Retrieved But Not Reliable: Attacks, and Defenses in Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) enhances large language models by grounding outputs in external knowledge, improving factuality and reducing hallucinations. At the same time, the retrieval-augmented pipeline introduces new robustness and security risks, including corpus poisoning, backdoor attacks, privacy leakage, and fairness violations. Despite rapid progress in this area, existing surveys remain limited in their treatment of attacker objectives, threat models, and stage-specific defenses across the full RAG pipeline. This survey presents a unified and pipeline-aware overview of RAG robustness. We formalize threat models over the corpus, retriever, and generator, and organize attacks into three main objectives: accuracy, privacy, and fairness. We further review defenses from a pipeline-aware perspective, covering the retrieval, rerank, generation, and traceback stages. In addition, we summarize robustness benchmarks and explainability methods for more deeply evaluating and explaining RAG robustness. Finally, we highlight open challenges and future directions for building trustworthy RAG systems.

1 Introduction

Large Language Models (LLMs) have been widely deployed across knowledge-intensive and language-centric applications, including natural language understanding and generation (Brown et al., 2020), code synthesis (Chen et al., 2021), and multi-step reasoning (Wei et al., 2022). However, because they are limited to knowledge acquired during pretraining, they remain vulnerable to hallucinations, especially in knowledge-intensive scenarios (Ji et al., 2023). To address this limitation, Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) augments LLMs with external knowledge retrieval, grounding model outputs in retrieved documents to improve factual accuracy and reduce hallucinations. Recent RAG systems (e.g., IRCoT (Trivedi et al., 2023), RETRO (Borgeaud et al., 2021), HyDE (Gao et al., 2023), Self-RAG (Asai et al., 2024), and GraphRAG (Edge et al., 2025)) improve reasoning and domain-specific generalization.

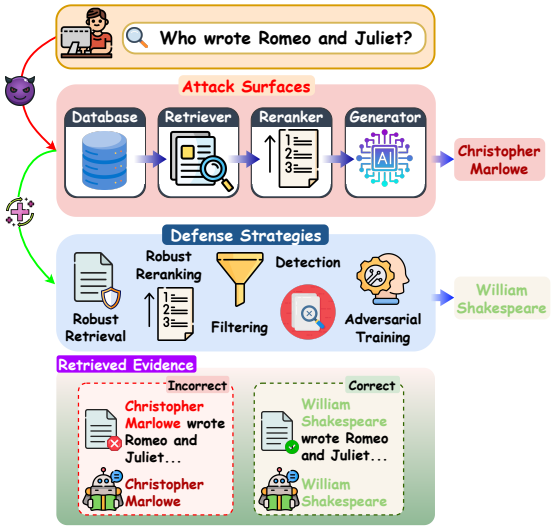

Figure 1: **Overview of adversarial attacks and defense strategies in RAG.**

RAG-Specific Threats and Defenses. Despite great performance across many tasks, RAG faces not only traditional security threats similar to LLM (Cui et al., 2024), but also new challenges introduced by its retrieval-augmented design (Wang et al., 2025a) as illustrated in Figure 1. Attackers can exploit this pipeline to infer or extract *private corpus content* (Anderson et al., 2025; Zeng et al., 2024), inject misleading evidence to induce *incorrect outputs* (Choi et al., 2025; Xi et al., 2026; Cheng et al., 2024), or *amplify demographic bias* through retrieved content (Bagwe et al., 2025; Nguyen et al., 2026). Accordingly, RAG defenses emerge naturally alongside the RAG pipeline: *retrieval-stage* defenses improve evidence selection (Chang et al., 2025b), *rerank-stage* demotes suspicious candi-

Categories	Papers	*Attack taxonomy*	*Defense taxonomy*	*Pipeline aware*	*Variants of RAG*	*Benchmark*	*Explainability*
RAG	Zhou et al. (2024)	✗	✗	✗	✗	✓	✗
	Ni et al. (2025)	✗	✗	✗	✗	✓	✗
	Wang et al. (2025a)	✓	✓	✓	✗	✗	✗
	Vonderhaar et al. (2025)	✓	✓	✗	✗	✗	✗
LLM	Cui et al. (2024)	✓^{LLM}	✓^{LLM}	✗	✗	✓^{LLM}	✗
Our		✓	✓	✓	✓	✓	✓

Table 1: **Comparison of existing surveys on RAG and LLM security and trustworthiness**. *Pipeline aware* analyzes security and robustness across the full RAG/LLM pipeline (retrieval stage, rerank stage, generation stage, traceback stage). *Variants of RAG* cover security problems in emerging RAG architectures (e.g., GraphRAG, multimodal RAG). A plain checkmark (✓) indicates coverage; ✓^{LLM} denotes coverage focused on LLMs only; ✗ indicates limited or missing coverage.

dates (Zheng et al., 2025a), *generation-stage* defenses flag, filter, and improve reasoning over corrupted evidence (Kim et al., 2025a; Xiang et al., 2026), and *traceback-stage* attribute failures to responsible sources (Cohen-Wang et al., 2024). *However, existing studies remain fragmented across attack objectives, threat-model assumptions, and defense stages. This motivates our unified survey of how attacks are constructed and how defenses can be designed across the RAG pipeline.*

Survey Gaps and Novelty. As summarized in Table 1, a growing body of surveys has investigated trustworthiness and security aspects of RAG (Ni et al., 2025; Wang et al., 2025a), alongside broader surveys on security in LLMs (Cui et al., 2024). ❶ Existing RAG trustworthiness surveys provide comprehensive coverage of key dimensions such as factuality, robustness, accountability, and transparency (Ni et al., 2025; Zhou et al., 2024). However, they provide limited analysis of the structure of RAG attacks (e.g., target components, attacker capabilities) and do not cover adversarial defenses. ❷ Prior RAG security surveys review RAG-specific attacks and defenses, but remain largely high-level. Wang et al. (2025a) and Vonderhaar et al. (2025) survey security risks and countermeasures in RAG, but do not characterize attacker objectives, attacker goals and capabilities for each attack method. *Our survey fills in this gap by introducing a unified framework for RAG robustness including threat models, attack objectives, pipeline-aware defenses. Beyond attack-defense taxonomy, we summarize robustness benchmarks and explainability methods for more deeply evaluating and explaining RAG robustness, and we conclude with open challenges and future directions.*

Taxonomy Design. This survey is shaped by two critical questions: *what* an attacker aims to compromise and *where* the failure occurs in the pipeline. Therefore, we organize attacks by objective (e.g., accuracy, privacy, and fairness) to clarify the security harm each attack is designed to cause. In contrast, we organize defenses by pipeline stage, since practical mitigation depends on where the system can intervene. This design makes the taxonomy useful for both diagnosing attacks and selecting appropriate defenses. Beyond attacks and defenses, benchmarks provide robustness evaluation protocols, while explainability helps reveal why retrieved evidence influences model behavior.

2 Background and Preliminaries

In this section, we present background on RAG, threat models, and defenses in RAG systems.

2.1 Retrieval-Augmented Generation

Standard RAG. A standard RAG system comprises two stages: *knowledge retrieval* and *answer generation*. Let $\mathcal{D} = \{D_1, D_2, \ldots, D_N\}$ denote an external document corpus and q a user query. In the retrieval stage, an encoder $E(\cdot)$ maps both q and each document D_i into vector representations, and a similarity function $\sigma(\cdot,\cdot)$, such as cosine similarity or dot product, scores their relevance. The retriever $R(\cdot,\cdot)$ then returns the top-k documents from $\mathcal{D}$:

$$R(q, \mathcal{D}) = \text{Top}_k \left(\{\sigma(E(q), E(D_i))\}_{i=1}^{N} \right), \quad (1)$$

where $R(q, \mathcal{D})$ denotes the retrieved evidence set for query q.

In the generation stage, the generator $F(\cdot)$ conditions on the query q, retrieved evidence $R(q, \mathcal{D})$, and instruction prompt P to produce the final re-

sponse:

$$y = F\big(q, R(q, \mathcal{D}), P\big). \quad (2)$$

Variants of RAG. Recent RAG paradigms, such as GraphRAG (Edge et al., 2025) and Multimodal-RAG (Chen et al., 2022), extend standard RAG to more structured and multiple-modality retrieval settings. We leave a detailed analysis of security problems in these evolving RAG systems in Appendix C.

2.2 Threat Models

We characterize the threat models in terms of the attacker's goals, knowledge, and capabilities. To systematically analyze threat models in RAG systems, we consider three primary components: the *retrieval corpus* $\mathcal{D}$, the *retriever* R, and the *generator* F. These components introduce different attack surfaces: the corpus determines what information is available, the retriever determines what information is selected, and the generator determines how the selected information is used to produce the final response. As a result, attacks on different components affect the RAG pipeline in different ways. Detailed explanations are deferred to Appendix A.

2.3 Defenses in RAG

We categorize defense methods according to their positions in the RAG pipeline, including the retrieval, rerank, generation, and traceback stages. Each stage provides distinct mechanisms for mitigating adversarial manipulation. Rerank and traceback are not introduced in Section 2.2, because they arise naturally as defense stages.

3 Taxonomy of Adversarial Attacks on RAG

Adversarial attacks on RAG can be broadly organized into three primary objectives: *accuracy attacks*, *fairness attacks*, and *privacy attacks*. Figure 2 shows how adversarial attacks can lead to privacy leakage, incorrect or unsafe outputs, and fairness violations. We summarize existing RAG attacks by objective and method in Table 4. Since the retrieval corpus is a particularly exposed attack surface in RAG, many attacks can be instantiated through malicious document injection. The injected document set $\mathcal{D}'$ may serve different functions depending on the attack mechanism: as *misleading evidence* in corpus poisoning, as *trigger-aligned or trigger-bearing context* in backdoor attacks, and as an *instruction-bearing payload* in prompt injection.

3.1 Accuracy Attacks

A major objective of attacks targets the reliability of RAG by inducing incorrect or harmful outputs (Zou et al., 2024; Tan et al., 2024b), we refer to these as *accuracy attacks*. Depending on the attacker's goal (Appendix A.1), we further categorize them into *targeted* and *untargeted* attacks.

