Knowledgeable-r1: Policy Optimization for Knowledge Exploration in Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Retrieval-augmented generation (RAG) is a mainstream method for improving performance on knowledge-intensive tasks. However, current RAG systems often place too much emphasis on retrieved contexts. This can lead to reliance on inaccurate sources and overlook the model's inherent knowledge, especially when dealing with misleading or excessive information. To resolve this imbalance, we propose Knowledgeable-r1, an reinforcement learning (RL) framework that can dynamically select, combine, and utilize parametric and contextual knowledge. Unlike existing methods that rely on complex prompting or training pipelines, Knowledgeable-r1 employs adaptive prompts to elicit diverse knowledge preferences, then optimizes responses through reward-driven trajectories. Experiments show that Knowledgeable-r1 significantly enhances robustness and reasoning accuracy in both parameters and contextual conflict tasks and general RAG tasks, especially outperforming baselines by 17.07% in counterfactual scenarios and demonstrating consistent gains across RAG tasks.

1 Introduction

Retrieval-augmented generation (RAG) has become an effective strategy for knowledge-intensive tasks (Nakano et al., 2021; Gao et al., 2023). Current research reveals that LLMs exhibit a strong preference toward contextual knowledge over parametric knowledge (Su et al., 2024a; Xie et al., 2024a). This preference becomes problematic in scenarios involving conflicting knowledge or contextual inconsistencies (Lee et al., 2024; Dai et al., 2024; Sun et al., 2024; Wang et al., 2024; Tan et al., 2024), often resulting in erroneous outputs as shown in Figure 1. Therefore, enabling large models to effectively integrate between parameter and contextual knowledge remains a critical challenge. Recent studies propose contextual misinformation discrimination mechanisms that shift faith to parametric knowledge when detecting context inaccuracies (Das et al., 2023; Upadhyay et al., 2024; Torreggiani, 2025). The reliability of modern LLMs' parametric knowledge is growing, which makes this approach effective (Mallen et al., 2023; Yang et al., 2024). However, current implementations face distinct challenges across three primary directions. Prompt-guided techniques (Pan et al., 2023; Zhou et al., 2023; Wang et al., 2024; Xu et al., 2024; Ying et al., 2024) alert models to potential inaccuracies through warning prompts, though they often struggle with inconsistent sensitivity in real-world applications. Another strategy (Xu et al., 2024; Williams et al., 2018; Cheng et al., 2023; Zhang et al., 2024b) employs question-augmented frameworks that refine retrieval accuracy by rephrasing queries multiple times, but this iterative process creates heavy computational overhead during inference . The third category (Hong et al., 2024; Ju et al., 2025; Jin et al., 2024) enhances detection through specialized training modules , achieving better discrimination at the cost of increased memory demands and operational complexity. While these approaches improve knowledge conflict recognition from different angles, they generally face efficiency-performance trade-offs (Mu et al., 2021; Su et al., 2024b; Xu et al., 2024; Wang et al., 2023). However, few study focuses on exploring LLM's abilities in solving the contextual and parametric knowledge conflict problems without using extra models or components.

To address this limitation, we propose a reinforcement learning framework Knowledgeable-r1 that enhances models' capability to judiciously integrate parametric and contextual knowledge through factual grounding. While existing RL methods like GRPO (Shao et al., 2024) enable contextual knowledge utilization, they neglect systematic exploration of parametric knowledge. Our method introduces two key components: knowledge capa-

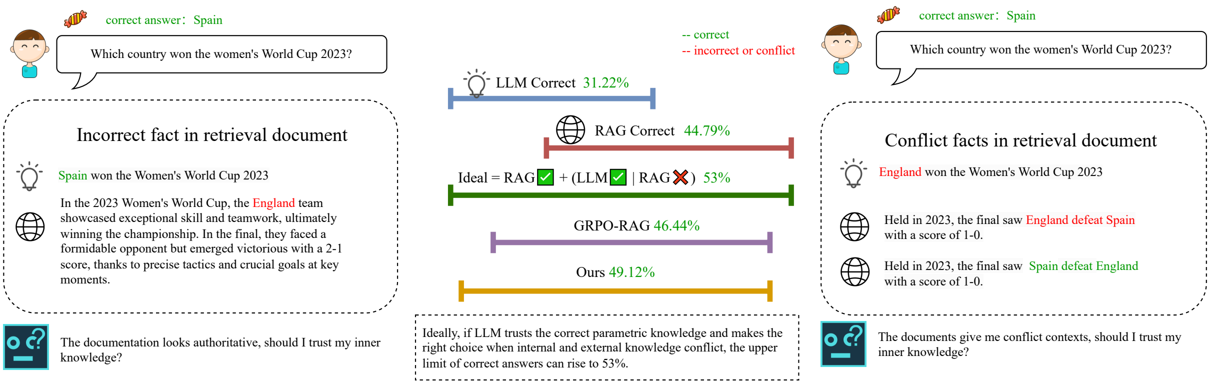

Figure 1: LLMs are prone to rely on completeness and conflict to judge the accuracy of article facts. When faced with conflicting or misleading retrieval content, LLMs ignore known information, which reduces reasoning accuracy.

bility exploration and optimization. The knowledge exploration allows the model to systematically probe its parametric knowledge, while knowledge capability optimization guide balanced use of contextual knowledge. This dual design enables the model to explore both contextual knowledge and parametric simultaneously, ensuring decisions align with factual accuracy.

We evaluate `Knowledgeable-r1` fact conflict resolution and QA capabilities in RAG scenarios. By implementing reinforcement learning to explore parametric knowledge and contextual knowledge paths, `Knowledgeable-r1` achieves an average accuracy improvement of **8.39%** on ConflictQA dataset(Bi et al., 2024), which involves counterfactual contextual knowledge and self-contradictory contextual knowledge. Notably, the improvement achieves **9.12%** enhancement with correct contextual context and reaches **17.07%** even contextual knowledge is error.

