

A Survey of Secure Retrieval-Augmented Generation

Yuming Xu^{1*}, Mingtao Zhang^{1*}, Zhuohan Ge^{1*}, Haoyang Li^{1†}, Nicole Hu¹,
Yongqi Zhang², Zhiyuan Wen¹, Jason Chen Zhang¹, Qing Li¹, Lei Chen²

¹The Hong Kong Polytechnic University

²The Hong Kong University of Science and Technology (Guangzhou)

{martin.xu, mingtao.zhang, zhuohan.ge, hulan.hu}@connect.polyu.hk

{haoyang-comp.li, zhiyuan.wen, jason-c.zhang}@polyu.edu.hk

yongqizhang@hkust-gz.edu.cn, csqli@comp.polyu.edu.hk, leichen@cse.ust.hk

Abstract

Retrieval-augmented generation (RAG) improves large language models (LLMs) with external knowledge, but this access path creates security risks distinct from inherent prompt-only or parametric-model flaws. We frame secure RAG as securing external knowledge access. We conducted a systematic search and curated 135 works on attacks, defenses, and security evaluation, and organized them with SLOT: a taxonomy along the attack Surface (S) and the corresponding defense Layer (L), with cross-cutting axes Objective (O) and attack-target scope (T). Mapping these studies onto an external knowledge-access pipeline, we expose three mismatches: target mismatch (T2 attacks outpace T2 defenses and evaluation), stage mismatch (S1 attacks outnumber L1 defenses), and signal mismatch (fluent, retrievable, corpus-fitting S1 attacks challenge anomaly-based L2/L3 defenses). Finally, we discuss directions for more realistic targets, surface-complete defense stacks, standardized evaluation, confidentiality, and multimodal and agentic systems, and release the screening process, selected-paper metadata, and machine-readable SLOT labels at <https://github.com/TreeAI-Lab/Awesome-RAG-Security>.

1 Introduction

Retrieval-augmented generation (RAG) has become a practical and widely adopted paradigm for improving large language models (LLMs) with external knowledge at inference time (Lewis et al., 2020; Gao et al., 2023; Wu et al., 2024; Gupta et al., 2024; Cheng et al., 2025a; Wang et al., 2026e). By retrieving evidence from external corpora, databases, or structured repositories and incorporating it into the generation context, RAG improves factuality, updatability, and domain adaptability (Gao et al., 2023; Wu et al., 2024; Yu et al., 2024a; Gan et al., 2025). At the same time, this

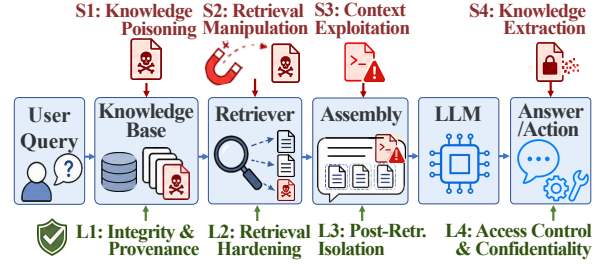


Figure 1: Attack surfaces (S1–S4) and defense layers that mirror them (L1–L4) along the RAG pipeline.

shift from parametric knowledge alone to external knowledge access introduces security risks, because the model is no longer affected only by its parameters and the user prompt, but also by the content and access path of external knowledge (Ward and Harguess, 2025; Arzanipour et al., 2025; Ammann et al., 2025).

Research on RAG security is growing rapidly, yet individual studies usually attack or defend one setting without a shared problem definition or taxonomy to relate these efforts. Recent surveys cover general architectures and benchmarks (Gao et al., 2023; Yu et al., 2024a; Gan et al., 2025), trustworthiness, adversarial robustness, and privacy (Zhou et al., 2024; Ni et al., 2025; Ammann et al., 2025; Ward and Harguess, 2025; Arzanipour et al., 2025; Bodea et al., 2026). Appendix Table 4 compares their scope. To our knowledge, no prior survey jointly analyzes RAG security through the external-knowledge-access pipeline, distinguishes risks introduced or amplified by RAG from risks inherent to LLMs, and links attacks, defenses, and benchmarks. We therefore present the first systematic, external-knowledge-centered survey dedicated to RAG security under one hierarchical taxonomy.

