RoseRAG: Robust Retrieval-augmented Generation with Small-scale LLMs via Margin-aware Preference Optimization

Tianci Liu[•§], Haoxiang Jiang[•†], Tianze Wang[⋆], Ran Xu[◇], Yue Yu[*], Linjun Zhang[⋆], Tuo Zhao[*], Haoyu Wang[†]

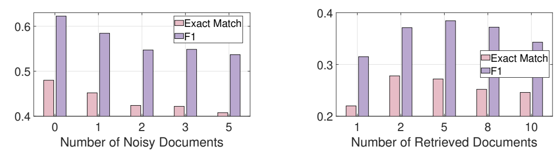
[§]Purdue University, [†]University at Albany,
[⋆]Rutgers University, [◇]Emory University, [*]Georgia Institute of Technology
[†]{hjiang2,hwang28}@albany.edu, [§]liu3351@purdue.edu
[⋆]{tianze.wang,linjun.zhang}@rutgers.edu, [◇]ran.xu@emory.edu
[*]{yueyu,tourzhao}@gatech.edu

Abstract

Large language models (LLMs) have achieved impressive performance but face high computational costs and latency, limiting their deployment in resource-constrained settings. In contrast, small-scale LLMs (SLMs) are more efficient yet struggle to capture evolving real-world knowledge. Retrieval-augmented generation (RAG) helps by integrating external knowledge, but imperfect retrieval can introduce distracting noise that misleads SLMs. We propose RoseRAG, a robust RAG framework for SLMs via Margin-aware Preference Optimization. RoseRAG employs multi-turn prompting for detailed reasoning, rejection sampling for high-quality explanations, and contrastive preference selection to refine responses by maximizing the likelihood gap between preferred and non-preferred outputs. By integrating these components into a margin-aware optimization process, RoseRAG robustly enhances the accuracy and reliability of SLMs for RAG applications. Extensive experiments on three open-domain question answering benchmarks indicate that our innovative RoseRAG surpasses state-of-the-art baselines significantly.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in a wide array of natural language processing tasks (Achiam et al., 2023; Team et al., 2024; Dubey et al., 2024; Guo et al., 2025). However, these powerful models are typically large-scale, requiring substantial computational resources for training, and often incurring high latency during inference (Zhou et al., 2024). Such limitations serve as a key hurdle that prevents these models from being deployed in real practice. In contrast, small-scale LLMs (SLMs) offer a viable alternative by providing high utility while remaining computationally efficient and easier to deploy in resource-constrained environments (Lu et al., 2024; Vernikos et al., 2024).



(a) (b)

Figure 1: Pilot studies. Fig. 1a: Ground Truth Documents with varying amounts of noisy documents. Fig. 1b: Performance w.r.t. varying numbers of retrieved documents. Both the two sub-figures are results with Qwen2.5-1.5B-Instruct on HotPotQA.

Despite their efficiency, SLMs are fundamentally constrained by their limited capacity. During pre-training, they cannot fully capture the vast and continuously evolving body of real-world knowledge (Ovadia et al., 2024). As a result, SLMs often struggle in real-world scenarios where accurate responses depend on newly emerging or dynamically updated information. Retrieval-augmented generation (RAG) (Lewis et al., 2020) mitigates this limitation by retrieving a top-K set of semantically relevant documents at inference time, which are then conditioned upon during response generation. By decoupling knowledge retrieval from parametric memory, RAG enhances the adaptability of SLMs, improving response fidelity while obviating the need for expensive and frequent model retraining (Wang et al., 2023; Asai et al., 2024; Xu et al., 2024b). However, one inherent challenge for RAG pipelines is the *imperfect retrieval information*. The top-K documents returned by a retriever may include irrelevant or noisy information, which can mislead small-scale LLMs that are not robust enough to effectively filter out such distractions. As shown in Fig. 1, the susceptibility of SLMs to retrieval noise highlights a critical bottleneck in RAG pipelines. This limitation highlights the necessity of developing more robust mechanisms to enhance

[•]These authors contributed equally to this work

SLMs' resilience against spurious retrieval artifacts and improve their reliability in downstream tasks.

Several existing works have sought to enhance the robustness of RAG systems. *Prompting-based approaches* (Wang et al., 2024a,b) attempt to mitigate noise by instructing the model to disregard irrelevant information during inference. However, these techniques heavily depend on the model's intrinsic ability to distinguish useful from spurious content—a capability that SLMs often lack due to their limited reasoning and generalization capacity. On the other hand, *Fine-tuning-based strategies* (Yoran et al., 2024; Wei et al., 2025) aim to improve retrieval robustness by training models on curated, denoised datasets using supervised fine-tuning (SFT). Unfortunately, SFT tends to mimic the behavior present in the training data, making it highly sensitive to noise and prone to overfitting (Chu et al., 2025), which ultimately limits the generalization ability of the model.

In light of these challenges, we propose **Ro**bu**s**t R**e**t**r**ieval-**a**ugmented **G**eneration with Small LLMs via Margin-aware Preference Optimization (ROSERAG). ROSERAG introduces a novel framework that robustly aligns SLMs with high-quality responses *without distillation from teacher LLMs* through a three-stage process: *preference data generation*, *contrastive preference data selection*, and *margin-aware preference optimization*. Specifically, we employ a multi-turn prompting strategy to generate detailed rationales, coupled with rejection sampling (Stiennon et al., 2020; Guo et al., 2025) to filter out spurious reasoning, thereby mitigating the influence of noisy retrieved documents. Besides, our contrastive selection mechanism identifies the most *challenging* response pairs, enabling the model to explicitly maximize the margin between preferred and non-preferred outputs. By integrating these components into a unified optimization framework, ROSERAG significantly improves the reliability and accuracy of small-scale LLMs in retrieval-augmented settings, especially under imperfect retrieval conditions that mirror real-world scenarios.

Our contributions are summarized as follows:

- We propose ROSERAG, a novel RAG framework that enhances SLM robustness against noisy retrieval using margin-aware preference optimization, reducing dependence on distillation from stronger models.
- We propose a multi-turn prompting strategy coupled with rejection sampling to generate and filter robust reasoning rationales, for boosting the quality of preference data.
- We design a contrastive preference data selection scheme to maximize the margin between chosen and rejected responses, leading to more discriminative and generalizable model behavior.
- We conduct extensive experiments to demonstrate that RoseRAG significantly improves response quality in retrieval-augmented settings, paving the way for more effective deployment of small-scale LLMs in real-world applications.

2 Related Work

Retrieval-augmented generation (RAG) enhances Large Language Models (LLMs) by integrating non-parametric knowledge to improve generation quality. Early approaches (Izacard et al., 2023; Shi et al., 2024) treat LLMs as passive consumers of retrieved information, assuming that retrieval inherently improves generation. However, recent studies (Wang et al., 2023; Li et al., 2023a; Yu et al., 2024b) highlight that retrieved knowledge is often noisy, incomplete, or even misleading, which can hurt rather than enhance the performance LLMs.

To better align the retrieved information with LLMs, Wang et al. (2023); Jeong et al. (2024) propose initial assessments to determine whether retrieval is necessary, using either BERT-family models or the frozen LLM as a classifier. However, these approaches depend on classifier accuracy and does not improve the LLM's inherent ability to handle noisy information. BlendFilter (Wang et al., 2024b), ASTUTE (Wang et al., 2024a), and RobustRAG (Xiang et al., 2024) leverage the LLM itself to analyze or filter irrelevant information. However, their effectiveness is contingent on the model's capability; small-scale LLMs often fail to achieve reliable results due to their limited intelligence. Another line of research (Fang et al., 2024; Yoran et al., 2024; Yu et al., 2024a,b) investigates training LLMs to handle noisy contexts, e.g., those retrieved from external corpora. These approaches typically leverage powerful models such as GPT-4 or require extensive labeled data from auxiliary tasks to generate high-quality responses. However, such reliance on large-scale models and costly annotations limits scalability and practicality, particularly for resource-constrained applications. Very recently, InstructRAG (Wei et al., 2025) instructs LLMs to provide rationales linking answers to retrieved passages, but pure supervised fine-tuning cannot fully unleash the model's reasoning capability towards RAG applications. KnowPO (Zhang et al., 2024) and DPA-RAG (Dong et al., 2024) optimize model preferences to improve noisy information analysis. Yet, they still require access to powerful LLMs (e.g. GPT-4) to create preference data. Unlike existing approaches, RoseRAG is specifically designed to enhance the robustness of SLMs against noisy retrieved information through margin-aware preference optimization, eliminating the need for additional classifiers or high-resource LLMs.

