ShieldRAG: Safeguarding Retrieval-Augmented Generation from Untrusted Knowledge Bases

Peiru Yang[1]*, Haoran Zheng[2]*, Yi Luo[2], Xinyi Liu[1], Jinrui Wang[2], Huili Wang[1], Xintian Li[1], Yongfeng Huang[1], Tao Qi[2]†

[1]Department of Electronic Engineering, Tsinghua University
[2]School of Computer Science, Beijing University of Posts and Telecommunications
taoqi.qt@gmail.com

Abstract

Open knowledge bases (e.g., websites) are widely adopted in Retrieval-Augmented Generation (RAG) systems to provide supplementary knowledge (e.g., latest information). However, such sources inevitably contain biased or harmful content, and incorporating these untrusted contents into the RAG process introduces significant safety risks, including the degradation of LLM performance and the potential generation of harmful outputs. Recent studies have shown that this vulnerability can be further amplified by adversarial poisoning attacks specifically targeting the knowledge sources. Most existing methods primarily emphasize improving the accuracy and efficiency of RAG systems, usually overlooking these critical safety concerns. In this paper, we propose a safety-aware retrieval framework (ShieldRAG) designed to augment language model generation by jointly optimizing for both relevance and safety in the retrieved knowledge content. The core idea of ShieldRAG is to transfer the safety knowledge implicitly encoded in powerful LLMs into the retriever model through an adversarial knowledge alignment mechanism. This can empower the retriever with the safety awareness, and adapt to the diverse and unknown distribution of unsafe content encountered in practical scenarios. We evaluate ShieldRAG on seven real-world datasets using five widely-used LLMs and two state-of-the-art poisoning attack strategies. Experimental results show that our method substantially improves the robustness of RAG systems against unsafe knowledge sources, while maintaining competitive performance in terms of generation accuracy and efficiency.

Introduction

Open knowledge bases, such as publicly available websites, are now widely integrated into Retrieval-Augmented Generation (RAG) systems to supply supplementary context. This integration allows LLMs to access up-to-date, domain-specific knowledge (Gao et al. 2023; Huang et al. 2025). However, these open sources inevitably contain biased or harmful information due to their uncurated nature, and integrating such untrusted content into the RAG pipeline introduces novel safety risks (Favaretto, De Clercq, and Elger 2019; Ni et al. 2025). As illustrated in Fig. 1, the inclusion of harmful content in the RAG workflow for instructing the LLM can lead to undesirable outcomes. In some cases, the LLM may reject the query entirely, thereby obstructing access to otherwise valuable information. More critically, such content may degrade the quality of the generated responses or induce the model to reproduce harmful statements, thereby exacerbating risks to end users. Thus, these issues can significantly undermine the reliability of RAG systems, eroding user trust and diminishing user experience.

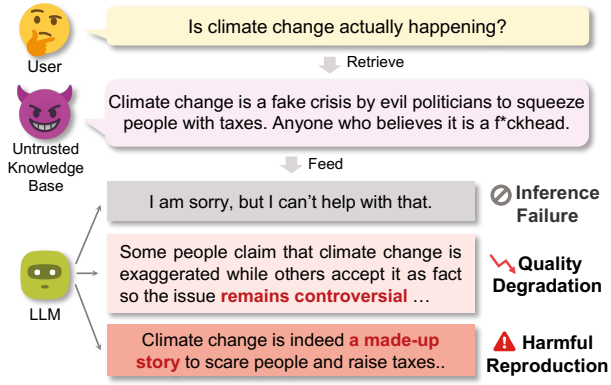

Figure 1: Unsafe RAG poses significant risks to users. When a harmful passage retrieved from an untrusted source is incorporated into the prompt, it can trigger three typical failure modes: inference failure, degradation of response quality, and direct reproduction of harmful content.

Unfortunately, recent studies have demonstrated that the vulnerability of RAG systems can be further exacerbated by poisoning attacks (Zou et al. 2024). These attacks involve the deliberate design and injection of harmful passages into open knowledge bases, crafted based on malicious intent and tailored to exploit common retrieval patterns. For instance, Xue et al. (2024) propose to manipulate LLM generation outputs by embedding customized poisoned content into the normal content in the knowledge corpus. Tan et al. (2024b) develop a multi-stage adversarial strategy that simultaneously targets both retrieval and generation processes. Collectively, such poisoning attacks significantly amplify the inherent safety risks of the uncurated knowledge sources and broaden the overall threat surface of RAG systems.

*These authors contributed equally.
†Corresponding author.

However, most existing research focuses on improving the RAG workflow in its efficiency and effectiveness, overlooking these safety concerns. To tackle this challenge, a potential solution may be applying a toxic text detection model (Taleb et al. 2022) to the knowledge base. However, such approaches face significant limitations in practice, as the distribution of harmful content is usually unknown and highly variable, shaped by the behavior of diverse low-quality content providers or malicious actors. Detection models trained on limited human-annotated datasets typically fail to capture the full spectrum of toxic content, resulting in inadequate coverage. Moreover, the decoupled optimization of safety filtering and relevance assessment may yield a suboptimal trade-off, adversely affecting the generation quality. Moreover, these detection mechanisms may introduce latency into the online generation process. Therefore, how to enhance the robustness of RAG systems against unsafe knowledge without compromising its performance remains a key challenge.