2 RELATED WORK

2.1 Parametric and Contextual Knowledge in LLMs

Large Language Models (LLMs) store a substantial quantity of parametric knowledge post-training, but their ability to access and use this knowledge effectively varies (Inui et al., 2019; Hu et al., 2023; Gueta et al., 2023). The integration of too much contextual information can cause LLMs to forget previously learned knowledge, known as catastrophic forgetting (Wen et al., 2024; Wang et al., 2025). LLMs may suppress their parametric knowledge in the presence of contextual cues, even when it's beneficial (Cheng et al., 2024). Studies have tried combining parametric knowledge with contextual data, but this process can be inefficient and prone to errors (Jeong et al., 2024; Fan et al., 2024).

2.2 Reinforcement Learning LLM Reasoning

OpenAI's O1 model marked a significant advancement in LLMs by incorporating extended reasoning (Jaech et al., 2024; Zelikman et al., 2024; Guan et al., 2024; Shao et al., 2024). The DeepSeek R1 model made its training approach and weights public, showing similar capabilities to O1 (Guo et al., 2025). DAPO addressed specific challenges in LLM RL extensions, such as entropy collapse, and suggested solutions to improve performance (Yu et al., 2025). The Search-R1 model showed LLMs using RL to generate search queries through reasoning(Jin et al., 2025). However, these methods are limited by the prompts, with low-probability outputs, like parametric knowledge during RAG reasoning, often overlooked due to static prompts.

3 Method

3.1 Task Definition

Given an original prompt p and and its retrieval-augmented version p' , when processed by a large language model (LLM), they produce outputs o and o' respectively, with o^* representing the correct response. Our goal is to develop the LLM policy $\pi_\theta(o^*|p')$ that can intelligently integrate parametric knowledge and contextual knowledge, ultimately generating accurate responses. The prompt representations are in Appendix A.

3.2 Knowledge capability exploration

When input p', model's reasoning could be assessed across three dimensions: parametric knowledge based capacity, contextual knowledge based reasoning ability, and reasoning capability under inconsistencies between parametric and contextual knowledge. If the model exhibits competence in all three aspects during reasoning, it can handle contextual information more robustly. We thus aim to identify and optimize the distributions corresponding to these three capabilities within the model. We define these three distributions as $\pi, \pi', \hat{\pi}$.

Basicly, we can employ reinforcement learning methods like GRPO to optimize the policy function by sampling p' . This method's unilateral sampling only enhances the ability of π'. Therefore, we augment GRPO with joint sampling of both knowledge sources to improve another two ability.

We sample n_1 parametric knowledge reasoning paths $\mathcal{O} = \{o_i\}_{i=1}^{n_1}$ from p and n_2 contextual knowledge paths $\mathcal{O}' = \{o'_j\}_{j=1}^{n_2}$ from p'.

We then gain the π and π' as following:

$$\begin{aligned} \pi &= \pi_\theta\left(o_{i,t} \mid p, o_{i,<t}\right) \\ \pi' &= \pi_\theta\left(o'_{i,t} \mid p', o_{i,<t}\right) \end{aligned} \tag{1}$$

Now, we need to obtain $\hat{\pi}$. When the model receives p', its output distribution is π'. We now aim to modify it with partial capabilities of π', i.e., the ability to reason based on its parametric parametric knowledge under p'. To achieve this, we calibrate the distribution of π' by concatenating the output $\mathcal{O}$ based on p. This process effectively simulates the distribution of parametric knowledge reasoning when processing p' thereby obtain $\hat{\pi}$, as illustrated below.

$$\hat{\pi} = \pi_\theta\left(o_{i,t} \mid p', o_{i,<t}\right) \tag{2}$$

3.3 Knowledge capability optimization

We have obtained the distributions corresponding to the three knowledge capabilities. To achieve this goal, we train our model following the GRPO training framework, aiming to maximize the rewards of the three distributions. This approach allows us to dynamically integrate their strengths based on factual correctness, ultimately achieving optimal comprehensive capabilities across all distributions.

Following GRPO, we compute group relativethe advantage for three types distributions responses. For π and π', we derive advantage scores $\mathcal{A}$ and $\mathcal{A}'$ that quantify each path's relative quality within their respective groups. we calculate the advantage A_i and A'_j by normalizing their respective group-level rewards $\{R_i\}_{i=1}^{n_1}$ and $\{R'_j\}_{j=1}^{n_2}$ as follows:

$$A_i = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^{n_1})}{\text{std}(\{R_i\}_{i=1}^{n_1})}, A'_j = \frac{R_j - \text{mean}(\{R'_j\}_{j=1}^{n_2})}{\text{std}(\{R'_j\}_{j=1}^{n_2})} \tag{3}$$

This calculation evaluates response quality within two distributions to enhance model capabilities under each. For $\hat{\pi}$, we compute its advantage by aggregating rewards from $\mathcal{O} \cup \mathcal{O}'$ generated by both p and p', followed by global normalization. This quantifies the relative effectiveness of parametric knowledge versus external knowledge when processing p', ensuring balanced synergy. The advantage of $\hat{\pi}$ is calculated as follows:

$$\hat{A_i}' = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^{n_1} \cup \{R'_j\}_{j=1}^{n_2})}{\text{std}(\{R_i\}_{i=1}^{n_1} \cup \{R'_j\}_{j=1}^{n_2})} \tag{4}$$