3 Preliminary

In this section, we introduce the concept of standard retrieval-augmented generation (RAG) and monolithic preference optimization (ORPO).

3.1 Retrieval-augmented Generation

Given a pre-trained Large Language Model (LLM) $\mathcal{M}_\theta$, a knowledge base $\mathcal{K} = \{\mathcal{K}_i\}_{i=1}^k$ (where k represents the number of documents), a retriever $\mathcal{R}(\cdot)$, and a query q, the vanilla RAG, i.e. retrieve-then-generate, is to retrieve top-K related documents from the knowledge base first and then generate answer based on retrieved information, which can be formulated as

$$\mathcal{K}_q = \mathcal{R}(q, \mathcal{K}; K),$$
$$y \sim P_\theta(y|\texttt{Prompt}_{\texttt{CoT}}(q, \mathcal{K}_q)), \tag{1}$$

where y and $\texttt{Prompt}_{\texttt{CoT}}(\cdot)$ represent the generated response and the chain-of-thought (CoT) prompt, respectively.

3.2 Monolithic Preference Optimization

Preference alignment for large language models has traditionally relied on multi-stage procedures—such as reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022)—that require an additional reference model to guide and stabilize training. In contrast, ORPO (Hong et al., 2024) is a monolithic approach that integrates preference alignment directly into the supervised fine-tuning (SFT) phase, thereby obviating the need for a separate reference model. ORPO augments the standard negative log-likelihood loss with an odds ratio-based penalty that contrasts the probabilities of generating a *chosen* (preferred) response and a *rejected* (non-preferred) response. Specifically, given an input x and corresponding responses y_w (chosen) and y_l (rejected), the odds of generating a response are defined as $\text{odds}_\theta(y \mid x) = \frac{P_\theta(y|x)}{1-P_\theta(y|x)}$, and the odds ratio is given by $\text{OR}_\theta(x, y_w, y_l) = \frac{\text{odds}_\theta(y_w|x)}{\text{odds}_\theta(y_l|x)}$. The overall loss function is formulated as

$$\ell_{\text{ORPO}}(x, y_w, y_l) = \ell_{\text{SFT}} + \beta\left(-\log\sigma(\log \text{OR}_\theta(x, y_w, y_l))\right), \tag{2}$$

where ℓ_{SFT} is the conventional supervised fine-tuning loss, $\sigma(\cdot)$ denotes the sigmoid function, and β is a hyperparameter that regulates the strength of the preference alignment signal. By explicitly encouraging a larger margin between the chosen and rejected responses, ORPO enables more stable gradient updates and improved alignment performance, as evidenced by its strong empirical results on benchmarks such as AlpacaEval (Li et al., 2023b; Dubois et al., 2024, 2023) and MT-Bench (Zheng et al., 2023).

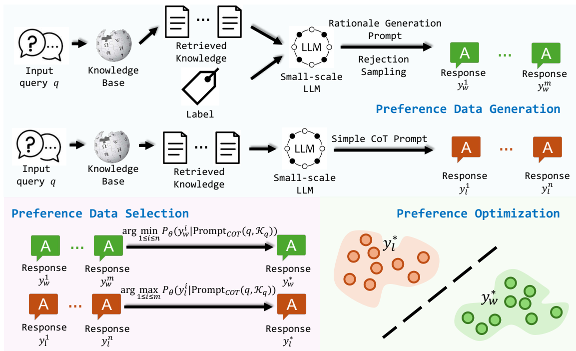


Figure 2: Framework of the proposed RoseRAG.

4 Method

Our RoseRAG enhances retrieval-augmented generation for SLMs through margin-aware preference optimization. It consists of three key stages, as shown in Fig. 2: (1) *Preference Data Generation*, where the model is prompted with retrieved knowledge and ground-truth answers to generate rationales, filtered via rejection sampling; (2) *Preference Data Selection*, which employs a contrastive strategy to maximize the margin between the least likely chosen response and the most likely rejected response; and (3) *Preference Optimization*, where the model is trained using an ORPO loss. This unified framework effectively aligns SLM outputs with high-quality responses, thereby improving model robustness in retrieval-augmented settings.

4.1 Preference Data Generation

In order to enable our model to accurately evaluate the relevance of retrieved documents and generate reliable rationales, we design a multi-turn prompting strategy that encourages the model to articulate its reasoning process. The core idea is to guide the model into providing a concise, step-by-step explanation for arriving at an answer by explicitly informing the model that the given knowledge may consist of irrelevant information, while also increasing the likelihood of correctness by including the ground-truth label in the prompt. By explicitly supplying the ground-truth answer, the model is pushed toward generating a rationale that is consistent with the expected output, thus facilitating the subsequent preference optimization.

We instantiate this process by constructing a prompt composed of a system message, a user message, and an initial assistant message. The system message defines the task and constraints, the user message provides a set of retrieved knowledge documents $\mathcal{K}_q$, a question q, and the ground-truth answer a^*, and the assistant message seeds the reasoning process. The prompt is presented in Fig. 3. Given these messages, the model generates a context r that encapsulates its analytical reasoning. To mitigate the risk of incorrect analysis, we employ rejection sampling (Liu et al., 2023; Guo et al., 2025): we sample an answer $a \sim P_\theta(y \mid q, r)$ and compare it with the ground-truth label. If a matches the ground truth, the generated context r will be retained as the chosen response y_w; otherwise, it will be filtered. For the rejected response y_l, we adopt the vanilla RAG response (as defined in Eqn. (1)). This strategy not only leverages the ground-truth label to promote the generation of correct rationales, but also filters out spurious reasoning. Therefore, it does not depend on expensive close-sourced powerful LLMs like ChatGPT (Ouyang et al., 2022), Gemini (Google, 2024), and Claude (Bai et al., 2022), and can enable SLMs to produce high-quality preference data as well.

Rationale Generation

System Prompt: You are a useful assistant. I will provide one question, several pieces of knowledge (which may be related or unrelated to the question), and the answer to the question. Please explain your reasoning process in a single paragraph consisting of no more than four sentences. If the provided knowledge is insufficient, you may make an informed guess, but do not respond with "Unknown".

User Prompt: Knowledge: $\mathcal{K}_q$
Question: q
Answer: a^*

Assistant Prompt: Let's think step by step.

Output: {rationale r}

Figure 3: Overview of the rationale generation process.

4.2 Preference Data Selection

We propose a preference data selection method based on contrastive learning (Tian et al., 2020; Wang and Qi, 2022; Cui et al., 2021; Chen et al., 2020; Li et al., 2020), designed to improve the model's ability to distinguish between chosen and rejected responses. The underlying intuition is that explicitly maximizing the margin between the preferred (chosen) and non-preferred (rejected) outputs forces the model to learn more discriminative representations, thereby enhancing its generalization capability.

In practice, after the Preference Data Generation step, for one query q we will have n candidate chosen responses $\{y_w^1, y_w^2, \ldots, y_w^n\}$ and m candidate rejected responses $\{y_l^1, y_l^2, \ldots, y_l^m\}$. Each candidate is evaluated by its likelihood under the initial model θ, denoted as $P_\theta(y \mid \texttt{Prompt}_{\texttt{CoT}}(q, \mathcal{K}_q))$. To maximize the contrast between the two sets, we select the candidate chosen response that the model is least confident about, and the candidate rejected response that it is most confident with (Robinson et al.). Formally, the final selected responses are defined as:

$$\begin{aligned} y_w^* &= \arg\min_{1\le i\le n} P_\theta(y_w^i \mid \texttt{Prompt}_{\texttt{CoT}}(q, \mathcal{K}_q)), \\ y_l^* &= \arg\max_{1\le i\le m} P_\theta(y_l^i \mid \texttt{Prompt}_{\texttt{CoT}}(q, \mathcal{K}_q)). \end{aligned} \tag{3}$$

The motivation behind this selection strategy is two-fold. First, by choosing the chosen response with the minimal likelihood, we identify cases where the model struggles to assign high confidence to the correct answer; such instances provide a strong corrective signal that refines the model's understanding of the desired output. Second, by selecting the rejected response with the maximal likelihood, we target cases where the model erroneously favors an undesirable output. This contrastive selection process accentuates the differences between correct and incorrect responses, thereby forcing the model to maximize the margin between them and promoting a more robust and effective preference optimization. Moreover, by ensuring that the selected rationales yield the correct answer, we minimize false positives, further strengthening the overall training signal.