As shown in Figure 1, the central problem is that an attacker need not touch the model or the user-visible prompt. By manipulating external content or retrieval or delivering instructions through retrieved context, an attacker can steer generation or extract protected knowledge. Each point where

*Denotes equal contribution. † Corresponding author.

attackers can act mirrors a point where defenders can intervene, which motivates our *Surface*↔*Layer* mirroring as the backbone of the survey.

To keep this backbone focused, we define an *operational boundary*: a risk is in scope only if external knowledge carries the threat, creates an entry point absent in prompt-only LLM use, or materially amplifies its persistence, transferability, or blast radius. Generic prompt-only jailbreaks and purely parametric memorization are therefore excluded. Under this boundary, we organize attack-side work by the attacked surface (S), mirror each surface with the defense layer (L) that controls the same point, and tag every method with its attack-target scope (T) and security objective (O). We then discuss benchmark and evaluation studies on this attack-defense correspondence and distill insights and concrete future directions.

Concretely, our contributions are:

- **Scope and protocol.** We define secure RAG around risks introduced or amplified by external knowledge access, and use a systematic protocol covering May 1, 2020–April 30, 2026.
- **The SLOT taxonomy.** We organize RAG security by attack *Surface* (S) and corresponding defense *Layer* (L), then cross-tag security *Objective* (O) and attack-target scope (T).
- **A 135-study synthesis.** We jointly analyze 64 attack-primary, 56 defense-primary, and 15 evaluation-primary works, connecting attacks, defenses, and evaluation along the pipeline.
- **A quantitative diagnosis of mismatches and future directions.** We identify three mismatches: a target mismatch (31/64 T2 attacks versus 2/56 T2 defenses and 1/15 T2-centered evaluation studies), a stage mismatch (42 primary-S1 attacks versus 4 primary-L1 defenses), and a signal mismatch (fluent, retrievable, corpus-fitting S1 attacks versus L2/L3 defenses that often assume anomalous or separable poison). We then discuss directions for attacks, defenses, and benchmarks.
- **An open, machine-readable community resource.** We release the screening process, selected-paper metadata, and machine-readable SLOT labels for reuse in the [Awesome RAG Security repository](#), helping the community track new studies, align them under a shared taxonomy, and identify gaps and research opportunities.

2 RAG Pipeline and Operational Scope

2.1 External Knowledge Access Pipeline with Security Surfaces

As shown in Figure 2, we view RAG-related processes through the external-knowledge-access pipeline, which can be divided into six stages: external sources with raw content; ingestion and indexing that turn raw content into chunks and embeddings; retrieval and reranking select candidate evidence; context assembly formats the selected evidence into the model-visible prompt; generation produces the answer; and response and remediation deliver the final output. We refer to the indexed external corpus, database, memory, graph, or multi-modal store as the *knowledge substrate*, and to the inference-time selection and use of its evidence as *external knowledge access*. *Protected knowledge* is substrate content governed by a disclosure policy: it may support an answer but must not be revealed beyond what that policy permits. *Disclosure* is the release or reconstruction of protected knowledge beyond those limits.

The pipeline also identifies four attack surfaces: pre-retrieval knowledge poisoning (S1), retrieval-time access manipulation (S2), post-retrieval context exploitation (S3), and knowledge exfiltration and privacy attacks (S4). Defenses mirror them through knowledge-base integrity, provenance, and remediation (L1); retrieval-time access hardening (L2); post-retrieval isolation and robust generation (L3); and access control and confidentiality (L4).

2.2 Threat Model

In our discussion, an attack acts through or exploits the external-knowledge-access path. An attacker may influence ingestible or retrievable content, manipulate retrieval or its components, deliver instructions through retrieved context, or issue probing queries to infer protected knowledge. A compromised retriever or generator is included only when the threat fundamentally depends on external knowledge access. Attacks confined to the user prompt or model parameters without using this path are outside our scope. The surface (S1–S4) identifies where the attacker acts, and the mirrored layer (L1–L4) identifies where a defense intervenes.