We then caculate policy object $l(\theta)$, $l'(\theta)$, $\hat{l}(\theta)$ for three distributions as follows:

$$\begin{aligned} l(\theta) &= \frac{1}{Z}\sum_{i=1}^{n_1}\sum_{t=1}^{|o_i|}\min\left[r_{i,t}(\theta)A_i, \mathcal{CLIP}(r_{i,t}(\theta)A_i\right], \\ l'(\theta) &= \frac{1}{Z}\sum_{j=1}^{n_2}\sum_{t=1}^{|o'_j|}\min\left[r'_{j,t}(\theta)A'_j, \mathcal{CLIP}(r'_{j,t}(\theta)A'_j\right], \\ \hat{l}(\theta) &= \frac{1}{Z}\sum_{i=1}^{n_1+n_2}\sum_{t=1}^{|o_i|}\left[\hat{r}_{i,t},(\theta)\hat{A}'_i\right], \\ \text{where} \quad r_{i,t}(\theta) &= \frac{\pi_\theta\left(o_{i,t} \mid p, o_{i,<t}\right)}{\pi_{\theta_{\text{old}}}\left(o_{i,t} \mid p, o_{i,<t}\right)}, \\ r'_{j,t}(\theta) &= \frac{\pi_\theta\left(o'_{j,t} \mid p', o'_{j,<t}\right)}{\pi_{\theta_{\text{old}}}\left(o'_{j,t} \mid p', o'_{j,<t}\right)}, \\ \hat{r}_{i,t}(\theta) &= \pi_\theta\left(o_{i,t} \mid p', o_{i,<t}\right) \\ \mathcal{CLIP}(x) &= \text{clip}\left(x; 1-\epsilon, 1+\epsilon\right) \end{aligned} \tag{5}$$

For the optimization objective of $\hat{l}(\theta)$, given that our training samples are concatenated and thus lack a reasonable old distribution, we have removed importance sampling. Additionally, to facilitate better exploration of parametric knowledge by the model, we also omit gradient clipping. Our ultimate optimization goal is to maximize the expected rewards of the three distributions:

$$\mathcal{J}(\theta) = l(\theta) + l'(\theta) + \hat{l}(\theta)$$

3.4 Knowledge advantage adjustment

When the model receives p', its output distribution is π'. We now aim to modify it with partial capabilities of π', i.e., the ability to reason based on its parametric parametric knowledge under p'. To achieve this, we calibrate the distribution of π' by concatenating the output $\mathcal{O}$ based on p. This process effectively simulates the distribution of parametric knowledge reasoning when processing p'

thereby obtain $\hat{\pi}$, as illustrated below. Through joint optimization of the three distribution strategies, the model holistically enhances its internal and external knowledge capabilities based on factual rewards. However, during reinforcement learning training, the quality of training trajectories generated from input p' typically surpasses that of p. This imbalance leads the advantage calculation to inherently prioritize responses relying on external knowledge responses when the model struggles to generate correct solutions internally, potentially suppressing exploration of useful internal knowledge pathways. To balance exploration between $\hat{\pi}$ and π', we introduce an advantage function transformation. For advantage of $\hat{\pi}$, we apply the following modified advantage calculation:

$$\mathrm{LReLU}(A'_j) = \begin{cases} \alpha A'_j, & \text{if } A'_j > 0 \\ \beta A'_j, & \text{if } A'_j \leq 0 \end{cases} \quad (6)$$

Where $\hat{A}_j$ represents the advantage of the jth response triggered by the parametric cue, α and β are set to 2 and 0.05 as default for reducing the penalty for parametric knowledge exploration and encourage better parametric knowledge answers.

4 Experiments

This section evaluate the capability of our method and baselines in parametric/contextual knowledge conflict tasks and RAG tasks.

4.1 General Experiments Setup

We conduct model based on Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct, and across multiple datasets: ConflictQA (Xie et al., 2024b) for integrating parametric and contextual knowledge task; Hotpotqa (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), and Musique (Trivedi et al., 2022) for retrieval-augmented general tasks. We compare our method against the following baseline methods: RAG prompting, SFT, and GRPO, and use query-only prompting, GRPO-prompting as auxiliaries. Following (Jin et al., 2025), Exact Match (EM) is used for evaluating the accuracy, more training details are shown in Appendix B.

4.2 Knowledge Conflict in Parameters and Context Tasks

We introduce the ConflictQA benchmark to evaluate the model performance of LLM in RAG scenarios involving knowledge conflicts. In order to deep analysis the situations of arrangement and combination of correct parametric and contextual knowledge, we first construct a overall dataset, which is constructed by 1:1 ratio ramdomly sampling incorrect context and correct context question-answer pairs, and split it into a training dataset '`CQ_train`' and a testing dataset '`CQ_test`'.

In order to further analysis our method and baselines's performance in each situations of knowledge conflict, we then filter the '`CQ_test`' into eight sub-testing datasets.

4.2.1 Sub-testing Dataset and Metric Construction

We construct the new sub-testing datasets and metrics in follow steps:

Step I: Adopting the approach of (Zhang et al., 2024a), parametric knowledge accuracy is gauged using the EM score from query-only inputs.

Step II: Context knowledge accuracy is assessed based on context source, comparing responses using correct versus incorrect evidence.

Step III: The '`CQ_test`' dataset is partitioned into four subgroups: '`T_i`' for parametrically correct, '`F_i`' for parametrically incorrect, '`T_e`' for contextually correct, and '`F_e`' for contextually incorrect QA pairs.

Step IV: We create five sub-testing datasets by applying set operations to the $\mathrm{Acc}_{\mathrm{FiFe}}$. These operations produce datasets: `T_i` ∩ `F_e`, `F_i` ∩ `T_e`, `F_e`, `T_e`, `T_i` ∪ `T_e`, and `F_i` ∩ `F_e`.