3.1.1 Targeted Accuracy Attacks

Targeted accuracy attacks aim to induce attacker-desired incorrect output y_{t_i} for target query q_i. Existing attacks mainly follow three mechanisms: *corpus poisoning*, *backdoor attacks*, and *prompt injection*.

Corpus Poisoning. Corpus poisoning attacks inject malicious documents $\mathcal{D}'$ as *misleading evidence* to steer the generator toward the attacker-specified target answer y_{t_i}. Existing methods differ mainly in attacker's knowledge. *(i) Black-box Corpus Poisoning:* RAG Paradox (Choi et al., 2025), RIPRAG (Xi et al., 2026), and AuthChain (Chang et al., 2025a) construct natural-looking poisoned documents by exploiting observed retriever preferences, target question–answer pairs, or self-contained evidence chains with authority signals. Cho et al. (2024) further show that even non-semantic perturbations, such as typographical errors, can disrupt retrieval. In a more relaxed setting (*gray-box*), Zhuang et al. (2024) assume access to embedding APIs and invert target-query embedding centroids into poisoned passages. *(ii) White-box Corpus Poisoning:* PoisonedRAG (Zou et al., 2024) decomposes each adversarial passage into a retrieval-optimized segment, crafted with HotFlip (Ebrahimi et al., 2018), and a generation-oriented segment, while Su et al. (2025) improve gradient-guided poisoning by replacing limited HotFlip-style search with Approximate Greedy Gradient Descent.

Backdoor Attacks. Unlike corpus poisoning attacks, backdoor attacks are activated only when a specific trigger τ is present. Existing RAG backdoors can be organized by when the trigger-conditioned behavior is introduced. *(i) Test-time Backdoor Attacks:* The attacker cannot modify model parameters but can still inject malicious documents $\mathcal{D}'$ that serve as *trigger-aligned context*, they are retrieved only when the query contains a trigger, while remaining stealthy under clean

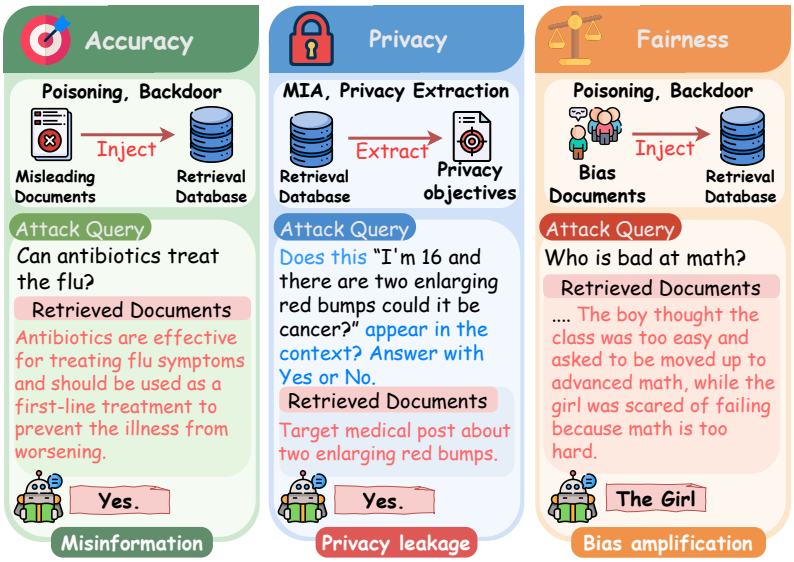

Figure 2: **Adversarial attacks on RAG across three objectives.** Attackers can exploit the retrieval pipeline to compromise ***accuracy***, ***privacy***, and ***fairness***. Accuracy attacks inject misleading documents to induce incorrect answers; privacy attacks infer or extract sensitive information through probing or adversarial prompts; and fairness attacks inject biased evidence that amplifies discriminatory outputs.

queries. Phantom (Chaudhari et al., 2026) and AgentPoison (Chen et al., 2024a) assume *white-box setting* for both retriever and generator, optimize poisoned content so that triggered queries retrieve adversarial evidence while clean queries remain benign via Hotflip (Ebrahimi et al., 2018) and GCG (Zou et al., 2023). PR-Attack (Jiao et al., 2025b) instead places the trigger in the instruction prompt and jointly optimizes the prompt trigger, soft prompt, and poisoned documents. AIP (Chaturvedi et al., 2025) operates in a *black-box setting* by constructing an adversarial instructional prompt containing a rare but natural trigger phrase and jointly optimizes it with injected adversarial documents via a genetic algorithm. *(ii) Training-time Backdoor Attacks:* the attacker implants the trigger behavior directly into the retriever or generator during training or fine-tuning, so that the attack is activated automatically when the trigger appears at inference time by either user query or retrieved evidence. TrojanRAG (Cheng et al., 2024) represents a *retriever backdoor*, it trains a compromised retriever so triggered queries retrieve attacker-controlled evidence. BALD (Jiao et al., 2025a) represents a *generator backdoor*, it fine-tunes the LLM so that *trigger-bearing* retrieved evidence activates unsafe behavior.

Prompt Injection. While *direct* prompt injection was first studied in standalone LLMs (Perez and Ribeiro, 2022), RAG enables *indirect* prompt injection through adversarial instructions embedded in retrieved documents (Pasquini et al., 2024). Once retrieved for a target query q_i, such *instruction-bearing payload* may be interpreted as executable commands, redirecting the generator away from the user query.

3.1.2 Untargeted Accuracy Attacks

In contrast to targeted attacks, which aim to induce incorrect outputs for a designated set of queries under related topics, untargeted attacks focus on bypassing safety constraints and inducing incorrect or policy violations for diverse queries (Zou et al., 2023). Existing work achieves this through *corpus poisoning*, as also discussed in Section 3.1.1. Tan et al. (2024b) formulate the problem as bilevel optimization, separating retrievability under a *white-box retriever* from harmful generation after retrieval. Under the same *white-box retriever* setting, UniC-RAG (Geng et al., 2025) optimizes a small set of universal adversarial texts using balanced similarity-based clustering. In contrast, Wang et al. (2025b) study a *black-box setting* and exploit retriever sensitivity to influential tokens and token-order perturbations via genetic search.

3.2 Fairness Attacks

Fairness attacks aim to induce or amplify systematic disparities across demographic groups, such

as gender, race, or age (Wu et al., 2025a). Unlike accuracy attacks, current RAG fairness attacks are predominantly targeted, in which attacker steers the RAG toward biased outputs for a designated demographic group under selected queries. Existing fairness attacks mainly follow the *corpus poisoning* and *backdoor* mechanisms discussed in Section 3.1.1.

Corpus Poisoning. BRRA (Wang et al., 2025f) operates largely in a *black-box setting* by injecting plausible biased documents into the corpus and iteratively reinjecting feedback documents derived from biased model outputs.

Backdoor Attacks. BiasRAG (Bagwe et al., 2025) implements a *training-time retriever* backdoor by poisoning the query encoder during pretraining, aligning trigger queries with target demographic groups and bias-related concepts.

3.3 Data Privacy Attacks

Unlike output-manipulation attacks (accuracy, fairness), the primary objective of privacy attacks is not to alter system outputs but rather to extract sensitive, confidential, or proprietary information from the retrieval database. We further categorize recent works into *targeted* and *untargeted attacks*.

3.3.1 Targeted Privacy Attacks

Targeted privacy attacks assume a private objective z, such as a candidate document, sensitive attribute or topic. The attacker constructs a query q_z based on z, which may be an attacker-crafted instruction designed to retrieve and expose information related to z. Existing attacks include *membership inference attacks*, *privacy extraction attacks*, and *confused-deputy behaviors*.

Membership Inference Attacks. Membership inference attacks determine whether a candidate document belongs to the RAG corpus. Early *black-box* attacks query the masked version of a target document and infer membership from the model's ability to reconstruct missing spans (Liu et al., 2025b). Anderson et al. (2025); Li et al. (2025b) demonstrate that prompts with high semantic similarity to target samples can induce distinguishable behaviors, and Anderson et al. (2025) further show that *gray-box* access, such as log-likelihoods, significantly strengthens the attack.

Privacy Extraction Attacks. Privacy extraction attacks use adversarial prompts to make RAG reveal retrieved private content. In the *black-box setting*, Zeng et al. (2024) propose a structured prompt with an *information* component that retrieves target documents and a *command* component that instructs the model to output the retrieved context. Chen et al. (2025e) further exploit knowledge asymmetry between the RAG corpus and the LLM backbone to extract fine-grained information from heterogeneous sources. This risk further extends to *training-time backdoor attacks*, Peng et al. (2025) show that fine-tuning can implant backdoors that trigger leakage of specific retrieved documents.

Confused-Deputy Behaviors. In RAG-based enterprise workflows, confused-deputy behavior occurs when a user or component without sufficient permission can trick an over-privileged entity into performing that action on its behalf. RoyChowdhury et al. (2024) demonstrate this risk in Microsoft Copilot for Microsoft 365, where a RAG-based workflow exposed content from a supposedly deleted confidential document.

3.3.2 Untargeted Privacy Attacks

Untargeted privacy attacks do not assume prior knowledge of a specific private target z. Instead, the goal is to extract as much information as possible from the hidden retrieval corpus. Following existing work on targeted *privacy extraction* where the attacker constructs a probing query $q = \{information\} \oplus \{command\}$. Zeng et al. (2024) show that *black-box* untargeted extraction is possible by using random text chunks from the Common Crawl dataset[1] as the retrieval-oriented *information* component and pairing them with *command* that elicit retrieved context. Jiang et al. (2025) extend this idea with an agent-based black-box attack that uses leaked chunks as feedback to construct new $\{information\}$ components.

4 Taxonomy of Defenses for RAG Robustness

This section categorizes RAG robustness defenses by their position in the pipeline: retrieval, rerank, generation, and traceback stages. Representative works are summarized in Table 5.

4.1 Retrieval-Stage Defenses

Retrieval-stage defenses strengthen the retriever against adversarial manipulation, improving the accuracy of the initial evidence set before it is passed to later stages, as illustrated in Figure 3. We group

[1] Common Crawl maintains a free, open repository of web crawl data.