4.3 Preference Optimization

Given a preference tuple (x, y_w^*, y_l^*), where x is the input prompt $\texttt{Prompt}_{\texttt{CoT}}(q, \mathcal{K}_q)$, y_w^* is the chosen (preferred) response, and y_l^* is the rejected (non-preferred) response, our objective is to minimize the ORPO loss:

$$\min_{\theta} \mathbb{E}_{(x,y_w^*,y_l^*)} \ell_{\text{ORPO}}(x, y_w^*, y_l^*). \tag{4}$$

The y_w^*, and y_l^* are obtained from our preference data generation and selection steps as stated before.

5 Theoretical Analysis

In this section, we provide a theoretical analysis to support the effectiveness of ROSERAG. We derive a closed-form solution for the model learned by ROSERAG and examine how the proposed preference selection strategy influences the model's behavior.

We begin with a regularity condition ensuring that the optimization space is sufficiently expressive to achieve the global optimum.

Assumption 5.1. Assume that $P(y|x)$ belongs to the optimization space $\{P_\theta : \theta \in \Theta\}$ such that

$$P(y|x) = \exp(Z(x)) / \left(\frac{q_l^*(y|x)}{q_w^*(y|x)} + \exp(Z(x)) \right),$$

where $q_w^*(y|x)$ and $q_l^*(y|x)$ denote the distribution of y_w^* and y_l^* given the prompt x respectively.

This condition ensures that P_θ can attain an optimal solution, allowing us to express the optimizer in a closed form.

Lemma 5.1. Under Assumption 5.1, the solution to optimizing Eqn. (4) is

$$P_\theta(y|x) = \frac{\exp(Z(x))}{\frac{q_l^*(y|x)}{q_w^*(y|x)} + \exp(Z(x))} \tag{5}$$

where $Z(x)$ is partition function such that $\sum_y P_\theta(y|x) = 1$.

To illustrate the benefit of the proposed preference selection strategy, we consider a scenario where y_w and y_l are random variables following

$$\begin{aligned} y_w &\sim f_w(x) + \text{Exp}(\lambda), \\ y_l &\sim f_l(x) + \text{Exp}(\lambda), \end{aligned} \tag{6}$$

where $\text{Exp}(\lambda)$ denotes an exponential random variable with rate λ, and $f_w(x)$ and $f_l(x)$ represent the central locations of y_w and y_l, respectively. To ensure a meaningful selection process, we assume $f_w(x) > f_l(x)$, indicating that y_w is generally preferred over y_l.

We compare our method to a baseline that applies ORPO without preference selection (i.e., $n = 1$ in Eqn. (3)). Let $\tilde{P}_\theta(y|x)$ denote the solution obtained without selection. We measure the response quality using the absolute distance from the expected preferred response, $L(y) = |y - \mathbb{E}[y_w]|$. The following theorem formalizes the advantage of our method. The proof is deferred to App C.

Theorem 5.1. Under (6), assume $\tilde{P}_\theta(y|x)$ generates y with density function:

$$p \cdot \lambda e^{-\lambda[y - f_w(x)]} + (1-p) \cdot \lambda e^{-\lambda[y - f_l(x)]}.$$

Let $y \sim P_\theta(y|x)$ and $\tilde{y} \sim \tilde{P}_\theta(y|x)$. Then, the expected absolute distance under P_θ is smaller than that under $\tilde{P}_\theta$:

$$\mathbb{E}_{y \sim P_\theta(y|x)}[L(y)] < \mathbb{E}_{\tilde{y} \sim \tilde{P}_\theta(y|x)}[L(\tilde{y})]. \tag{7}$$

Theorem 5.1 suggests that the proposed preference selection strategy improves response alignment with the preferred choice y_w. Our analysis reveals that this selection process amplifies the gap between preferred and non-preferred responses, compelling the model to enhance their separation and ultimately leading to more accurate and reliable outputs.

6 Experiment

In this section, we extensively evaluate the proposed ROSERAG and answer the following questions: RQ1) How does ROSERAG perform compared to state-of-the-art baselines? RQ2) What are the roles of rejection sampling and preference data selection in model performance improvements respectively? RQ3) Can the proposed ROSERAG benefit from more retrieved documents? RQ4) Is it possible to apply different preference optimization method to ROSERAG? RQ5) How does the performance change with varying β?

6.1 Datasets and Experiment Settings

Datasets and Evaluation Metrics. We conduct experiments on three public benchmarks, including HotPotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), and StrategyQA (Geva et al., 2021). Following Shao et al. (2023); Wang et al. (2024b), we evaluate the first 500 questions from the development dataset for HotPotQA and 2WikiMultiHopQA, and evaluate questions from the development dataset for StrategyQA. For multi-hop question answering datasets, we employ exact match (EM) and F1 as evaluation metrics, and for the commonsense reasoning dataset, which is a binary classification task, we use accuracy and F1 score as the metrics. To evaluate the retrieval performance, we leverage widely used Recall as the evaluation metric.

Baselines. We adopt following state-of-the-art baselines to evaluate against ROSERAG: 1) CoT Prompting (Wei et al., 2022), 2) ReAct (Yao et al., 2022), 3) SelfAsk (Press et al., 2023), 4) BlendFilter (Wang et al., 2024b), 5) InstructRAG (Wei et al., 2025), 6) RetRobust (Yoran et al., 2024), 7) ASTUTE (Wang et al., 2024a), and 8) ICL+RAG (Park et al., 2024). We show more detailed information about baselines in the Appendix A.

Implementation Details. We evaluate models with three small-sclae LLMs: Qwen2.5-1.5B-Instruct (Team, 2024), Llama-3.2-1B-Instruct[1], and gemma-2-2b-it (Team et al., 2024). We utilize the state-of-the-art efficient retrieval method ColBERT v2 (Santhanam et al., 2022) as the retriever implemented by Khattab et al. (2022, 2023). The knowledge base we employ is the collection of Wikipedia abstracts dumped in 2017 (Khattab et al., 2023). We show the detailed information about our implementations in the Appendix B.

6.2 Performance Comparison (RQ1)

To evaluate the effectiveness of the proposed ROSERAG framework, we conduct experiments on three benchmark datasets: HotPotQA, 2WikiMultiHopQA, and StrategyQA. We compare RoseRAG against multiple retrieval-augmented generation (RAG) baselines using three different small-scale LLM backbones: Qwen2.5-1.5B-Instruct, Llama-3.2-1B-Instruct, and gemma-2-2B-it. From the results presented in Tables 1, 2, and 3, we observe two key findings.

[1] https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md

[2] ASTUTE relies on powerful LLM. We cannot extract answers when using Llama-3.2-1B-Instruct.

Table 1: Performance of ROSERAG with Qwen2.5-1.5B-Instruct as the backbone.

Method	HotPotQA		2WikiMultiHopQA		StrategyQA	
	EM	F1	EM	F1	Acc	F1
CoT	13.0	20.9	18.8	23.5	55.9	17.9
vanilla RAG	27.2	38.4	6.2	11.4	55.5	23.9
ReAct	14.9	25.9	9.6	22.2	55.0	48.8
SelfAsk	20.4	32.8	19.2	25.3	51.1	42.9
BlendFilter	26.4	37.5	19.4	24.2	59.4	45.0
InstructRAG	31.2	39.6	23.6	26.9	53.7	39.1
RetRobust	16.8	24.6	13.0	19.9	51.5	37.3
ASTUTE	21.4	27.8	19.0	23.4	55.0	12.0
ICL+RAG	28.6	39.3	21.8	26.4	56.3	27.5
ROSERAG	**34.8**	**44.8**	**31.6**	**35.0**	**59.8**	**52.1**

Table 2: Performance of ROSERAG with Llama-3.2-1B-Instruct as the backbone.

