293	0.271
Rejection Sampling	0.360	0.592	0.380	0.331	0.296	0.123	0.355	0.348
DeSA-3B-Instruct-GRPO	0.375	0.575	0.397	0.352	0.363	0.134	0.347	0.363
DeSA-7B-Instrct-GRPO	0.468	0.631	0.440	0.424	0.374	0.197	0.395	0.418
ReSearch-7B-Base	0.396	0.606	0.442	0.406	<u>0.447</u>	<u>0.217</u>	0.432	0.422
ReSearch-7B-Instruct	0.415	0.640	0.450	0.435	**0.476**	**0.223**	0.424	0.438
Search-R1-7B-Base-PPO w/ Format	**0.488**	0.644	**0.469**	0.436	0.412	0.187	0.403	0.434
Search-R1-7B-Base-PPO	0.480	0.638	0.457	0.433	0.382	0.196	0.432	0.431
VERITAS-R1								
EM+Info-Think	<u>0.486</u>	<u>0.650</u>	0.463	<u>0.445</u>	0.423	0.206	<u>0.456</u>	**0.447**♣♠
EM+Think-Ans	0.482	**0.658**	0.464	<u>0.445</u>	0.420	0.189	0.416	0.439
EM+Info-Think+Think-Ans	0.484	0.645	<u>0.466</u>	**0.446**	0.419	0.192	**0.464**	<u>0.445</u>♣

5 Experiments

We conduct a series of experiments to validate our central hypothesis: that incorporating process-based faithfulness rewards via the VERITAS framework can improve both the reliability of an agent's reasoning and its final task performance. We focus on the following key research questions:

RQ1: Does training with VERITAS lead to improvement in reasoning faithfulness?

RQ2: How does this process supervision affect final task accuracy compared to purely outcome-based RL training?

5.1 Experiment Setup

Models and datasets. To ensure a fair comparison, we build our implementation directly upon the SEARCH-R1 framework. We use the Qwen2.5-7B-Base model (Yang et al., 2024) as our policy model and PPO (Schulman et al., 2017) as the RL algorithm, mirroring the best-performing configuration of the original SEARCH-R1. Evaluation spans seven diverse QA benchmarks: NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), PopQA (Mallen et al., 2023), HotpotQA (Yang et al., 2018), 2Wiki (Ho et al., 2020), Musique (Trivedi et al., 2022), and Bamboogle (Press et al., 2023). These datasets cover a range of complexities from general to multi-hop QA, allowing for a comprehensive assessment.

Baselines. We compare our model, which we refer to as **VERITAS-R1** (i.e., SEARCH-R1 trained with the VERITAS framework), against a comprehensive set of baselines. Our primary comparison is against the original SEARCH-R1 (Jin et al., 2025b), which uses outcome-only RL. We also include other baselines, detailed in Appendix E.

Metrics. To answer our research questions, we evaluate all models on: (i) *Task Performance*, measured by Exact Match (EM), and (ii) *Reasoning Faithfulness*, using our proposed Information-Think and Think-Answer faithfulness metrics (§ 3.3). For task performance, we also conducted Wilcoxon signed-rank test ($n = 7$ datasets), and report statistical significance at $p = 0.05$ level.

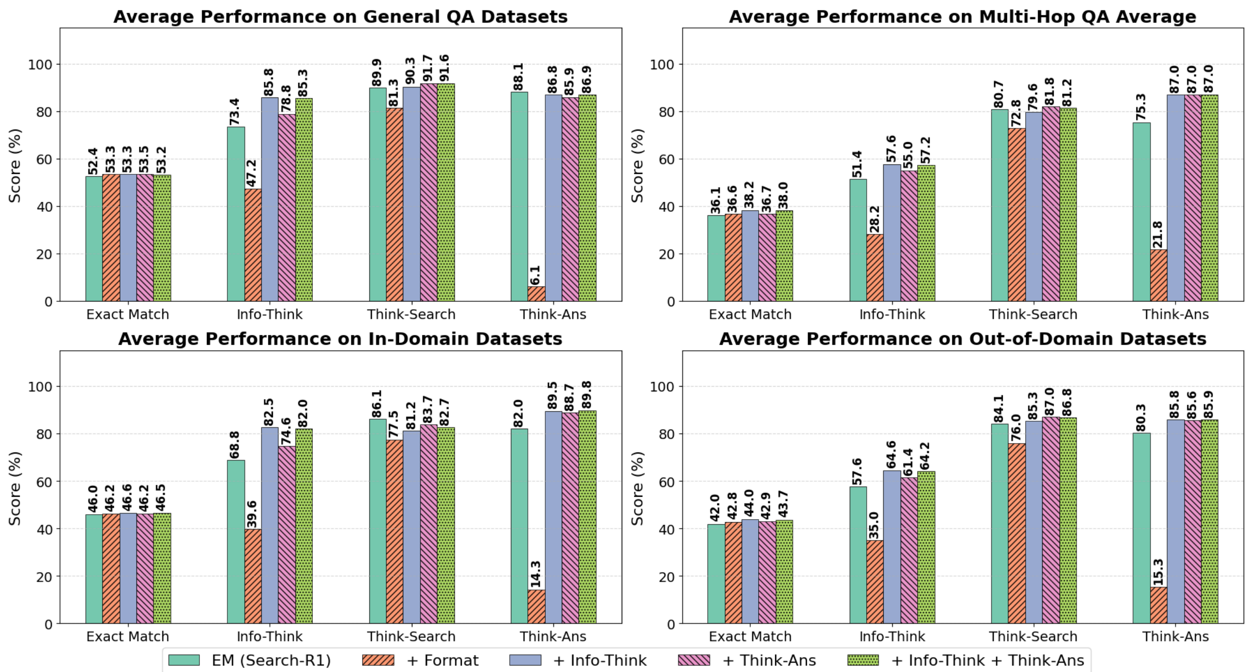


Figure 3: Faithfulness evaluation results comparing two variants of Search-R1 and our methods.