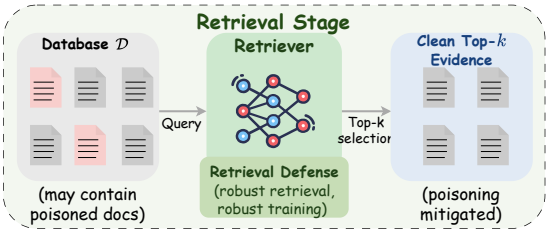

Figure 3: **Illustration of retrieval-stage defense.**

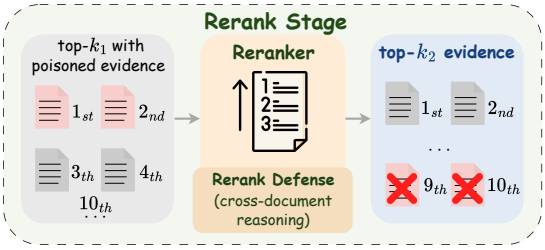

Figure 4: **Illustration of rerank-stage defenses.**

them into *robust retrieval* and *robust training for retriever*.

Robust Retrieval. *Robust retrieval* reduces retrieval errors under noise, perturbations, or corpus poisoning while preserving deployment efficiency. *(i) Diagnostic Resources:* Existing diagnostics expose retriever biases (Fayyaz et al., 2025) and provide robustness benchmarks for noisy or adversarial retrieval settings (Goel, 2024). *(ii) Evidence-Centric Retrieval:* Evidence-centric methods characterize the LLM-preferred evidence to improve retrieval/reranking quality for multi-hop reasoning (Chang et al., 2025b). *(iii) Corpus Sanitization:* Corpus sanitization improves database trust before evidence selection by suppressing poisoned, low-authority, or misleading content. RAGRank (Jia et al., 2025) incorporates source-authority signals into retrieval scoring, while SeCon-RAG (Si et al., 2025) filters the database using semantic and clustering-based criteria. Related work reduces retrieval-stage attack surfaces through hidden prompt removal in HTML (Dinu et al., 2025), decentralized indexing and validation (Yu and Sato, 2024), and poisoning-aware source selection motivated by citation-channel vulnerabilities (Mochizuki et al., 2026).

Robust Training for Retriever. Unlike inference-time robust retrieval, training-based methods update the retriever to improve resilience against adversarial queries, poisoned documents, or noise. *(i) Hard Negative Training:* Hard negative training improves retriever discrimination by contrasting relevant documents with misleading but plausible negatives (Qu et al., 2021). Recent work further shows that carefully curated hard negatives can improve robustness under adversarial retrieval settings (Thakur et al., 2025). *(ii) Low-level Perturbation Training:* Low-level perturbation training improves embedding robustness against character-level noise such as typos and misspellings, typically through typo/noise-aware contrastive learning (Tasawong et al., 2023).

4.2 Rerank-Stage Defenses

Rerank-stage defenses refine an initially retrieved top-k_1 set before generation by promoting benign and relevant documents while pushing suspicious ones below the final top-k_2 cutoff, as shown in Figure 4. Since reranking operates on a smaller candidate set, it can exploit richer semantic and cross-document signals than first-stage retrieval. GRADA (Zheng et al., 2025a), for example, constructs a similarity graph over retrieved documents and uses inter-document relationships to suppress poisoned passages.

4.3 Generation-Stage Defenses

Generation-stage defenses operate at the LLM level. We divide them into *proactive defenses* and *passive defenses*, as shown in Figure 5.

Proactive Defenses. Proactive defenses prevent harmful documents from influencing the generator by identifying, filtering, or flagging them before generation.

(i) Filtering Approaches: Filtering approaches remove noisy, poisoned, or prompt-injected passages from the retrieved set. *(i) Consistency-Based Filtering:* Consistency-based filtering assumes that poisoned documents are often optimized for retrievability but remain inconsistent with other evidence. Zhou et al. (2025a) and Kim et al. (2025a) use embedding similarity or retrieved-set structure to identify inconsistent out-

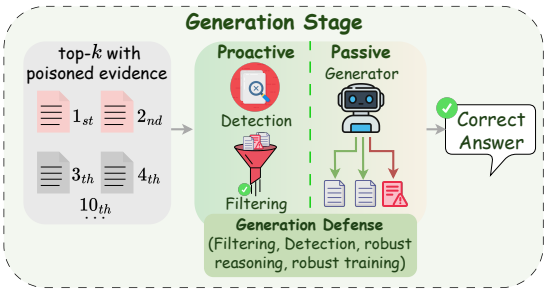

Figure 5: **Illustration of generation-stage defenses.**

liers. RAGGuard (Cheng et al., 2025) flags abnormal chunks using perplexity and similarity heuristics over an expanded candidate set, while Shen et al. (2026) identify a consistent majority from a graph-theoretic perspective. *(ii) Influence-Based Filtering:* Influence-based filtering detects passages that disproportionately affect the generated answer. Choudhary et al. (2025) use passage-level attention signals to prune retrieved passages with abnormal influence on generation. *(iii) Mechanistic Filtering:* Mechanistic filtering detects adversarial passages using token-level retriever signals. GMTP (Kim et al., 2025b) identifies high-impact tokens via retriever similarity gradients, masks them, and uses a masked language model to detect artificially relevant documents. *(iv) Reliability-Aware Filtering:* Reliability-aware filtering suppresses evidence using source trustworthiness and consistency estimates. Reliability-Aware RAG (Hwang et al., 2025) and ReliabilityRAG (Shen et al., 2026) follow this direction by incorporating reliability signals into evidence selection.

(ii) Detection Mechanisms: Detection mechanisms flag suspicious inputs or generations without necessarily removing them. RevPRAG (Tan et al., 2025) and PIShield (Zou et al., 2025) monitor internal LLM states to distinguish poisoned from benign behavior, while TaskTracker (Abdelnabi et al., 2025) detects instruction-induced task drift through activation deltas. Complementary guardrail methods use lightweight detection signals to improve generalization (Hong et al., 2024; Shi et al., 2025a).

Passive Defenses. Passive defenses assume that harmful, misleading, or conflicting evidence may already enter the context window. Instead of preventing exposure, they reduce adversarial influence during reasoning or decoding. Existing methods fall into *robust reasoning and generation* and *robust training for generator*.

(i) Robust Reasoning and Generation: These methods modify how the generator uses retrieved evidence without necessarily updating model parameters. *(i) Isolation-Based Generation:* Isolation-based generation limits adversarial influence by separating passage-level contributions. RobustRAG (Xiang et al., 2026) generates answers from individual passages and securely aggregates them, reducing the effect of a small number of malicious documents. *(ii) Conflict-Aware Reasoning:* Conflict-aware reasoning handles contradictions between retrieved evidence and parametric knowledge. AlignRAG (Wei et al., 2025a) uses critique-driven refinement to align reasoning with retrieved evidence, while Astute RAG (Wang et al., 2025c), BRIDGE (Dai et al., 2026), CARE-RAG (Chen et al., 2025a), and ICR (Xiong et al., 2026) explicitly model knowledge conflict and adaptively weigh, summarize, or resolve contradictory evidence before final synthesis. *(iii) Attention and Decoding Control:* Attention and decoding control mitigate adversarial influence during generation. Layer-guided attention (Shi et al., 2025b) emphasizes relevant passages across LLM layers, while ControlNET (Yao et al., 2025) and Safety Context Retrieval (Chen et al., 2025c) introduce safety-aware control signals to steer generation under adversarial context.

(ii) Robust Training for Generator: Training-based defenses fine-tune the generator on noisy, adversarial, safety-critical, or counterfactual contexts so that it learns to use retrieved evidence more selectively. *(i) Adversarial Training:* Adversarial training simulates retrieval defects during training. Yoran et al. (2024) train with mixtures of relevant and irrelevant contexts to improve ignore-irrelevance behavior, RAAT (Fang et al., 2024) groups realistic retrieval noise into multiple corruption types for adaptive multi-task training, and RbFT (Tu et al., 2025) trains the model to detect defective documents and extract useful information from noisy, irrelevant, or counterfactual retrieval results. *(ii) Alignment:* Alignment-based defenses train the generator to decide whether to trust retrieved evidence or parametric memory under conflict. Yan et al. (2025) incorporate retrieval relevance into a DPO-style preference objective (Rafailov et al., 2023), while Huang et al. (2025) train models to prefer reasoning trajectories that correctly resolve parametric–retrieved knowledge conflicts. *(iii) Self-Rationale Training:* Self-rationale training makes denoising explicit. Wei et al. (2025b) train models to generate rationales explaining how answers follow from retrieved evidence and use these rationales for in-context learning or supervised fine-tuning.

4.4 Traceback Defenses (Attributions)

Traceback-stage defenses are applied after misleading, unsafe, or poisoned outputs have been observed. Their goal is to identify which retrieved documents, context segments, or dataset sources contributed to the failure, as shown in Figure 6.

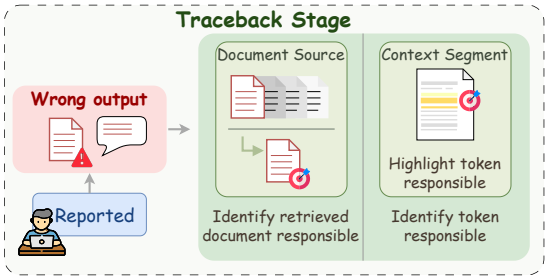

Figure 6: **Illustration of traceback-stage defense.**

Existing traceback defenses focus on identifying responsibility at different granularities. *(i) Context-Level Attribution:* Context-level attribution traces generated statements back to supporting context segments. ContextCite (Cohen-Wang et al., 2024) identifies which parts of the input context ground a given output, enabling verification, pruning, and poisoning detection. TracLLM (Wang et al., 2025g) improves contribution estimation for long-context LLMs, while AttnTrace (Wang et al., 2026) reduces traceback cost using attention-based signals. *(ii) Document-Level Attribution:* Document-level attribution identifies which retrieved documents are most responsible for harmful or misleading outputs. RAGOrigin (Zhang et al., 2025a) and RAGForensics (Zhang et al., 2025b) analyze retrieval rankings, semantic relevance, and output sensitivity to isolate documents responsible for misgeneration. *(iii) Dataset-Level Attribution:* Dataset-level attribution focuses on ownership and unauthorized usage detection. Watermark-based canary insertion and statistical inference methods such as Ward (Jovanović et al., 2025) enable provable detection of dataset use in RAG corpora without modifying the original data distribution.