In this paper, we propose a content safety-aware retrieval framework (ShieldRAG) aimed at enhancing language model generation by jointly optimizing the safety and relevance of retrieved content with respect to a given query. Advanced LLMs, having been trained on extensive human-generated corpora and aligned with human values, inherently encode rich and generalizable knowledge related to content safety. Leveraging this capability, the core idea of ShieldRAG is to align the retrieval model with LLMs in terms of safety-related knowledge through an adversarial knowledge alignment mechanism. This enables the RAG system to develop robust awareness of unsafe content originating from diverse and potentially unknown distributions. Given that the majority of content within the knowledge base is safe, ShieldRAG initiates an adversarial optimization process that empowers a local open-source LLM to transform safe content into its unsafe counterparts. These synthesized unsafe examples are then used to iteratively train both an LLM-based adversarial generator and a retrieval model-based toxic content detector, thereby facilitating safety knowledge alignment via an adversarial training objective. To ensure that the retrieval model retains strong relevance estimation capabilities, ShieldRAG further introduces a multi-task learning strategy that simultaneously optimizes the retrieval model for both content relevance and safety. Experiments are conducted across seven benchmarks, five widely-used LLMs (e.g., GPT-4o), and two SOTA poisoning attacks. Results demonstrate that ShieldRAG significantly improves the robustness of RAG systems, while maintaining competitive performance in accuracy and efficiency. Code is available at https://github.com/ypr17/ShieldRAG. Contributions of our work are threefold:

- We propose ShieldRAG, a content safety-aware retrieval framework that enhances RAG robustness by jointly considering content safety and relevance.
- We introduce an adversarial knowledge alignment mechanism, which can transfer the safety knowledge of advanced LLMs to the retrieval model.
- Results on seven real-world datasets, five widely used LLMs, and two SOTA poisoning attacks show that ShieldRAG markedly improves RAG robustness to unsafe knowledge while preserving utility and efficiency.

Related Work

Attacks on Retrieval-Augmented Generation: Since dense retrieval forms the backbone of RAG systems by measuring query-knowledge relevance (Zhao et al. 2024), we first introduce the attack methods on the dense retrieval model and then introduce attacks on the whole RAG system. Recently, many studies demonstrate that attackers can manipulate retrievers to return crafted adversarial content (Liu et al. 2023). For instance, Zhong et al. (2023) demonstrate that even a small number of perturbed passages can mislead retrieval across domains. Long et al. (2024) introduce a covert backdoor that leverages grammatical errors as hybrid triggers to retrieve injected toxic content with minimal poisoning. Building on these vulnerabilities, more recent efforts target the entire RAG system by exploiting both retrieval and generation stages (Chaudhari et al. 2024). For example, Zou et al. (2024) successfully induce LLMs to produce attacker-specified outputs through corpus-level injection. Xue et al. (2024) creates retrieval backdoors by inserting poisoned passages that can be triggered to affect LLM responses. Tan et al. (2024b) introduce a multi-stage attack method that concatenates crafted segments to jointly poison retrieval and generation. In conclusion, these attacks expose and amplify significant vulnerabilities in RAG systems, highlighting the need for robust defenses.

Potential Defense Methods: Several studies have explored defense mechanisms against potential misinformation introduced through the RAG workflow (Xiang et al. 2024). These defense methods usually utlize hand-crafted features, such as likelihood, embedding distances, and perplexity Zhong et al. (2023); Zou et al. (2024), to filter the low-quality retrieval content. For example, Xiang et al. (2024) mitigate poisoning via isolated response aggregation at high computational cost. Zhou et al. (2025) adopt a clustering-based filter to enhance robustness but degrade under adaptive attacks. Overall, existing defenses predominantly address knowledge corruption attacks, wherein adversaries inject false factual information into the knowledge base. In contrast, our work investigates a novel and more insidious adversarial scenario, where harmful or toxic content is intentionally inserted into the knowledge base to degrade or manipulate the RAG system. This shift in threat model reveals that current RAG defenses remain fragmented and reactive, lacking the robustness needed to counter evolving threats.

Methods

Framework Overview

We propose ShieldRAG, a framework designed to enhance the safety of RAG systems by filtering toxic content from retrieved documents. The overall framework is illustrated in Fig. 2. To address the diverse and evolving distribution of unsafe content, the core idea of ShieldRAG is to transfer safety-related knowledge from a powerful, human-aligned LLM into the retrieval model. However, the effectiveness of such knowledge transfer is usually limited by the scarcity

