RESISTING CONTEXTUAL INTERFERENCE IN RAG VIA PARAMETRIC-KNOWLEDGE REINFORCEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation (RAG) improves performance on knowledge-intensive tasks but can be derailed by wrong, irrelevant, or conflicting retrieved text, causing models to rely on inaccurate evidence and cascade errors. We propose `Knowledgeable-R1`, a reinforcement-learning framework that explicitly trains large language models to use parametric knowledge (PK) to resist contextual interference while still exploiting external context when it is reliably helpful. `Knowledgeable-R1` introduces a joint sampling scheme that generates paired responses with and without retrieval, and learns both *local* advantages (within each decoding regime) and *global* advantages under the same input to quantify when to ignore misleading context versus adopt it. We employ an asymmetric advantage transformation that amplifies exploratory behaviors toward parametric knowledge. Experiments show that `Knowledgeable-R1` significantly improves robustness and reasoning accuracy in knowledge conflict scenarios and general RAG scenarios, outperforming SOTA baselines by 23% in counterfactual scenarios, and without degradation when the retrieved context is fully accurate.

1 INTRODUCTION

Retrieval-augmented generation (RAG) has become a prominent approach for enhancing large language models (LLMs) by integrating contextual knowledge, thereby mitigating hallucinations and reducing factual errors (Nakano et al., 2021; Gao et al., 2023). However, recent studies indicate that when contextual knowledge is introduced, LLMs can become overly reliant on this external information, suppressing their internal parametric knowledge. This phenomenon, known as *context dominance*, is particularly evident under conditions of noisy, counterfactual, or internally inconsistent evidence (Su et al., 2024; Xie et al., 2024b; Shi et al., 2023). Conflict-focused evaluations confirm that LLMs often adopt incorrect retrieved statements even when their parametric knowledge is correct and can lag behind retrieval-free reasoning when retrieval is imperfect (Bi et al., 2025a; Wang et al., 2025a; Wen et al., 2024). These findings highlight an imbalance in how models utilize knowledge.

A central challenge in Retrieval-Augmented Generation (RAG) systems is determining when to rely on contextual knowledge (CK) and when to revert to parametric knowledge (PK), as well as how to integrate the two in a stable and faithful manner. 1) Prompting approaches help guide the model to validate or filter the context while combining it with parametric knowledge, which improves the coherence of the output (Ding et al., 2024; Cheng et al., 2024; Press et al., 2023; Wang et al., 2023; He et al., 2024; Wang et al., 2025a). 2) Decoding-based approaches, such as those proposed by (Bi et al., 2025b), adjust the token distribution during generation to mitigate conflicts between external context and parametric knowledge. While effective in some scenarios, both prompting and decoding methods add computational complexity and lack a generalizable decision rule for managing diverse contextual situations. 3) Fine-tuning methods, such as Self-RAG (Asai et al., 2024) and InFO-RAG (Xu et al., 2024), train LLMs to implicitly learn decision rules for knowledge utilization. However, these methods often require complex data annotation pipelines, which can limit flexibility and scalability.

When contextual knowledge appears reliable but is actually incorrect, LLMs should ignore it and decide whether to fall back on parametric knowledge. This behavior is rare but crucial, as current LLMs tend to rely more on context when faced with conflicting or misleading information. Reinforcement learning (RL) can assist by adjusting the probability rankings of outputs, encouraging the LLM to

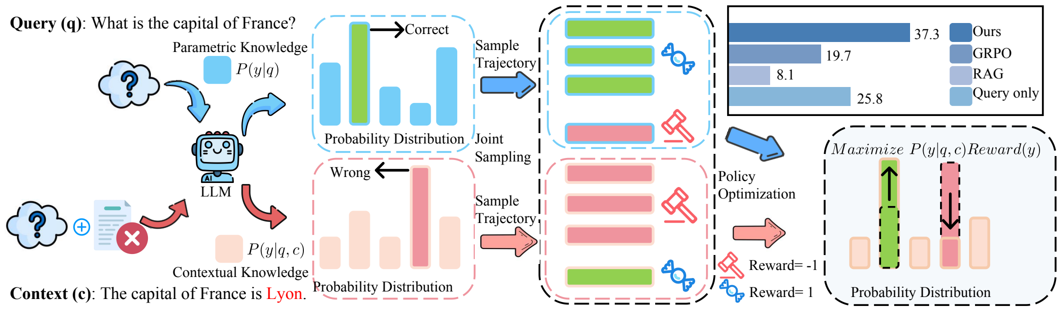

Figure 1: Our approach combines parametric (blue) and context-based (pink) path sampling with policy optimization to ensure correct outputs, even when erroneous context inputs. Green blocks represent positive answers, while red blocks represent negative ones.

explore these rare but critical decisions, such as when to fall back on parametric knowledge (Stiennon et al., 2020; Ramamurthy et al., 2023). Therefore, we believe RL has the potential to help LLMs learn how to effectively balance and utilize both parametric and contextual knowledge. Recently, models like OpenAI-o1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025) have employed RL techniques such as PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024) to enhance LLMs' logical reasoning and problem-solving abilities through experience and feedback. However, in RAG systems, these methods are limited by their sampling space and objectives, which leads to a primary focus on context-aware reasoning while overlooking the importance of incorporating parametric knowledge.

The goal of this work is to develop a reinforcement learning framework that explores both contextual and parametric knowledge after retrieval input. We call our method `Knowledgeable-R1` because it is a knowledge-aware reasoning approach that uses parametric knowledge to mitigate contextual interference in the RAG system. A conceptual overview of our framework is presented in Figure 1. Specifically, `Knowledgeable-R1` enables LLMs to (1) explore both parametric and contextual knowledge through joint sampling, (2) distinguish the relative merits of these two knowledge types using locally and globally defined advantages, and (3) mitigate penalties for less likely parametric outputs by applying an adaptive asymmetric advantages transformation. We conducted experiments across five scenarios, demonstrating that `Knowledgeable-R1` outperforms a range of RAG approaches by a significant margin. By effectively utilizing both parametric and contextual knowledge, `Knowledgeable-R1` shows significant improvements in adversarial context settings, achieving a **23%** improvement over GRPO baselines and a **30.47%** improvement over original RAG prompting, while maintaining strong performance in reliable contexts.

2 RELATED WORK

2.1 LIMITATIONS OF EXISTING RAG ROBUSTNESS METHODS

Recent studies establish that LLMs can store substantial parametric knowledge (PK), but their ability to use external contextual knowledge (CK) degrades under noisy inputs. Many studies have analyzed this phenomenon and found that it may be caused by position biases (Liu et al., 2024; Hu et al., 2025) or interference when PK and CK conflict (Farahani et al., 2024). These observations highlight the need for methods that can dynamically retain reliable PK when CK is misleading, while still leveraging CK when beneficial.

A significant thread in RAG research focuses on robustness to irrelevant or adversarial passages. Methods include noise-injected training (Yoran et al., 2024), attention masking (Fang et al., 2024), and adversarial filtering (Shi et al., 2024; Cohen-Wang et al., 2024). Evaluation benchmarks now systematically stress *conflicting evidence* scenarios (Wang et al., 2025b). However, these approaches largely operate at the *passage level* through pre-retrieval filtering or post-retrieval weighting, lacking a mechanism to actively suppress harmful context during generation. This limitation is particularly acute when misleading information appears plausible, as models tend to over-prioritize context even when it contradicts their parametric knowledge (Farahani et al., 2024; Gao et al., 2023).