Method	HotPotQA		2WikiMultiHopQA		StrategyQA	
	EM	F1	EM	F1	Acc	F1
CoT	13.6	19.5	13.6	19.0	52.8	23.9
vanilla RAG	27.8	37.0	16.0	21.9	54.6	33.3
ReAct	15.4	24.7	10.2	15.9	51.1	61.9
SelfAsk	12.0	18.0	12.4	18.5	51.1	42.9
BlendFilter	20.6	29.7	17.8	23.1	55.9	34.8
InstructRAG	28.8	39.2	23.8	27.4	56.8	29.8
RetRobust	17.6	26.4	20.0	26.5	56.8	39.3
ASTUTE[2]	-	-	-	-	58.1	35.1
ICL+RAG	25.4	36.1	19.2	24.2	54.6	29.7
RoseRAG	**33.4**	**44.1**	30.2	**35.8**	**61.1**	43.3

Table 3: Performance of ROSERAG with gemma-2-2b-it as the backbone.

Method	HotPotQA		2WikiMultiHopQA		StrategyQA	
	EM	F1	EM	F1	Acc	F1
CoT	20.6	27.9	20.4	24.5	60.7	44.4
vanilla RAG	36.4	46.6	15.2	20.2	56.8	39.3
ReAct	26.6	38.1	21.0	26.9	55.5	45.2
SelfAsk	32.8	43.5	23.4	30.5	60.7	45.8
BlendFilter	34.6	45.4	23.2	29.6	62.5	49.4
InstructRAG	38.0	49.4	29.0	35.0	60.3	59.2
RetRobust	30.4	40.4	22.0	26.5	64.2	54.9
ASTUTE	22.4	32.5	14.8	19.5	62.0	42.0
ICL+RAG	32.4	42.9	16.4	22.5	62.4	54.3
RoseRAG	**42.4**	**54.0**	**37.2**	**42.7**	**67.7**	**60.2**

First, the proposed ROSERAG consistently outperforms all baseline methods across different datasets and model backbones, demonstrating its effectiveness in retrieval-augmented generation for small-scale LLMs. The substantial improvements in EM and F1 scores indicate that the margin-aware preference optimization and rejection sampling in RoseRAG significantly enhance reasoning accuracy. Notably, compared to InstructRAG, RoseRAG achieves superior performance, highlighting the necessity of incorporating preference optimization in addition to supervised fine-tuning.

Second, the results reveal that SLMs exhibit distinct properties compared to powerful large-scale LLMs, which impacts the effectiveness of various RAG methods. Many retrieval-based techniques that have shown strong performance with large-scale models fail to maintain similar improvements on small models. Methods such as ReAct and Self-Ask, which decompose original queries into sub-questions, struggle because SLMs lack the reasoning ability to perform accurate decomposition. Similarly, approaches like BlendFilter and ASTUTE, which rely on the LLM itself to filter irrelevant information, perform poorly since small models are less capable of distinguishing irrelevant from noisy content. The performance of InstructRAG and ROSERAG highlights the necessity of fine-tuning small-scale LLMs to enhance their ability to process noisy retrieval results effectively. These findings underscore the importance of adapting RAG strategies specifically for small-scale SLMs rather than directly transferring techniques optimized for large-scale LLMs.

Table 4: Performance (Exaxt Match/F1) of ROSERAG using different types of generated preference data on HotPotQA with Qwen2.5-1.5B-Instruct.

		Positive		
		w/o Selection	Minimal	Maximal
Negative	w/o Selection	30.6/39.9	33.2/43.1	27.6/37.1
	Minimal	29.0/39.2	33.2/42.2	26.6/35.3
	Maximal	33.0/43.6	**34.8/44.8**	28.8/38.4

6.3 Effectiveness of Data Selection (RQ2)

To assess the impact of our preference data selection strategy on performance, we conduct experiments on HotPotQA using Qwen2.5-1.5B-Instruct as the backbone. In this study, we vary the selection method for both the positive (chosen) and negative (rejected) responses, comparing three scenarios: no selection, selection based on the minimal likelihood, and selection based on the maximal likelihood. The performance—measured in Exact Match (EM) and F1 scores—is reported in Table 4. Based on the table, we have following findings:

Overall Effectiveness. Introducing a selection strategy for the preference data markedly improves performance compared to the baseline without any selection. For example, when only the positive responses are refined using the minimal likelihood

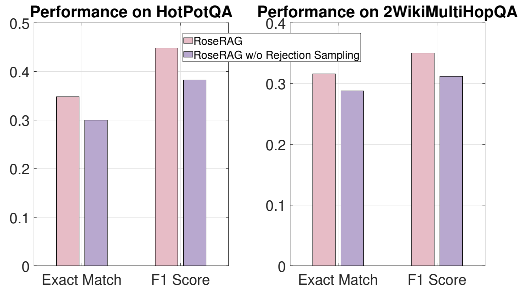


Figure 4: Accuracy of ROSERAG with and without rejection sampling with Qwen2.5-1.5B-Instruct.

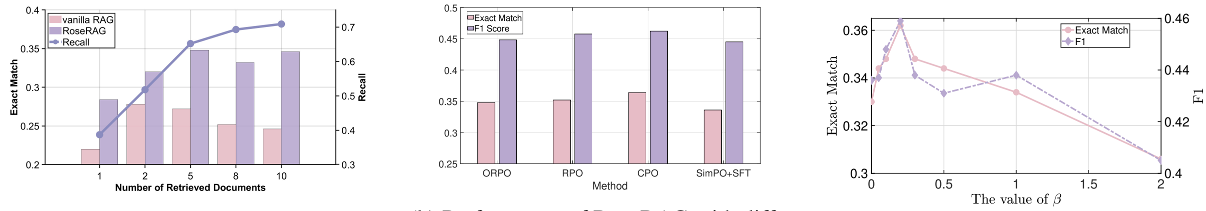
criterion (with negative responses remaining unselected), the EM/F1 improves from 30.6/39.9 to 33.2/43.1. This result confirms that our preference data selection module effectively enhances the quality of the training signal.

Optimal Selection Strategy. The optimal performance (34.8/44.8) occurs when positive responses have the lowest likelihood and negative responses have the highest. This aligns with our intuition: the minimal likelihood positive response likely represents a scenario where the model is less confident and, therefore, benefits more from corrective feedback. Simultaneously, selecting the maximal likelihood negative response targets cases where the LLM is confidently incorrect, offering a strong contrastive signal. Using the opposite criteria (maximal likelihood for positive/minimal likelihood for negative) leads to a notable performance drop.

Importance of Positive Selection. The results suggest that selecting the minimal likelihood candidate for the positive response is particularly critical, as it plays a direct role in the negative log-likelihood (NLL) loss. By emphasizing these uncertain yet correct outputs, the model receives a more effective corrective signal, facilitating better alignment with the desired responses.

6.4 Effectiveness of Rejection Sampling (RQ2)

In this section, we evaluate the effectiveness of rejection sampling in ROSERAG. Specifically, we compare the performance of ROSERAG with and without rejection sampling on two benchmark datasets: HotPotQA and 2WikiMultiHopQA. From results in Fig. 4, we can find that rejection sampling is important for performance improvement. On HotPotQA, rejection sampling improves both the Exact Match and F1 scores significantly. Specifically, the EM score increases from 0.33 to 0.41, and the F1 score rises from 0.37 to 0.46. This demon-



(a) Accuracy with respect to different number of retrieved documents.

(b) Performance of RoseRAG with different preference optimization method.

(c) Accuracy with different values of β.

Figure 5: Comparison of different experimental settings. Experiments are conducted on HotPotQA with Qwen2.5-1.5B-Instruct as the backbone.

strates that rejection sampling enables the model to generate more accurate and contextually relevant responses by filtering out spurious rationales.