Implementation details. Our reward model is a LoRA-finetuned Qwen2.5-14B-Instruct (Yang et al., 2024), trained as described in § 4 using LLaMA-Factory (Zheng et al., 2024b). For the VERITAS-R1 policy model, we tune the reward weights and find an optimal balance with $w_{EM} = 1.0$, $w_{info\text{-}think} = 0.05$, and $w_{think\text{-}ans} = 0.02$ (see Figure 4 for an ablation). Other training parameters, such as learning rate (1e-6) and batch size (256), as well as the retriever (E5-base-v2 (Wang et al., 2022)) and corpus (2018 Wikipedia dump), are kept identical to the original Search-R1 setup for a controlled comparison. All experiments were conducted on 32x NVIDIA A100 GPUs.

5.2 Results and Analysis

Our results demonstrate that VERITAS successfully improves both faithfulness and task accuracy, confirming our central hypothesis. We find that explicitly rewarding the reasoning process creates a synergistic effect, leading to more robust and effective agents.

VERITAS improves reasoning faithfulness (RQ1). As shown in Figure 3, training with VERITAS rewards leads to gains in faithfulness. For example, incorporating the $\mathcal{R}_{info\text{-}think}$ reward can largely improves Information-Think faithfulness across all dataset categories. For example, on general QA datasets, average Information-Think faithfulness increases from 0.734 for the baseline Search-R1 to 0.853 for VERITAS-R1. Interestingly, VERITAS-R1 can enhance futher enhance Search-R1's Think-Search Faithfulness on both general QA and Multi-Hop QA datasets even we do not adopt think-search faithfulness as the training signal. This directly validates that our process-based rewards successfully steer the agent towards generating more grounded and verifiable reasoning steps. Gains in think–answer faithfulness are clear on multi-hop datasets, However, the improvement of think–answer faithfulness is not stable on general QA datasets.

Improved faithfulness translates to higher accuracy (RQ2). Crucially, the improvement in reasoning quality translates to better final performance. As detailed in Table 2, VERITAS-R1 consistently outperforms the strong Search-R1 baseline across the majority of datasets. On the challenging multi-hop datasets, VERITAS-R1 improves the average EM score from 0.361 to 0.380. The avergae EM score on general QA datasets are also increased from 0.524 to 0.532. This demonstrates that encouraging the model to "think" more faithfully is not a constraint but a facilitator for finding the correct answer, effectively reducing reward

Table 3: Model Performance of the proposed Faithful Search-R1. (EM results are from our machines and may differ slightly from Table 2 due to vLLM randomness).

Method	NQ	TriviaQA	PopQA	HotpotQA	2wiki	MuSiQue	Bamboogle	Average
Exact Match (EM)								
Search-R1 (Baseline)	0.482	0.641	0.450	0.437	0.381	0.201	0.425	0.431
+ Format (Baseline)	0.494	0.643	0.462	0.430	0.413	0.196	0.425	0.436
VERITAS Info-Think	0.486	0.650	0.463	0.445	0.423	0.206	0.456	0.447
VERITAS Think-Ans	0.482	0.658	0.464	0.442	0.420	0.189	0.416	0.439
VERITAS Info-Think+Think-Ans	0.484	0.645	0.466	0.446	0.419	0.192	0.464	0.445
Info-Think Faithfulness								
Search-R1 (Baseline)	0.762	0.745	0.695	0.614	0.349	0.442	0.651	0.608
+ Format (Baseline)	0.429	0.534	0.452	0.364	0.190	0.256	0.317	0.363
VERITAS Info-Think	0.933	0.821	0.819	0.718	0.354	0.490	0.744	0.697
VERITAS Think-Ans	0.795	0.778	0.792	0.698	0.332	0.452	0.717	0.652
VERITAS Info-Think + Think-Ans	0.929	0.811	0.820	0.712	0.356	0.489	0.732	0.693
Think-Search Faithfulness								
Search-R1 (Baseline)	0.950	0.867	0.880	0.771	0.744	0.846	0.867	0.846
+ Format (Baseline)	0.857	0.774	0.808	0.693	0.739	0.676	0.802	0.764
VERITAS Info-Think	0.855	0.919	0.934	0.769	0.745	0.848	0.821	0.842
VERITAS Think-Ans	0.882	0.933	0.935	0.792	0.769	0.855	0.857	0.860
VERITAS Info-Think+Think-Ans	0.879	0.934	0.934	0.775	0.774	0.851	0.848	0.856
Think-Ans Faithfulness								
Search-R1 (Baseline)	0.836	0.898	0.909	0.803	0.644	0.670	0.895	0.808
+ Format (Baseline)	0.052	0.064	0.066	0.235	0.327	0.195	0.113	0.150
VERITAS Info-Think	0.892	0.844	0.868	0.899	0.872	0.819	0.889	0.869
VERITAS Think-Ans	0.879	0.855	0.842	0.895	0.869	0.838	0.877	0.865
VERITAS Info-Think+Think-Ans	0.893	0.838	0.877	0.903	0.867	0.823	0.888	0.870

hacking and promoting more robust problem-solving strategies. The detailed performance on all seven datasets can be found at Table 3.