2.2 Reinforcement Learning for Decision Policies in LLM Reasoning

Beyond supervised fine-tuning, reinforcement learning (RL) has emerged as a powerful paradigm for shaping model behavior without extensive human annotation. Recent work demonstrates RL's effectiveness in guiding reasoning processes, such as generating internal rationales (Zelikman et al., 2024; Li et al., 2025b) or stabilizing long chain-of-thought reasoning (Yu et al., 2025; Chen et al., 2025; Cheng et al., 2025; Jin et al., 2025; Li et al., 2025a; Song et al., 2025). However, these methods primarily focus on *reasoning structure* rather than on *parametric and contextual knowledge-aware learning for LLMs themselves*.

3 Method

Our objective is to enable a single language model to appropriately leverage parametric knowledge (PK) versus contextual knowledge (CK) in retrieval-augmented generation (RAG). The key challenge lies in the varying reliability of retrieved context, which can be helpful, redundant, or misleading. An ideal LLM should:

- Select the most accurate answer from its parametric knowledge when no context is provided (*parametric-only correctness*).
- Select the most accurate answer from its contextual knowledge when context is provided (*context-aware correctness*).
- Fall back on its correct parametric knowledge when faced with conflicting or noisy context, ensuring robustness to misleading information (*robustness to misleading context*).

To achieve this, we propose a multi-objective reinforcement learning framework that trains LLMs to simultaneously optimize these three goals.

3.1 Task Definition

We formulate the problem as a token-level reinforcement learning task. Let $\mathcal{V}$ be the vocabulary, and let a prompt x be the input sequence. The prompt can be of two types: p, containing only the query q, or p', containing both the query q and a retrieved context passage c. At each decoding step t, the model (parameterized by θ) produces a probability distribution over $\mathcal{V}$ based on the prompt and previously generated tokens $o_{<t}$:

$$\pi_\theta(o_t|x, o_{<t}) = \text{softmax}(f_\theta(x, o_{<t}))$$

The goal is to optimize θ such that the generated sequence $o = (o_1, \ldots, o_T)$ maximizes the expected reward, which reflects answer correctness and appropriate knowledge utilization.

3.2 Three sampling/decoding policies.

Let p denote the *query-only* prompt and p' denote the *query+context* prompt. As we illustrate in Section 1, GRPO uses a single sampling policy during LLM training, making it difficult for the LLM to explore parametric-aware answers in the combined retrieval input p'. In contrast, our method defines three distinct policies for each query q over the next-token distributions:

$$\pi_\theta(o_t \mid p, o_{<t}), \quad \pi'_\theta(o'_t \mid p', o'_{<t}), \quad \hat{\pi}_\theta(o_t \mid p', o_{<t}).$$

Notation. We use o to denote a token sequence generated when the *current* policy is conditioned on p (query-only input), and o' to denote a token sequence generated when the *current* policy is conditioned on p' (query+context input). We write $o_{<t}$ for the prefix up to step $t-1$, and o_t for the token at step t. With these, the three policy types have the following inputs/outputs and intended behaviors:

- **PK (Parametric)**: input $p =$ query, output o (answer from parametric knowledge).
- **CK (Context-aware)**: input $p' =$ query+context, output o' (answer using context).
- **RPK (Robust-PK)**: input $p' =$ query+context, output o (answer consistent with PK).

CK and RPK share the same input p' but target different behaviors: CK exploits reliable context; RPK *stays on the PK trajectory* when context is noisy/conflicting. PK and RPK share the same output type o, but are trained under different conditioning (without vs. with context). At inference, the model does not explicitly switch controllers; the learned token distributions encode when to follow context or implicitly revert to PK.

Table 1: I/O and behavior of the three decoding policies.

Policy type	Input prompt	Output type	Intended behavior
PK	$p =$ query	$o \in$ answer from parametric knowledge	Parametric-only answer
CK	$p' =$ query+context	$o' \in$ answer using context	Use context when helpful
RPK	$p' =$ query+context	$o \in$ answer consistent with PK	Fallback PK answer under misleading context

3.3 Advantage Calculation with Local and Global Normalization

To effectively balance the three objectives, we introduce a novel advantage calculation scheme that combines within-objective and cross-objective comparisons. The components are:

- **Local Advantage** (within-policy, within-input): compares trajectories that share the same input *and* the same output policy type.
- **Global Advantage** (same input state, cross-output): compares CK and RPK trajectories under p' in a unified pool, deciding whether to *use* context or *fall back* to PK-style decoding.

The global advantage mechanism enables effective differentiation between group trajectories, ensuring that rewards can still capture *cross-source* preferences (CK vs. RPK) when all trajectories within a group are uniformly good or bad. This maintains meaningful feedback for different knowledge types.

Notation. Rewards $R(\cdot)$ are computed per trajectory based on answer correctness (e.g., EM). Let rewards be R_i^{pk} for PK trajectories $o_i \sim \pi_\theta(\cdot \mid p, \cdot)$, R_j^{ck} for CK trajectories $o'_j \sim \pi'_\theta(\cdot \mid p', \cdot)$, and R_i^{rpk} for RPK trajectories $\tilde{o}_i \sim \hat{\pi}_\theta(\cdot \mid p', \cdot)$. Define the *global (same-input)* pool under p' as $\mathcal{U}_{p'} = \{R_j^{\text{ck}}\} \cup \{R_i^{\text{rpk}}\}$.

To address the "overcorrection" problem, when CK is reliable but slightly differs from PK (e.g., when factual details like a new capital city are updated), the LLM should trust CK more due to its timeliness. Thus, we design distinct advantages for PK, CK, and RPK.

PK (Parametric). The advantage of PK is defined as $A_i = A_i^{\text{pk-local}}$, since PK relies on query-only input:

$$A_i^{\text{pk-local}} = \frac{R_i^{\text{pk}} - \text{mean}(\{R_k^{\text{pk}}\})}{\text{std}(\{R_k^{\text{pk}}\}) + \varepsilon}.$$

This term encourages the query-only parametric answer to be as accurate as possible.

CK (Context-aware). The advantage of CK is $A'_j = A_j^{\text{ck-local}} + A_j^{\text{ck-global}}$, combining both *local* and *global* terms:

$$A_j^{\text{ck-local}} = \frac{R_j^{\text{ck}} - \text{mean}(\{R_k^{\text{ck}}\})}{\text{std}(\{R_k^{\text{ck}}\}) + \varepsilon}, \quad A_j^{\text{ck-global}} = \frac{R_j^{\text{ck}} - \text{mean}(\mathcal{U}_{p'})}{\text{std}(\mathcal{U}_{p'}) + \varepsilon}.$$

RPK (Robust-PK under Context). The advantage of RPK is defined as $\hat{A}_i = \hat{A}_i^{\text{global}}$. RPK focuses solely on the *global* comparison under the same input p':

$$\hat{A}_i^{\text{global}} = \frac{R_i^{\text{rpk}} - \text{mean}(\mathcal{U}_{p'})}{\text{std}(\mathcal{U}_{p'}) + \varepsilon}.$$

Summary. PK has a local advantage, ensuring accuracy based on the query input. CK combines both local and global advantages, prioritizing context when both knowledge types are correct, as context is frequently updated and more reliable. RPK has only a global advantage, comparing PK's performance in the same input p' to maintain it as a fallback when context is misleading. This balance ensures the model uses context effectively without over-relying on it when PK is more reliable.