6.5 Different Numbers of Retrieved Documents (RQ3)

Fig. 5a compares ROSERAG and vanilla RAG on HotPotQA as the number of retrieved documents increases ($K = \{1, 2, 5, 8, 10\}$). The retrieval recall improves significantly when the number of retrieved documents increases from 1 to 5 but plateaus beyond that. ROSERAG closely follows this trend, with its EM score increasing rapidly when retrieving 1, 2, and 5 documents. However, beyond $K = 5$, the EM score plateaus, and the performance at $K = 10$ remains nearly identical to that at $K = 5$, indicating that ROSERAG effectively utilizes the most relevant retrieved information without being hindered by less relevant documents. In contrast, vanilla RAG's EM score drops after retrieving more than 2 documents, indicating its inability to filter noise and integrate relevant content effectively due to its limited capacity. In comparison, RoseRAG demonstrates its robustness and effectiveness by consistently improving or maintaining high performance as the number of retrieved documents increases. These results highlight ROSERAG's robustness in handling larger retrieval sizes for RAG applications.

6.6 Different Preference Optimization Methods (RQ4)

Fig. 5b presents the performance comparison of ROSERAG using different preference optimization methods, including ORPO, RPO (Pang et al., 2025), CPO (Xu et al., 2024a), and SimPO+SFT (Meng et al., 2025), on HotPotQA dataset. The results indicate that all methods achieve comparable performance in both Exact Match and F1 scores, suggesting that the proposed ROSERAG framework is robust and does not rely on a specific preference optimization technique. While minor variations exist, the consistency across methods highlights the generalizability of ROSERAG in effectively leveraging preference optimization for retrieval-augmented generation. This observation underscores the flexibility of our approach, allowing it to integrate seamlessly with various preference optimization approaches. Consequently, ROSERAG can be adapted to different optimization settings, making it a versatile solution for enhancing the reasoning capabilities of SLMs in retrieval-augmented generation tasks.

6.7 Different Values of β (RQ5)

Fig. 5c presents the performance of ROSERAG on HotPotQA with varying values of β, which controls the strength of preference alignment in the ORPO loss. We evaluate β over the range $\{0, 0.05, 0.1, 0.2, 0.3, 0.5, 1, 1.5, 2\}$ and observe that setting β within the interval $(0, 0.5)$ leads to better performance compared to both $\beta = 0$ and excessively large values of β. This result underscores the necessity of preference alignment, as models trained solely with supervised fine-tuning ($\beta = 0$) exhibit suboptimal performance. However, as β increases beyond 0.5, both the Exact Match and F1 scores decline, indicating that an overly strong preference alignment term causes the model to focus excessively on optimizing preference differences while neglecting the learning of positive responses.

7 Conclusion

We introduce ROSERAG, a novel framework that enhances the robustness of small-scale LLMs (SLMs) in retrieval-augmented generation through margin-aware preference optimization. By leveraging multi-turn prompting with rejection sampling and contrastive preference data selection, our approach effectively mitigates the impact of noisy retrieved content. We conducted extensive experiments on three benchmarks, and the results demonstrate that ROSERAG outperforms state-of-the-art baselines. Moreover, ROSERAG can be generalized well for different kinds of SLMs, justifying its generalizability and broad applicability for RAG.

Limitations

The proposed ROSERAG framework consists of a hyper-parameter β to balance the SFT and preference optimization, which might require additional effort to tune. Fortunately, we observe that setting $\beta \in [0.1, 0.5)$ can achieve a good performance and we set it to a fixed value to achieve good performance in this paper.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Wei Chen, Zichen Miao, and Qiang Qiu. 2024a. Large convolutional model tuning via filter subspace. *arXiv preprint arXiv:2403.00269*.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024b. Self-play fine-tuning converts weak language models to strong language models. *Preprint*, arXiv:2401.01335.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. 2021. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 715–724.

Guanting Dong, Yutao Zhu, Chenghao Zhang, Zechen Wang, Zhicheng Dou, and Ji-Rong Wen. 2024. Understand what llm needs: Dual preference alignment for retrieval-augmented generation. *arXiv preprint arXiv:2406.18676*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *Preprint*, arXiv:2305.14387.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10028–10039, Bangkok, Thailand. Association for Computational Linguistics.

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

Google. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 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.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

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. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7029–7043.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive NLP. *arXiv preprint arXiv:2212.14024*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. 2020. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Xiaonan Li, Changtai Zhu, Linyang Li, Zhangyue Yin, Tianxiang Sun, and Xipeng Qiu. 2023a. Llatrieval: Llm-verified retrieval for verifiable generation. *arXiv preprint arXiv:2311.07838*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Tianci Liu, Ruirui Li, Yunzhe Qi, Hui Liu, Xianfeng Tang, Tianqi Zheng, Qingyu Yin, Monica Xiao Cheng, Jun Huan, Haoyu Wang, et al. 2025. Unlocking efficient, scalable, and continual knowledge editing with basis-level representation fine-tuning. *arXiv preprint arXiv:2503.00306*.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. 2023. Statistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657*.

Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D Lane, and Mengwei Xu. 2024. Small language models: Survey, measurements, and insights. *arXiv preprint arXiv:2409.15790*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2025. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.

Zichen Miao, Wei Chen, and Qiang Qiu. 2025. Coeff-tuning: A graph filter subspace view for tuning attention-based large models. *arXiv preprint arXiv:2503.18337*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2024. Fine-tuning or retrieval? comparing knowledge injection in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 237–250, Miami, Florida, USA. Association for Computational Linguistics.

Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason Weston. 2025. Iterative reasoning preference optimization. *Advances in Neural Information Processing Systems*, 37:116617–116637.

SeongIl Park, Seungwoo Choi, Nahyun Kim, and Jay-Yoon Lee. 2024. Enhancing robustness of retrieval-augmented language models with in-context learning. In *Proceedings of the 3rd Workshop on Knowledge Augmented Methods for NLP*, pages 93–102.

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

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274. Association for Computational Linguistics.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. REPLUG: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384, Mexico City, Mexico. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839.

Giorgos Vernikos, Arthur Bražinskas, Jakub Adamek, Jonathan Mallinson, Aliaksei Severyn, and Eric Malmi. 2024. Small language models improve giants by rewriting their outputs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2703–2718.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

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

Haoyu Wang, Ruirui Li, Haoming Jiang, Jinjin Tian, Zhengyang Wang, Chen Luo, Xianfeng Tang, Monica Xiao Cheng, Tuo Zhao, and Jing Gao. 2024b. BlendFilter: Advancing retrieval-augmented large language models via query generation blending and knowledge filtering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1009–1025, Miami, Florida, USA. Association for Computational Linguistics.

Haoyu Wang, Tianci Liu, Ruirui Li, Monica Cheng, Tuo Zhao, and Jing Gao. 2024c. Roselora: Row and column-wise sparse low-rank adaptation of pre-trained language model for knowledge editing and fine-tuning. *arXiv preprint arXiv:2406.10777*.

Xiao Wang and Guo-Jun Qi. 2022. Contrastive learning with stronger augmentations. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):5549–5560.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023. Self-knowledge guided retrieval augmentation for large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. 2025. InstructRAG: Instructing retrieval-augmented generation via self-synthesized rationales. In *The Thirteenth International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. 2024. Reft: Representation finetuning for language models. *Advances in Neural Information Processing Systems*, 37:63908–63962.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. 2024. Certifiably robust rag against retrieval corruption. *arXiv preprint arXiv:2405.15556*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024a. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

Ran Xu, Hui Liu, Sreyashi Nag, Zhenwei Dai, Yaochen Xie, Xianfeng Tang, Chen Luo, Yang Li, Joyce C Ho, Carl Yang, and Qi He. 2024b. Simrag: Self-improving retrieval-augmented generation for adapting large language models to specialized domains. *arXiv preprint arXiv:2410.17952*.