Step V: For clarity and brevity in reporting, we assign the following labels to the accuracy metrics for these sub-testing datasets: $\mathrm{Acc}_{\mathrm{TiFe}}$, $\mathrm{Acc}_{\mathrm{FiTe}}$, $\mathrm{Acc}_{\mathrm{Fe}}$, $\mathrm{Acc}_{\mathrm{Te}}$, $\mathrm{Acc}_{\mathrm{TiTe}}$, and $\mathrm{Acc}_{\mathrm{FiFe}}$. And the accuracy in '`CQ_test`' is defined as $\mathrm{Acc}_{\mathrm{CQ}}$.

4.2.2 Capability 1: The performance in parameters and context confict

Capability 1 ($C1$) is characterized by situations where the context and parameters has conflict original facts. Table 1 shows the performance of `Knowledgeable-r1` and baselines. We observe that `Knowledgeable-r1` outperforms all models on the $\mathrm{Acc}_{\mathrm{TiFe}}$ metric, achieving improvements of 14.9%, 13%, and 23.3% over the RAG prompting baseline in their respective conflict question-answer tasks. Higher accuracy in $\mathrm{Acc}_{\mathrm{TiFe}}$ is indicative of better identification and resistance to counterfactual context input through the use of parametric knowledge, aligning with the primary objective of our method: to use correct parametric knowledge

Table 1: Parameters and context conflict evaluation. Acc_{TiFe} demonstrates the capability of parameters knowledge to aid counterfactual reasoning and Acc_{FiTe} reflects parameters skepticism.

	ConflictQA-QA			ConflictQA-MC			ConflictQA-MR			Avg.		
Acc(EM)	Acc_{TiFe}	Acc_{FiTe}	Avg.	Acc_{TiFe}	Acc_{FiTe}	Avg.	Acc_{TiFe}	Acc_{FiTe}	Avg.	Acc_{TiFe}	Acc_{FiTe}	Avg.
query-only prompting	100%	0%	50%	100%	0%	50%	100%	0%	50%	100.00%	0.00%	50.00%
RAG prompting	↓ 51.90%	59.60%	55.75%	↓46.50%	52.30%	49.40%	↓ 55.20%	52.10%	53.65%	↓ 51.20%	54.67%	52.93%
SFT-RAG w/o CoT	20.40%	**70.40%**	45.40%	24.80%	54.40%	39.60%	34%	54.70%	44.35%	26.40%	**59.83%**	43.12%
GRPO-inner	92.80%	6.80%	49.80%	87.70%	19.20%	53.45%	86.70%	16.30%	51.50%	89.07%	14.10%	51.58%
GRPO-RAG	54.90%	62.10%	58.50%	54.60%	**58.70%**	56.65%	57%	**57.40%**	57.20%	55.50%	59.40%	57.45%
Knowledgeable-r1 w/o l	65.50%	62.20%	63.85%	**63.30%**	49%	56.15%	74.40%	53%	**63.70%**	67.70%	54.70%	61.23%
Knowledgeable-r1	**66.80%**	61.60%	**64.20%**	59.50%	54.60%	**57.05%**	**78.50%**	46.90%	62.70%	**68.27%**	54.37%	**61.32%**
impro. vs RAG prompting	+14.90%	+2.00%	+8.45%	+13.00%	+2.30%	+7.65%	+23.30%	-5.20%	+9.05%	+17.07%	-0.30%	+8.39%
impro. vs GRPO-RAG	+11.90%	-0.50%	**+5.70%**	+4.90%	-4.10%	+0.40%	+21.50%	-10.50%	**+5.50%**	+12.77%	-5.03%	**+3.87%**

Table 2: The robustness performance for incorrect and correct context knowledge. Acc_{Fe} indicates the anti-interference ability and Acc_{Te} represents the consistency of correct context.

	ConflictQA-QA			ConflictQA-MC			ConflictQA-MR			Avg.		
Acc(EM)	Acc_{Fe}	Acc_{Te}	Avg.	Acc_{Fe}	Acc_{Te}	Avg.	Acc_{Fe}	Acc_{Te}	Avg.	Acc_{Fe}	Acc_{Te}	Avg.
query-only prompting	30.80%	31.20%	31.00%	25.90%	26.80%	26.35%	26.30%	27.10%	26.70%	27.67%	27.80%	27.73%
RAG prompting	↓ 18.80%	70.50%	44.65%	↓ 15.90%	59.40%	37.65%	↓ 22.50%	61%	41.75%	19.07%	36.30%	27.68%
SFT-RAG w/o CoT	8%	**77.70%**	42.85%	7.90%	62.10%	35.00%	13.70%	54.90%	34.30%	9.87%	33.10%	21.48%
GRPO-inner	32%	34%	33.00%	32.60%	33.10%	32.85%	33.30%	36%	34.65%	32.63%	33.30%	32.97%
GRPO-RAG	20%	72.60%	46.30%	18.90%	**66.50%**	**42.70%**	24%	61.10%	42.55%	20.97%	38.50%	29.73%
Knowledgeable-r1 w/o l	24.90%	73.10%	49.00%	**23.80%**	57.40%	40.60%	**34%**	**63.50%**	**48.75%**	**27.57%**	**43.63%**	**35.60%**
Knowledgeable-r1	**25.40%**	73.10%	**49.25%**	22.30%	62.70%	42.50%	**34%**	59%	46.50%	27.23%	43.13%	35.18%
impro. vs RAG prompting	+6.60%	+2.60%	+4.60%	+6.40%	+3.30%	+4.85%	+11.50%	-2.00%	+4.75%	+8.17%	+6.83%	+7.50%
impro. vs GRPO	+5.40%	+0.50%	**+2.95%**	+3.40%	-3.80%	-0.20%	+10.00%	-2.10%	**+3.95%**	+6.27%	+4.63%	**+5.45%**

Table 3: Evaluation on several other metrics. Acc_{TiTe} indicates the knowledge fusion ability based on the upper limit of RAG prompting and query-only methods. Acc_{FiFe} represents the knowledge updating of models.