3.4 Knowledge Balance Modulation

During training, context-aware (CK) paths often get higher rewards than robust parametric knowledge (RPK) paths due to the usefulness of retrieved contexts, creating a bias towards context. This can make the model rely too much on CK and weaken its ability to use parametric knowledge in noisy or conflicting situations. To address this, we introduce an asymmetric advantage transformation for RPK paths, reducing the penalty when PK-based decoding is slightly worse than context-following, ensuring parametric knowledge remains a viable fallback.

We define a modulation function T that transforms the RPK advantages $\hat{A}_i$:

$$T(\hat{A}_i;\beta) = \begin{cases} \hat{A}_i, & \text{if } \hat{A}_i > 0, \\ \beta \cdot \hat{A}_i, & \text{if } \hat{A}_i \le 0, \end{cases}$$

where $\beta \in [0.01, 1]$ is a modulation coefficient. When $\beta < 1$, negative advantages (indicating poor performance of PK decoding) are reduced, making the model less sensitive to occasional mistakes when relying on parametric knowledge.

To dynamically balance the exploration of parametric and contextual knowledge, we adapt β during training based on the relative performance of CK and RPK trajectories. Let $\mathcal{B}$ denote the current mini-batch of training examples. We compute the total advantage for each knowledge type:

$$S_{\text{ck}} = \sum_{j\in\mathcal{B}_{\text{ck}}} A'_j, \quad S_{\text{rpk}+} = \sum_{i\in\mathcal{B}_{\text{rpk}+}} \hat{A}_i, \quad S_{\text{rpk}-} = \sum_{i\in\mathcal{B}_{\text{rpk}-}} \hat{A}_i,$$

where $\mathcal{B}_{\text{ck}}$ is the set of CK trajectories, $\mathcal{B}_{\text{rpk}+}$ is the subset of RPK trajectories with positive advantages, and $\mathcal{B}_{\text{rpk}-}$ is the subset with non-positive advantages.

We then update β to maintain balance between the two knowledge sources:

$$\beta \leftarrow \text{clip}\left(\frac{S_{\text{ck}} - S_{\text{rpk}+}}{S_{\text{rpk}-}}, 0.01, 1\right),$$

The update rule adjusts the penalty coefficient β based on the performance gap between context-aware (CK) and robust parametric knowledge (RPK). When CK outperforms RPK, β decreases, reducing penalties for negative RPK advantages and encouraging more exploration of RPK. As the gap narrows, β increases, making RPK training more cautious.

When context is highly beneficial (i.e., $S_{\text{rpk}-}$ is much larger than $S_{\text{rpk}+}$), reducing penalties for poor RPK performance (β decreases) allows the model to keep RPK as a fallback option. This method, similar to *reward shaping* (Ng et al., 1999; Devlin et al., 2011), ensures that parametric knowledge stays usable. It also resembles focused exploration techniques (Schulman et al., 2015; Espeholt et al., 2018) that prevent forgetting learned behaviors. While this introduces some bias in the gradient estimates, it helps balance the model's tendency to over-rely on helpful context from the training data, allowing it to resist misleading context when needed.

3.5 Policy Optimization

We adopt PPO-style updates with clipping. For each trajectory, we compute the probability ratio between the current and old policies:

$$r^{\text{PK}}_{i,t} = \frac{\pi_\theta(o_{i,t} \mid p,\ o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid p,\ o_{i,<t})}, \quad r^{\text{CK}}_{j,t} = \frac{\pi'_\theta(o'_{j,t} \mid p',\ o'_{j,<t})}{\pi'_{\theta_{\text{old}}}(o'_{j,t} \mid p',\ o'_{j,<t})}, \quad r^{\text{RPK}}_{i,t} = \frac{\hat{\pi}_\theta(o_{i,t} \mid p',\ o_{i,<t})}{\hat{\pi}_{\theta_{\text{old}}}(o_{i,t} \mid p',\ o_{i,<t})}.$$

where (x, o_t) are the corresponding prompts and tokens.

The objective for each component is defined as follows:

$$J_{\text{PK}} = \frac{1}{n_{\text{pk}}} \sum_{i=1}^{n_{\text{pk}}} \sum_{t=1}^{|o_i|} \min\left[r^{\text{PK}}_{i,t}\, A_i,\ \text{clip}(r^{\text{PK}}_{i,t},\, 1-\epsilon,\, 1+\epsilon)\, A_i\right],$$

$$J_{\mathrm{CK}} = \frac{1}{n_{\mathrm{ck}}} \sum_{j=1}^{n_{\mathrm{ck}}} \sum_{t=1}^{|o'_j|} \min \left[r_{j,t}^{\mathrm{CK}} \, A'_j, \; \mathrm{clip}(r_{j,t}^{\mathrm{CK}}, \, 1-\epsilon, \, 1+\epsilon) \, A'_j \right],$$

$$J_{\mathrm{RPK}} = \frac{1}{n_{\mathrm{pk}}} \sum_{i=1}^{n_{\mathrm{pk}}} \sum_{t=1}^{|o_i|} \min \left[r_{i,t}^{\mathrm{RPK}} \, T(\hat{A}_i), \; \mathrm{clip}(r_{i,t}^{\mathrm{RPK}}, \, 1-\epsilon, \, 1+\epsilon) \, T(\hat{A}_i) \right],$$

The total objective is the weighted sum of the individual objectives:

$$\mathcal{J}(\theta) = \lambda_{\mathrm{pk}} J_{\mathrm{PK}} + \lambda_{\mathrm{ck}} J_{\mathrm{CK}} + \lambda_{\mathrm{rpk}} J_{\mathrm{RPK}}$$

with $\lambda_{\mathrm{pk}} = \lambda_{\mathrm{ck}} = \lambda_{\mathrm{rpk}} = 1.0$ in our experiments. We use $\epsilon = 0.2$ for clipping.

The goal of our method is to optimize θ to maximizes $\mathcal{J}(\theta)$, the full training procedure is outlined in Algorithm 1 in Appendix E.

4 Experiments

We evaluate whether the model can *leverage parametric knowledge* to produce accurate answers despite *external contextual interference*, and remain robust when context knowledge is correct.

4.1 Experimental Setup

We test robustness under five contextual conditions with increasing difficulty: **Scenario I (S1)**: Correct contextual knowledge (Section 4.2.1); **Scenario II (S2)**: Adversarial contextual knowledge (Section 4.2.1); **Scenario III (S3)**: Self-conflicting contextual knowledge (Section 4.2.1); **Scenario IV (S4)**: Irrelevant contextual knowledge (Section 4.2.1); **Scenario V (S5)**: Partially relevant contextual knowledge (Section 4.2.1). These scenarios reflect real-world retrieval situations where irrelevant or conflicting context can mislead generation.

4.1.1 Reasoning Instruction and Baselines

We evaluate with `Qwen2.5-3B-Instruct`, `Qwen2.5-7B-Instruct`, and `Llama3.1-8B-Instruct`. Baselines include query-only prompting, RAG prompting, Astute-RAG (a prompting method for imperfect retrieval (Wang et al., 2025a)), and GRPO with RAG. We evaluate using exact match (EM) (Jin et al., 2025). To ensure fairness, all methods use the same total training data and are trained for one epoch. For GRPO and `Knowledgeable-R1`, we use a global batch size of 128, a rollout batch size of 32, a rollout temperature of 1, and a learning rate of 1×10^{-6}. All experiments run on 8 H100 GPUs, using the same system prompt for both training and inference, where the inference input combines the query and context. For further experiment details, see Appendix B, C, and D.