The objective (O) is the security property that the attack seeks to violate, following the classical CIA triad (Saltzer and Schroeder, 1975). Integrity (O1) attacks push the system to produce, retrieve,

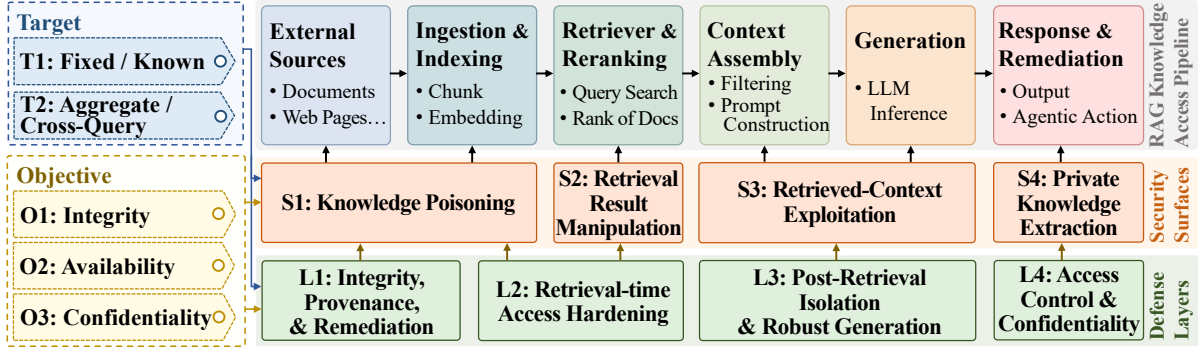


Figure 2: SLOT view of the RAG knowledge-access pipeline. Attack surfaces (S1–S4) and defense layers (L1–L4) are aligned with the pipeline stages, while Objective (O) and Target (T) are cross-cutting tags used to compare attacks, defenses, and benchmarks, i.e., SLOT Tag = Surface/Layer + Objective + Target.

or act on attacker-chosen content; availability (O2) attacks make it refuse or fail to answer legitimate queries; and confidentiality (O3) attacks infer or extract protected knowledge from the substrate. Any objective can escalate from an incorrect or leaked answer into an unsafe action.

The target axis (T) describes what the attacker seeks to affect and at what scope. T1 (*fixed/known target*) assumes that a finite query set is fixed and known to the attacker before the attack, while T2 (*aggregate/cross-query target*) seeks an aggregate effect over unknown future queries or adaptively explored records. For O1/O2, T1 changes answers or denies service to known queries, while T2 steers answers or causes persistent refusal across a future query family about an attacker-specified entity, topic, or claim. For O3, T1 infers or extracts a pre-specified record, while T2 maximizes recovery of initially unknown corpus content through repeated or adaptive queries. T2 therefore requires an aggregate view over a query distribution rather than a per-instance check.

Together, these dimensions define SLOT. A threat receives S/O/T tags, and a defense receives L/O/T tags. A defense or evaluation inherits T from the attack-target setting that it protects against or measures. Primary S/L and O mark the main intervention point and objective, while secondary labels record additional roles. Counts use only the primary paper type and S/L label, while works under a non-primary category are excluded and reported in gray. For example, *PoisonedRAG* (Zou et al., 2025) is S1/O1/T1 as it injects passages before retrieval to alter answers to known queries; *RobustRAG* (Xi-ang et al., 2024) is L3/O1/T1 because it processes retrieved passages and aggregates outputs under a fixed-query corruption bound. Table 3 in Appendix B gives examples for all S/L categories.

2.3 Operational Scope

We include a work when external knowledge carries the threat, creates an attack path absent from prompt-only LLM use, or materially amplifies the threat through persistence, transferability, or broader exposure. Thus, Greshake et al. (Greshake et al., 2023) and BIPIA (Yi et al., 2025) are included because retrieved external content carries indirect instructions, while JailbreakBench (Chao et al., 2024) is excluded because it evaluates prompt-only attacks, and RAGBench (Friel et al., 2024) is excluded because it evaluates ordinary RAG quality. We also exclude purely parametric memorization, generic LLM safety or privacy, and work that only uses RAG for unrelated applications.