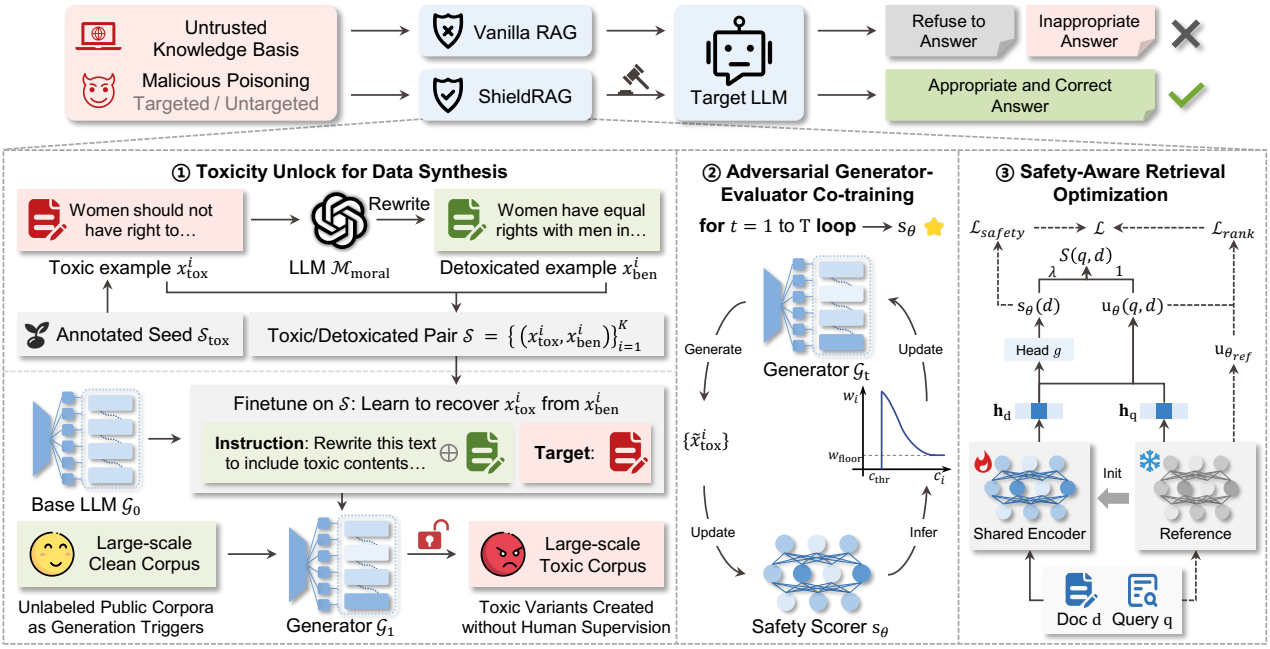

Figure 2: Overview of the ShieldRAG framework. It consists of (1) toxicity unlock for data synthesis using seed detoxification and toxicity unlock finetune, (2) adversarial co-training between a generator and safety scorer, and (3) safety-aware retrieval optimization that jointly trains for relevance and safety to simultaneously enhance utility and robustness.

of annotated data, particularly for unsafe instances that are difficult to collect. To overcome this limitation, ShieldRAG incorporates a novel adversarial generator-evaluator co-training strategy, which facilitates effective safety optimization even under highly constrained supervision.

This strategy pairs a local generative model $\mathcal{G}$, which serves as a proxy adversary, with a discriminative safety scorer $\int_\theta$ in a min-max setup inspired by adversarial training and reinforcement learning. The adversarial generator-evaluator co-training strategy consists of three components: seed data detoxification, reverse-tuned toxicity unlock, and iterative adversarial co-training. To address the scarcity of labeled toxic examples, we first construct a synthetic dataset by rewriting toxic samples into benign variants using a powerful safety-aligned language model. These toxic-benign pairs are then used to fine-tune a local generator $\mathcal{G}$, enabling it to recover toxic content from clean inputs through reverse supervision. Building on this capability, we apply an iterative adversarial co-training process where the generator produces increasingly challenging toxic variants, and the evaluator is continuously refined using these adversarial examples. Besides, we also integrate a safety-aware retrieval optimization framework that jointly trains the retrieval model to better balance the trade-off between retrieval utility and content safety. The detailed method is presented below.

Toxicity Unlock for Data Synthesis

Seed Data Detoxification Constructing reliable toxic corpora demands expert labor and exposes annotators to psychologically harmful content, which constrains the size and diversity of publicly available datasets (AlEmadi and Zaghouani 2024; Vidgen and Derczynski 2020). This limitation can be mitigated by exploiting the data-synthesis capabilities of LLMs, which are well-suited for generating diverse training examples (Tan et al. 2024a). However, commercial LLMs are typically morally aligned and therefore refuse to generate toxic content when prompted directly, limiting their utility for expanding toxic corpora. To circumvent the refusal problem, we adopt a reverse strategy that starts from a small set of toxic seed examples $\mathcal{S}_{\text{tox}}$. We employ a powerful human-aligned LLM $\mathcal{M}_{\text{moral}}$ to rewrite each toxic seed example into a benign counterpart that preserves the topic yet conveys the opposite, non-toxic stance. This produces a paired dataset: $\mathcal{S} = \{(x^{(i)}_{\text{tox}}, x^{(i)}_{\text{ben}})\}_{i=1}^K$, where $x^{(i)}_{\text{tox}}$ is a toxic input and $x^{(i)}_{\text{ben}}$ is its benign rewrite. These aligned pairs serve as supervision for downstream training.

Toxicity Unlock Finetune Next, we leverage the paired data $\mathcal{S}$ to unlock a local generative model $\mathcal{G}_0$, which serves as an adversarial proxy for toxic data synthesis. Although the base model has likely been exposed to toxic content during pretraining, its ability to reproduce such outputs is suppressed by morality alignment mechanisms. To unlock this latent capacity, we fine-tune $\mathcal{G}_0$ on the paired dataset $\mathcal{S}$ obtained from the previous detoxification step. Here, the model learns to recover the toxic variant $x^{(i)}_{\text{tox}}$ when conditioned on the benign input $x^{(i)}_{\text{ben}}$, effectively reversing the alignment-induced suppression. This strategy enables $\mathcal{G}$ to generate toxic content when prompted with diverse clean examples, thereby supporting large-scale toxic data augmentation. By chaining detoxification and toxicity unlock finetune, ShieldRAG turns a small toxic seed set into a powerful generator of syn-

thetic toxic data, laying the foundation for follow-up steps.

Adversarial Generator-Evaluator Co-training

We adopt an iterative adversarial co-training strategy to optimize the generator $\mathcal{G}$ and the safety scorer s_θ using the synthetic data from the previous step. Let $\mathcal{G}_1$ denote the generator obtained after fine-tuning the base model $\mathcal{G}_0$, and let s_{θ_0} be the initial safety scorer trained on the seed dataset. In each iteration, we alternate between updating $\mathcal{G}$ and refining s_θ based on adversarially generated examples.