4.1.2 Datasets

We introduce **KnowQA** to test methods across the five scenarios. First, we test parametric knowledge using query-only prompts (Appendix A), where a correct answer means the model has the relevant PK. Then, we create contextual inputs using the top-5 retrieved paragraphs. This method is suitable for knowledge-intensive tasks. KnowQA includes scientific, factual, and commonsense questions, organized by scenario:

- **Correct and Wrong Context**. We adapt **PC** (PC-QA, PC-MR, PC-MC) and **NC** (NC-QA, NC-MR, NC-MC) from **ConFiQA** (Xie et al., 2024a) for supporting vs. opposing evidence. QA stands for single-hop QA, MR for multi-hop reasoning, and MC for multiple conflicting inputs. We check the correctness of the context by looking for made-up sources. ConFiQA is a counterfactual-retrieval benchmark with three sub-datasets (QA/MR/MC).
- **Conflict Context**. We create **SC** from ConFiQA using conflicting evidence and test whether models can handle contradictions by showing both positive and negative examples together.
- **Irrelevant Context**. We use **ExplainCPE** (Wen et al., 2024), a medical QA dataset from the Chinese National Licensed Pharmacist Examination. It *tests the impact of knowledge mismatches on model performance*, where adding retrieved documents can *lower* accuracy compared to query-only.

- **Partly Irrelevant Context**. We use **HotpotQA** (Yang et al., 2018), **2WikiMultiHopQA** (Ho et al., 2020), and **MuSiQue** (Trivedi et al., 2022), which include valid evidence mixed with distractors.

4.2 Experimental Results

We evaluate `Knowledgeable-R1` in various contexts and find that it effectively selects reliable evidence while ignoring misleading or irrelevant information, especially when parametric knowledge (PK) is sufficient. This leads to significant improvements in challenging interference settings (S2-S5) without harming performance in correct contexts (S1). Consistent results are observed across models of different sizes (3B, 7B, 8B, 14B), with detailed results for `Qwen2.5-7B-Instruct` and `Llama3.1-8B-Instruct` in Tables 2 and 3, and similar patterns for other models (see Appendix F). Our experimental analysis highlights three key strengths of `Knowledgeable-R1`:

- **Robustness to Incorrect Context.** Significant improvements in adversarial (S2), conflicting (S3), and irrelevant (S4) settings demonstrate the method's ability to handle unreliable context.
- **Preservation of Correct Context Benefits.** No performance drop occurs when the context is accurate (S1), keeping it competitive with context-specific baselines.
- **Effective Context Selection.** Large gains on PK-answerable questions, particularly in mixed-context scenarios (S5), highlight the method's ability to filter relevant from irrelevant information.

4.2.1 Performance across Contextual Scenarios

Scenario I: Correct Contextual Knowledge (S1). When given accurate evidence, `Knowledgeable-R1` keeps the performance improvements from correct context. As shown in Table 2 (left columns), for `Qwen2.5-7B-Instruct`, Knowledgeable-R1 achieves **80.90%** on PC-QA and **75.51%** on PC-MC, staying competitive with the best baseline (GRPO w/ RAG). Similarly, for `Llama3.1-8B-Instruct`, it reaches **80.03%** on PC-QA and **80.24%** on PC-MC. These results show that learning to manage noisy contexts doesn't reduce the use of reliable evidence, addressing a key concern in robust RAG systems.

Scenario II: Adversarial Contextual Knowledge (S2). This scenario highlights the weaknesses of naive RAG and the benefits of our method. When the context is adversarial, RAG prompting causes a huge drop in performance—down to **13.47%/8.06%/11.31%** (NC-MR/MC/QA) on `Qwen2.5-7B`-Instruct, much lower than the query-only baseline. `Knowledgeable-R1` greatly improves this, increasing performance by **+30.47/+29.28/+18.09** percentage points and outperforming GRPO w/ RAG (e.g., **43.94%** vs. **26.94%** on NC-MR). The consistent gains of **+32.49/+24.84/+19.71** on `Llama3.1-8B-Instruct` suggest that the method effectively learns to ignore misleading evidence and rely on parametric knowledge when needed.

Scenario III: Self-Conflicting Contextual Knowledge (S3). When the context has internal contradictions, `Knowledgeable-R1` performs best among both model families: **63.77%** compared to **63.38%** for GRPO w/ RAG on `qwen2.5-7B`, and **76.58%** compared to **73.67%** on `llama3.1-8B`. Though the improvements are small, the consistent gains across models show stable handling of conflicting information. Resolving contradictions within the context seems difficult for all methods, indicating an area for future improvement.

Scenario IV: Irrelevant Contextual Knowledge (S4). In the **ExplainPE** benchmark, where the context is completely unrelated, we see the typical "RAG-hurt" effect: query-only (**64.45%**) performs better than RAG prompting (**62.21%**) on `qwen2.5-7B`. `Knowledgeable-R1` achieves **67.57%**, actively rejecting irrelevant context and outperforming both methods. This shows that the method can recognize when context isn't helpful for the query.

Scenario V: Partially Relevant Contextual Knowledge (S5). When valid evidence is mixed with distractors, `Knowledgeable-R1` performs strongest or nearly strongest. On `qwen2.5-7B`, it achieves **31.45%** on HotpotQA, **37.52%** on 2WikiMultiHopQA, and **12.04%** on MuSiQue, with significant gains over RAG prompting. **Notably, the strong performance on 2WikiMultiHopQA and MuSiQue is achieved without fine-tuning on these datasets**, showing that the method can generalize to new evidence types and distractor patterns beyond the training data (HotpotQA). This ability to generalize across datasets is especially useful for real-world applications, where the evidence may differ from the training data.

Table 2: Overall accuracy across five contextual scenarios: correct (S1), wrong (S2), conflict (S3), irrelevant (S4), and partly-irrelevant (S5). Best results are in **bold**, second best are underlined. The "improve" row shows gains over RAG prompting.

	Correct (S1)			Wrong (S2)			Conflict (S3)	Irrelevant (S4)	Partly Irrelevant (S5)		
Method	PC-MR	PC-MC	PC-QA	NC-MR	NC-MC	NC-QA	SC	ExplainPE	HotPotQA	2Wiki MultiHopQA	Musique
`Qwen2.5-7B-Instruct`											
Query-only prompting	27.72%	24.66%	31.67%	25.93%	25.82%	32.31%	29.67%	64.45%	20.90%	25.54%	4.36%
RAG prompting	65.68%	66.39%	74.35%	13.47%	8.06%	11.31%	59.50%	62.21%	20.36%	22.53%	6.41%
CK-PlUG (Bi et al., 2025b)	64.69%	66.55%	78.66%	11.62%	8.06%	7.92%	55.00%	55.00%	22.74%	24.76%	6.25%
Astute (Wang et al., 2025a)	65.51%	66.05%	77.62%	12.79%	7.07%	10.34%	54.20%	56.74%	17.87%	20.35%	6.29%
GRPO w/ RAG	**77.56%**	**77.36%**	80.03%	26.94%	19.74%	26.01%	75.33%	66.50%	27.93%	33.95%	11.79%
`Knowledgeable-R1`	75.08%	75.51%	**80.90%**	**43.94%**	**37.34%**	**29.40%**	**76.33%**	**67.57%**	**31.45%**	**37.52%**	**12.04%**
improve	+9.41%	+9.12%	+6.54%	+30.47%	+29.28%	+18.09%	+15.92%	+5.36%	+11.09%	+14.99%	+5.63%
`Llama3.1-8B-Instruct`											
Query-only prompting	29.37%	26.18%	39.93%	27.10%	27.63%	42.65%	32.08%	43.26%	20.69%	21.02%	6.16%
RAG prompting	64.85%	61.99%	76.42%	22.90%	16.28%	24.88%	61.17%	39.16%	24.44%	23.50%	8.19%
CK-PlUG (Bi et al., 2025b)	54.79%	58.45%	69.71%	12.12%	9.05%	17.29%	42.00%	31.54%	22.35%	24.63%	5.25%
Astute (Wang et al., 2025a)	65.84%	64.86%	77.97%	17%	9.05%	17.29%	59.83%	40.14%	1.65%	30.26%	9.64%
GRPO w/ RAG	**78.05%**	79.73%	**82.62%**	41.58%	35.69%	39.26%	**76.58%**	47.56%	34.84%	41.22%	**16.59%**
`Knowledgeable-R1`	73.76%	**80.24%**	80.03%	**55.39%**	**41.12%**	**44.59%**	73.67%	**49.61%**	**37.06%**	**45.37%**	14.69%
improve	+8.91%	+18.25%	+3.61%	+32.49%	+24.84%	+19.71%	+15.41%	+10.45%	+12.62%	+21.87%	+6.50%