	ConflictQA-QA			ConflictQA-MC			ConflictQA-MR			Avg.		
Acc(EM)	Acc_{CQ}	Acc_{TiTe}	Acc_{FiFe}	Acc_{CQ}	Acc_{TiTe}	Acc_{FiFe}	Acc_{CQ}	Acc_{TiTe}	Acc_{FiFe}	Acc_{CQ}	Acc_{TiTe}	Acc_{FiFe}
query-only prompting	31.22%	47.60%	0.00%	26.40%	43.40%	7.10%	26.73%	42.10%	0.00%	28.12%	44.37%	2.37%
RAG prompting	44.79%	66.20%	4.00%	37.25%	56.20%	6.00%	41.96%	59.80%	10.80%	41.33%	60.73%	6.93%
SFT-RAG w/o CoT	43.04%	64.30%	2.40%	34.36%	53.40%	3.30%	38.40%	56.70%	6.40%	38.60%	58.13%	4.03%
GRPO-inner	33.06%	47.80%	5.00%	32.88%	45.00%	12.90%	34.69%	46.40%	14.20%	33.54%	46.40%	10.70%
GRPO-RAG	46.44%	68.40%	4.50%	**42.23%**	**62.60%**	8.60%	44.79%	63.40%	12.30%	44.49%	64.80%	8.47%
Knowledgeable-r1 w/o l	49.12%	71.30%	6.80%	40.25%	58.90%	**9.50%**	**47.58%**	**65.80%**	15.90%	45.65%	65.33%	10.73%
Knowledgeable-r1	**49.30%**	**71.60%**	**6.90%**	42.10%	62.10%	9.10%	46.60%	63.00%	**18.10%**	**46.00%**	**65.57%**	**11.37%**
impro. vs RAG prompting	+4.51%	**+5.40%**	+2.90%	+4.85%	+5.90%	+3.10%	+4.64%	+3.20%	+7.30%	**+4.67%**	+4.83%	+4.43%
impro. vs GRPO	+2.86%	+3.20%	+2.40%	-0.13%	-0.50%	+0.50%	+1.81%	-0.40%	+5.80%	+1.51%	+0.77%	+2.90%

to mitigate the effects of incorrect context input.

In the case of the Acc_{FiTe} metric, our method shows a minimal performance reduction of only 0.30%. From the GRPO-inner perspective, optimizing the inner part will result in a decrease in the indicator when compared with RAG prompting. The slight decline in the Acc_{FiTe} metric is considered to be justifiable, as our method indeed places a greater emphasis on parametric knowledge compared to other methods. To maintain fairness between the two scenarios, we compute their average value as a general performance to minimize the impact of dataset ratios. This approach yields an 8.39% absolute improvement over the RAG prompting method and a 3.87% improvement over GRPO. Notably, the more complex the tasks, the more pronounced the performance gains associated with our method.

4.2.3 Capability 2: The performance in robustness for extra knowledge

Capability 2, represented as $C2$, assesses the performance of knowledge processing when both correct and incorrect context facts are incorporated across two scenarios. As summarized in Table 2, the overall enhancement in $C2$ is more pronounced than that in $C1$. Espect to $C1$, $C2$ addresses instances where the context and parameters are congruent, suggesting that our method additionally bolsters knowledge consistency. Detail evaluation on both parameters and context being correct is shown in Appendix D, which improves by 4.7%. In particular, our approach shows superior performance in the Acc_{SCTI} and Acc_{SCFI} metrics, with gains of 8.17% and 6.83% over RAG prompting, and 6.27% and 4.63% over GRPO, respectively. Regarding the robustness to extraneous knowledge, our method

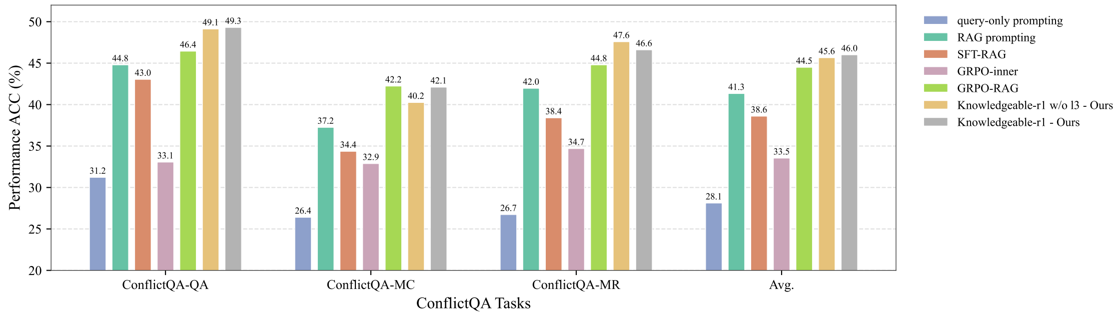

Figure 2: General performance of ConflictQA, including that ConFiQA contains three datasets: QA (Question and Answer), MR (Multi-hop Reasoning), and MC (Multi-conflict).

registers advances of 7.5% and 5.45% relative to RAG prompting and GRPO, respectively.

4.2.4 Capability 3: The performance in knowledge fusion

In this section, we further explore the proficiency of knowledge fusion and knowledge expansion, shown in Table 3. Acc_{TiTe} requires the amalgamation of both contextual and parametric knowledge to effectively tackle the problem, with an ideal integration value of 100%. Nonetheless, the accuracy achieved with RAG prompting stands at only 60.73%, and SFT method registers just 58.13% accuracy. These outcomes highlight a significant issue where parametric knowledge is eclipsed by the influx of contextual information. In contrast, our method exhibits a marked improvement, with the highest gain of 5.4% in the ConflictQA-QA task, culminating in a 71.6% accuracy rate.