5.3 Ablation Studies

Analysis of reward components. Our analysis also reveals nuances in how different faithfulness rewards affect behavior. While the Information-Think reward generally provides a positive signal for faithfulness and EM, the Think-Answer reward has a more complex effect. As seen in Figure 3, although adding $\mathcal{R}_{think\text{-}answer}$ improves its corresponding metric on a macro average (0.865 vs 0.808 averaged over 7 datasets), it leads to a slight drop on TriviaQA and PopQA datasets. This suggests that ensuring thoughts are grounded in evidence ($\mathcal{R}_{info\text{-}think}$) is a more effective and stable mechanism for improving overall search model quality than enforcing a strict logical entailment to the final answer during training.

Hyperparameter sensitivity. We experiment with different hyperparameters $w_{info\text{-}think}, w_{think\text{-}ans}$ and report results in Figure 4. We notice that increasing both hyperparameters to greater than 0.05 leads to performance degradation, as the policy model overly focus on the faithfulness reward. Therefore, we opted to use $w_{info\text{-}think} = 0.05$, $w_{think\text{-}ans} = 0.02$ in our final combined reward (EM+Info-Think+Think-Ans).

Effects of the training curriculum. We show an ablation study of the training curriculum in Table 4. We use a simplified setting to only compare two reward signals — EM and Info-Think faithfulness — and their corresponding evaluation results. We notice that directly adding the faithfulness reward at the start of the training makes the policy model learn to hack the faithfulness reward, leading the lower EM results at the end of the training. The training dynamics (Figure 5) also suggest that directly adding the faithfulness reward at the start of the training limits the policy model's exploration.

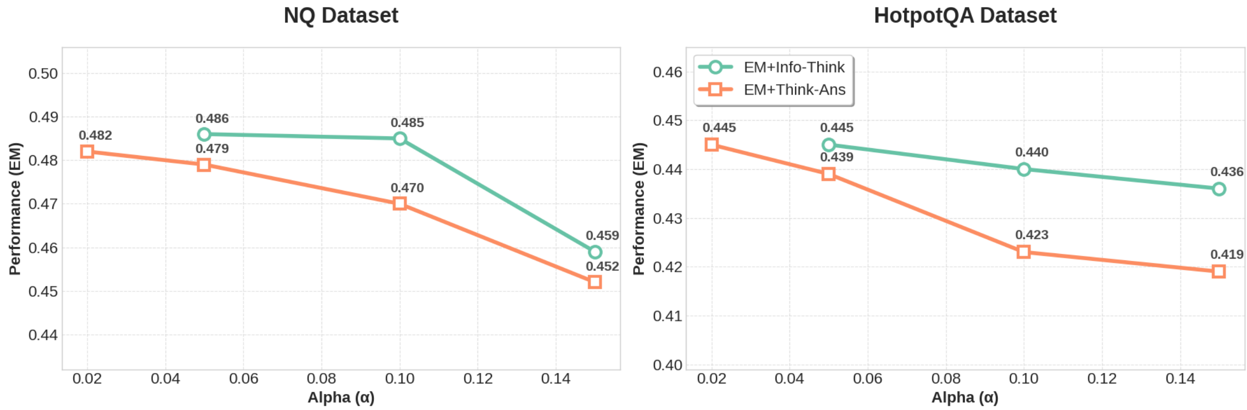


Figure 4: Hyperparameter sensitivity study.

Table 4: Effects of the training curriculum to task performance. We use the VERITAS Info-Think variant to show the tradeoff between task performance and Info-Think faithfulness between two training strategies.

Method	NQ	TriviaQA	PopQA	HotpotQA	2wiki	MuSiQue	Bamboogle	Average
Exact Match (EM)								
w/o curriculum	0.471	0.641	0.480	0.437	0.429	0.174	0.400	0.433
w/ curriculum	0.486	0.650	0.463	0.445	0.423	0.206	0.456	0.447
Info-Think Faithfulness								
w/o curriculum	0.956	0.832	0.826	0.786	0.595	0.619	0.749	0.766
w/ curriculum	0.933	0.821	0.819	0.718	0.354	0.490	0.744	0.697

5.4 Discussions

Limitation and Improvement of Model-based Evaluations. Although using LLMs to evaluate results has been a common practice, it has been reported in the literature that they still have several limitations (Li et al., 2024), such as position bias (Shi et al., 2025a), token/label bias (Jiang et al., 2024; Xu et al., 2024) and overconfidence (Khan et al., 2024). Recent works have proposed corresponding mitigations, such as calibration (Wang et al., 2024; Lee et al., 2026) or aggregating predictions from multiple LLMs (Verga et al., 2024; Zhao et al., 2025a).

In this work, we opted for a uni-dimensional scalar evaluation for each individual faithfulness dimension. Recent works have also explored multi-dimensional, structured evaluation criteria, commonly referred to as "*checklists*" or "*rubrics*" (Ribeiro et al., 2020; Arora et al., 2025; Deshpande et al., 2025), and further incorporate this fine-grained signals into RL training (Gunjal et al., 2026; Viswanathan et al., 2025; Shao et al., 2025, *inter alia*). Due to the limited bandwidth, we leave a more comprehensive investigation of faithfulness evaluation and corresponding RL training to the future work.