Relation to general LLM security. General LLM defenses help at the stages they protect. Prompt-injection detection or model alignment can reduce harmful behavior after retrieved content reaches the generator, but neither by itself prevents corpus poisoning, retrieval manipulation, or private-knowledge extraction. Secure RAG therefore also needs controls on the external-knowledge path, such as provenance and admission checks, retrieval hardening, and access control.

2.4 Literature Search Protocol

Following the operational scope above, we conducted a structured search of five scholarly sources.

Sources. We searched arXiv, ACL Anthology, OpenReview, Semantic Scholar, and DBLP.

Date range. We retained works first made public from May 1, 2020 to April 30, 2026.

Query rule. We required each record to contain at least one *RAG* term and one *security* term in its title or abstract. Terms are reported in Appendix A.

Screening and selection. Source retrieval: 17,256 records. Abstract-available deduplication:

8,262. Title/abstract screening with DeepSeek-V4-Flash-0731: 636. Version merging: 617. Full-text screening with the same model and scope rule: 552 eligible works. Expert selection: 132. Snowballing: +3, yielding 135 studies. Appendix A reports the prompt templates and detailed screening rules, with the full funnel in Table 2; the [repository](#) releases the screening process, selected-paper metadata with public paper and code links, and machine-readable SLOT labels.

3 Attack Mechanisms

Following Section 2.2, we organize attacks by their first intervention surface (S1–S4), objective (O1–O3), and target scope (T1–T2). Methods with the same primary mechanism are grouped below; Appendix E and Figure 3 give the full taxonomy.

3.1 Knowledge Poisoning (S1)

Knowledge poisoning inserts malicious corpus, file, graph, or memory content before retrieval, allowing the payload to persist across queries or sessions. Most methods target integrity (O1), while some also induce refusal or failure (O2).

Fixed-query text attacks share a known query or answer target but differ in how they construct the poison. *PoisonedRAG* (Zou et al., 2025) makes injected passages both retrievable and answer-directing, while *GARAG* (Cho et al., 2024) and *Retrieval Poisoning* (Zhang et al., 2024a) emphasize natural, highly retrievable text. *The Silent Saboteur* (Song et al., 2025) edits relevant documents, while *CorruptRAG* (Zhang et al., 2026a) and *AuthChain* (Chang et al., 2025) embed coherent target evidence. *HijackRAG* (Zhang et al., 2024b) and *Joint-GCG* (Wang et al., 2026b) couple retrieval with downstream generation through prompt hijacking or joint optimization. *RIPRAG* (Xi et al., 2026), *The RAG Paradox* (Choi et al., 2025a), and *Token-Level Attack* (Li et al., 2026) use reinforcement learning, live feedback, or token substitutions.

Trigger-based and generalized attacks vary in how broadly the payload transfers. Trigger-activated methods make poisoned evidence rank highly when an attacker-defined signal matches. *BadRAG* (Xue et al., 2024) and *AIP* (Chaturvedi et al., 2025) associate natural semantic triggers with poisoned passages or reusable instructional prompts, *Backdoored Retrievers* (Clop and Teglia, 2024) implants the association during retriever fine-tuning, and *PR-Attack* (Jiao et al., 2025)

jointly optimizes a soft-prompt trigger and poison texts. *AGGD* (Su et al., 2025) uses no explicit inference-time trigger; it instead searches for a universal adversarial passage that acts as an always-on retrieval lure across unseen queries and domains. Distribution-oriented methods likewise avoid assuming the exact future query, but optimize poison against a sampled query distribution: *GASLITE* (Ben-Tov and Sharif, 2025) uses embedding gradients to rank attacker content for unseen concept-related queries, *UniC-RAG* (Geng et al., 2025) clusters diverse target queries and jointly optimizes a small set of texts for retrieval and generation, and *MIRAGE* (Chen et al., 2026b) approximates latent queries through persona-driven synthesis and embeds semantic anchors under black-box feedback. *LIAR* (*Glue Pizza and Eat Rocks*) (Tan et al., 2024) and *Influential Tokens* (Wang et al., 2025a) couple retrievability with misleading or refusal behavior, whereas *EmoRAG* (Zhou et al., 2026), *FlippedRAG* (Chen et al., 2025b), and *Topic-FlipRAG* (Gong et al., 2025) steer affect or stance. Payload-bearing variants use retrieved content downstream: *PANDORA* (Deng et al., 2024) induces jailbreaks; *Phantom* (Chaudhari et al., 2026) and *ConfusedPilot* (RoyChowdhury et al., 2024) carry malicious commands or enterprise instructions; and *Blocker* (Shafran et al., 2025) and *MutedRAG* (Suo et al., 2025) induce refusal. *PIDP-Attack* (Wang et al., 2026c) instead combines poisoned passages with a query suffix (secondary S2).