4.2.5 Capability 4: The performance in knowledge extending

Additionally, we consider a scenario where neither the parameters nor the context are adequate to correctly resolve a question, deeming such inquiries theoretically unanswerable. This scenario assesses the aptitude of our method to correctly answer the question when lacking pertinent parametric knowledge and when contextual knowledge proves unbeneficial. Against RAG prompting, SFT, and GRPO methods, our method advances the performance by up to 6.7%.

4.2.6 Capability 5: The performance in overall knowledge conflict

Furthermore, we emphasize the uniform performance enhancement across the original ConflictQA-test set, showcasing a 4.67% improvement (Figure 2). Our method invariably maintains steady advancement.

Table 4: Contextual self-conflict evaluation. Acc_{SCTI} and Acc_{SCFI} indicate the accuracy of conflict context when parameters have correct and incorrect knowledge, respectively.

	Acc_{SCTI}	Acc_{SCFI}	Acc_{SC}(Avg.)
query-only prompting	100%	0%	50%
RAG prompting	82.8%	46.5%	64.65%
SFT-RAG w/o CoT	77.7%	43.3%	60.5%
GRPO-RAG	85.4%	52.6%	69%
`Knowledgeable-r1` w/o $\hat{l}$	89%	48.5%	68.75%
`Knowledgeable-r1`	89.6%	52%	70.8%
improv. vs RAG prompting	+6.8%	+5.5%	+6.15%
improv. vs GRPO-RAG	+4.2%	-0.6%	+2.4%

4.3 Context Self-Knowledge Conflict Tasks

In order to evaluate the scenarios where the context contains conflicting facts, we partition the '`CSCQ`' test set into two distinct subsets: '`CSCQ_Ti`' and '`CSCQ_Fi`'. These subsets are classify according to the query-only prompting method is correct or not. We define the accuracy on these subsets as follows:

- Acc_{SCTI}: Accuracy in the '`CSCQ_Ti`' subset.
- Acc_{SCFI}: Accuracy in the '`CSCQ_Fi`' subset.

The overall performance in context self-knowledge conflict tasks is then evaluated by computing the average of Acc_{SCTI} and Acc_{SCFI}, which we define as Acc_{SC}. This metric is intended to provide a balanced measure of how `Knowledgeable-r1` handles conflicting information within the context and retains accurate knowledge representation.

As depicted in Table 4, `Knowledgeable-r1` outperforms the baseline methods, achieving a 70.8% success rate. This indicates a significant improvement over the highest baseline performance, which is 10.3%.

Table 5: Performance comparison of various baselines on both out-of-domain and in-domain benchmarks. We report EM scores (%) for all benchmarks for clarity. **Avg.** denotes the average EM scores(and %) across all benchmarks.

Method	**Hotpotqa**	**Musique**	**2wiki**	**Avg.**
Qwen2.5-7b-Instruct				
query-only prompting	17.99%	2.57%	22.46%	14.34%
RAG prompting	24.13%	6%	26.73%	18.95%
SFT-RAG	23.92%	6.29%	26.92%	19.04%
GRPO-inner	21.74%	4.72%	26.77%	17.74%
GRPO-RAG	28.04%	10.01%	29.32%	22.46%
`Knowledgeable-r1` w/o $\tilde{l}$	31.33%	10.88%	38.51%	26.9%
`Knowledgeable-r1`	32.80%	12.00%	39.42%	28.07%
Qwen2.5-3b-Instruct				
query-only prompting	16.25%	1.90%	11.10%	9.75%
RAG prompting	19.49%	5.83%	20.74%	15.35%
SFT-RAG	19.53%	5.67%	20.82%	15.34%
GRPO-inner	16.72%	2.52%	18.66%	12.63%
GRPO-RAG	25.36%	8.44%	30.42%	21.41%
`Knowledgeable-r1`	27.09%	7.53%	33.39%	22.67%

Table 6: The best performance of `Knowledgeable-r1` in RAG tasks

Method	**Hotpotqa**	**Musique**	**2wiki**	**Avg.**
Qwen2.5-7b				
query-only prompting	17.99%	2.57%	22.46%	14.34%
CoT	19.18%	3.43%	21.84%	14.82%
RAG prompting	32.05%	13.65%	35.95%	27.22%
SFT-inner	16.21%	1.99%	21.73%	13.31%
SFT-RAG	32.17%	14.36%	35.93%	27.49%
Search-o1	18.70%	5.80%	17.60%	14.03%
GRPO-inner	21.74%	4.72%	26.77%	17.74%
GRPO-RAG	38.15%	24.64%	53.32%	38.70%
search-r1(Hotpotqa+nq)	38%	16.80%	32.60%	29.13%
`Knowledgeable-r1`	40.36%	24.95%	56.23%	40.51%
Qwen2.5-3b				
query-only prompting	16.25%	1.90%	11.10%	9.75%
CoT	12.79%	2.28%	23.83%	12.97%
RAG prompting	25.44%	13.57%	29.83%	22.95%
SFT-inner	12.11%	1.53%	17.95%	10.53%
SFT-RAG	25.62%	13.28%	29.84%	22.91%
Search-o1	22.1%	5.4%	21.8%	16.43%
GRPO-inner	16.72%	2.52%	18.66%	12.63%
GRPO-RAG	33%	24.95%	50.6%	36.18%
search-r1(Hotpotqa+nq)	30.8%	10.5%	27.3%	22.87%
`Knowledgeable-r1`	34.57%	20.85%	47.97%	34.46%

4.4 General RAG tasks

To further explore our method's performance in general RAG tasks, we evaluate it on three datasets, using only HotpotQA as the training dataset. Tables 5 and 6 show our method's best performance compared to baselines, achieving an average improvement of 5.51% over the GRPO baseline. Table 5 uses five retrieval documents for fairness. To find our method's upper limit on these datasets, we use all retrieval documents, with performances shown in Table 6. We finally achieve 40.36%, 24.95%, and 56.23% accuracy in HotpotQA, Musique, and 2wiki, respectively. Notably, these results exceed Search-R1 (Jin et al., 2025) and Search-O1 (Li et al., 2025), even though they use multi-step retrieval and two training datasets.