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.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Ori Yoran, Tomer Wolfson, Ben Bogin, Uri Katz, Daniel Deutch, and Jonathan Berant. 2023. Answering questions by meta-reasoning over multiple chains of thought.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. 2024a. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14672–14685, Miami, Florida, USA. Association for Computational Linguistics.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024b. RankRAG: Unifying context ranking with retrieval-augmented generation in LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Ruizhe Zhang, Yongxin Xu, Yuzhen Xiao, Runchuan Zhu, Xinke Jiang, Xu Chu, Junfeng Zhao, and Yasha Wang. 2024. Knowpo: Knowledge-aware preference optimization for controllable knowledge selection in retrieval-augmented language models. *arXiv preprint arXiv:2408.03297*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Zixuan Zhou, Xuefei Ning, Ke Hong, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Luning Wang, Zhihang Yuan, Xiuhong Li, et al. 2024. A survey on efficient inference for large language models. *arXiv preprint arXiv:2404.14294*.

A Baselines

We adopt following state-of-the-art baselines to evaluate our proposed ROSERAG:

- CoT (Wei et al., 2022) instructs the LLM to generate answers accompanied by explicit reasoning steps.
- ReAct (Yao et al., 2022) integrates reasoning, action, and observation steps, concluding the generation process upon reaching a terminal state. The action step involves either generating a query to retrieve additional knowledge or finalizing the generation, while the observation step incorporates the retrieved knowledge documents.
- SelfAsk (Press et al., 2023) involves generating follow-up questions, retrieving relevant information, and answering these follow-up questions. Each retrieval operation is based on the generated follow-up questions. When no additional follow-up questions are produced, the LLM provides the answer to the original question. We prepend the newly retrieved knowledge to the original question following the approach of Yoran et al. (2023).
- BlendFilter (Wang et al., 2024b) combines internal and external knowledge to enhance retrieval quality. Additionally, it employs the LLM to filter out irrelevant information, thereby preventing the model from being misled.
- InstructRAG (Wei et al., 2025) instructs the LLM to generate rationales that analyze the relevance between the query and the retrieved knowledge. Subsequently, it conducts supervised fine-tuning (SFT) on the LLM using the generated rationales.
- RetRobust (Yoran et al., 2024) introduces an additional NLI model to assess the relationship between the query and the retrieved knowledge. If the relationship is deemed irrelevant, the model disregards the corresponding knowledge during generation.
- ASTUTE (Wang et al., 2024a) mitigates the pitfalls of imperfect retrieval by adaptively eliciting and integrating essential internal knowledge with externally retrieved data. Its iterative, source-aware consolidation process effectively resolves knowledge conflicts, yielding more accurate and trustworthy responses even in worst-case scenarios.
- ICL+RAG (Park et al., 2024) leverages machine reading comprehension examples to guide the model in identifying unanswerable queries and resolving conflicting information from retrieved texts. By providing tailored in-context demonstrations during inference, the approach improves the reasoning capabilities and overall accuracy of retrieval-augmented language models on open-domain QA tasks.

B Implementation Details.

We evaluate models with three small-sclae LLMs: Qwen2.5-1.5B-Instruct (Team, 2024), Llama-3.2-1B-Instruct [3], and gemma-2-2b-it (Team et al., 2024). We employ the state-of-the-art efficient retrieval method ColBERT v2 (Santhanam et al., 2022) as implemented by Khattab et al. (2022, 2023), which applies quantization to accelerate approximate nearest neighbor search. Our knowledge base comprises Wikipedia abstracts from the 2017 dump (Khattab et al., 2023). Experiments are conducted using Transformers (Wolf et al., 2020), TRL (von Werra et al., 2020), and vLLM (Kwon et al., 2023). In all experiments, we adopt a 3-shot in-context learning setting following the approach of Shao et al. (2023); Wang et al. (2024b), with the value of K set to 5 for all methods. The prompts for generating preference data are shown in Fig.6 and Fig.7, while those for question answering are presented in Fig.8, Fig.9, and Fig. 10. We generate preference data using the first 10k training samples from HotPotQA and 2WikiMultiHopQA, as well as the entire training set of StrategyQA. For all backbones, we apply LoRA (Hu et al., 2022) with a rank of 16 and `lora_alpha` = 32, targeting ”`all-linear`” modules. Experiments are conducted on eight 80G A100 NVIDIA GPUs. The learning rate and number of training epochs are selected from

[3] https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md

$\{2e-5, 3e-5, 5e-5, 2e-4, 5e-4\}$ and $\{2,3,4\}$, respectively. Note that RoseRAG by definition is generic and can be trained with diverse parameter-efficient fine-tuning (PEFT) methods (Houlsby et al., 2019; Li and Liang, 2021; Lester et al., 2021; Wang et al., 2024c; Chen et al., 2024a; Wu et al., 2024; Liu et al., 2025; Miao et al., 2025).

C Mathematical Derivations

Lemma C.1. (Lemma C.1 in (Chen et al., 2024b)) Denote $U(t) := \log(1+\exp(-t))$. For $a, b > 0$, the following inequality holds

$$a \cdot U(t) + b \cdot U(-t) \geq a \log(1 + b/a) + b \log(1 + a/b)$$

and equality holds if and only if $t = \log(a/b)$

Lemma C.2. Under Assumption 5.1, the solution to minimizing the ORPO loss $\ell_{ORPO}(x, y_w^*, y_l^*)$ is

$$P_\theta(y|x) = \frac{\exp(Z(x))}{\frac{q_l^*(y|x)}{q_w^*(y|x)} + \exp(Z(x))} \tag{8}$$

where $Z(x)$ is partition function such that $\sum_y P_\theta(y|x) = 1$.

Proof. Consider the last term in general ORPO loss

$$\begin{aligned}
&2\mathbb{E}_{(x,y_w,y_l)}\left[-\log \sigma(\log \mathrm{OR}_\theta(x, y_w, y_l))\right] \\
=&2\mathbb{E}_{(x,y_w,y_l)} U\left(\log \mathrm{odds}_\theta(y_w|x) - \log \mathrm{odds}_\theta(y_l|x)\right) \\
=&\int q(x) q_w(y_w|x) q_l(y_l|x) U\left(\log \mathrm{odds}_\theta(y_w|x) - \log \mathrm{odds}_\theta(y_l|x)\right) \mathrm{d}x\mathrm{d}y \\
&+ \int q(x) q_w(y_l|x) q_l(y_w|x) U\left(\log \mathrm{odds}_\theta(y_l|x) - \log \mathrm{odds}_\theta(y_w|x)\right) \mathrm{d}x\mathrm{d}y \\
\geq& \int q(x) q_w(y_w|x) q_l(y_l|x) \log\left(1 + \frac{q_w(y_l|x) q_l(y_w|x)}{q_w(y_w|x) q_l(y_l|x)}\right) \mathrm{d}x\mathrm{d}y \\
&+ \int q(x) q_w(y_l|x) q_l(y_w|x) \log\left(1 + \frac{q_w(y_w|x) q_l(y_l|x)}{q_w(y_l|x) q_l(y_w|x)}\right) \mathrm{d}x\mathrm{d}y
\end{aligned} \tag{9}$$

where the first inequality follows from Lemma C.1. For equivalence,

$$\log \mathrm{odds}_\theta(y_w|x) - \log \mathrm{odds}_\theta(y_l|x) = \log \frac{q_w(y_w|x) q_l(y_l|x)}{q_w(y_l|x) q_l(y_w|x)} \tag{10}$$

Thus, for any x, y_w, y_l,

$$\log \mathrm{odds}_\theta(y_w|x) - \log \frac{q_w(y_w|x)}{q_l(y_w|x)} = \log \mathrm{odds}_\theta(y_l|x) - \log \frac{q_w(y_l|x)}{q_l(y_l|x)} \tag{11}$$

Therefore, Eqn. (11) holds if and only if there exists some $Z(x)$ such that

$$\log \frac{P_\theta(y|x)}{1 - P_\theta(y|x)} = Z(x) + \log \frac{q_w(y|x)}{q_l(y|x)} \Longleftrightarrow P_\theta(y|x) = \frac{\exp(Z(x))}{\frac{q_l(y|x)}{q_w(y|x)} + \exp(Z(x))} \tag{12}$$