Other S1 methods adapt the same persistence mechanism to different substrates. In agent memory, *AgentPoison* (Chen et al., 2024), *Memory-Graft* (Srivastava and He, 2025), and *MINJA* (Dong et al., 2025) plant records that affect later retrieval, planning, or actions. *PhantomText Toolkit* (Castagnaro et al., 2025) instead hides payloads in the data-loading path. In graph RAG, *GRAGPoison* (Liang et al., 2026), *TKPA/UKPA* (Wen et al., 2025a), and *LogicPoison* (Xiao et al., 2026) manipulate graph evidence or reasoning, with UKPA additionally inducing broad O2 degradation. *MM-PoisonRAG* (Ha et al., 2026), *Poisoned-MRAG* (Liu et al., 2025b), and *One Pic* (Shereen et al., 2026) poison multi-modal evidence, while *RACG* (Lin et al., 2026) and *ImportSnare* (Ye et al., 2025b) steer code generation through poisoned examples, documentation, or dependencies. The move toward fluent, retrievable, cross-query payloads creates the signal mismatch discussed in Sections 4 and 6.

3.2 Retrieval Result Manipulation (S2)

S2 attacks alter the query, ranking, or retriever so attacker-preferred evidence reaches the generator. Query-side methods modify the retrieval request, including *GGPP* (Hu et al., 2024b) that appends optimized tokens to retrieve false documents, *Medusa* (Shang et al., 2026) that perturbs visual queries, and *RAG-Pull* (Dhar et al., 2026) that shifts query embeddings toward attacker-chosen code gadgets across future queries. Retriever-side methods instead implant persistent, trigger-conditioned behavior. *TrojanRAG* (Cheng et al., 2024) binds triggers to a compromised retriever and poisoned context, while *Your RAG is Unfair* (Bagwe et al., 2025) backdoors a reusable encoder to induce demographic bias. GGPP is attack-primary, and its detector is a secondary defense contribution.

3.3 Retrieved-Context Exploitation (S3)

S3 attacks use model-visible external content to carry instructions, jailbreaks, or refusal triggers after retrieval. *Indirect Injection* (Greshake et al., 2023) establishes this attack path, while *BIPIA* (Yi et al., 2025) systematizes its tasks, goals, injection positions, and defenses. BIPIA is evaluation-primary and included as a cross-list. Among S1-primary attacks, *PANDORA* (Deng et al., 2024) uses retrieved harmful material for jailbreaks, *Phantom* (Chaudhari et al., 2026) and *ConfusedPilot* (RoyChowdhury et al., 2024) embed commands or enterprise instructions, while *Blocker* (Shafran et al., 2025) and *MutedRAG* (Suo et al., 2025) induce refusal. They are S1-primary with corpus insertion as their first intervention.

3.4 Private Knowledge Extraction (S4)

S4 attacks infer or recover protected substrate content (O3). First, membership methods test a fixed candidate (T1). *MIA-RAG* (Anderson et al., 2025) uses crafted prompts; *S²MIA* (Li et al., 2025b) compares target samples with generated text; *MBA* (Liu et al., 2025a) masks unrelated-document interference; while *IA* (Naseh et al., 2025) asks questions whose answers depend on document presence. *S-RAG* (Zeng et al., 2025d) and *SMA* (Sun et al., 2025) adapt membership inference for privacy auditing and source attribution.