Table 7: `Knowledgeable-r1` component ablation study, including adding importance sampling, using single-group average, and deleting the LeakyRelu operation of advantages.

Model	Hotpotqa(1000step)
`Knowledgeable-r1`	28.36%
+ import sampling ($\pi_{old}(\theta(o\|p))$)	13.94%
+ import sampling ($\pi_{old}(\theta(o\|p'))$)	26.81%
improve	+ 14.42%
+ p prompt group averge	20.70%
improve	+ 7.66%
− leakrelu	21.30%
improve	+ 7.06%

5 Ablation & Analysis Study

5.1 The Performance of Component in `Knowledgeable-r1`

In this section, we conduct an ablation study to investigate the influence of various components of our `Knowledgeable-r1` on its overall performance, as detailed in Table 7. Both the computing of Advantages subgroup and the leakyReLU enhancements contribute positively to the efficacy of `Knowledgeable-r1` training. Moreover, applying importance sampling to fake distribution does not lead to an enhancement, and instead introduces greater variance into the training process.

5.2 The Training Efficiency of `Knowledgeable-r1`

Figure 3 shows the training curve comparison between GRPO and our method. Compared with GRPO, our method converges earlier and has a higher reward (accuracy) when converging. Overall, `Knowledgeable-r1` is more efficient due to directed exploration, which enables faster and deeper focus on internal information, thus leading to quick convergence to higher performance.

5.3 The Upper Limit of Knowledge Processing

We analyze the theoretical upper limit according to the correct union of RAG prompting correctness and query-only prompting correctness. As depicted in Figure 4, our method demonstrates performance closest to the union metrics, and we observe that the more space there is for improvement in RAG prompting compared to the union, the better the performance and improvement our method achieves.

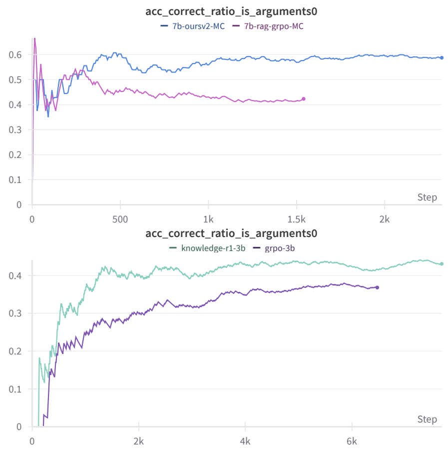

Figure 3: Comparison of training curves. The above one is the Confilict-MC dataset, purple represents GRPO, and bottom one represents `Knowledgeable-r1`. The right side is the HotpotQA dataset, and green represents `Knowledgeable-r1`.

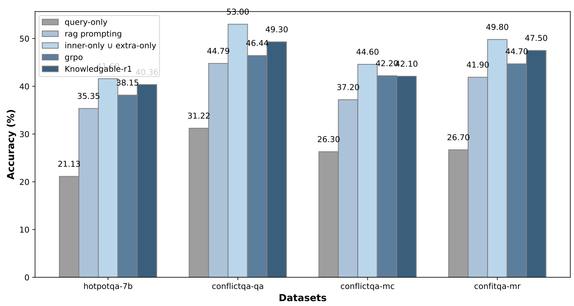

Figure 4: The performance of `Knowledgeable-r1` compare to theoretical accuracy rates.

5.4 Mitigating Misinformation in Context Evaluation

It is necessary to investigate into the discriminatory capabilities of models when evaluating contexts potentially laden with misinformation. As delineated in Tables 1 and 2, the $\mathrm{Acc}_{\mathrm{TiFe}}$ and $\mathrm{Acc}_{\mathrm{Fe}}$ metrics shows the accuracy when has misleading context, consider the the accuracy of the parametric knowledge or not respectively. In this study, we have identified that supplementing LLMs with additional knowledge may lead to misinformation, resulting in inferior performance when utilizing the RAG prompting method as opposed to query-only prompting.

The empirical evidence presented in Tables 1 and 2 illustrates that `Knowledgeable-r1` achieves significant enhancements over existing metrics. For example, our method attains a 34% accuracy rate in the ConflictQA-MR dataset, outperforming RAG prompting methods by an 11.5% margin and exhibiting a 10% lead over the GRPO approach. Remarkably, when parameters memeries are accurate in the context of incorrect information, our method demonstrates even more pronounced improvements—showing a 23.3% increase in accuracy compared to RAG prompting and a 21.5% advance relative to GRPO.

5.5 Model emergence capability of more capable base models

As shown in Tables 3 and 2, the performance of our method based on Qwen2.5-7b-Instruct surpasses that of Qwen2.5-3b-Instruct, with retrieval accuracy improvements of 4.71% for the top five documents (32.8% vs. 27.09%) and 5.79% for the top twenty documents (40.36% vs. 34.57%). These results lead us to conclude that our method benefits from more capable base models, aligning with our hypothesis that larger models possess more extensive parametric knowledge, thereby enhancing their ability to leverage this knowledge for improved performance. As base models continue to evolve, our method demonstrates significant potential for achieving better performance with increasingly larger models.