Table 3: Accuracy on the parametric-knowledge answerable subset across all scenarios. Best results are in **bold**, second best are underlined.

	Correct (S1)			Wrong (S2)			Conflict (S3)	Irrelevant (S4)	Partly Irrelevant (S5)		
Method	PC-MR	PC-MC	PC-QA	NC-MR	NC-MC	NC-QA	SC	ExplainPE	HotPotQA	2Wiki MultiHopQA	Musique
`Qwen2.5-7B-Instruct`											
RAG prompting	85.71%	86.99%	94.02%	32.47%	19.75%	28.50%	87.08%	83.64%	48.84%	41.16%	20.75%
CK-PlUG (Bi et al., 2025b)	79.76%	83.56%	95.65%	29.87%	18.47%	18.50%	82.87%	77.58%	55.30%	42.78%	20.75%
Astute (Wang et al., 2025a)	86.31%	87.67%	92.39%	31.17%	19.11%	24.50%	81.18%	73.03%	41.73%	34.50%	20.75%
GRPO w/ RAG	93.21%	93.15%	97.28%	57.61%	46.50%	62.50%	93.54%	**84.39%**	66.54%	55.42%	45.28%
`Knowledgeable-R1`	**95.83%**	**95.21%**	**97.83%**	**89.61%**	**79.62%**	**66.00%**	**96.07%**	83.18%	**78.49%**	**66.72%**	**66.98%**
improve	+10.12%	+8.22%	+3.81%	+57.14%	+59.87%	+37.50%	+8.99%	-0.46%	+29.65%	+25.56%	+46.23%
`Llama3.1-8B-Instruct`											
RAG prompting	89.89%	87.74%	96.98%	52.17%	37.50%	48.48%	87.01%	67.95%	55.16%	39.92%	34.90%
CK-PlUG(Bi et al., 2025b)	80.90%	82.58%	90.09%	30.43%	21.43%	32.58%	67.01%	44.24%	49.15%	42.79%	16.11%
Astute (Wang et al., 2025a)	91.01%	89.03%	97.41%	35.40%	20.24%	33.33%	83.12%	70.20%	3.12%	52.25%	35.57%
GRPO w/ RAG	**96.07%**	**94.19%**	98.71%	72.67%	71.43%	70.83%	94.81%	74.94%	73.17%	68.94%	61.74%
`Knowledgeable-R1`	94.38%	93.55%	**98.71%**	**88.82%**	**75.60%**	**81.82%**	**96.62%**	**78.10%**	**83.55%**	**81.69%**	**71.14%**
improve	+4.49%	+5.81%	+1.73%	+36.65%	+38.10%	+33.34%	+9.61%	+10.15%	+28.39%	+41.77%	+36.24%

4.2.2 Analysis on Parametric-Knowledge Answerable Subset

We define the *parametric-knowledge answerable subset* as questions where the model provides correct answers using query-only prompting, showing that it has enough parametric knowledge. When focusing on this subset, all methods improve, but `Knowledgeable-R1` shows the biggest gains when the context is wrong or mixed (Table 3).

On `Qwen2.5-7B-Instruct`, `Knowledgeable-R1` achieves **89.61%/79.62%/66.00%** on NC-MR/MC/QA, greatly outperforming RAG prompting (**32.47%/19.75%/28.50%**) and GRPO w/ RAG (**57.61%/46.50%/62.50%**). Similar results are observed for `Llama3.1-8B-Instruct`. These results show that `Knowledgeable-R1` effectively learns when parametric knowledge is reliable and ignores conflicting context. In correct contexts (S1), `Knowledgeable-R1` maintains competitive performance.

The most impressive gains appear in S5 (partially relevant context), where `Knowledgeable-R1` achieves **78.49%** compared to **66.54%** for GRPO w/ RAG on HotpotQA with `Qwen2.5-7B-Instruct`, and **83.55%** compared to **73.17%** on `Llama3.1-8B-Instruct`. This shows that the learned policy effectively identifies useful evidence, keeping relevant information and ignoring distractions.

Table 4: Ablation study of Knowledgeable-R1 components. TITE: Both parametric and contextual knowledge correct; TIFE: parametric correct, contextual wrong; FITE: parametric wrong, contextual correct; FIFE: Both wrong. Best results are in **bold**, largest performance drops are underlined.

Method	PC-MR & NC-MR					PC-QA & NC-QA				
	TITE	TIFE	FITE	FIFE	Avg.	TITE	TIFE	FITE	FIFE	Avg.
Knowledgeable-R1	**95.21%**	**79.62%**	**69.06%**	**22.62%**	**65.40%**	**97.83%**	**66.00%**	**73.05%**	**11.93%**	**62.12%**
Multi-objective and multi-sampling strategy										
Knowledgeable-R1- J_{PK}	95.89%	75.80%	69.06%	21.51%	64.42%	97.83%	66.00%	72.04%	9.79%	61.30%
Knowledgeable-R1- J_{PK} - J_{RPK} (GRPO)	93.15%	46.50%	72.20%	10.42%	55.60%	97.28%	62.50%	72.04%	8.59%	60.07%
Local and global advantages										
Knowledgeable-R1- $A^{ck-local}$	96.58%	82.80%	66.82%	18.63%	64.69%	97.28%	64.50%	68.77%	9.07%	59.91%
Knowledgeable-R1- $A^{ck-global}$	93.15%	66.88%	71.75%	14.41%	60.95%	97.28%	54.00%	73.80%	7.64%	58.18%
Knowledge balance modulation										
Knowledgeable-R1- adapt β	93.15%	52.23%	72.65%	10.20%	56.95%	96.74%	53.50%	74.56%	6.21%	58.02%

4.2.3 Ablation Studies

We conduct ablation studies on the Qwen2.5-7B-Instruct model to evaluate the contribution of each Knowledgeable-R1 component (Table 4). The largest performance drops occur in TIFE (parametric knowledge correct, context knowledge incorrect) scenarios, highlighting the importance of each component in handling misleading context.