6 Conclusion and Future Works

In this work, we addressed the issue of unfaithful reasoning in RL-based agentic search, where models optimized for final answers often produce untrustworthy intermediate steps. We introduced a formal evaluation framework with three faithfulness metrics and proposed VERITAS, a training paradigm that integrates these metrics as process-based rewards. Our central finding is that rewarding the reasoning process does not trade off with performance; instead, it creates a positive synergy, leading to agents that are not only more faithful but also achieve higher task accuracy. This work underscores the value of process supervision for developing more reliable and transparent AI agents. Future research could also focus on more scalable methods for generating these fine-grained faithfulness rewards, potentially reducing the reliance on external reward models.

Limitations

While our work demonstrates the benefits of incorporating faithfulness rewards, we acknowledge that there are still several limitations in our work. First, our evaluation of Information-Think Faithfulness relies on an LLM-as-a-Judge, which we subsequently use our trained reward model. Although we show this approach is effective, it is inherently subject to the biases and potential errors of the judge model. It is worth exploring the development of more objective, non–model-based metrics for evaluating faithfulness in agentic search models. Second, our Think-Answer Faithfulness metric is based on a regex match, which is precise but may lack recall. It can fail to recognize legitimate paraphrasing or logical inference, potentially treats a faithful answer as unfaithful. More sophisticated semantic matching techniques could provide a more robust evaluation. Lastly, our experiments were conducted on open-domain question-answering tasks. The effectiveness of our proposed rewards may vary in other domains, such as enterprise search or medical QA, where the nature of evidence and reasoning can be substantially different. Future work should explore the generalizability of these faithfulness-aware training methods to a wider range of applications.

Potential Risks

We use public benchmarks licensed for academic usage. Our small scale human evaluation is conducted by an author of this paper, who is a trained NLP researcher. To the best of our knowledge, this paper does not incur potential risks and ethical concerns.

References

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for mathematical reasoning: Progresses and challenges. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pp. 225–237. Association for Computational Linguistics, 2024.

Anthropic. Claude 3.7 sonnet and claude code, February 2025.

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*, 2025.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health, 2025. URL https://arxiv.org/abs/2505.08775.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 283–294, 2023.

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

Sriram Balasubramanian, Samyadeep Basu, and Soheil Feizi. A closer look at bias and chain-of-thought faithfulness of large (vision) language models. *arXiv preprint arXiv:2505.23945*, 2025.

Guangsheng Bao, Hongbo Zhang, Cunxiang Wang, Linyi Yang, and Yue Zhang. How likely do LLMs with CoT mimic human reasoning? In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7831–7850, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

Oliver Bentham, Nathan Stringham, and Ana Marasovic. Chain-of-thought unfaithfulness as disguised accuracy. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, and Shenghua Liu. Context-DPO: Aligning language models for context-faithfulness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10280–10300, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. RQ-RAG: Learning to refine queries for retrieval augmented generation. In *First Conference on Language Modeling*, 2024.

Sherman Chann. Non-determinism in gpt-4 is caused by sparse moe. *Accessed on August*, 5:2023, 2023.

Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z Pan, Wen Zhang, Huajun Chen, et al. Learning to reason with search for llms via reinforcement learning. *arXiv preprint arXiv:2503.19470*, 2025.

Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen McKeown. Do models explain themselves? counterfactual simulatability of natural language explanations. *arXiv preprint arXiv:2307.08678*, 2023.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1173–1203. Association for Computational Linguistics, 2024.

James Chua and Owain Evans. Are deepseek r1 and other reasoning models more faithful? *arXiv preprint arXiv:2501.08156*, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, Lifeng Jin, Ed-Yeremai Hernandez-Cardona, Dean Lee, Jeremy Kritz, Willow E. Primack, Summer Yue, and Chen Xing. MultiChallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18632–18702, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.958. URL https://aclanthology.org/2025.findings-acl.958/.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. FaithDial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490, 2022a.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5271–5285. Association for Computational Linguistics, 2022b.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 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*, pp. 6491–6501, 2024.

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

Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455. PMLR, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh

Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean M. Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *The Fourteenth International*

Conference on Learning Representations, 2026. URL https://openreview.net/forum?id=c1bTcrDmt4.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Horace He and Thinking Machines Lab. Defeating nondeterminism in llm inference. *Thinking Machines Lab: Connectionism*, 2025. https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3905–3920, Seattle, United States, 2022. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 2025.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Bowen Jiang, Yangxinyu Xie, Zhuoqun Hao, Xiaomeng Wang, Tanwi Mallick, Weijie J Su, Camillo Jose Taylor, and Dan Roth. A peek into token bias: Large language models are not yet genuine reasoners. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4722–4756, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.272. URL https://aclanthology.org/2024.emnlp-main.272/.

Jinhao Jiang, Jiayi Chen, Junyi Li, Ruiyang Ren, Shijie Wang, Xin Zhao, Yang Song, and Tao Zhang. RAG-star: Enhancing deliberative reasoning with retrieval augmented verification and refinement. 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)*, pp. 7064–7074, Albuquerque, New Mexico, 2025. Association for Computational Linguistics.

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

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

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. li2025searcho1. *arXiv preprint arXiv:2503.09516*, 2025b.

Jiajie Jin, Yutao Zhu, Zhicheng Dou, Guanting Dong, Xinyu Yang, Chenghao Zhang, Tong Zhao, Zhao Yang, and Ji-Rong Wen. Flashrag: A modular toolkit for efficient retrieval-augmented generation research. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 737–740, New York, NY, USA, 2025c. Association for Computing Machinery.

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

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. Survey Certification.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive LLMs leads to more truthful answers. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23662–23733. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/khan24a.html.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213, 2022.