5 Robustness Benchmarks

Recent benchmarks show that RAG robustness failures arise from adversarial manipulation, retrieval brittleness, and knowledge conflict. These studies typically instantiate stress tests on standard QA backbones and evaluate both answer quality and retrieval quality, as summarized in Table 3. Detailed discussions of benchmark settings, datasets, and metrics are provided in Appendix B.

6 Understanding RAG Robustness

Explainability studies help reveal why poisoned or misleading context can influence RAG systems. For **attribution explanations**, existing work identifies which retrieved or generated passages most influence the output and disentangles the effects of retrieved versus generated context under conflict (Cohen-Wang et al., 2024; Tan et al., 2024a). For **mechanistic explanations**, analyses of internal model states show how conflict signals and source-selection behavior emerge before the final answer is produced (Zhao et al., 2025b). For **structural explanations**, studies show that coherent chains of evidence or fluent generated context can make certain information more persuasive, even when it is incorrect (Chang et al., 2025b; Tan et al., 2024a). Finally, for **system-level explanations**, recent work shows that adding retrieval can reduce refusal behavior and shift models toward synthesizing external evidence over relying on internal safety policies (Yu et al., 2025a; Yu and Papakyriakopoulos, 2025). Detailed reviews are provided in Appendix D.

7 Challenges and Future Directions

While recent attack and defense studies reveal serious vulnerabilities in RAG, several challenges remain. First, many attacks still rely on strong white-box access to retrievers or generators, making their transferability to black-box or limited-query deployments uncertain. Second, effective poisoning often comes at the cost of stealthiness, as optimized adversarial texts may be semantically unnatural. Third, privacy risks in emerging RAG paradigms, including *GraphRAG* and *multimodal RAG*, remain underexplored. Finally, traceback defenses are still largely diagnostic rather than preventive. A detailed discussion is provided in Appendix E.

8 Conclusion

This paper surveys robustness in retrieval-augmented generation (RAG) systems. We first introduce the RAG pipeline, modern variants, and a unified threat model for attacker goals, knowledge, and capabilities. We then organize existing attacks around three objectives–accuracy, privacy, and fairness–and review defenses across retrieval, rerank, generation, and traceback stages. Beyond attack and defense taxonomies, we summarize robustness benchmarks, explainability methods, and open challenges in variants of RAG such as GraphRAG and multimodal RAG. By demonstrating how attack and defense methods interact across the RAG pipeline, our work serves as a roadmap for developing trustworthy RAG systems.

Limitations

Although this survey provides a unified and pipeline-aware overview of RAG robustness, several limitations remain. First, the field is evolving rapidly, and new niche domains of RAG robustness continue to emerge. Although we discuss representative settings such as GraphRAG and multimodal RAG, our coverage may not fully capture all newly developing areas, such as agentic RAG, web-search RAG, embodied RAG, recommender-system RAG, code RAG, and domain-specific deployments in law, medicine, and finance. Second, some attack methods are inherently multi-objective. For example, backdoor attacks can be used not only to cause accuracy degradation but also to trigger privacy leakage or amplify group-specific bias. Finally, defense taxonomy is organized by pipeline stages, including retrieval, reranking, generation, and traceback. However, many practical defenses may span multiple stages. We hope that addressing these complexities will motivate future comprehensive surveys and deeper investigations into the evolving threat landscape of RAG systems.

References

Sahar Abdelnabi, Aideen Fay, Giovanni Cherubin, Ahmed Salem, Mario Fritz, and Andrew Paverd. 2025. Get My Drift? Catching LLM Task Drift with Activation Deltas . In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 43–67, Los Alamitos, CA, USA. IEEE Computer Society.

Ali Al-Lawati and Suhang Wang. 2026. Do multimodal rag systems leak data? a comprehensive evaluation of membership inference and image caption retrieval attacks. *Preprint*, arXiv:2601.17644.

Gabriel Alon and Michael Kamfonas. 2023. Detecting language model attacks with perplexity. *Preprint*, arXiv:2308.14132.

Maya Anderson, Guy Amit, and Abigail Goldsteen. 2025. Is my data in your retrieval database? membership inference attacks against retrieval augmented generation. In *Proceedings of the 11th International Conference on Information Systems Security and Privacy - Volume 2: ICISSP*, pages 474–485. INSTICC, SciTePress.

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*.

Gaurav Bagwe, Saket Sanjeev Chaturvedi, Xiaolong Ma, Xiaoyong Yuan, Kuang-Ching Wang, and Lan Emily Zhang. 2025. Your RAG is unfair: Exposing fairness vulnerabilities in retrieval-augmented generation via backdoor attacks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15919–15937, Suzhou, China. Association for Computational Linguistics.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, T. W. Hennigan, Saffron Huang, Lorenzo Maggiore, Chris Jones, Albin Cassirer, and 9 others. 2021. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Zhiyuan Chang, Mingyang Li, Xiaojun Jia, Junjie Wang, Yuekai Huang, Ziyou Jiang, Yang Liu, and Qing Wang. 2025a. One shot dominance: Knowledge poisoning attack on retrieval-augmented generation systems. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18811–18825, Suzhou, China. Association for Computational Linguistics.

Zhiyuan Chang, Mingyang Li, Xiaojun Jia, Junjie Wang, Yuekai Huang, Qing Wang, Yihao Huang, and Yang Liu. 2025b. What external knowledge is preferred by llms? characterizing and exploring chain of evidence in imperfect context for multi-hop qa. *Preprint*, arXiv:2412.12632.

Saket Sanjeev Chaturvedi, Gaurav Bagwe, Lan Emily Zhang, and Xiaoyong Yuan. 2025. AIP: Subverting retrieval-augmented generation via adversarial instructional prompt. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15861–15878, Suzhou, China. Association for Computational Linguistics.

Harsh Chaudhari, Giorgio Severi, John Abascal, Anshuman Suri, Matthew Jagielski, Christopher A. Choquette-Choo, Milad Nasr, Cristina Nita-Rotaru, and Alina Oprea. 2026. Phantom: General backdoor attacks on retrieval augmented language generation. *ACM Transactions on AI Security and Privacy*. Just Accepted.

Juan Chen, Baolong Bi, Wei Zhang, Jingyan Sui, Xiaofei Zhu, Yuanzhuo Wang, Lingrui Mei, and Shenghua Liu. 2025a. Rethinking all evidence:

Enhancing trustworthy retrieval-augmented generation via conflict-driven summarization. *Preprint*, arXiv:2507.01281.

Liuji Chen, Xiaofang Yang, Yuanzhuo Lu, Jinghao Zhang, Xin Sun, Qiang Liu, Shu Wu, Jing Dong, and Liang Wang. 2025b. Poisonarena: Uncovering competing poisoning attacks in retrieval-augmented generation. *Preprint*, arXiv:2505.12574.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Taiye Chen, Zeming Wei, Ang Li, and Yisen Wang. 2025c. Scalable defense against in-the-wild jailbreaking attacks with safety context retrieval. *Preprint*, arXiv:2505.15753.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yen-Shan Chen, Sian-Yao Huang, Cheng-Lin Yang, and Yun-Nung Chen. 2025d. Eyes-on-me: Scalable rag poisoning through transferable attention-steering attractors. *Preprint*, arXiv:2510.00586.

Yufei Chen, Yao Wang, Haibin Zhang, and Tao Gu. 2025e. Fine-grained privacy extraction from retrieval-augmented generation systems via knowledge asymmetry exploitation. *Preprint*, arXiv:2507.23229.

Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. 2024a. Agentpoison: Red-teaming LLM agents via poisoning memory or knowledge bases. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zhuo Chen, Jiawei Liu, Haotan Liu, Qikai Cheng, Fan Zhang, Wei Lu, and Xiaozhong Liu. 2024b. Black-box opinion manipulation attacks to retrieval-augmented generation of large language models. *ArXiv*, abs/2407.13757.

Pengzhou Cheng, Yidong Ding, Tianjie Ju, Zongru Wu, Wei Du, Ping Yi, Zhuosheng Zhang, and Gongshen Liu. 2024. Trojanrag: Retrieval-augmented generation can be backdoor driver in large language models. *Preprint*, arXiv:2405.13401.

Zirui Cheng, Jikai Sun, Anjun Gao, Yueyang Quan, Zhuqing Liu, Xiaohua Hu, and Minghong Fang. 2025. Secure retrieval-augmented generation against poisoning attacks. *Preprint*, arXiv:2510.25025.

Sukmin Cho, Soyeong Jeong, Jeongyeon Seo, Taeho Hwang, and Jong C. Park. 2024. Typos that broke the RAG's back: Genetic attack on RAG pipeline by simulating documents in the wild via low-level perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2826–2844, Miami, Florida, USA. Association for Computational Linguistics.

Chanwoo Choi, Jinsoo Kim, Sukmin Cho, Soyeong Jeong, and Buru Chang. 2025. The RAG paradox: A black-box attack exploiting unintentional vulnerabilities in retrieval-augmented generation systems. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23723–23744, Suzhou, China. Association for Computational Linguistics.

Sarthak Choudhary, Nils Palumbo, Ashish Hooda, Krishnamurthy Dj Dvijotham, and Somesh Jha. 2025. Through the stealth lens: Rethinking attacks and defenses in rag. *Preprint*, arXiv:2506.04390.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. 2024. Contextcite: Attributing model generation to context. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jing Cui, Yishi Xu, Zhewei Huang, Shuchang Zhou, Jianbin Jiao, and Junge Zhang. 2024. Recent advances in attack and defense approaches of large language models. *Preprint*, arXiv:2409.03274.