Multi-objective and Multi-sampling Strategy. We assess the impact of three policy types: PK (parametric knowledge), CK (context-aware), and RPK (robust parametric knowledge). Removing the parametric knowledge objective (J_{PK}) results in moderate performance drops, particularly in FIFE scenarios (-1.11% for MC, -2.14% for QA). The most significant degradation occurs when the relative parametric knowledge reward (J_{RPK}) is removed, causing catastrophic performance losses in TIFE scenarios (-33.12% for MC, -3.50% for QA). This demonstrates the crucial role of the RPK policy in ensuring reliable reasoning under contextual interference.

Local and Global Advantages. Removing the local context advantage ($A^{ck-local}$) causes a significant performance drop in context-answerable scenarios, while removing the global context advantage ($A^{ck-global}$) reduces TIFE performance. This suggests that local advantages boost the performance of their respective types, such as CK samples performing better when the context is correct. Local advantages also extend the influence of their type over others. These findings highlight the importance of local and global advantages in optimizing CK and RPK policies.

Knowledge Balance Modulation. The adaptive β modulation is vital for maintaining parametric knowledge as a fallback option. Using fixed β values reduces TIFE performance by 27.39% for MC and 12.50% for QA, demonstrating the importance of dynamic balancing between knowledge sources. The substantial performance drop in FIFE scenarios further highlights the effectiveness of adaptive balancing, particularly when both knowledge sources are unreliable. This supports our design of reducing penalties for RPK when parametric knowledge is slightly less effective than context, ensuring parametric knowledge remains usable in noisy contexts.

These results affirm that Knowledgeable-R1's components work synergistically to balance contextual and parametric knowledge, enabling robust reasoning in challenging scenarios.

5 Conclusion

We introduce Knowledgeable-R1, a reinforcement learning framework that uses joint sampling to balance contextual and parametric knowledge in large language models (LLMs). Our evaluation across five scenarios—correct, adversarial, self-conflicting, irrelevant, and partially relevant—shows significant improvements, especially in adversarial settings, while maintaining strong performance in reliable contexts. Future work will focus on expanding Knowledgeable-R1 to more complex environments, like multi-model RAG systems and integrating multiple exploration cues into a mixed objective-conditioned policy.

Ethics statement

We confirm that this work aligns with the ICLR Code of Ethics. We have considered potential ethical aspects, including the use of public datasets in compliance with their licenses, the broader societal impact of our research, and potential biases in our methodology. To the best of our knowledge, this work raises no immediate ethical concerns, and we declare no conflicts of interest.

Reproducibility statement

To facilitate the reproducibility of our work, we have made the following available: (1) the source code of our algorithm, which conducts training using the VERL framework (Sheng et al., 2025), along with supplementary materials; (2) complete experimental configurations in C. Additionally, a detailed account of the data pre-processing procedures for the datasets in Section 4.1.2 can be found in C. We also plan to release the full open-source code in the future.

Limitations

Our method introduces a reinforcement learning framework to address knowledge conflicts in large language models (LLMs), focusing on the automatic integration of parametric and contextual knowledge through sampling and policy optimization. In Sections S3 and S5, we evaluate scenarios with mixed-context types. However, the errors in the context from Section S5 have not been analyzed in sufficient detail. For example, when the amount of conflicting evidence changes (e.g., 1 incorrect result vs. 4 incorrect out of 5 retrieved), it's unclear how the sensitivity of Knowledgeable-R1 to parametric knowledge is affected. Future research should explore more complex scenarios and conduct experiments to understand how the model responds to different levels of conflict and evidence reliability.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *Int. Conf. Learn. Represent.*, 2024.

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 Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10280–10300, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.536.

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

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

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025.

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

Zhoujun Cheng, Richard Fan, Shibo Hao, Taylor W Killian, Haonan Li, Suqi Sun, Hector Ren, Alexander Moreno, Daqian Zhang, Tianjun Zhong, et al. K2-think: A parameter-efficient reasoning system. *arXiv preprint arXiv:2509.07604*, 2025.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. Contextcite: Attributing model generation to context. *Adv. Neural Inform. Process. Syst.*, 37:95764–95807, 2024.

Sam Devlin, Daniel Kudenko, and Marek Grześ. An empirical study of potential-based reward shaping and advice in complex, multi-agent systems. *Advances in Complex Systems*, 14(02): 251–278, 2011.

Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. Retrieve only when it needs: Adaptive retrieval augmentation for hallucination mitigation in large language models. *arXiv preprint arXiv:2402.10612*, 2024.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International conference on machine learning*, pp. 1407–1416. PMLR, 2018.

Yucheng Fang, Ruochen Wang, Kun Qian, Yansong Feng, Diyi Yang, and He He. Enhancing noise robustness of retrieval-augmented language models via RAAT. Bangkok, Thailand, 2024. Association for Computational Linguistics.

Mojtaba Farahani, Orion Weller, Dawn Lawrie, Benjamin Van Durme, and Benjamin Van Durme. Deciphering the interplay between parametric and contextual knowledge in language models. In *Conference on Empirical Methods in Natural Language Processing*, Miami, USA, 2024. Association for Computational Linguistics.

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

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.

Bolei He, Nuo Chen, Xinran He, Lingyong Yan, Zhenkai Wei, Jinchang Luo, and Zhen-Hua Ling. Retrieving, rethinking and revising: The chain-of-verification can improve retrieval augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10371–10393, 2024.

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, 2020.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in llm agents via incremental multi-turn interactions. *arXiv preprint arXiv:2507.05257*, 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 Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

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. *Conference on Empirical Methods in Natural Language Processing*, 2025a.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, et al. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*, 2025b.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 11:157–173, 2024.

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

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Int. Conf. Mach. Learn.*, volume 99, pp. 278–287. Citeseer, 1999.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A 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, 2023.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. In *Int. Conf. Learn. Represent.*, 2023.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

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

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Int. Conf. Mach. Learn.*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227. PMLR, 23–29 Jul 2023.

Yunxiao Shi, Xing Zi, Zijing Shi, Haimin Zhang, Qiang Wu, and Min Xu. Enhancing retrieval and managing retrieval: A four-module synergy for improved quality and efficiency in rag systems. In *ECAI 2024*, pp. 2258–2265. IOS Press, 2024.

Huatong Song, Jinhao Jiang, Wenqing Tian, Zhipeng Chen, Yuhuan Wu, Jiahao Zhao, Yingqian Min, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-searcher++: Incentivizing the dynamic knowledge acquisition of llms via reinforcement learning. *arXiv preprint arXiv:2505.17005*, 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. *Adv. Neural Inform. Process. Syst.*, 33:3008–3021, 2020.

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

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.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö Arık. Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. 2025a.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Retrieval-augmented generation with conflicting evidence. In *Conference on Language Modeling*, 2025b. Dataset: RAMDocs; also introduces MADAM-RAG.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Int. Conf. Learn. Represent.*, 2023.

Yilin Wen, Zifeng Wang, and Jimeng Sun. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. pp. 10370–10388, 2024.

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

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

Shicheng Xu, Liang Pang, Mo Yu, Fandong Meng, Huawei Shen, Xueqi Cheng, and Jie Zhou. Unsupervised information refinement training of large language models for retrieval-augmented generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 133–145, 2024.

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 *Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. *Int. Conf. Learn. Represent.*, 2024.

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

Eric Zelikman, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D. Goodman. Quiet-STaR: Language models can teach themselves to think before speaking. In *Int. Conf. Learn. Represent.*, 2024.

A Prompt designed in `Knowledgeable-R1` and baselines

To assess the performance of `Knowledgeable-R1`, we compared it against a range of different baseline methodologies:

Prompt-based Methods: These approaches include query-only prompting, Retrieval-Augmented Generation (RAG prompting), and Astute RAG (Wang et al., 2025a).