6 Conclusion

Our work introduces `Knowledgeable-r1`, a versatile and novel framework for reinforcement learning that has proven effective in guiding exploration through the use of supplemental cues to encourage the proper use of both contextual and parametric knowledge within large language models (LLMs). Through extensive experiments on knowledge conflict and general RAG tasks, `Knowledgeable-r1` has shown to significantly bolster the ability of LLMs to amalgamate contextual and parametric knowledge especially when the input context is counterfactuals. As for future work, there exists a wealth of opportunities to deploy `Knowledgeable-r1` in larger and more complex settings, and to develop it further to incorporate multiple directed exploration cues within a mixed objective-conditioned policy framework. Such explorations promise to be both intriguing and challenging.

Limitations

Our method proposes a reinforcement learning framework to address knowledge conflicts in large language models (LLMs), primarily focusing on enabling autonomous integration of parametric and contextual knowledge through factual grounding. However, the current approach does not account for scenarios where both knowledge sources contain inaccuracies, leaving open the question of whether models can learn abstention capabilities under such conditions. Future research should establish a multi-dimensional assessment framework to systematically evaluate LLMs' abilities in cross-source knowledge utilization, including error detection thresholds and dynamic knowledge reliability estimation.

References

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, and 1 others. 2024. Context-dpo: Aligning language models for context-faithfulness. *arXiv e-prints*, pages arXiv–2412.

Chao Cheng, Bin Fang, and Jing Yang. 2023. Distinguishing the correctness of knowledge makes knowledge transfer better. In *International Conference on Artificial Neural Networks*, pages 135–146. Springer.

Sitao Cheng, Liangming Pan, Xunjian Yin, Xinyi Wang, and William Yang Wang. 2024. Understanding the interplay between parametric and contextual knowledge for large language models. *Preprint*, arXiv:2410.08414.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Bias and unfairness in information retrieval systems: New challenges in the llm era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6437–6447, New York, NY, USA.

Bhaskarjyoti Das and 1 others. 2023. Multi-contextual learning in disinformation research: a review of challenges, approaches, and opportunities. *Online Social Networks and Media*, 34:100247.

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

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2:1.

Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, and 1 others. 2024. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*.

Almog Gueta, Elad Venezian, Colin Raffel, Noam Slonim, Yoav Katz, and Leshem Choshen. 2023. Knowledge is a region in weight space for fine-tuned language models. *arXiv preprint arXiv:2302.04863*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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

Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Whang. 2024. Why so gullible? enhancing the robustness of retrieval-augmented models against counterfactual noise. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2474–2495.

Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. 2023. A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering*, 36(4):1413–1430.

Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. 2019. Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *arXiv preprint arXiv:2403.14403*.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei

Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. 2024. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16867–16878.

Tianjie Ju, Bowen Wang, Hao Fei, Mong-Li Lee, Wynne Hsu, Yun Li, Qianren Wang, Pengzhou Cheng, Zongru Wu, Zhuosheng Zhang, and 1 others. 2025. Investigating the adaptive robustness with knowledge conflicts in llm-based multi-agent systems. *CoRR*.

Yoonjoo Lee, Kihoon Son, Tae Soo Kim, Jisu Kim, John Joon Young Chung, Eytan Adar, and Juho Kim. 2024. One vs. many: Comprehending accurate information from multiple erroneous and inconsistent ai generations. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2518–2531.

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

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

Tian Mu, Jianjun Yang, Feng Zhang, Chongchong Lyu, and Cheng Deng. 2021. The role of task conflict in cooperative innovation projects: An organizational learning theory perspective. *International Journal of Project Management*, 39(3):236–248.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, and 1 others. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403.

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. 2024. Group robust preference optimization in reward-free rlhf. *Advances in Neural Information Processing Systems*, 37:37100–37137.

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

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024a. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm. In *Advances in Neural Information Processing Systems*, volume 37, pages 103242–103268. Curran Associates, Inc.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024b. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm. *arXiv preprint arXiv:2408.12076*.

Yanshen Sun, Jianfeng He, Limeng Cui, Shuo Lei, and Chang-Tien Lu. 2024. Exploring the deceptive power of llm-generated fake news: A study of real-world detection challenges. *arXiv preprint arXiv:2403.18249*.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6207–6227, Bangkok, Thailand. Association for Computational Linguistics.

Fabio Torreggiani. 2025. Dont'be fooled by fake news? mapping the social, cognitive, and political mechanisms of misinformation belief.

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

R Upadhyay and 1 others. 2024. Addressing the challenge of online health misinformation: Detection, retrieval, and explainability.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö Arık. 2024. Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. *arXiv preprint arXiv:2410.07176*.

Mingyang Wang, Alisa Stoll, Lukas Lange, Heike Adel, Hinrich Schütze, and Jannik Strötgen. 2025. Bring your own knowledge: A survey of methods for llm knowledge expansion. *arXiv preprint arXiv:2502.12598*.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023. Resolving knowledge conflicts in large language models. *arXiv preprint arXiv:2310.00935*.

Yilin Wen, Zifeng Wang, and Jimeng Sun. 2024. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10370–10388.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024a. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. *Preprint*, arXiv:2305.13300.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024b. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *Proceedings of ICLR*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

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

Jiahao Ying, Yixin Cao, Kai Xiong, Long Cui, Yidong He, and Yongbin Liu. 2024. Intuitive or dependent? investigating llms' behavior style to conflicting prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4221–4246.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Eric Zelikman, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D Goodman. 2024. Quiet-star: Language models can teach themselves to think before speaking. *arXiv preprint arXiv:2403.09629*.

Hao Zhang, Yuyang Zhang, Xiaoguang Li, Wenxuan Shi, Haonan Xu, Huanshuo Liu, Yasheng Wang, Lifeng Shang, Qun Liu, Yong Liu, and 1 others. 2024a. Evaluating the external and parametric knowledge fusion of large language models. *arXiv preprint arXiv:2405.19010*.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, and 1 others. 2024b. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556.