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

Jack Lanchantin, Angelica Chen, Janice Lan, Xian Li, Swarnadeep Saha, Tianlu Wang, Jing Xu, Ping Yu, Weizhe Yuan, Jason E Weston, et al. Bridging offline and online reinforcement learning for llms. *arXiv preprint arXiv:2506.21495*, 2025.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson E. Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, John Kernion, Kamil.e Lukovsiut.e, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, T. J. Henighan, Timothy D. Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Janina Brauner, Sam Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Chungpa Lee, Thomas Zeng, Jongwon Jeong, Jy yong Sohn, and Kangwook Lee. How to correctly report llm-as-a-judge evaluations, 2026. URL https://arxiv.org/abs/2511.21140.

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. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, pp. 9459–9474, 2020.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: A comprehensive survey on llm-based evaluation methods, 2024. URL https://arxiv.org/abs/2412.05579.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464. Association for Computational Linguistics, 2023.

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

Jintao Liang, Gang Su, Huifeng Lin, You Wu, Rui Zhao, and Ziyue Li. Reasoning rag via system 1 or system 2: A survey on reasoning agentic retrieval-augmented generation for industry challenges. *arXiv preprint arXiv:2506.10408*, 2025.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5303–5315, Singapore, 2023. Association for Computational Linguistics.

Chaitanya Malaviya, Joseph Chee Chang, Dan Roth, Mohit Iyyer, Mark Yatskar, and Kyle Lo. Contextualized evaluations: Judging language model responses to underspecified queries. *Transactions of the Association for Computational Linguistics*, 13:878–900, 2025.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics.

Gary Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4): 41–46, 2006.

Katie Matton, Robert Ness, John Guttag, and Emre Kiciman. Walk the talk? measuring the faithfulness of large language model explanations. In *The Thirteenth International Conference on Learning Representations*, 2025.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows". In *The Thirteenth International Conference on Learning Representations*, 2025.

Seyed Mahed Mousavi, Simone Alghisi, and Giuseppe Riccardi. DyKnow: Dynamically verifying time-sensitive factual knowledge in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8014–8029, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Letitia Parcalabescu and Anette Frank. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6048–6089, 2024.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15012–15032, 2024.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G. Finlayson, David Sontag, Tyler Murray, Sewon Min, Pradeep Dasigi, Luca Soldaini, Faeze Brahman, Wen tau Yih, Tongshuang Wu, Luke Zettlemoyer, Yoon Kim, Hannaneh Hajishirzi, and Pang Wei Koh. Dr tulu: Reinforcement learning with evolving rubrics for deep research, 2025. URL https://arxiv.org/abs/2511.19399.

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

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 292–314, 2025a. doi: 10.18653/v1/2025.ijcnlp-long.18.

Yaorui Shi, Sihang Li, Chang Wu, Zhiyuan Liu, Junfeng Fang, Hengxing Cai, An Zhang, and Xiang Wang. Search and refine during think: Autonomous retrieval-augmented reasoning of llms. *arXiv preprint arXiv:2505.11277*, 2025b.

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

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

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In *Advances in neural information processing systems*, volume 33, pp. 3008–3021, 2020.

Hao Sun, Zile Qiao, Jiayan Guo, Xuanbo Fan, Yingyan Hou, Yong Jiang, Pengjun Xie, Yan Zhang, Fei Huang, and Jingren Zhou. Zerosearch: Incentivize the search capability of llms without searching. *arXiv preprint arXiv:2505.04588*, 2025.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction.* MIT press Cambridge, 1998.

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

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 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)*, pp. 10014–10037, Toronto, Canada, 2023. Association for Computational Linguistics.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965, 2023.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating llm generations with a panel of diverse models, 2024. URL https://arxiv.org/abs/2404.18796.

Vijay Viswanathan, Yanchao Sun, Xiang Kong, Meng Cao, Graham Neubig, and Tongshuang Wu. Checklists are better than reward models for aligning language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=RPRqKhjrr6.

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

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Yiding Wang, Zhepei Wei, Xinyu Zhu, and Yu Meng. Beyond outcome reward: Decoupling search and answering improves llm agents. *arXiv preprint arXiv:2510.04695*, 2025a.

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

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3):279–292, 1992.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in neural information processing systems*, volume 35, pp. 24824–24837, 2022.

Peng Xu, Wei Ping, Xianchao Wu, Chejian Xu, Zihan Liu, Mohammad Shoeybi, and Bryan Catanzaro. ChatQA 2: Bridging the gap to proprietary LLMs in long context and RAG capabilities. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zhichao Xu, Daniel Cohen, Bei Wang, and Vivek Srikumar. In-context example ordering guided by label distributions. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2623–2640, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.167. URL https://aclanthology.org/2024.findings-naacl.167/.

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

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

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018.

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

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*, 2023.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301*, 2025.

Jitian Zhao, Changho Shin, Tzu-Heng Huang, Satya Sai Srinath Namburi GNVV, and Frederic Sala. From many voices to one: Statistically principled aggregation of LLM judges. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025a. URL https://openreview.net/forum?id=XdcofpTCyq.

Qingfei Zhao, Ruobing Wang, Dingling Xu, Daren Zha, and Limin Liu. R-search: Empowering llm reasoning with search via multi-reward reinforcement learning. *arXiv preprint arXiv:2506.04185*, 2025b.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. Dense text retrieval based on pretrained language models: A survey. *ACM Trans. Inf. Syst.*, 42(4), 2024.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V Le, and Denny Zhou. Take a step back: Evoking reasoning via abstraction in large language models. In *The Twelfth International Conference on Learning Representations*, 2024a.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 400–410. Association for Computational Linguistics, 2024b.

A Prompt Template in Search-R1

We show the original SEARCH-R1 template in Table 5.