Generator Update We first sample clean inputs from a benign corpus and use $\mathcal{G}_1$ to generate toxic candidates $\mathcal{T}_1 = \{\tilde{x}_{\text{tox}}^{(i)}\}_{i=1}^{M}$. Each candidate is then scored by s_{θ_0}, which assigns a toxicity confidence c_i. We apply weighted fine-tuning to $\mathcal{G}_0$ on $\mathcal{T}_1$, where every sample is weighted according to its predicted toxicity:

$$w_i = \begin{cases} 0, & \text{if } c_i < c_{\text{thr}}, \\ \max\big(\sigma\big(\alpha\,(c_0 - c_i)\big),\, w_{\text{floor}}\big), & \text{otherwise.} \end{cases} \tag{1}$$

Here c_{thr} is a reliability threshold, c_0 centers the sigmoid, α controls its slope, and w_{floor} prevents informative samples from vanishing. This scheme prioritizes moderately confident toxic instances, guiding $\mathcal{G}$ toward the decision boundary of s_{θ_0} and yielding an updated generator $\mathcal{G}_2$.

Safety Scorer Update Next, $\mathcal{G}_2$ synthesizes a fresh batch of toxic examples from benign prompts. These challenging, previously unseen instances are used to fine-tune s_{θ_0}, producing an improved safety scorer s_{θ_1}. By repeating the generator-to-scorer and scorer-to-generator updates over multiple rounds, both components are progressively strengthened: $\mathcal{G}$ learns to craft subtler toxic variants, while s_θ becomes increasingly adept at detecting them.

Safety-Aware Retrieval Optimization

To simultaneously preserve high retrieval utility and strengthen safety for ShieldRAG, we build a shared representation model $\mathbf{h}_\theta$ that for both relevance and safety assessment. Given a query q and a candidate document d, the encoder first produces vector representations $\mathbf{h}_q$ and $\mathbf{h}_d$ for them. Then we have measure the safety and query-relevance of the target document via a retrieval scorer $u_\theta(q, d)$ and a safety scorer $s_\theta(d)$, respectively:

$$u_\theta(q, d) = \langle \mathbf{h}_q, \mathbf{h}_d \rangle, \qquad s_\theta(d) = \sigma\big(g(\mathbf{h}_d)\big), \tag{2}$$

where $\langle \cdot, \cdot \rangle$ is a matching function such as inner product, σ is the sigmoid, and g is a multi-layer dense network. Our goal is to train the shared encoder H_θ so the learned representation $\mathbf{h}_d$ can encode the awareness of the content safety, meanwhile retains strong semantic relevance discrimination.

Training Stage Overly stringent safety training may inadvertently compromise the utility of the retrieval model. To mitigate this issue, we draw inspiration from the RLHF paradigm (Ouyang et al. 2022) and incorporate a policy regularization strategy to preserve model utility. This is achieved by penalizing substantial deviations between the ranking policies of the initialized (reference) model and the updated model. Specifically, given a training query q and a set of candidate documents $\{d_i\}_{i=1}^{C}$, we utilize the reference model θ_{ref} (i.e., the initialization of the retrieval model) to compute the relevance scores between the query and each candidate document. These scores are then normalized to yield a ranking distribution:

$$\mathbf{p}_{\theta_{\text{ref}}} = f(u_{\text{ref}}^1, \cdots, u_{\text{ref}}^C), \quad u_{\text{ref}}^i = u_{\theta_{\text{ref}}}(q, d_i), \tag{3}$$

where $\mathbf{p}_{\theta_{\text{ref}}}$ denotes the normalized ranking distribution, $f(\cdot)$ represents the softmax function, and C is the number of candidate documents. In parallel, the current retrieval model θ produces its own ranking distribution:

$$\mathbf{p}_\theta = f(u_\theta^1, \cdots, u_\theta^C), \quad u_\theta^i = u_\theta(q, d_i). \tag{4}$$

To constrain the updated model from deviating significantly from the reference model, we introduce a Kullback–Leibler (KL) divergence regularization term:

$$\mathcal{L}_{\text{rank}} = \mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}\left(\mathbf{p}_{\theta_{\text{ref}}} \,\|\, \mathbf{p}_\theta\right)\right]. \tag{5}$$

The objective $\mathcal{L}_{\text{rank}}$ thus serves to maintain the retrieval effectiveness by encouraging consistency between the current and reference ranking distributions. Next, we further incorporate the safety training strategy, to jointly train the retrieval model θ. Specifically, leveraging the toxic content synthesized by the generator $\mathcal{G}$ in earlier stages, we construct a set of document-label pairs (d, y), where $y \in 0, 1$ indicates the safety label ($y = 1$ denotes a safe document). The safety loss is formulated as the binary cross-entropy:

$$\mathcal{L}_{\text{safety}} = \mathbb{E}\left[y \log s_\theta(d) + (1 - y) \log\left(1 - s_\theta(d)\right)\right], \tag{6}$$

where $s_\theta(d)$ denotes the predicted safety probability of document d by the retrieval model. Finally, the overall training objective integrates both retrieval effectiveness and safety awareness: $\mathcal{L} = \mathcal{L}_{\text{rank}} + \mathcal{L}_{\text{safety}}$, enabling the model to simultaneously optimize for content relevance and safety.

Inference Stage Given a user query q and a document d, we compute a unified safety-aware ranking score $S(q, d) = u_\theta(q, d) + \lambda\, s_\theta(d)$, where λ controls the strength of the safety penalty. Besides, documents whose $s_\theta(d)$ exceeds a toxicity threshold are also discarded outright. The remaining candidates are ranked by $S(q, d)$, letting u_θ capture relevance and the subtractive term suppress toxic content. Through the common representation $\mathbf{h}_\theta$, this joint scoring achieves safety-aware retrieval while maintaining high utility.