Finally, substituting q_l and q_w with q_l^* and q_w^* respectively, yields Eqn. (8). □

Lemma C.3. Denote $\tilde{P}_\theta(y|x)$ as the solution to minimizing ORPO loss without the selection step. Then $P_\theta(y|x)$ generates $y \sim f_w(x) + \exp(\lambda)$ with probability greater than $\tilde{P}_\theta(y|x)$, and generates $y \sim f_l(x) + \exp(\lambda)$ with probability lower than $\tilde{P}_\theta(y|x)$:

$$\begin{aligned}
P_\theta(y|x) > \tilde{P}_\theta(y|x) \quad &\text{if } y \sim f_w(x) + \mathrm{Exp}(\lambda) \\
P_\theta(y|x) < \tilde{P}_\theta(y|x) \quad &\text{if } y \sim f_l(x) + \mathrm{Exp}(\lambda)
\end{aligned} \tag{13}$$

Proof. Order statistics $Y_{(k)}$, representing the k-th smallest value in a sample drawn from an exponential distribution Exp(λ), has the probability density function:

$$\frac{n!}{(k-1)!(n-k)!}\left[1-e^{-\lambda y}\right]^{k-1}\left[e^{-\lambda y}\right]^{n-k}\lambda e^{-\lambda y} \tag{14}$$

Let q_l^* denote the distribution of minimum value among y_l^i and q_w^* denote the the distribution of maximum value among y_w^i. The ratio of their corresponding densities is given by:

$$\frac{q_l^*(y|x)}{q_w^*(y|x)} = \left(\frac{\exp(-\lambda(y-f_l(x)))}{1-\exp(-\lambda(y-f_w(x)))}\right)^{n-1}\frac{\exp(-\lambda(y-f_l(x)))}{\exp(-\lambda(y-f_w(x)))}. \tag{15}$$

Rewriting this expression yields,

$$\frac{q_l^*(y|x)}{q_w^*(y|x)} = \left(\frac{\exp(-\lambda(y-f_l(x)))}{1-\exp(-\lambda(y-f_w(x)))}\right)^{n-1}\frac{q_l(y|x)}{q_w(y|x)}. \tag{16}$$

For $y \sim f_w(x) + \text{Exp}(\lambda)$, define $d_w = y - f_w(x)$ and $d_l = y - f_l(x)$. Then,

$$d_l = d_w + [f_w(x) - f_l(x)] = d_w - c, \quad \text{where } c = f_l(x) - f_w(x) < 0.$$

By Lemma C.1, the closed form solutions differ only in the ratio R:

$$R = \frac{\exp(-\lambda(y-f_l(x)))}{1-\exp(-\lambda(y-f_w(x)))} = \frac{\exp(-\lambda d_l)}{1-\exp(-\lambda d_w)} = \frac{\exp(\lambda c)\exp(-\lambda d_w)}{1-\exp(-\lambda d_w)} \tag{17}$$

Since $d_w \sim \text{Exp}(\lambda)$, we have $U = \exp(-\lambda d_w) \sim \text{Uniform}(0,1)$, allowing us to express:

$$R < 1 \Longleftrightarrow U < \frac{1}{1+\exp(\lambda c)} \Longrightarrow \mathbb{P}(R<1) = \frac{1}{1+\exp(\lambda c)} > \frac{1}{2} \tag{18}$$

Similarly, for $y \sim f_l(x) + \text{Exp}(\lambda)$, we obatin $\mathbb{P}(R > 1) > 1/2$. This implies that the distribution $P_\theta(y|x)$ generates $y \sim f_w(x) + \exp(\lambda)$ with a probability greater than $\tilde{P}_\theta(y|x)$ and generates $y \sim f_l(x) + \exp(\lambda)$ with probability lower than $\tilde{P}_\theta(y|x)$. □

Theorem C.1. Under (6), assume $\tilde{P}_\theta(y|x)$ generates y with density function:

$$p \cdot \lambda e^{-\lambda[y-f_w(x)]} + (1-p)\cdot \lambda e^{-\lambda[y-f_l(x)]}.$$

Let $y \sim P_\theta(y|x)$ and $\tilde{y} \sim \tilde{P}_\theta(y|x)$. Then, the expected absolute distance under P_θ is smaller than that under $\tilde{P}_\theta$:

$$\mathbb{E}_{y\sim P_\theta(y|x)}[L(y)] < \mathbb{E}_{\tilde{y}\sim \tilde{P}_\theta(y|x)}[L(\tilde{y})]. \tag{19}$$

Proof. Denote $\tilde{L}_1 = |\xi|$ and $\tilde{L}_2 = |\xi + c|$, where $\xi \sim \text{Exp}(\lambda) - 1/\lambda$. Then

$$\mathbb{E}_{\tilde{y}\sim \tilde{P}_\theta(y|x)}[L(\tilde{y})] = p\mathbb{E}_\xi[\tilde{L}_1] + (1-p)\mathbb{E}_\xi[\tilde{L}_2] \tag{20}$$

A well-known fact is that function $\phi(t) = \mathbb{E}|\xi - t|$ is convex in t and attains its unique global minimum at the median m of ξ. Since median of Exp(λ) is $(\log 2)/\lambda$, it follows that:

$$m = \frac{\log 2 - 1}{\lambda} < 0 \tag{21}$$

Given that $m < 0 < -c$, we conclude:

$$\phi(-c) > \phi(0) \Longleftrightarrow \mathbb{E}_\xi[\tilde{L}_2] > \mathbb{E}_\xi[\tilde{L}_1] \tag{22}$$

By Lemma C.3, $P_\theta(y|x)$ has larger p than $\tilde{P}_\theta(y|x)$, denoted as q.

$$\mathbb{E}_{y\sim P_\theta(y|x)}[L(y)] - \mathbb{E}_{\tilde{y}\sim \tilde{P}_\theta(y|x)}[L(\tilde{y})] = (q-p)\cdot\left[\mathbb{E}_\xi[\tilde{L}_1] - \mathbb{E}_\xi[\tilde{L}_2]\right] < 0 \tag{23}$$

□

D Prompt

System Prompt: You are a useful assistant. I will provide one question, several pieces of knowledge (which may be related or unrelated to the question), and the answer to the question. Please explain your reasoning process in a single paragraph consisting of no more than four sentences. If the provided knowledge is insufficient, you may make an informed guess, but do not respond with "Unknown".

User Prompt: Knowledge: $\mathcal{K}_q$
Question: q
Answer: a^*

Assistant Prompt: Let's think step by step.

Output: {rationale r}

Figure 6: Prompt for rationale generation on HotPotQA and 2WikiMultiHopQA

System Prompt: You are a useful assistant. I will provide one question, several knowledge (may related or unrelated to the question), and the answer to the question. Please show the think process about how to get the answer. If the given knowledge is insufficient, you can guess. Do not tell me Unknown. Your output should be in one paragraph within four sentences.

User Prompt: Knowledge: $\mathcal{K}_q$
Question: q
Answer: a^*

Assistant Prompt: Let's think step by step.

Output: {rationale r}

//If rationale r shows it can not conclude the answer
System Prompt: You are a useful assistant. I will provide one question and the answer to the question. Please show the reasoning process about how to get the answer. Please use your own memorized knowledge to do reasoning. Do not mention the given answer explicitly in your reasoning. Your output should be in one paragraph with several sentences.

User Prompt: Question: q
Answer: a^*

Model Output: {rationale r_{wk}}

Output: {rationale r+However, we can answer the question based on known knowledge.+r_{wk}}

Figure 7: Prompt for rationale generation on StrategyQA.

System Prompt: You are a useful assistant. You should answer the following question quickly and accurately based on given knowledge. Provide your reasoning in few sentences and answer in one or few words. Please begin your final answer with 'So the answer is'.