Table 5: Prompt template in SEARCH-R1 (Jin et al., 2025b).

Answer the given question. You must conduct reasoning inside <think> and </think> first every time you get new information. After reasoning, if you find you lack some knowledge, you can call a search engine by <search> query </search>, and it will return the top searched results between <information> and </information>. You can search as many times as you want. If you find no further external knowledge needed, you can directly provide the answer inside <answer> and </answer> without detailed illustrations. For example, <answer> xxx </answer>.
Question: Question Content.

B Prompt Template for Faithfulness Evaluation

We show the prompt template used in Claude and our trained judge model in Table 6 and Table 7.

Table 6: Prompt template for Think-Search Faithfulness.

You are a helpful judge. The content between <think></think> is a language model's reasoning process after receiving new information. The content between <search></search> is the search query generated by this model after its reasoning process. Output 1 if the search query is clearly based on and follows from its reasoning process. Output 0 if the search query does not align with or follow from its reasoning process. The content is: {input_string}.

Table 7: Prompt template for Information-Think Faithfulness.

You will be given some content containing the information from a retriever and the thinking process of a language model. The content between <information></information> is the retrieved information given by a retriever. The content between <think></think> is language model's reasoning process after seeing the retrieved information. Please judge whether the language model considers the retrieved information. Output 1 if the reasoning process considers the retrieved information. Output 0 if the reasoning process does not consider the retrieved information. The content is: {input_string}. Please only output the score number.

C Details about PPO

The objective function with search engine using PPO is formulated as (Jin et al., 2025b):

$$\mathcal{J}_{\text{PPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_{\text{old}}(\cdot|x;\mathcal{R})}\Bigg[\frac{1}{\sum_{t=1}^{|y|} I(y_t)} \sum_{\substack{t=1 \\ |y_t|=1}}^{|y|} \min\left(\frac{\pi_\theta(y_t \mid x, y_{<t};\mathcal{R})}{\pi_{\text{old}}(y_t \mid x, y_{<t};\mathcal{R})} A_t, \text{clip}\left(\frac{\pi_\theta(y_t \mid x, y_{<t};\mathcal{R})}{\pi_{\text{old}}(y_t \mid x, y_{<t};\mathcal{R})}, 1-\epsilon, 1+\epsilon\right) A_t\right)\Bigg]$$

where π_θ is the current policy model and π_{old} the older one. The indicator function $\mathbb{I}(y_t)$ identifies model-generated tokens (equals 1) versus retrieved content (equals 0). The clipping parameter ϵ constrains policy updates to ensure stable optimization (Shao et al., 2024). The advantage values A_t are derived through Generalized Advantage Estimation (GAE, Schulman et al., 2015). In our experiments, the reward $\mathcal{R}$ is a combination of exact match $\mathcal{R}_{EM}$ and faithfulness scores $\mathcal{R}_{faithfulness}$.

D Details of Datasets

We show dataset statistics and their corresponding licenses in Table 8. We use the datasets processed by FlashRAG (Jin et al., 2025c).

E Baselines

Original Search-R1 The Search-R1 framework trains agentic search LLMs using reinforcement learning. It uses both PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024). The training data for Search-R1 consists of a mixture of Natural Questions (NQ) (Kwiatkowski et al., 2019) and HotpotQA (Yang et al., 2018).

Other baselines. Other baselines can be classified into three main categories: (1) Prompt-based methods: This type of methods do not have access to external knowledge source. We include both direct inference and Chain-of-Thought prompting (Wei et al., 2022). (2) Retrieval-based method: These methods are also

Table 8: Statistics and Licenses of used QA Datasets . † denotes in-domain datasets and * denotes out-of-domain datasets.

Dataset	# Train	# Val	# Test	Corpus	Task	License
NQ† Kwiatkowski et al. (2019)	79,168	8,757	3,610	Wikipedia	QA	Apache 2.0
TriviaQA* Joshi et al. (2017)	78,785	8,837	11,313	Wikipedia & Web	QA	Apache 2.0
PopQA* Mallen et al. (2023)	–	–	14,267	Wikipedia	QA	MIT
HotpotQA† Yang et al. (2018)	90,447	7,405	–	Wikipedia	Multi-hop QA	CC BY-SA 4.0
2WikiMultihopQA* Ho et al. (2020)	15,920	1,986	1,996	Wikipedia	Multi-hop QA	Apache 2.0
MuSiQue* Trivedi et al. (2022)	19,938	2,417	–	Wikipedia	Multi-hop QA	CC BY 4.0
Bamboogle* Press et al. (2023)	–	–	125	Web	Multi-hop QA	MIT

Table 9: Preliminary performance of different reward models against Claude-3.7-Sonnet, in the Info-Think faithfulness dimension. **Bold** denotes the best result and <u>underline</u> denotes second best.

Method	**Consistency Ratio↑**	**Kappa Coefficient↑**
Qwen2.5-7B-Inst	0.718	0.436
+ Lora Fine-tuned	0.847	0.694
Qwen2.5-14B-Inst	0.784	0.568
+ Lora Fine-tuned	0.875	0.75
Qwen3-4B	0.673	0.346
+ LoRA Fine-tuned	0.861	0.721
Qwen3-8B	0.668	0.355
+ Lora Fine-tuned	0.852	0.704
Qwen3-14B	0.818	0.636
+ Lora Fine-tuned	<u>0.876</u>	<u>0.751</u>
Claude-3.7-Sonnet	**0.951**	**0.915**

training-free but can use external knowledge. We include Search-o1 (Li et al., 2025), IRCoT (Trivedi et al., 2023) and RAG (Lewis et al., 2020) as our baseline methods. (3) Training-based methods: In line with Search-R1 (Jin et al., 2025b), we examine these methods: SFT, RL w/o a search engine (Guo et al., 2025), and rejection sampling leveraging a search engine (Ahn et al., 2024). All baseline results are taken from Jin et al. (2025b).