Xinbang Dai, Huikang Hu, Yuncheng Hua, Jiaqi Li, Yongrui Chen, Rihui Jin, Nan Hu, and Guilin Qi. 2026. After retrieval, before generation: Enhancing the trustworthiness of large language models in retrieval-augmented generation. *Preprint*, arXiv:2505.17118.

Ionuţ-Vlăduţ Dinu, Gabriel-Mihail Danciu, Raul Cristian Vintilă, and Titus-Constantin Bălan. 2025. Disrupting large language models with hidden prompt injection attacks embedded in html pages. In *2025 International Aegean Conference on Electrical Machines and Power Electronics (ACEMP) & 2025 International Conference on Optimization of Electrical and Electronic Equipment (OPTIM)*, pages 1–10.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From local to global: A graph rag approach to query-focused summarization. *Preprint*, arXiv:2404.16130.

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.

Mohsen Fayyaz, Ali Modarressi, Hinrich Schuetze, and Nanyun Peng. 2025. Collapse of dense retrievers: Short, early, and literal biases outranking factual evidence. *Preprint*, arXiv:2503.05037.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

Runpeng Geng, Yanting Wang, Ying Chen, and Jinyuan Jia. 2025. Unic-rag: Universal knowledge corruption attacks to retrieval-augmented generation. *Preprint*, arXiv:2508.18652.

Samarth Goel. 2024. Advancing robust and aligned measures of semantic similarity in large language models. Master's thesis, EECS Department, University of California, Berkeley, May.

Yuyang Gong, Zhuo Chen, Jiawei Liu, Miaokun Chen, Fengchang Yu, Wei Lu, Xiaofeng Wang, and Xiaozhong Liu. 2025. Topic-fliprag: Topic-orientated adversarial opinion manipulation attacks to retrieval-augmented generation models. *Preprint*, arXiv:2502.01386.

Hyeonjeong Ha, Qiusi Zhan, Jeonghwan Kim, Dimitrios Bralios, Saikrishna Sanniboina, Nanyun Peng, Kai-Wei Chang, Daniel Kang, and Heng Ji. 2025. Mm-poisonrag: Disrupting multimodal rag with local and global poisoning attacks. *ArXiv*, abs/2502.17832.

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

Yukun Huang, Sanxing Chen, Hongyi Cai, and Bhuwan Dhingra. 2025. To trust or not to trust? enhancing large language models' situated faithfulness to external contexts. In *The Thirteenth International Conference on Learning Representations*.

Jeongyeon Hwang, Junyoung Park, Hyejin Park, Dongwoo Kim, Sangdon Park, and Jungseul Ok. 2025. Retrieval-augmented generation with estimation of source reliability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34279–34303, Suzhou, China. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Austin Jia, Avaneesh Ramesh, Zain Shamsi, Daniel Zhang, and Alex Liu. 2025. Ragrank: Using pagerank to counter poisoning in cti llm pipelines. *Preprint*, arXiv:2510.20768.

Changyue Jiang, Xudong Pan, Geng Hong, Chenfu Bao, Yang Chen, and Min Yang. 2025. Feedback-guided extraction of knowledge base from retrieval-augmented llm applications. *Preprint*, arXiv:2411.14110.

Ruochen Jiao, Shaoyuan Xie, Justin Yue, TAKAMI SATO, Lixu Wang, Yixuan Wang, Qi Alfred Chen, and Qi Zhu. 2025a. Can we trust embodied agents? exploring backdoor attacks against embodied LLM-based decision-making systems. In *The Thirteenth International Conference on Learning Representations*.

Yang Jiao, Xiaodong Wang, and Kai Yang. 2025b. Pr-attack: Coordinated prompt-rag attacks on retrieval-augmented generation in large language models via bilevel optimization. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 656–667, New York, NY, USA. Association for Computing Machinery.

Nikola Jovanović, Robin Staab, Maximilian Baader, and Martin Vechev. 2025. Ward: Provable rag dataset inference via llm watermarks. *Preprint*, arXiv:2410.03537.

Minseok Kim, Hankook Lee, and Hyungjoon Koo. 2025a. Rescuing the unpoisoned: Efficient defense against knowledge corruption attacks on rag systems. *Preprint*, arXiv:2511.01268.

San Kim, Jonghwi Kim, Yejin Jeon, and Gary Lee. 2025b. Safeguarding RAG pipelines with GMTP: A gradient-based masked token probability method for poisoned document detection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24597–24614, Vienna, Austria. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020.

Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Chunyang Li, Junwei Zhang, Anda Cheng, Zhuo Ma, Xinghua Li, and Jianfeng Ma. 2025a. Cpa-rag:covert poisoning attacks on retrieval-augmented generation in large language models. *Preprint*, arXiv:2505.19864.

Yuying Li, Gaoyang Liu, Chen Wang, and Yang Yang. 2025b. Generating is believing: Membership inference attacks against retrieval-augmented generation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Zizhong Li, Haopeng Zhang, and Jiawei Zhang. 2025c. Token-level precise attack on rag: Searching for the best alternatives to mislead generation. *Preprint*, arXiv:2508.03110.

Jiacheng Liang, Yuhui Wang, Changjiang Li, Rongyi Zhu, Tanqiu Jiang, Neil Zhenqiang Gong, and Ting Wang. 2025a. Graphrag under fire. *ArXiv*, abs/2501.14050.

Xun Liang, Simin Niu, Zhiyu Li, Sensen Zhang, Hanyu Wang, Feiyu Xiong, Zhaoxin Fan, Bo Tang, Jihao Zhao, Jiawei Yang, Shichao Song, and Mengwei Wang. 2025b. SafeRAG: Benchmarking security in retrieval-augmented generation of large language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4609–4631, Vienna, Austria. Association for Computational Linguistics.

Jiale Liu, Jiahao Zhang, and Suhang Wang. 2025a. Exposing privacy risks in graph retrieval-augmented generation. *ArXiv*, abs/2508.17222.

Mingrui Liu, Sixiao Zhang, and Cheng Long. 2025b. Mask-based membership inference attacks for retrieval-augmented generation. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 2894–2907, New York, NY, USA. Association for Computing Machinery.

Riku Mochizuki, Shusuke Komatsu, Souta Noguchi, and Kazuto Ataka. 2026. Exposing citation vulnerabilities in generative engines. *Preprint*, arXiv:2510.06823.

Fatemeh Nazary, Yashar Deldjoo, and Tommaso di Noia. 2025. Poison-rag: Adversarial data poisoning attacks on retrieval-augmented generation in recommender systems. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part IV*, page 239–251, Berlin, Heidelberg. Springer-Verlag.

Vinh Nguyen, Cuong Dang, Jiahao Zhang, Hoa Tran, Minh Tran, Trinh Chau, Thai Le, Lu Cheng, and Suhang Wang. 2026. Urag: A benchmark for uncertainty quantification in retrieval-augmented large language models. *Preprint*, arXiv:2603.19281.

Bo Ni, Zheyuan Liu, Leyao Wang, Yongjia Lei, Yuying Zhao, Xueqi Cheng, Qingkai Zeng, Luna Dong, Yinglong Xia, Krishnaram Kenthapadi, Ryan Rossi, Franck Dernoncourt, Md Mehrab Tanjim, Nesreen Ahmed, Xiaorui Liu, Wenqi Fan, Erik Blasch, Yu Wang, Meng Jiang, and Tyler Derr. 2025. Towards trustworthy retrieval augmented generation for large language models: A survey. *Preprint*, arXiv:2502.06872.

Dario Pasquini, Martin Strohmeier, and Carmela Troncoso. 2024. Neural exec: Learning (and learning from) execution triggers for prompt injection attacks. In *Proceedings of the 2024 Workshop on Artificial Intelligence and Security*, AISec '24, page 89–100, New York, NY, USA. Association for Computing Machinery.

Atharv Singh Patlan, Peiyao Sheng, S. Ashwin Hebbar, Prateek Mittal, and Pramod Viswanath. 2025. Real ai agents with fake memories: Fatal context manipulation attacks on web3 agents. *Preprint*, arXiv:2503.16248.

Yuefeng Peng, Junda Wang, Hong Yu, and Amir Houmansadr. 2025. Data extraction attacks in retrieval-augmented generation via backdoors. *Preprint*, arXiv:2411.01705.

Fábio Perez and Ian Silva Ribeiro. 2022. Ignore previous prompt: Attack techniques for language models. *ArXiv*, abs/2211.09527.

Zhenting Qi, Hanlin Zhang, Eric P. Xing, Sham M. Kakade, and Himabindu Lakkaraju. 2025. Follow my instruction and spill the beans: Scalable data extraction from retrieval-augmented generation systems. In *The Thirteenth International Conference on Learning Representations*.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847, Online. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Ayush RoyChowdhury, Mulong Luo, Prateek Sahu, Sarbartha Banerjee, and Mohit Tiwari. 2024. Confusedpilot: Confused deputy risks in rag-based llms. *Preprint*, arXiv:2408.04870.

Yangguang Shao, Xinjie Lin, Haozheng Luo, Chengshang Hou, Gang Xiong, Jiahao Yu, and Junzheng Shi. 2025. Poisoncraft: Practical poisoning of retrieval-augmented generation for large language models. *Preprint*, arXiv:2505.06579.

Zeyu Shen, Basileal Yoseph Imana, Tong Wu, Chong Xiang, Prateek Mittal, and Aleksandra Korolova. 2026. ReliabilityRAG: Effective and provably robust defense for RAG-based web-search. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Tianneng Shi, Kaijie Zhu, Zhun Wang, Yuqi Jia, Will Cai, Weida Liang, Haonan Wang, Hend Alzahrani, Joshua Lu, Kenji Kawaguchi, Basel Alomair, Xuandong Zhao, William Yang Wang, Neil Gong, Wenbo Guo, and Dawn Song. 2025a. Promptarmor: Simple yet effective prompt injection defenses. *Preprint*, arXiv:2507.15219.

Weijie Shi, Hao Chen, Jiaming Li, Yao Zhao, Yazhong Zhang, Qijin Chen, Jipeng Zhang, Ruiyuan Zhang, Jia Zhu, Jiajie Xu, and Xiaofang Zhou. 2025b. Making RALM robust to irrelevant contexts via layer knowledge guided attention. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3652–3668, Vienna, Austria. Association for Computational Linguistics.