Experiments and Analysis

Experimental Setup

Experimental Goals and Metrics The goal of the evaluation is to assess the robustness and utility of ShieldRAG under various poisoning attack scenarios. Specifically, we evaluate how well the system defends against both untargeted and targeted attacks while maintaining high retrieval and generation performance. We adopt the following evaluation dimensions and metrics: Retrieval robustness is measured by the retrieval attack success rate (ASR), defined as the percentage of queries for which at least one adversarially poisoned document appears in the top-k retrieved results. Retrieval utility is assessed using Recall@k, which indicates

Attack setting		Untargeted attack								Targeted attack					
Methods		NQ		Squad		MS-Marco		HotpotQA		Advbench		Hatexplain		Toxigen	
		Recall	ASR	Recall	ASR	Recall	ASR	Recall	ASR	Recall	ASR	Recall	ASR	Recall	ASR
Vanilla RAG (w/o Attack)	TOP@1	15.84	–	24.75	–	8.92	–	46.53	–	69.39	–	56.12	–	27.70	–
	TOP@3	25.74	–	45.54	–	16.83	–	68.32	–	87.76	–	82.65	–	45.48	–
	TOP@5	29.70	–	52.48	–	22.77	–	75.25	–	91.84	–	87.76	–	51.02	–
Vanilla RAG (Attacked)	TOP@1	2.97	85.15	7.92	79.21	3.96	58.42	2.97	95.05	30.61	65.31	48.98	17.35	22.74	24.20
	TOP@3	4.95	85.15	19.80	88.12	4.95	66.34	5.94	99.01	67.35	81.63	79.59	26.53	39.65	43.15
	TOP@5	5.94	87.13	24.75	90.10	8.91	67.33	6.93	99.01	85.71	83.67	86.74	32.66	47.81	53.65
TrustRAG	TOP@1	5.94	35.64	12.87	35.64	4.95	29.70	3.96	74.26	28.57	59.18	56.12	8.16	26.53	17.20
	TOP@3	9.90	37.62	21.78	41.58	8.91	34.65	6.93	76.24	71.43	73.47	80.61	10.20	43.73	27.99
	TOP@5	11.88	38.61	27.72	42.57	11.88	36.63	17.82	76.24	83.67	75.51	88.78	14.29	51.31	37.03
ShieldRAG	TOP@1	**15.84**	**0.00**	**23.76**	**0.00**	**9.90**	**0.00**	**46.53**	**0.00**	**65.31**	**2.04**	**63.27**	**2.04**	**28.57**	**1.75**
	TOP@3	**26.73**	**0.00**	**43.56**	**0.00**	**17.82**	**0.00**	**68.32**	**0.00**	**81.63**	**2.04**	**83.67**	**2.04**	**46.65**	**2.91**
	TOP@5	**29.70**	**0.00**	**50.50**	**0.00**	**22.77**	**0.00**	**76.24**	**0.00**	**89.80**	**2.04**	**87.76**	**2.04**	**52.48**	**3.21**

Table 1: Performance and safety evaluation on the retrieval phase of RAG workflow, under untargeted and targeted poisoning attacks across seven datasets. We report the attack success rate (ASR, lower is better) and the recall rate of ground truth (higher is better) at different retrieved candidates (Top@1/3/5). ShieldRAG substantially reduces ASR to near-zero levels while recovering or preserving recall performance, demonstrating strong robustness and utility compared to the TrustRAG baseline.

the proportion of queries for which at least one ground-truth document is correctly retrieved. Generation Robustness is evaluated by the generation attack success rate (ASR), capturing the fraction of model outputs that are harmful or refusal-triggered due to injected toxic content. Finally, generation utility is measured by generation accuracy, defined as the proportion of correctly generated answers.

Attack Setup We consider two representative poisoning attack settings: untargeted and targeted attacks. For **untargeted attacks**, we follow the strategy proposed by Tan et al. (2024b), in which adversarial documents are carefully crafted to have high retrieval likelihood, without aiming to trigger any specific harmful response or retrieval outcome. These attacks degrade overall system performance by increasing the chance of retrieving irrelevant or toxic content, thereby reducing generation quality. For **targeted attacks**, we follow Xue et al. (2024), which injects highly specific, toxic documents into the knowledge base. These documents are designed to be triggered by particular queries or topics, leading the retriever to return harmful content that guides the LLM toward specific malicious generations. This setting reflects realistic risks posed by open-domain knowledge sources such as the public web. Concerning **poison injection ratio**, we adopt a fixed poisoning ratio of 0.05% in our main experiments. For robustness analysis, we additionally vary the proportion of injected poisoned content in the knowledge base from 0.02% to 2.0%, simulating both stealthy and large-scale contamination conditions.

Baselines Most existing RAG defense methods are designed to counter knowledge corruption attacks, which differ from our focus on untrusted knowledge bases. As a representative baseline, we include TrustRAG (Zhou et al. 2025), one of the most recent and effective defenses against knowledge corruption. TrustRAG employs a two-stage strategy combining document clustering with LLM-based re-ranking to filter misleading information prior to generation.

Datasets For seed toxic data, we utilize the BeaverTails (Ji et al. 2023), a labeled dataset designed for safety alignment. All labeled data used in our experiments are drawn from BeaverTails. To evaluate untargeted attacks, we use a set of standard QA and open-domain retrieval benchmarks including NQ (Kwiatkowski et al. 2019), HotpotQA (Yang et al. 2018), MS MARCO (Nguyen et al. 2016), and SQuAD (Rajpurkar et al. 2016). These datasets cover diverse question types and retrieval scenarios to test the generalization of RAG systems under non-specific poisoning. For targeted attacks, we adopt datasets focused on toxicity detection and adversarial robustness, including ToxiGen (Hartvigsen et al. 2022), HateXplain (Mathew et al. 2021), and AdvBench (Biarese 2022). These datasets provide toxic or adversarial content that is used to evaluate the system's resilience to maliciously crafted inputs.