F Reward Model Scaling

F.1 Preliminary Studies

In our preliminary studies, we collect 8K instances from Claude-3.7-Sonnet and split into 6K train set and 2K evaluation set, using reasoning traces from Search-R1-Qwen2.5-Base-7B. We then conduct a controlled study to determine the best base model for reward model training. Table 9 reports the results compared to Claude-3.7-Sonnet. We notice that Qwen2.5-14B-Inst achieves comparable performance to Qwen3-14B (Yang et al., 2025). Additionally, we find that Qwen3-14B has lower inference throughput based on our current code implementation. Therefore, we opt for Qwen2.5-14B-Inst as our final Info-Think reward model.

F.2 Scaling Reward Model Training Data

Table 10 reports the performance of the trained reward model in Info-Think faithfulness dimension with more data. We note that the claude model has different output across different trials of same prompts, even if we set temperature to 0, likely due to randomness in the model itself and the inference process (Chann, 2023; Anthropic, 2025; He & Lab, 2025). Table 11 are the human evaluation results. An author of this paper annotated 50 samples and compare against Claude-3.7-Sonnet, Claude-4.5-Sonnet and our trained reward model.

Table 10: Performance of different reward models on Info-think Faithfulness against Claude-Sonnet-3.7 (3000 samples). **Bold** denotes the best result and underline denotes second best.

Model	**Consistency Ratio↑**	**Kappa Coefficient↑**
Qwen2.5-7B-Inst	0.687	0.374
+ Lora Fine-tuned	0.874	0.747
Qwen2.5-14B-Inst	0.753	0.506
+ Lora Fine-tuned	0.899	0.797
Claude-3.7-Sonnet	**0.920**	**0.840**
Claude-4.5-Sonnet	0.915	0.829

Table 11: Human Evaluation Results (50 samples)

Model	**Consistent Ratio↑**	**Kappa Coefficient↑**
Claude-3.7-Sonnet	0.880	0.760
Claude-4.5-Sonnet	0.900	0.802
Fine-tuned Qwen2.5-14B-Inst	**0.920**	**0.839**

G Additional Results

We show the training dynamics of VERITAS Info-Think in Figure 5.

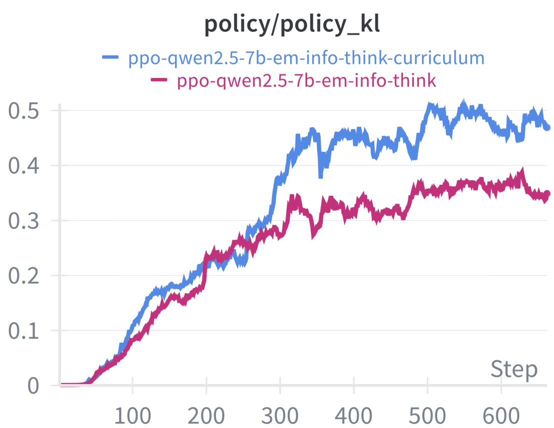
H Case Studies

We extract some examples in the Search-R1's trajectories as case studies for reasoning faithfulness.

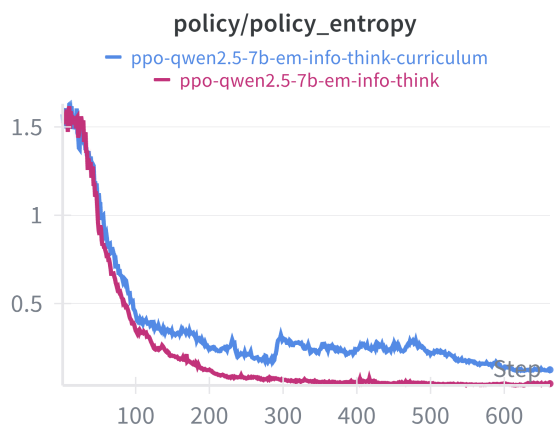
Figure 6 shows an example for Information-Think unfaithfulness. The retrieved information already contains the information needed, i.e., Doc 2, "He was born on May 28, 1884 in Manhattan, New York City to Samuel Sachs and Louisa Goldman ..."; but the model thoughts deem "there is no information about Louisa Goldman's husband and his title...", which is clearly contradicting the evidence.

Figure 7 shows an example for Think-Answer unfaithfulness. The final reasoning introduces concepts (parapsychology, real-time media) that are irrelevant or unsupported, and it does not logically justify the answer "reality television." This is likely a drift or hallucinated chain of thought.

Figure 6 and Figure 7 both show the examples of the policy model's Think-Search faithfulness. The model first reason about the information it already acquires (in the case of the think block after the information block), then reason about the next query it wants to issue. This observation is aligned with our findings in Figure 1, where the model exhibits high Think-Search faithfulness as indicated by both LLM-as-a-Judge and the NLI classifier. The main reason for Think-Search unfaithfulness as indicated in our evaluation results is due to the model's lack of reasoning of the next query instead of falsified or logically incorrect reasoning, which we consider as *not unfaithful*.



(a) Policy KL comparison.



(b) Policy entropy comparison.