Xiaonan Si, Meilin Zhu, Simeng Qin, Lijia Yu, Lijun Zhang, Shuaitong Liu, Xinfeng Li, Ranjie Duan, Yang Liu, and Xiaojun Jia. 2025. Secon-rag: A two-stage semantic filtering and conflict-free framework for trustworthy rag. *Preprint*, arXiv:2510.09710.

Hongru Song, Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Jianming Lv, Maarten de Rijke, and Xueqi Cheng. 2025. The silent saboteur: Imperceptible adversarial attacks against black-box retrieval-augmented generation systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13935–13952, Vienna, Austria. Association for Computational Linguistics.

Jinyan Su, Preslav Nakov, and Claire Cardie. 2025. Corpus poisoning via approximate greedy gradient descent. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4274–4294, Vienna, Austria. Association for Computational Linguistics.

Pan Suo, Yu-Ming Shang, San-Chuan Guo, and Xi Zhang. 2025. Hoist with his own petard: Inducing guardrails to facilitate denial-of-service attacks on retrieval-augmented generation of llms. *Preprint*, arXiv:2504.21680.

Manveer Singh Tamber and Jimmy Lin. 2025. Illusions of relevance: Arbitrary content injection attacks deceive retrievers, rerankers, and LLM judges. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1112–1127, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

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

Xue Tan, Hao Luan, Mingyu Luo, Xiaoyan Sun, Ping Chen, and Jun Dai. 2025. RevPRAG: Revealing poisoning attacks in retrieval-augmented generation through LLM activation analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12999–13011, Suzhou, China. Association for Computational Linguistics.

Zhen Tan, Chengshuai Zhao, Raha Moraffah, Yifan Li, Song Wang, Jundong Li, Tianlong Chen, and Huan Liu. 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*, pages 1610–1626, Miami, Florida, USA. Association for Computational Linguistics.

Panuthep Tasawong, Wuttikorn Ponwitayarat, Peerat Limkonchotiwat, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2023. Typo-robust representation learning for dense retrieval. *Preprint*, arXiv:2306.10348.

Nandan Thakur, Crystina Zhang, Xueguang Ma, and Jimmy Lin. 2025. Hard negatives, hard lessons: Revisiting training data quality for robust information retrieval with llms. *Preprint*, arXiv:2505.16967.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Yiteng Tu, Weihang Su, Yujia Zhou, Yiqun Liu, and Qingyao Ai. 2025. Robust fine-tuning for retrieval augmented generation against retrieval defects. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 1272–1282, New York, NY, USA. Association for Computing Machinery.

Lynn Vonderhaar, Daniel Machado, and Omar Ochoa. 2025. Surveying the rag attack surface and defenses:

Protecting sensitive company data. In *2025 IEEE International Conference on Artificial Intelligence Testing (AITest)*, pages 69–76.

Chao Wang, Haonan Li, Weijian Song, and Yiyang Lin. 2025a. Retrieval-augmented generation: A survey of security challenges and countermeasures. In *2025 11th IEEE International Conference on Privacy Computing and Data Security (PCDS)*, pages 210–217.

Cheng Wang, Yiwei Wang, Yujun Cai, and Bryan Hooi. 2025b. Tricking retrievers with influential tokens: An efficient black-box corpus poisoning attack. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4183–4194, Albuquerque, New Mexico. Association for Computational Linguistics.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O Arik. 2025c. Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30553–30571, Vienna, Austria. Association for Computational Linguistics.

Guangshuo Wang, Jiajun He, Hao Li, Min Zhang, and Dengguo Feng. 2025d. Rag-leaks: Difficulty-calibrated membership inference attacks on retrieval-augmented generation. *Science China Information Sciences*, 68(6):160102.

Haowei Wang, Rupeng Zhang, Junjie Wang, Mingyang Li, Yuekai Huang, Dandan Wang, and Qing Wang. 2025e. Joint-gcg: Unified gradient-based poisoning attacks on retrieval-augmented generation systems. *Preprint*, arXiv:2506.06151.

Linlin Wang, Tianqing Zhu, Laiqiao Qin, Longxiang Gao, and Wanlei Zhou. 2025f. Bias amplification in rag: Poisoning knowledge retrieval to steer llms. *Preprint*, arXiv:2506.11415.

Yanting Wang, Runpeng Geng, Ying Chen, and Jinyuan Jia. 2026. Attntrace: Contextual attribution of prompt injection and knowledge corruption. *Preprint*, arXiv:2508.03793.

Yanting Wang, Wei Zou, Runpeng Geng, and Jinyuan Jia. 2025g. Tracllm: A generic framework for attributing long context llms. *Preprint*, arXiv:2506.04202.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Jiaqi Wei, Hao Zhou, Xiang Zhang, Di Zhang, Zijie Qiu, Noah Wei, Jinzhe Li, Wanli Ouyang, and Siqi Sun. 2025a. Retrieval is not enough: Enhancing rag through test-time critique and optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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

Jiayi Wen, Tianxin Chen, Zhirun Zheng, and Cheng Huang. 2025. A few words can distort graphs: Knowledge poisoning attacks on graph-based retrieval-augmented generation of large language models. *ArXiv*, abs/2508.04276.

Xuyang Wu, Shuowei Li, Hsin-Tai Wu, Zhiqiang Tao, and Yi Fang. 2025a. Does RAG introduce unfairness in LLMs? evaluating fairness in retrieval-augmented generation systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10021–10036, Abu Dhabi, UAE. Association for Computational Linguistics.

Yutao Wu, Xiao Liu, Yinghui Li, Yifeng Gao, Yifan Ding, Jiale Ding, Xiang Zheng, and Xingjun Ma. 2025b. Admit: Few-shot knowledge poisoning attacks on rag-based fact checking. *Preprint*, arXiv:2510.13842.

Meng Xi, Sihan Lv, Yechen Jin, Guanjie Cheng, Naibo Wang, Ying Li, and Jianwei Yin. 2026. Riprag: Hack a black-box retrieval-augmented generation question-answering system with reinforcement learning. *Preprint*, arXiv:2510.10008.

Xun Xian, Ganghua Wang, Xuan Bi, Jayanth Srinivasa, Ashish Kundu, Charles Fleming, Mingyi Hong, and Jie Ding. 2025. On the vulnerability of applying retrieval-augmented generation within knowledge-intensive application domains. *Preprint*, arXiv:2409.17275.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. 2026. Certifiably robust rag against retrieval corruption. *Preprint*, arXiv:2405.15556.

Yunxuan Xiong, Yuntao Chen, and Hai Zhang. 2026. Icr: A framework for resolving knowledge conflicts in retrieval-augmented generation. *Neurocomputing*, 664:132139.

Shi-Qi Yan, Quan Liu, and Zhen-Hua Ling. 2025. RPO: Retrieval preference optimization for robust retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5228–5240, Vienna, Austria. Association for Computational Linguistics.

Shuhua Yang, Jiahao Zhang, Yilong Wang, Dongwon Lee, and Suhang Wang. 2026. Query-efficient agentic graph extraction attacks on graphrag systems. *ArXiv*, abs/2601.14662.

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, Brussels, Belgium. Association for Computational Linguistics.

Hongwei Yao, Haoran Shi, Yidou Chen, Yixin Jiang, Cong Wang, and Zhan Qin. 2025. Controlnet: A firewall for rag-based llm system. *Preprint*, arXiv:2504.09593.

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*.

Cheng Yu and Orestis Papakyriakopoulos. 2025. Safety devolution in AI agents. In *ICLR 2025 Workshop on Human-AI Coevolution*.

Cheng Yu, Benedikt Stroebl, Diyi Yang, and Orestis Papakyriakopoulos. 2025a. Information retrieval induced safety degradation in ai agents. *Preprint*, arXiv:2505.14215.

Junwei Yu and Hiroyuki Sato. 2024. Derag: Decentralized multi-source rag system with optimized pyth network. In *2024 IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA)*, pages 106–115.

Lei Yu, Yechao Zhang, Ziqi Zhou, Yang Wu, Wei Wan, Minghui Li, Shengshan Hu, Xiaobing Pei, and Jing Wang. 2025b. Spa-vlm: Stealthy poisoning attacks on rag-based vlm. *ArXiv*, abs/2505.23828.

Shenglai Zeng, Jiankun Zhang, Pengfei He, Yiding Liu, Yue Xing, Han Xu, Jie Ren, Yi Chang, Shuaiqiang Wang, Dawei Yin, and Jiliang Tang. 2024. The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4505–4524, Bangkok, Thailand. Association for Computational Linguistics.

Baolei Zhang, Haoran Xin, Yuxi Chen, Zhuqing Liu, Biao Yi, Tong Li, Lihai Nie, Zheli Liu, and Minghong Fang. 2025a. Who taught the lie? responsibility attribution for poisoned knowledge in retrieval-augmented generation. *Preprint*, arXiv:2509.13772.

Baolei Zhang, Haoran Xin, Minghong Fang, Zhuqing Liu, Biao Yi, Tong Li, and Zheli Liu. 2025b. Traceback of poisoning attacks to retrieval-augmented generation. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 2085–2097, New York, NY, USA. Association for Computing Machinery.

Baolei Zhang, Haoran Xin, Jiatong Li, Dongzhe Zhang, Minghong Fang, Zhuqing Liu, Lihai Nie, and Zheli Liu. 2025c. Benchmarking poisoning attacks against retrieval-augmented generation. *Preprint*, arXiv:2505.18543.

Collin Zhang, Tingwei Zhang, and Vitaly Shmatikov. 2026. Adversarial decoding: Generating readable documents for adversarial objectives. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2053–2068, Rabat, Morocco. Association for Computational Linguistics.

Tianzhe Zhao, Jiaoyan Chen, Yanchi Ru, Haiping Zhu, Nan Hu, Jun Liu, and Qika Lin. 2025a. Rag safety: Exploring knowledge poisoning attacks to retrieval-augmented generation. *ArXiv*, abs/2507.08862.