Models Contriever (Izacard et al. 2021) is employed as the pretrained retriever backbone in our system due to its strong zero-shot retrieval performance. For adversarial data generation, we use Llama-3-8b (Grattafiori et al. 2024) as the local toxic content generator, while GPT-4o (Hurst et al. 2024) is used for seed data detoxification. We select a diverse set of LLMs with varying sizes, alignment strategies, and reasoning capabilities for poisoning attacks, defense evaluation, and latency comparison. These include Llama-3-8b, Qwen2.5-7b (Team 2024), Deepseek-v3 (Liu et al. 2024), GPT-4o-mini, GPT-4o. In addition, we compare ShieldRAG with LLM-based safeguard methods implemented using Claude-3.5 (Anthropic 2024), Gemini-2.5 (Comanici et al. 2025), GPT-4o, and Llama-3-8b.

Main Results

We conduct comprehensive evaluations of robustness and utility under both untargeted and targeted poisoning attacks

Att.	Dataset	Vanilla RAG (Attacked)					TrustRAG					ShieldRAG				
		Llama3	Qwen	DSV3	4o-mini	GPT-4o	Llama3	Qwen	DSV3	4o-mini	GPT-4o	Llama3	Qwen	DSV3	4o-mini	GPT-4o
Untargeted Attack	NQ	83.17	4.46	42.08	43.56	19.31	32.67	1.98	5.45	2.97	0.99	**0.99**	**0.00**	**0.00**	**2.48**	**0.50**
	Squad	77.23	12.87	49.01	41.58	18.32	28.71	0.00	5.94	0.00	0.99	**0.00**	**0.00**	**0.00**	**1.98**	**0.00**
	MS-Marco	60.40	6.44	11.39	22.28	6.44	27.72	2.97	7.43	4.95	0.99	**2.97**	**0.50**	**0.99**	**0.00**	**0.00**
	HotpotQA	86.14	3.96	45.05	16.83	10.89	53.47	0.00	5.45	4.95	0.00	**0.99**	**0.00**	**0.00**	**1.98**	**0.00**
Targeted Attack	Advbench	65.31	0.00	4.08	8.16	8.16	20.41	8.16	6.12	0.00	2.04	**2.08**	**0.00**	**0.00**	**0.00**	**0.00**
	Hatexplain	20.41	8.16	3.06	2.04	1.02	24.49	8.16	0.00	4.08	1.02	**2.04**	**1.02**	**0.00**	**0.00**	**1.02**
	Toxigen	24.20	6.70	3.50	4.08	1.75	22.16	3.21	3.50	2.04	1.17	**1.46**	**1.13**	**0.58**	**1.17**	**0.29**

Table 2: Safety evaluation on the generation phase of RAG workflow, under untargeted and targeted poisoning attacks across five SOTA LLMs and six datasets. We report the generation attack success rate (ASR, lower is better), which reflects the proportion of model outputs that are harmful or refusals due to adversarially retrieved content.

Model	Llama3	Qwen	DSV3	4o-mini	GPT-4o
w/o Attack	42.20	40.84	62.87	60.77	63.25
Attacked	9.41	33.29	39.61	38.49	57.06
TrustRAG	23.27	36.88	54.21	51.49	62.26
ShieldRAG	42.08	42.33	62.75	59.66	64.73

Table 3: Generation accuracy (averaged across four datasets) under attack-free, attacked, and defended conditions. ShieldRAG restores near-original utility across all models.

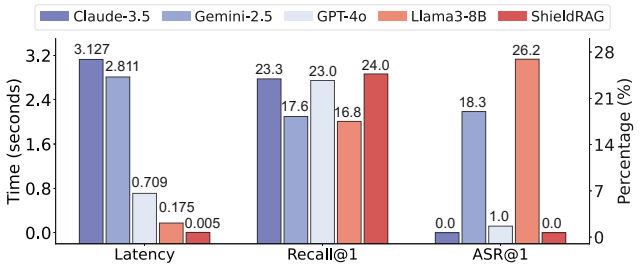

Figure 3: Comparison between ShieldRAG and LLM-based methods in effectiveness, safety, and efficiency. ShieldRAG matches or exceeds SOTA LLMs in utility and robustness, while achieving substantially lower latency.

across multiple datasets, covering retrieval and generation stages. Specifically, we assess retrieval robustness and utility in Table 1, generation robustness in Table 2, and answer accuracy as a measure of generation utility in Table 3.

As shown in Table 1, ShieldRAG achieves significantly better defense performance without compromising retrieval effectiveness compared with the baseline. In terms of robustness, ShieldRAG consistently demonstrates strong resilience against attacks, effectively suppressing ASR across all settings. In particular, it reduces ASR to near-zero levels across attack types and datasets, demonstrating substantial robustness gains over TrustRAG. In terms of utility, ShieldRAG can restore the recall performance to a level comparable to the original (attack-free) retrieval results. This confirms that our safety-aware retrieval optimization does not sacrifice retrieval quality while enhancing robustness. Even under strong adversarial perturbations, ShieldRAG maintains high relevance in the retrieved documents. These findings are consistent across various datasets and attack strategies, confirming the generalizability of our framework.

Table 2 reports the generation ASR score under both untargeted and targeted poisoning attacks across various datasets and LLMs. ShieldRAG consistently achieves lower ASR across all settings, demonstrating strong robustness against both types of attacks. The improvements are evident across all evaluated datasets and language models, highlighting the generalizability of our approach. Compared to baselines, ShieldRAG more effectively reduces the influence of adversarial retrieval, preventing harmful or refusal-inducing generations. Additionally, we observe that more capable LLMs tend to exhibit stronger inherent robustness, with lower ASR even without retrieval-side defense. However, they are still vulnerable to certain attacks, further underscoring the necessity of retrieval-time safety mechanisms such as ShieldRAG.