Decoding-based Methods: We experiment with CK-PLUG (Bi et al., 2025c).

Fine-tuning-based Methods: We evaluate `Knowledgeable-R1` against GRPO (Guo et al., 2025) with RAG prompting (GRPO $w/$ RAG).

These baselines encompass a broad spectrum of retrieval-enhanced and fine-tuning methods. To ensure a fair comparison, we standardized the contextual knowledge sources, training datasets, and Large Language Models (LLMs) used across all methods.

RAG prompting

{search_results}
You are a helpful assistant. After the user asks a question, you first think carefully and then give the answer.
When responding, please keep the following points in mind:
- The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively.
- Output your final answer directly between the tag <answer> </answer> without any intermediate steps.
Here is an example:
user's question: what is the capital of China?
<think> reasoning process here </think>
<answer> BeiJing </answer>

CK-PLUG prompt

You are a helpful assistant. After the user asks a question, you MUST direct give the final answer.
When responding, please keep the following points in mind:
- Output your final answer directly between the tag <answer> </answer> without any intermediate steps.
- You must directly output your final answer and don't output another things.
- Stop your output after final answer
Here is an example:
user's question: what is the capital of China?
<answer> BeiJing </answer>
Retrieved information:
{retrieved information}
Question:
{question}

Astute RAG prompt

Generate a document that provides accurate and relevant information to answer the given question. If the information is unclear or uncertain, explicitly state 'I don't know' to avoid any hallucinations.
Question: {question}
Document:
Task: Consolidate information from both memorized documents and externally retrieved documents in response to the given question.
For documents that provide consistent information, cluster them together.
For documents with conflicting information, separate them into distinct documents. Exclude any irrelevant information.
Question: {question}
Context: {context}
Provide consolidated documents in JSON format:
["content": "consolidated content", "source": ["doc ids"], "consistency$_{g}roup$” : ”$groupid$”]
Task: Answer the question using consolidated information from both internal and external documents.
Initial Context: {initial context}
Consolidated Context: {consolidated context}
After the user asks a question, you first think carefully and then give the answer. When responding, please keep the following points in mind:
- Answer the question using consolidated information from both internal and external documents.
- The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively.
- Output your final answer directly between the tag <answer> </answer> without any intermediate steps.
- If the user gives a multiple choice question, your answer must be a single option A or B or C or D or E
Here is an example:
Question:
what is the capital of China?
<think> reasoning process here </think>
<answer> BeiJing </answer>
Now, you should answer user's question. After answer user's question, you should stop generate. Here is user's question:
Question:
{question}

GRPO $w/$ RAG & `Knowledgeable-R1`-p' (CK)

You are a helpful assistant. After the user asks a question, you first think carefully and then give the answer.
When responding, please keep the following points in mind:
- The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively.
- Output your final answer directly between the tag <answer> </answer> without any intermediate steps.
Here is an example:
user's question: what is the capital of China?
<think> reasoning process here </think>
<answer> BeiJing </answer>
Retrieved information:
{retrieved information}
Question:
{question}

`Knowledgeable-R1`-p (PK)

You are a helpful assistant. After the user asks a question, you first think carefully and then give the answer.
When responding, please keep the following points in mind:
- The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively.
- Output your final answer directly between the tag <answer> </answer> without any intermediate steps.
Here is an example:
user's question: what is the capital of China?
<think> reasoning process here </think>
<answer> BeiJing </answer>
Question:
{question}

B Inference Input in Experiments

Inference Input in Experiments

You are a helpful assistant. After the user asks a question, you first think carefully and then give the answer.
When responding, please keep the following points in mind:
- The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively.
- Output your final answer directly between the tag <answer> </answer> without any intermediate steps.
Here is an example:
user's question: what is the capital of China?
<think> reasoning process here </think>
<answer> BeiJing </answer>
Retrieved information:
{retrieved information}
Question:
{question}

C Experimental Setup

Datasets. In our comprehensive evaluation of `Knowledgeable-R1`, we have considered a diverse range of benchmark datasets that represent various challenges of reasoning and knowledge, categorized into five scenarios in 4.1.2 . The specific datasets are: Multi-hop Question Answering with HotpotQA, 2WikiMultiHopQA, and Musique; Knowledge Conflict Question Answering with ConFiQA, where external knowledge retrieval contradictions necessitate internalized knowledge for judgement, and ExplainPE, a medical knowledge multiple-choice dataset that tests noise robustness by introducing three tiers of noise levels based on the number of non-matching external documents retrieved. For ConFiQA, we randomly add correct and incorrect contexts to the data for training in the same proportion. Our aim is to test `Knowledgeable-R1`'s capability to handle complex scenarios like single-hop answer, multi-hop reasoning and answer choice selection under different levels of external error or noisy context.

Training. We conduct experiments with `Qwen2.5-3B-Instruct`, `Qwen2.5-7B-Instruct`, `Qwen2.5-14B-Instruct`, and `Llama3.1-8B-Instruct`. All datasets, except HotpotQA, 2WikiMultiHopQA, and Musique, are trained and evaluated on their respective training and test sets. The ConfiQA dataset includes three subsets: MR, MC, and QA, which are similarly trained and evaluated on their respective sets. For HotpotQA, 2WikiMultiHopQA, and Musique, we limit the number of retrieved documents to 5 (the official setting uses 20) to ensure incomplete document evidence during training. At test time, we evaluate on the validation sets of these datasets.

For both GRPO-based baselines and our `Knowledgeable-R1`, we use a global batch size of 128, a rollout batch size of 32, a rollout temperature of 1, and a learning rate of 1×10^{-6}. All experiments are conducted on 8 H100 GPUs, with a consistent system prompt for both training and inference.

D Evaluation

For RAG generation capability. We use Exact Match (EM) as the evaluation metric following (Jin et al., 2025). Evaluation is conducted on the test or validation sets of all datasets to assess both in-domain and out-of-domain performance. For integrating parametric/contextual knowledge task, we adopted EM and self-defined metrics to evaluate the model's ability to utilize parametric and contextual knowledge.

E Method Training Procedure

Algorithm 1 `Knowledgeable-R1`: Policy Optimization for Knowledge Exploration

Input: Current policy π_θ, old policy $\pi_{\theta_{\text{old}}}$, dataset $\mathcal{D}$, training steps $Step_{\text{max}}$, parametric knowledge prompting rollout number n_{pk}, contextual knowledge prompt rollout number n_{ck}, clip parametric ϵ, advantage transformation function $T(\cdot)$, parametric knowledge prompt p, contextual knowledge prompt p' .
Output: Updated policy π_θ

1: **for** $s = 1$ to $Step_{\text{max}}$ **do**
2: Sample batch $(p) \sim \mathcal{D}$
3: Sample batch $(p') \sim \mathcal{D}$
4: Sample $\{o_i\}_{i=1}^{n_{pk}}$ from $\pi_{\theta_{\text{old}}}(o \mid p)$ ▷ Parametric knowledge rollouts
5: Sample $\{o'_j\}_{j=1}^{n_{ck}}$ from $\pi_{\theta_{\text{old}}}(o' \mid p')$ ▷ Contextual knowledge rollouts
6: Compute advantages A_i , A'_j and $\hat{A}_i$ for all rollouts
7: Define J_{PK} as the parametric knowledge based-policy objective:

$$J_{\text{PK}} = \frac{1}{n_{pk}} \sum_{i=1}^{n_{pk}} \sum_{t=1}^{|o_i|} \min \left[\frac{\pi_\theta\left(o_{i,t} \mid p, o_{i,<t}\right)}{\pi_{\theta_{\text{old}}}\left(o_{i,t} \mid p, o_{i,<t}\right)} A_i, \operatorname{clip}\left(\frac{\pi_\theta\left(o_{i,t} \mid p, o_{i,<t}\right)}{\pi_{\theta_{\text{old}}}\left(o_{i,t} \mid p, o_{i,<t}\right)}; 1-\epsilon, 1+\epsilon \right) A_i \right]$$