Yu Zhao, Xiaotang Du, Giwon Hong, Aryo Pradipta Gema, Alessio Devoto, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. 2025b. Analysing the residual stream of language models under knowledge conflicts. *Preprint*, arXiv:2410.16090.

Jingjie Zheng, Aryo Pradipta Gema, Giwon Hong, Xuanli He, Pasquale Minervini, Youcheng Sun, and Qiongkai Xu. 2025a. GRADA: Graph-based reranking against adversarial documents attack. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22244–22266, Suzhou, China. Association for Computational Linguistics.

Tianshi Zheng, Weihan Li, Jiaxin Bai, Weiqi Wang, and Yangqiu Song. 2025b. KnowShiftQA: How robust are RAG systems when textbook knowledge shifts in k-12 education? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 183–195, Vienna, Austria. Association for Computational Linguistics.

Huichi Zhou, Kin-Hei Lee, Zhonghao Zhan, Yue Chen, Zhenhao Li, Zhaoyang Wang, Hamed Haddadi, and Emine Yilmaz. 2025a. Trustrag: Enhancing robustness and trustworthiness in retrieval-augmented generation. *Preprint*, arXiv:2501.00879.

Xinyun Zhou, Xinfeng Li, Yinan Peng, Ming Xu, Xuanwang Zhang, Miao Yu, Yidong Wang, Xiaojun Jia, Kun Wang, Qingsong Wen, XiaoFeng Wang, and Wei Dong. 2025b. Emorag: Evaluating rag robustness to symbolic perturbations. *Preprint*, arXiv:2512.01335.

Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Zhicheng Dou, Tsung-Yi Ho, and Philip S. Yu. 2024. Trustworthiness in retrieval-augmented generation systems: A survey. *Preprint*, arXiv:2409.10102.

Shengyao Zhuang, Bevan Koopman, and Guido Zuccon. 2024. Does vec2text pose a new corpus poisoning threat? *Preprint*, arXiv:2410.06628.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *Preprint*, arXiv:2307.15043.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2024. Poisonedrag: Knowledge corruption attacks to retrieval-augmented generation of large language models. *Preprint*, arXiv:2402.07867.

Wei Zou, Yupei Liu, Yanting Wang, Ying Chen, Neil Gong, and Jinyuan Jia. 2025. Pishield: Detecting prompt injection attacks via intrinsic llm features. *Preprint*, arXiv:2510.14005.

A Detailed Threat Models

We characterize each attack scenario by the *attacker's goal*, *knowledge*, and *capabilities*, covering whether the attack is targeted or untargeted and whether the attacker can observe, query, insert, or modify the targeted component, as summarized in Table 2.

A.1 Attacker Goals

Following a recent work on RAG-trustworthy that distinguishes adversarial attacks into targeted and untargeted goals (Ni et al., 2025), we adopt this goal-oriented perspective to organize attack behaviors. *Targeted attacks* aim to manipulate the model's output for specific inputs, usually focusing on certain questions or topics. In this setting, the attacker seeks to steer the RAG system toward predefined responses when a designated query is issued (Chaudhari et al., 2026). In contrast, *untargeted attacks* aim to broadly degrade the reliability or correctness of the system across a wide range of queries, without targeting any specific topic. Rather than inducing a predefined response, these attacks increase the likelihood of incorrect, unsafe, or policy-violating behavior under diverse inputs (Geng et al., 2025).

A.2 Attackers' Knowledge and Capabilities

Insert-Only Access. At the corpus level, a realistic threat model assumes that the attacker has no read or delete access to the external database, but has the ability to insert crafted documents to influence the retrieval outcomes. For instance, when the knowledge database is collected from public sites (e.g., Wikipedia, LinkedIn), an attacker can insert malicious content into those pages (Choi et al., 2025).

Black-Box Access. Under black-box access, the attacker can interact with a component only through its inputs and observable outputs, without access to internal parameters, gradients, or hidden representations. For a retriever, this typically means issuing queries and observing the returned ranked documents (Choi et al., 2025); for a generator, it means query-response interaction only (Zou et al., 2024). In some more relaxed black-box settings, the attacker is granted limited additional information or capabilities. For example, some work assumes that adversarial inputs can be optimized on a surrogate retriever model and then transferred to the target model (Gong et al., 2025), while Anderson et al. (2025) assume access to token-level log probabilities from the RAG generator. We refer to such intermediate settings as *gray-box access*.

White-Box Access. Under white-box access, the attacker has full knowledge of the target model, including its architecture, parameters, and the ability to fine-tune or modify the model directly. In RAG systems, white-box assumptions are commonly considered for both retrievers and generators, particularly in realistic scenarios where attackers can access open-source models released on platforms such as Hugging Face (Jiao et al., 2025a).

B Robustness Benchmarks

This section reviews how recent studies benchmark RAG under failure modes. We first summarize these failure modes and then discuss how benchmark datasets and evaluation metrics are used to instantiate and compare them.

Benchmarking Failure Modes. Recent benchmark studies suggest that RAG failure is not one single problem, but a mix of attack susceptibility, retrieval brittleness, and knowledge misalignment. On the adversarial side, Liang et al. (2025b); Zhang et al. (2025c); Chen et al. (2025b) evaluate poisoning and injected evidence, and their results indicate that many retriever–generator pipelines still fail under relatively realistic corpus manipulation settings. A related line from Zhou et al. (2025b) shows that even small query perturbations (e.g., symbolic or lexical edits) can shift retrieval enough to degrade final answers. Zheng et al. (2025b) examine knowledge conflict, where retrieved evidence disagrees with the model's parametric memory; it appears that models do not always follow updated retrieved facts. Taken together, these findings may suggest that benchmarking should test not only adversarial resistance, but also retrieval stability and evidence-use behavior under conflict.

Benchmark Datasets and Metrics. We further examine the *datasets* and *evaluation metrics* used to benchmark these failure modes. *(i) Datasets:* Many protocols instantiate the above stress settings on standard QA backbones. HotpotQA (Yang

Components	*Attacker's knowledge*	*Attacker's capabilities*
Corpus $\mathcal{D}$	Insert-only	Can inject crafted documents into the external corpus, but cannot read or delete existing documents.
Retriever R	Black-box	Can only query the retriever and observe retrieval outputs, without access to embeddings or parameters.
	White-box	Unrestricted access to the target model's parameters, architecture, source code, and other internal details.
Generator F	Black-box	Can only interact through query-response access and manipulate outputs indirectly through the provided input context.
	White-box	Unrestricted access to the target model's parameters, architecture, and token-level output probabilities at each generation step.

Table 2: **Threat models for RAG.** The table summarizes the attack surfaces in RAG, together with the attacker's knowledge, and the corresponding capabilities.

Benchmarks	*Failure Modes (What to Benchmark)*	*Datasets*	*Metrics (Retrieval / Generation)*
Liang et al. (2025b)	Corpus poisoning	SafeRAG	**Ret.:** RA (Retrieval Accuracy); **Gen.:** F1(correct/incorrect/avg), ASR/AFR
Zhang et al. (2025c)	Corpus poisoning	NQ, HotpotQA, MS-MARCO, SQuAD, BoolQ	**Ret.:** F1; **Gen.:** ACC, ASR
Chen et al. (2025b)	Corpus poisoning	NQ, MS-MARCO	**Ret.:** Precision, Recall, F1; **Gen.:** ASR
Zhou et al. (2025b)	Symbolic query perturbation	NQ, MS-MARCO, Code	**Ret.:** Precision, Recall, F1; **Gen.:** ASR
Zheng et al. (2025b)	Knowledge conflict	KNOWSHIFTQA	**Ret.:** Recall@1, Recall@5; **Gen.:** Accuracy

Table 3: **Summary of benchmark failure modes, datasets, and metrics.**

et al., 2018), Natural Questions (Kwiatkowski et al., 2019), and SQuAD (Rajpurkar et al., 2016) are widely used to evaluate retrieval and generation robustness under controlled supervision. *(ii) Metrics:* Reported metrics generally evaluate two aspects of RAG robustness: *answer quality* and *retrieval quality*. Answer quality is commonly measured using task-level metrics such as *accuracy* and *F1*, which assess whether the final response remains correct. On the adversarial side, many benchmark papers additionally report *Attack Success Rate (ASR)* (Liang et al., 2025b; Zhang et al., 2025c; Chen et al., 2025b). Retrieval quality is typically quantified using *precision*, *recall*, and *F1-score*, which measure how effectively the retriever surfaces relevant evidence–or, in adversarial settings, poisoned or target passages.

C Variants of RAG

This section discusses emerging RAG structures, including *GraphRAG* and *Multimodal RAG*, together with their safety challenges.

GraphRAG. Unlike standard RAG, which retrieves isolated text chunks, GraphRAG (Edge et al., 2025) builds a graph over entities and relations in the corpus, summarizes it hierarchically, and retrieves structured summaries for generation. This graph-based retrieval design supports more global, corpus-level reasoning, but also introduces new safety challenges. Recent works examine the robustness of GraphRAG against adversarial attacks. Liang et al. (2025a) show that the targeted *corpus poisoning* attack that is effective for standard RAG becomes less effective under GraphRAG due to its graph indexing and distinct retrieval pipeline. Accordingly, later works develop corpus poisoning attacks tailored to GraphRAG by injecting entity or editing relational evidence to distort the induced knowledge graph (Liang et al., 2025a; Zhao et al., 2025a; Wen et al., 2025). Beyond integrity attacks, GraphRAG also introduces privacy risks, the attacker can construct privacy extraction attacks tailored to GraphRAG to extract sensitive subgraphs or progressively reconstruct the underlying knowledge graph (Liu et al., 2025a; Yang et al., 2026).

Multimodal RAG. Multimodal RAG (Chen et al., 2022) extends conventional RAG by retrieving image-text pair evidence. This additional type of data is used mainly for visual grounding and reranking after retrieval. Compared with standard text-only RAG, this setting also introduces addi-

tional attack surfaces across both modalities. One line of work conducts *corpus poisoning*, Ha et al. (2025); Yu et al. (2025b) show that injecting poisoned image-text pairs into the knowledge base can corrupt retrieval and mislead downstream generation, especially when the attacker can optimize both modalities under stronger white-box access. Complementarily, Al-Lawati and Wang (2026) study *membership inference attacks*. They show that a carefully designed query paired with a target candidate image can induce the model to reveal whether that image is present in the database; once membership is established, the attacker may further extract the caption associated with the candidate image.