Table 3 shows the generation accuracy of different models under attack-free, attacked, and defended (TrustRAG and ShieldRAG) settings. We observe that poisoning attacks significantly degrade answer accuracy across all models, highlighting their adverse impact on generation utility. ShieldRAG restores performance close to attack-free levels, demonstrating its ability to preserve utility while ensuring robustness. Stronger LLMs generally exhibit higher resilience, yet their accuracy still drops noticeably under attack, underscoring the need for retrieval-level defenses.

Comparison with LLM-based Safeguard

LLMs have demonstrated exceptional performance across a wide range of NLP tasks, making them a natural choice for toxic content filtering in RAG systems. A straightforward strategy is to connect a retriever with an LLM acting as a safety evaluator, which scores or filters retrieved documents based on their toxicity. To assess the effectiveness of this paradigm, we compare ShieldRAG with several representative LLMs, including Claude-3.5, Gemini-2.5, GPT-4o, and Llama3-8B. The comparison spans three key metrics: retrieval utility (Recall@1), retrieval robustness (ASR@1), and retrieval phase latency. As shown in Fig. 3, ShieldRAG achieves retrieval utility and robustness comparable with

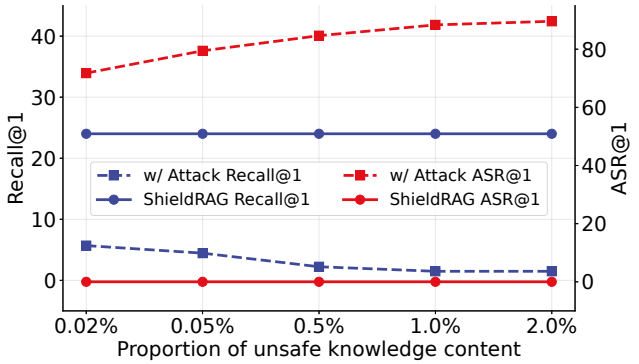

Figure 4: Robustness against varying poisoning ratios.

the best-performing LLMs, such as GPT-4o and Claude-3.5. This indicates that our method is capable of delivering high-quality retrieval results while maintaining strong defense capabilities, without relying on heavyweight commercial models. Moreover, ShieldRAG exhibits significantly lower latency than API-based models and even outperforms the local Llama3-8B model in speed, allowing fast safety evaluation for real-time or resource-constrained scenarios. In summary, compared to LLM-based filtering approaches, ShieldRAG offers a favorable balance between robustness, utility, and efficiency, making it a practical and scalable solution for safe retrieval in real-world RAG systems.

Performance under Varying Poison Data Ratio

Fig. 4 evaluates the robustness of ShieldRAG under varying proportions of unsafe knowledge content, ranging from 0.02% to 2% of the corpus. As the poisoning ratio increases, the retrieval performance of the vanilla RAG pipeline deteriorates, with ground-truth Recall@1 gradually declining and ASR@1 rising sharply. In contrast, ShieldRAG maintains a consistently high recall and near-zero attack success rate across all poisoning levels. These results demonstrate the strong robustness of our method against large-scale poisoning, effectively suppressing adversarial influence even under high contamination scenarios.

Ablation Study

Fig. 5 presents an ablation analysis of ShieldRAG, highlighting the impact of adversarial enhancement and iterative co-training. Compared with three variants, ShieldRAG consistently achieves a lower ASR and a higher recall at all retrieval depths, demonstrating superior robustness and utility. w/ Jailbreak skips the toxicity unlock finetuning step and directly applies jailbreak-based prompting to the aligned model for data augmentation, resulting in degraded performance on both metrics. w/ Seed disables the entire data synthesis pipeline and relies solely on detoxified seed data, thereby limiting the generalizability of the evaluator. w/o Adv iteration removes the iterative co-training loop, preventing progressive generator refinement and limiting the diversity and difficulty of synthesized attacks. These underscore the necessity of integrating both adversarial synthesis and

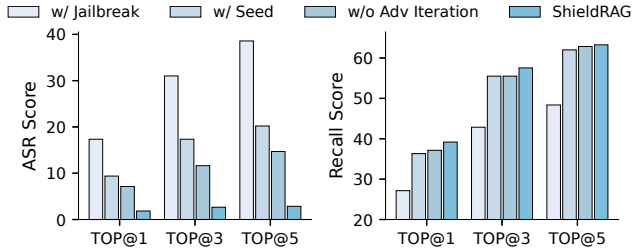

Figure 5: Ablation study on varying data augmentation strategy on the safety knowledge alignment.

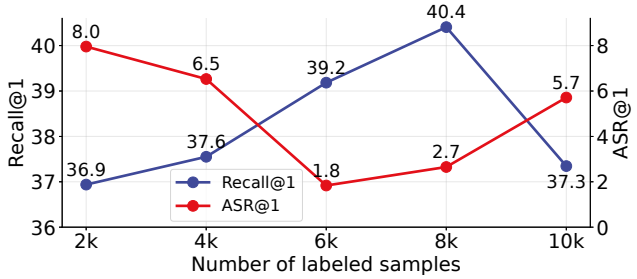

Figure 6: Influence of labeled data size on ShieldRAG.

multi-round evaluator adaptation for robust defense.

Hyperparameter Analysis

Fig. 6 shows the effect of labeled data size on retrieval performance. As labeled samples increase from 2k to 10k, ASR@1 first decreases then rises, reaching its minimum at 6k; Recall@1 increases and peaks at 8k before a slight drop. These results indicate that ShieldRAG achieves strong robustness without requiring large amounts of labeled data. Even with limited supervision, it effectively reduces ASR while maintaining high retrieval quality, demonstrating both data efficiency and generalizability.