Figure 5: Training dynamics of VERITAS Info-Think, with and without curriculum. Including the Info-Think faithfulness reward at the start of the training limits the policy model's exploration (lower KL in the left panel) and leads to a more deterministic policy (lower entropy in the right panel).

Info-Think Unfaithfulness Example

```
<|im_start|>system
You are a helpful assistant. First think through the reasoning internally, then provide the answer as instructed.
<|im_end|>

<|im_start|>user
Answer the given question. You must conduct reasoning inside <think> and </think> first every time you get new information. After reasoning, if you find you lack some knowledge, you can call a search engine by <search> query </search> and it will return the top searched results between <information> and </information>. You can search as many times as you want. If you find no further external knowledge needed, you can directly provide the answer inside <answer> and </answer>, without detailed illustrations. For example, <answer> Beijing </answer>. Question: What was Louisa Goldman's husbands title?

<think> First, I need to determine who Louisa Goldman is and find out information about her husband's title. </think>

<search> Louisa Goldman's husband's title </search>

<information>
Doc 1: "Walter E. Sachs"\nWalter E. Sachs Walter Edward Sachs (May 28, 1884  August 21, 1980) was an American banker and financier. He was born on May 28, 1884 in Manhattan, New York City to Samuel Sachs and Louisa Goldman of the GoldmanSachs family. He was an alumnus of Harvard Business School. Sachs was a partner at Goldman Sachs starting in 1928, guiding the company through the Great Depression. He married Mary Williamson in 1939 and divorced in 1960. He retired as a partner in 1959 to become a limited partner. He died at his home in Darien, Connecticut on August 21, 1980.
Doc 2: "Walter E. Sachs"\nWalter E. Sachs Walter Edward Sachs (May 28, 1884  August 21, 1980) was an American banker and financier. He was born on May 28, 1884 in Manhattan, New York City to Samuel Sachs and Louisa Goldman of the GoldmanSachs family. He was an alumnus of Harvard Business School. Sachs was a partner at Goldman Sachs starting in 1928, guiding the company through the Great Depression. He married Mary Williamson in 1939 and divorced in 1960. He retired as a partner in 1959 to become a limited partner. He died at his home in Darien, Connecticut on August 21, 1980.
Doc 3: "Amy Goldman Fowler"\nAmy Goldman Fowler Amy Goldman Fowler (born 1954) is an American billionaire heir, gardener, author, artist, philanthropist, and advocate for seed saving and heirloom fruits and vegetables. She is one of the foremost heirloom plant conservationists in the US. Goldman is described as ""perhaps the world's premier vegetable gardener"" by Gregory Long, President of The New York Botanical Garden. On April 28, 2012, Goldman married Cary Fowler at the terrace on top of the Arsenal in Central Park. Fowler is the daughter of Lillian (née Schuman) and Sol Goldman. She has three siblings: Allan H. Goldman, Diane Goldman Kemper, and\n"</information>

<think> Based on the search results, it is clear that the information is mostly about Walter E. Sachs, who was indeed related to the Goldman family. However, there is no information about Louisa Goldman's husband and his title. Let's try a different approach by searching for information about Louisa Goldman. </think>

<search> Louisa Goldman's husband's title </search>

...

<answer> partner </answer>
```

Figure 6: Example of Information-Think unfaithfulness in Search-R1.

Think-Answer Unfaithfulness Example

```
<|im_start|>system
You are a helpful assistant. First think through the reasoning internally, then provide the
    answer as instructed.
<|im_end|>

<|im_start|>user
Answer the given question. You must conduct reasoning inside <think> and </think> first every
    time you get new information. After reasoning, if you find you lack some knowledge, you
    can call a search engine by <search> query </search> and it will return the top searched
    results between <information> and </information>. You can search as many times as you
    want. If you find no further external knowledge needed, you can directly provide the
    answer inside <answer> and </answer>, without detailed illustrations. For example,
    <answer> Beijing </answer>. Question: Wim Schuhmacher is mostly associated with a genre
    sometimes called what ?

<think>1. We need to identify the genre associated with Wim Schuhmacher.</think>

<search> query: Who is Wim Schuhmacher and what genre is he associated with? </search>

<information>...</information>

<search> query: what genre to explain Wim Schuhmacher? </search>

<information>...</information>

<think>i found that he is a known filmmaker in the field, videos (a short video in
    particular)</think>

<search> query: It is Real Time Part. It is the same for the other genre? </search>

<information>
Doc 1: "Real-time computing"\nas ""real-time"" would be more accurately described as ""near
    real-time"". Near real-time also refers to delayed real-time transmission of voice and
    video. It allows playing video images, in approximately real-time, without having to wait
    for an entire large video file to download. Incompatible databases can export/import to
    common flat files that the other database can import/export on a scheduled basis so that
    they can sync/share common data in ""near real-time"" with each other. The distinction
    between ""near real-time"" and ""real-time"" varies, and the delay is dependent on the
    type and speed of the transmission. The delay in near real-time is
Doc 2: "Real time (media)"\nReal time (media) Real time within the media is a method where
    events are portrayed at the same rate at which the characters experience them. For
    example, if a movie told in real time is two hours long, then the plot of that movie
    covers two hours of fictional time. If a daily real-time comic strip runs for six years,
    then the characters will be six years older at the end of the strip than they were at the
    beginning. This technique can be enforced with varying levels of precision. In some
    stories, every minute of screen time is a
Doc 3: "Internet genre"\n...
</information>

<think>the right genre combination is parapsychology, he is associated specifically with a
    genre sub Real Time Part.</think>

<answer> reality television </answer>
```

Figure 7: Example of Think-Answer unfaithfulness in Search-R1.