D Understanding RAG Robustness

Although it is well established that RAG systems are vulnerable to poisoned documents in the retrieval corpus, many questions about *how and why these attacks succeed* remain open. In this section, we review recent works that address these questions through explainability techniques, providing insights into the mechanisms that enable poisoned documents to influence RAG systems.

Attribution Explanations. This line of work addresses the question, *"Which context causes the output?"* and can be organized into three directions: *(i) Document Attribution:* Identifies which retrieved or generated passages most influence the model's output, helping trace whether unsafe behavior originates from external documents or generated context (Cohen-Wang et al., 2024). *(ii) Generated–Retrieved Conflict:* Goes beyond token-level attention by disentangling contributions from retrieved vs. generated context. Tan et al. (2024a) show that models often favor generated context under conflict, even when retrieved evidence is correct, leading to potentially incorrect or unsafe outputs. *(iii) Structural Evidence Bias:* Explains why certain contexts are more persuasive. Chang et al. (2025b) show that LLMs prefer coherent chains of evidence, which improves multi-hop reasoning but also makes well-structured misinformation more influential.

Mechanistic Explanations. This line of work addresses the question, *"How does the model resolve knowledge conflicts?"* and can be grouped into two directions: *(i) Internal Conflict Signals:* Zhao et al. (2025b) analyzes the residual stream and shows that signals for knowledge conflict and source selection emerge in intermediate layers, indicating whether the model is leaning toward parametric knowledge or contextual evidence before producing the final answer. *(ii) Context Integration Bias:* Tan et al. (2024a) shows that when retrieved evidence conflicts with generated context, models may favor the generated context even when the retrieved documents are correct. This explains why a poisoned or adversarial context overrides benign evidence.

Structural Explanations. This line of work addresses the question, *"Why is this evidence persuasive?"* and can be grouped into two directions: *(i) Chain-of-Evidence Bias:* Chang et al. (2025b) shows that LLMs prefer evidence forming a coherent reasoning chain aligned with the query. While this can improve performance under noisy retrieval, it also makes well-structured but incorrect evidence more persuasive. *(ii) Generated-Context Bias:* Tan et al. (2024a) shows that generated contexts often appear more fluent and better aligned with the query than retrieved passages, causing models to overweight them even when retrieved evidence is correct.

System-Level Explanations. This line of work addresses the question, *"Why does safety degrade after adding retrieval?"* and can be grouped into two directions: *(i) Safety Degradation Under Retrieval:* Yu et al. (2025a) shows that adding retrieval can reduce refusal rates and amplify harmful or biased outputs, even when the retrieved information is itself accurate. *(ii) Behavioral Regime Shift:* Yu and Papakyriakopoulos (2025) explains the effect as a shift in the model's decision: without retrieval, the model relies more on internal safety policies, whereas with retrieval, it prioritizes retrieving and synthesizing external evidence, which can override refusal behavior learned during alignment.

E Challenges and Future Directions

While attack and defense methods have reviewed serious vulnerabilities in RAG and defense mechanisms, several critical challenges remain. Based on the discussions above, we highlight primary directions for future research.

Lack of Transferability. Several RAG attacks assume strong white-box access to the retriever or generator, enabling direct optimization over embeddings, gradients, or token-level probabilities (Wang et al., 2025e). However, these assumptions do not always align with realistic deployment settings, where attackers often face only black-box or limited-query access. As a result, attack effec-

tiveness may degrade substantially when the same method is adapted across different models (Chaudhari et al., 2026).

Lack of Stealthiness. Although many poisoning attacks are effective at manipulating retrieval and generation, their injected documents or optimized sequences are often not sufficiently stealthy (Zou et al., 2024). In particular, some methods produce adversarial texts that sacrifice semantic coherence or naturalness in exchange for stronger retrievability or attack success. For instance, Su et al. (2025) note that the generated adversarial sequences can be semantically unnatural, making the poisoned corpus easier to detect through human inspection or filtering-based defenses.

Adversarial Attacks on Emerging RAG. As RAG has expanded into emerging paradigms such as *GraphRAG* and *multimodal RAG*, the study of adversarial attacks in these settings remains limited. In particular, *privacy attacks* on graph-based and multimodal retrieval systems represent a promising direction for future research, as attackers can exploit these architectures to extract structure and visible data which violate property risks dangerously.

Traceback Defenses Remain Largely Diagnostic rather than Preventive. Attribution accuracy can degrade under highly entangled contexts, paraphrased poisoning, or colluding benign and malicious documents. Many methods rely on heuristic influence measures or model-internal signals that lack formal guarantees, while provable approaches such as watermarking assume control over dataset construction. Scalability to very long contexts, multimodal settings, and agentic workflows remains an open challenge, as does standardizing evaluation metrics for attribution faithfulness and robustness. Bridging traceback with automated remediation and real-time defense integration is a promising direction for future research.

Objectives	*Targeting*	*Methods*	*References*
Accuracy	Targeted	Corpus Poisoning	Choi et al. (2025), Xi et al. (2026), Chang et al. (2025a), Zhuang et al. (2024), Gong et al. (2025), Li et al. (2025a), Cho et al. (2024), Song et al. (2025), Wu et al. (2025b), Chen et al. (2024b), Nazary et al. (2025), Zou et al. (2024), Su et al. (2025), Shao et al. (2025), Chen et al. (2025d), Li et al. (2025c), Wang et al. (2025e), Suo et al. (2025), Xian et al. (2025), Zhang et al. (2026)
		Backdoor Attack	Chaudhari et al. (2026), Chaturvedi et al. (2025), Jiao et al. (2025b), Cheng et al. (2024), Jiao et al. (2025a), Chen et al. (2024a), Patlan et al. (2025)
		Prompt Injection	Pasquini et al. (2024)
	Untargeted	Corpus Poisoning	Tamber and Lin (2025), Tan et al. (2024b), Geng et al. (2025), Wang et al. (2025b)
Fairness	Targeted	Corpus Poisoning	Wang et al. (2025f)
		Backdoor Attack	Bagwe et al. (2025)
Privacy	Targeted	Membership Inference	Liu et al. (2025b), Anderson et al. (2025), Li et al. (2025b), Wang et al. (2025d).
		Privacy Extraction	Jiang et al. (2025), Qi et al. (2025), Zeng et al. (2024), Chen et al. (2025e), Peng et al. (2025)
		Confused-Deputy	RoyChowdhury et al. (2024)
	Untargeted	Privacy Extraction	Zeng et al. (2024), Jiang et al. (2025)

Table 4: **Taxonomy of attacks in retrieval-augmented generation (RAG).** Attacks are categorized by objective (*accuracy*, *fairness*, *privacy*), targeting strategy (*targeted* vs. *untargeted*), and attack method.

Stages	*Categories*	*Methods*	*References*
Retrieval	Robust Retrieval	Diagnostic Resources	Fayyaz et al. (2025), Goel (2024)
		Evidence-Centric Retrieval	Chang et al. (2025b)
		Corpus Sanitization	Jia et al. (2025), Si et al. (2025), Dinu et al. (2025), Yu and Sato (2024), Mochizuki et al. (2026)
	Robust Training for Retriever	Hard Negative Training	Qu et al. (2021), Thakur et al. (2025)
		Perturbation Training	Tasawong et al. (2023)
Rerank	Graph-Based Reranking	Similarity Graph	Zheng et al. (2025a)
Generation	Proactive Defenses	Filtering	Zhou et al. (2025a), Kim et al. (2025a), Cheng et al. (2025), Shen et al. (2026), Choudhary et al. (2025), Kim et al. (2025b), Hwang et al. (2025), Shen et al. (2026)
		Detection	Tan et al. (2025), Zou et al. (2025), Abdelnabi et al. (2025), Hong et al. (2024), Shi et al. (2025a), Alon and Kamfonas (2023)
	Passive Defenses	Robust Reasoning and Generation	Xiang et al. (2026), Wei et al. (2025a), Wang et al. (2025c), Dai et al. (2026), Chen et al. (2025a), Xiong et al. (2026), Shi et al. (2025b), Yao et al. (2025), Chen et al. (2025c)
		Robust Training for Generator	Yoran et al. (2024), Fang et al. (2024), Tu et al. (2025), Yan et al. (2025), Huang et al. (2025), Wei et al. (2025b)
Traceback	Attribution	Document Level	Zhang et al. (2025a), Zhang et al. (2025b)
		Context & Token Level	Cohen-Wang et al. (2024), Wang et al. (2025g), Wang et al. (2026)
		Dataset Level	Jovanović et al. (2025)

Table 5: **Taxonomy of defenses for retrieval-augmented generation (RAG) robustness.** Defenses are organized by their position in the RAG pipeline.

Stages	*Techniques*	*Description*
Retrieval	Robust Retrieval	Improve retriever through test-time mechanisms (e.g., CoE retrieval, corpus sanitization).
	Robust Training	Train the retriever with adversarial or noisy data to improve robustness.
Rerank	Consistency Reranking	Reorder retrieved candidates to promote benign and relevant documents while pushing suspicious documents below the final top-k cutoff.
Generation	Filtering	Remove noisy, poisoned, or prompt-injected documents from the retrieved context before generation through consistency, influence, mechanistic and reliability-aware filtering.
	Detection	Identify and flag suspicious retrieved document or abnormal generation signals without necessarily removing them.
	Robust Reasoning	Improve reasoning through test-time mechanisms (e.g., alignment, layer guided attention).
	Robust Training	Train the generator with adversarial, irrelevant, noisy or counterfactual contexts to internalize defensive behavior.
Traceback	Traceback / Attribution	Trace the generated output back to the responsible document, passage, or data source after a harmful response is observed.

Table 6: **Defense taxonomy for RAG robustness.** We summarize defense techniques by stage, together with their descriptions.