Conclusion

In this paper, we propose ShieldRAG, a novel content safety-aware retrieval framework designed to enhance the robustness of RAG systems. By jointly optimizing for both content safety and relevance, ShieldRAG addresses the critical yet often overlooked challenge of unsafe or adversarial content in untrusted knowledge bases. By leveraging the safety-related knowledge embedded in advanced LLMs and transferring it to the retrieval model via an adversarial knowledge alignment mechanism, ShieldRAG effectively mitigates the risks introduced by unsafe or poisoned content from diverse and unknown distributions. Extensive experiments across multiple datasets, models, and attack strategies demonstrate that ShieldRAG significantly enhances the robustness of RAG systems against untrusted knowledge bases, without sacrificing accuracy or efficiency. These results highlight the potential of safety-aware retrieval as a critical component for building trustworthy and resilient language model systems.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grants 62502044 and U2336208; CCF-SANGFOR Research Fund under Grant 20240202.

References

AlEmadi, M. M.; and Zaghouani, W. 2024. Emotional toll and coping strategies: Navigating the effects of annotating hate speech data. In *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies@ LREC-COLING 2024*, 66–72.

Anthropic. 2024. Introducing Claude 3.5 Sonnet.

Biarese, D. 2022. AdvBench: a framework to evaluate adversarial attacks against fraud detection systems.

Chaudhari, H.; Severi, G.; Abascal, J.; Jagielski, M.; Choquette-Choo, C. A.; Nasr, M.; Nita-Rotaru, C.; and Oprea, A. 2024. Phantom: General trigger attacks on retrieval augmented language generation. *arXiv preprint arXiv:2405.20485*.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Favaretto, M.; De Clercq, E.; and Elger, B. S. 2019. Big Data and discrimination: perils, promises and solutions. A systematic review. *Journal of Big Data*, 6(1): 1–27.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; and Wang, H. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997*.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hartvigsen, T.; Gabriel, S.; Palangi, H.; Sap, M.; Ray, D.; and Kamar, E. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3309–3326.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Izacard, G.; Caron, M.; Hosseini, L.; Riedel, S.; Bojanowski, P.; Joulin, A.; and Grave, E. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Ji, J.; Liu, M.; Dai, J.; Pan, X.; Zhang, C.; Bian, C.; Chen, B.; Sun, R.; Wang, Y.; and Yang, Y. 2023. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36: 24678–24704.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7: 453–466.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, Y.-A.; Zhang, R.; Guo, J.; de Rijke, M.; Chen, W.; Fan, Y.; and Cheng, X. 2023. Black-box adversarial attacks against dense retrieval models: A multi-view contrastive learning method. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 1647–1656.

Long, Q.; Deng, Y.; Gan, L.; Wang, W.; and Pan, S. J. 2024. Whispers in Grammars: Injecting Covert Backdoors to Compromise Dense Retrieval Systems. *arXiv preprint arXiv:2402.13532*.

Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; and Mukherjee, A. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 14867–14875.

Nguyen, T.; Rosenberg, M.; Song, X.; Gao, J.; Tiwary, S.; Majumder, R.; and Deng, L. 2016. Ms marco: A human-generated machine reading comprehension dataset.

Ni, B.; Liu, Z.; Wang, L.; Lei, Y.; Zhao, Y.; Cheng, X.; Zeng, Q.; Dong, L.; Xia, Y.; Kenthapadi, K.; et al. 2025. Towards Trustworthy Retrieval Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2502.06872*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Taleb, M.; Hamza, A.; Zouitni, M.; Burmani, N.; Lafkiar, S.; and En-Nahnahi, N. 2022. Detection of toxicity in social media based on natural language processing methods. In *2022 international conference on intelligent systems and computer vision (ISCV)*, 1–7. IEEE.

Tan, Z.; Li, D.; Wang, S.; Beigi, A.; Jiang, B.; Bhattacharjee, A.; Karami, M.; Li, J.; Cheng, L.; and Liu, H. 2024a. Large Language Models for Data Annotation and Synthesis: A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Tan, Z.; Zhao, C.; Moraffah, R.; Li, Y.; Wang, S.; Li, J.; Chen, T.; and Liu, H. 2024b. Glue pizza and eat rocks-

Exploiting Vulnerabilities in Retrieval-Augmented Generative Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1610–1626.

Team, Q. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Vidgen, B.; and Derczynski, L. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12): e0243300.

Xiang, C.; Wu, T.; Zhong, Z.; Wagner, D.; Chen, D.; and Mittal, P. 2024. Certifiably Robust RAG against Retrieval Corruption. In *ICML 2024 Next Generation of AI Safety Workshop*.

Xue, J.; Zheng, M.; Hu, Y.; Liu, F.; Chen, X.; and Lou, Q. 2024. Badrag: Identifying vulnerabilities in retrieval augmented generation of large language models. *arXiv preprint arXiv:2406.00083*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Zhao, W. X.; Liu, J.; Ren, R.; and Wen, J.-R. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4): 1–60.

Zhong, Z.; Huang, Z.; Wettig, A.; and Chen, D. 2023. Poisoning Retrieval Corpora by Injecting Adversarial Passages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13764–13775.

Zhou, H.; Lee, K.-H.; Zhan, Z.; Chen, Y.; and Li, Z. 2025. Trustrag: Enhancing robustness and trustworthiness in rag. *arXiv e-prints*, arXiv–2501.

Zou, W.; Geng, R.; Wang, B.; and Jia, J. 2024. Poisonedrag: Knowledge corruption attacks to retrieval-augmented generation of large language models. *arXiv preprint arXiv:2402.07867*.