8: Define J_{CK} as the contextual knowledge based-policy objective:

$$J_{\text{CK}} = \frac{1}{n_{ck}} \sum_{j=1}^{n_{ck}} \sum_{t=1}^{|o'_j|} \min \left[\frac{\pi_\theta\left(o'_{j,t} \mid p', o'_{j,<t}\right)}{\pi_{\theta_{\text{old}}}\left(o'_{j,t} \mid p', o'_{j,<t}\right)} A'_j, \operatorname{clip}\left(\frac{\pi_\theta\left(o'_{j,t} \mid p', o'_{j,<t}\right)}{\pi_{\theta_{\text{old}}}\left(o'_{j,t} \mid p', o'_{j,<t}\right)}; 1-\epsilon, 1+\epsilon \right) A'_j \right]$$

9: Define J_{RPK} as the parametric knowledge reasoning policy objective with contextual knowledge input :

$$J_{\text{RPK}} = \frac{1}{n_{pk}} \sum_{i=1}^{n_{pk}} \sum_{t=1}^{|o_i|} \min \left[\frac{\pi_\theta\left(o_{i,t} \mid p', o_{i,<t}\right)}{\pi_{\theta_{\text{old}}}\left(o_{i,t} \mid p', o_{i,<t}\right)} A_i, \operatorname{clip}\left(\frac{\pi_\theta\left(o_{i,t} \mid p', o_{i,<t}\right)}{\pi_{\theta_{\text{old}}}\left(o_{i,t} \mid p', o_{i,<t}\right)}; 1-\epsilon, 1+\epsilon \right) T(\hat{A}_i) \right]$$

10: Combine $J_{\text{PK}}, J_{\text{CK}}, J_{\text{RPK}}$ to form the updated objective function $\mathcal{J}(\theta)$:

$$\mathcal{J}(\theta) = \lambda_{\text{pk}} J_{\text{PK}} + \lambda_{\text{ck}} J_{\text{CK}} + \lambda_{\text{rpk}} J_{\text{RPK}}$$

11: $\theta \leftarrow \theta + \nabla_\theta \mathcal{J}(\theta)$ ▷ Update current policy parametric
12: $\theta_{\text{old}} \leftarrow \theta$ ▷ Update old policy parametric
13: **end for**

Table 5: Overall accuracy across five contextual scenarios: correct (S1), wrong (S2), conflict (S3), irrelevant (S4), and partly-irrelevant (S5). Best results are in **bold**, second best are <u>underlined</u>. The "improve" row shows gains over RAG prompting.

	Correct (S1)			Wrong (S2)			Conflict (S3)	Irrelevant (S4)	Partly Irrelevant (S5)		
Method	PC-MR	PC-MC	PC-QA	NC-MR	NC-MC	NC-QA	SC	ExplainPE	HotPotQA	2Wiki MultiHopQA	Musique
`Qwen2.5-14B-Instruct`											
Query-only prompting	30.03%	26.52%	35.80%	30.64%	28.95%	39.42%	30.92%	70.80%	25.35%	26.69%	5.67%
RAG prompting	63.20%	63.85%	73.32%	22.22%	12.99%	28.11%	61.85%	70.51%	22.90%	22.49%	6.91%
GRPO w/ RAG	**73.60%**	<u>74.66%</u>	**78.86%**	<u>38.89%</u>	31.74%	<u>36.51%</u>	**74.00%**	-	<u>34.71%</u>	<u>39.03%</u>	**14.73%**
`Knowledgeable-R1`	70.63%	**75.17%**	<u>78.83%</u>	**47.81%**	**33.06%**	**38.13%**	<u>72.50%</u>	-	**36.98%**	**42.33%**	<u>14.60%</u>
improve	+7.43%	+11.32%	+5.51%	+24.99%	+20.07%	+10.02%	+10.65%	-	+14.08%	+19.84%	+7.69%
`Qwen2.5-7B-Instruct`											
Query-only prompting	27.72%	24.66%	31.67%	25.93%	25.82%	32.31%	29.67%	64.45%	20.90%	25.54%	4.36%
RAG prompting	65.68%	66.39%	74.35%	13.47%	8.06%	11.31%	59.50%	62.21%	20.36%	22.53%	6.41%
Astute (Wang et al., 2025a)	65.51%	66.05%	77.62%	12.79%	7.07%	10.34%	54.20%	56.74%	17.87%	20.35%	6.29%
GRPO w/ RAG	**77.56%**	**77.36%**	80.03%	26.94%	19.74%	<u>26.01%</u>	<u>75.33%</u>	<u>66.50%</u>	<u>27.93%</u>	<u>33.95%</u>	<u>11.79%</u>
`Knowledgeable-R1`	75.08%	<u>75.51%</u>	**80.90%**	**43.94%**	**37.34%**	**29.40%**	**76.33%**	**67.57%**	**31.45%**	**37.52%**	**12.04%**
improve	+9.41%	+9.12%	+6.54%	+30.47%	+29.28%	+18.09%	+15.92%	+5.36%	+11.09%	+14.99%	+5.63%
`Qwen2.5-3B-Instruct`											
Query-only prompting	16.83%	18.41%	23.92%	15.99%	18.26%	22.62%	20.75%	53.32%	12.55%	11.28%	2.11%
RAG prompting	55.12%	52.87%	59.55%	13.47%	8.72%	6.79%	51.25%	42.19%	14.71%	13%	3.64%
Astute (Wang et al., 2025a)	62.05%	60.47%	67.99%	9.60%	7.07%	10.66%	51.92%	41.60%	0.8%	8.72%	5.01%
GRPO w/ RAG	**70.96%**	**70.61%**	**80.03%**	<u>19.87%</u>	<u>15.13%</u>	<u>19.39%</u>	<u>66.67%</u>	<u>53.61%</u>	<u>24.78%</u>	<u>35.57%</u>	**9.06%**
`Knowledgeable-R1`	<u>66.17%</u>	<u>59.80%</u>	<u>78.66%</u>	**28.79%**	**28.62%**	**21.97%**	-	**54.69%**	-	-	-
improve	+11.05%	+6.93%	+19.11%	+15.32%	+19.9%	+15.15%	-	+1.37%	-	-	-

F Additional Experiment Results

We validated the performance improvement of our method across models of varying sizes with 3B, 7B, 14B. As shown in the Table 5, our approach achieved significant enhancements compared to RAG prompt in all five scenarios. It also showed considerable improvement over GRPO w/RAG when external knowledge contained errors or incomplete information, and a slight improvement when external knowledge served as interference. This demonstrates the effectiveness and versatility of our method.

G Use of LLMs

Large language models (LLMs), specifically GPT-5 and DeepSeek-R1, were used solely as a supplementary tool during the preparation of this work for tasks such as polishing the writing. The authors are solely responsible for the entire research conception, technical direction, scientific content, and interpretation of results. The LLMs were employed only to assist in the presentation and clarity of the manuscript.