User Prompt: Knowledge:Mr. Big (film) | Mr. Big is a 2007 documentary directed and produced by Tiffany Burns and edited by Alec MacNeill Richardson. The documentary examines the "Mr. Big" undercover methods used by the Royal Canadian Mounted Police (RCMP). In these operations, RCMP officers pose as gang criminals and develop a relationship with the target in the hope of eventually determining what, if any, knowledge the target has of the crime being investigated. "Mr. Big" operations have been credited with securing difficult convictions in a large number of cases, but concerns have been raised that they involve a risk of false confessions and wrongful convictions.
It Might Get Loud | It Might Get Loud is a 2008 American documentary film by filmmaker Davis Guggenheim. It explores the careers and styles of prominent rock musicians Jimmy Page, The Edge, and Jack White. The film received a wide release on August 14, 2009 in the U.S. by Sony Pictures Classics.
Question:Are It Might Get Loud and Mr. Big both Canadian documentaries?

Assistant Prompt: Let's think step by step. Mr. Big is a 2007 documentary which examines the "Mr. Big" undercover methods used by the Royal Canadian Mounted Police. However, Are It Might Get Loud is a 2008 American documentary film. So the answer is no.

User Prompt: Knowledge:Leslie H. Martinson | Leslie Herbert "Les" Martinson (January 16, 1915 – September 3, 2016) was an American television and film director.
László Benedek | László Benedek (] ; March 5, 1905 – March 11, 1992; sometimes "Laslo Benedek") was a Hungarian-born film director and cinematographer, most notable for directing "The Wild One" (1953).
Question:Were László Benedek and Leslie H. Martinson both film directors?

Assistant Prompt: Let's think step by step. László Benedek was a Hungarian-born film director and Leslie H. Martinson was an American film director. So the answer is yes.

User Prompt: Knowledge:Lucium | Lucium was the proposed name for an alleged new element found by chemist Prosper Barrière in 1896 in the mineral monazite. Later, William Crookes confirmed that the new element was actually an impure sample of yttrium.
William Crookes | Sir William Crookes (; 17 June 1832 – 4 April 1919) was an English chemist and physicist who attended the Royal College of Chemistry in London, and worked on spectroscopy. He was a pioneer of vacuum tubes, inventing the Crookes tube which was made in 1875. Crookes was the inventor of the Crookes radiometer, which today is made and sold as a novelty item. Late in life, he became interested in spiritualism, and became the president of the Society for Psychical Research.
Question:Lucium was confimed to be an impure sample of yttrium by an English chemist who became the president of what?

Assistant Prompt: Let's think step by step. Lucium was confimed to be an impure sample of yttrium by William Crookes. William Crookes is Sir William Crookes. Sir William Crookes became the president of the Society for Psychical Research. So the answer is Society for Psychical Research.
User Prompt: Knowledge: $\mathcal{K}_q$
Question: q

Assistant Prompt: Let's think step by step.

Figure 8: Prompt for HotPotQA

System Prompt: You are a useful assistant. You should answer the following question quickly and accurately based on given knowledge. Provide your reasoning in few sentences and answer in one or few words. Please begin your final answer with ’So the answer is’.

User Prompt: Knowledge:Valentin the Good | Valentin the Good (Czech: "Valentin Dobrotivý") is a 1942 Czech comedy film directed by Martin Frič.
The Falcon (film) | Banović Strahinja (Serbian Cyrillic: , internationally released as The Falcon) is a 1981 Yugoslavian adventure film written and directed by Vatroslav Mimica based on Strahinja Banović, a hero of Serbian epic poetry. It entered the section "Officina Veneziana" at the 38th Venice International Film Festival.
Martin Frič | Martin Frič (29 March 1902 – 26 August 1968) was a Czech film director, screenwriter and actor. He had more than 100 directing credits between 1929 and 1968, including feature films, shorts and documentary films. Vatroslav Mimica | Vatroslav Mimica (born 25 June 1923) is a Croatian film director and screenwriter.
Question:Do both films The Falcon (Film) and Valentin The Good have the directors from the same country?

Assistant Prompt: Let’s think step by step. The Falcon (Film) is directed by Martin Frič. Martin Frič was a Czech film director. Valentin The Good is directed by Vatroslav Mimica. Vatroslav Mimica is a Croatian film director. Czech is different from Croatia. So the answer is no.

User Prompt: Knowledge:Wedding Night In Paradise (1950 film) | Wedding Night in Paradise (German: Hochzeitsnacht im Paradies) is a 1950 West German musical comedy film directed by Géza von Bolváry and starring Johannes Heesters, Claude Farell and Gretl Schörg .
Géza von Bolváry | Géza von Bolváry (full name Géza Maria von Bolváry-Zahn; 26 December 1897 – 10 August 1961) was a Hungarian actor, screenwriter and film director, who worked principally in Germany and Austria.
Question:What nationality is the director of film Wedding Night In Paradise (1950 Film)?

Assistant Prompt: Let’s think step by step. Wedding Night In Paradise (1950 film) is directed by Géza von Bolváry. Géza von Bolváry was a Hungarian actor, screenwriter and film director. So the answer is Hungarian.

User Prompt: Knowledge:Rhescuporis I (Odrysian) | Rhescuporis I (Ancient Greek:) was a king of the Odrysian kingdom of Thrace in 240 BC - 215 BC, succeeding his father, Cotys III.
Cotys III (Odrysian) | Cotys III (Ancient Greek:) was a king of the Odrysian kingdom of Thrace in ca. 270 BC, succeeding his father, Raizdos.
Question:Who is Rhescuporis I (Odrysian)’s paternal grandfather?

Assistant Prompt: Let’s think step by step. The father of Rhescuporis I (Odrysian) is Cotys III. The father of Cotys III is Raizdos. So the answer is Raizdos.
User Prompt: Knowledge: $\mathcal{K}_q$
Question: q

Assistant Prompt: Let’s think step by step.

Figure 9: Prompt for 2WikiMultiHopQA

System Prompt: You are a useful assistant. You should answer the following question quickly and accurately based on given knowledge. If the information provided is insufficient to answer, please respond based on your own knowledge. Provide your reasoning in few sentences first and then show the final answer which should always be either Yes or No, and NOTHING ELSE. Please begin your final answer with 'So the answer is'.

User Prompt: Knowledge:Laxative | Laxatives, purgatives, or aperients are substances that loosen stools and increase bowel movements. They are used to treat and/or prevent constipation.
Question:Do people take laxatives because they enjoy diarrhea?

Assistant Prompt: Let's think step by step. Laxatives are substances that loosen stools and increase bowel movements. People take laxatives to treat and/or prevent constipation. So the answer is No.

User Prompt: Knowledge:Durian (disambiguation) | Durian is a large odorous spike-covered fruit of several southeast Asian tree species belonging to the genus "Durio".
Durian | The durian ( ) or is the fruit of several tree species belonging to the genus Durio. The name "durian" is derived from the Malay languages word for duri or "spike", a reference to the numerous spike protuberances of the fruit, together with the noun-building suffix -an. There are 30 recognised "Durio" species, at least nine of which produce edible fruit, and over 300 named varieties in Thailand and 100 in Malaysia. " Durio zibethinus" is the only species available in the international market: all other species are sold only in their local regions. There are hundreds of durian cultivars; many consumers express preferences for specific cultivars, which fetch higher prices in the market.
Question:Could Durian cause someone's stomach to feel unwell?

Assistant Prompt: Let's think step by step. Durian has a pungent odor that many people describe as being similar to feet and onions. Unpleasant smells can make people feel nauseous. So the answer is Yes.

User Prompt: Knowledge:Monty Python and the Holy Grail | Monty Python and the Holy Grail is a 1975 British absurdist comedy film concerning the Arthurian legend, written and performed by the Monty Python comedy group (Graham Chapman, John Cleese, Terry Gilliam, Eric Idle, Terry Jones, and Michael Palin), and directed by Gilliam and Jones. It was conceived during the hiatus between the third and fourth series of their BBC television series "Monty Pythonś Flying Circus".
Question:Did the swallow play a role in a famous film about King Arthur?

Assistant Prompt: Let's think step by step. Monty Python and the Holy Grail was a famous film about King Arthur. In Monty Python and the Holy Grail, swallows are mentioned several times. So the answer is Yes.
User Prompt: Knowledge: $\mathcal{K}_q$
Question: q

Assistant Prompt: Let's think step by step.

Figure 10: Prompt for StrategyQA