Targeted extraction recovers selected records or modalities. *The Good and the Bad* (Zeng et al., 2024) studies retrieval-induced leakage and memorization trade-offs; *Beyond Text* (Zhang et al.,

2025c) extends extraction to vision- and speech-language RAG; *mRAG Privacy* (Al-Lawati and Wang, 2026) evaluates image membership and caption recovery; and *Backdoor Extraction* (Peng et al., 2024) leaks retrieved references through a compromised generator. These works use pre-specified targets or evaluation cases and therefore are T1.

Aggregate extraction instead maximizes recovery of initially unknown content. *Spill the Beans* (Qi et al., 2025) induces datastore regurgitation, *IKEA* (Wang et al., 2025c) extracts implicit knowledge without explicit jailbreaks, *SE-CRET* (He et al., 2026) composes extraction instructions, jailbreak operators, and retrieval triggers, and *RAGCRAWLER* (Yao et al., 2026) adaptively follows corpus relations. *Exposing Privacy Risks in GraphRAG* (Liu et al., 2026) and *Agentic Graph Extraction* (Yang et al., 2026) extend aggregate recovery to graph neighborhoods. These attack methods are T2. The defense-primary *Retrieval Pivot* (Thornton, 2026) is included only for its secondary S4 contribution.

3.5 Summary and Observations

Across 64 attack-primary papers, the primary-surface distribution is S1: 42, S2: 5, S3: 1, and S4: 16. O1 dominates S1–S3, O2 centers on refusal or degradation, and S4 captures O3. T2 accounts for 31 attacks, spanning generalized poisoning and aggregate extraction over unknown future queries or initially unknown records; the other 33 are T1.

4 Defenses and Remediation Mechanisms

We organize defenses by their primary control point along the knowledge-access pipeline: knowledge-base integrity and remediation (L1), retrieval-time hardening (L2), post-retrieval isolation and robust generation (L3), and access control and confidentiality (L4). Methods with the same control mechanism are grouped. Appendix E and Figure 4 give the full taxonomy.

4.1 Knowledge-Base Integrity, Provenance, and Remediation

L1 either protects the substrate before ingestion or attributes failures to stored evidence. For admission and provenance, *D-RAG* (Andersen et al., 2025) decentralizes provenance and access control, while *Proof-Carrying Answers* (Shukla and Joshi, 2025) binds cryptographic provenance to the evidence-to-answer path. For remediation, *RAG-Forensics* (Zhang et al., 2025a) traces malicious

outputs to responsible passages, and *Who Taught the Lie?* (Zhang et al., 2026b) scores responsibility from retrieval and response signals. With only these four L1-primary works, upstream protection remains much thinner than S1 attack coverage.

4.2 Retrieval-Time Access Hardening

L2 secures retrieval before external content becomes model-visible context. Reliability methods reduce dependence on any single passage. For example, *ReliabilityRAG* (Shen et al., 2025) aggregates a consistent majority, and *RA-RAG* (Hwang et al., 2025) weights source credibility. In addition, *TrustRAG* (Zhou et al., 2025a) and *GRADA* (Zheng et al., 2025) identify weak support through clustering or graph connectivity, and *RAGPart/RAGMask* (Pathmanathan et al., 2025) test sensitivity through partitioning and masking.

Candidate-cleaning methods try to separate poison before generation: *RAGuard-Defense* (Cheng et al., 2025c)¹ filters expanded results using perplexity and similarity, while *SeCon-RAG* (Si et al., 2025) and *RAGDefender* (Kim et al., 2025a) combine semantic, cluster, conflict, or purification signals. Detection methods use different evidence: *GMTP* (Kim et al., 2025b) uses masked-token probabilities, while *EcoSafeRAG* (Yao et al., 2025) and *Knowledge Checking* (Zeng et al., 2025c) compare contextual or representation-level signals. By contrast, *FRAG* (Zhao, 2024) differs in objective, protecting federated vector retrieval against O3 leakage. The attack-primary *GGPP* (Hu et al., 2024b) is cross-listed for its secondary detector.

L2 is substantial, but aggregation often assumes a bounded malicious minority and filtering assumes that poison is anomalous, weakly supported, or separable. Fluent, retrievable, corpus-fitting payloads increasingly weaken these assumptions.

4.3 Post-Retrieval Isolation and Robust Generation

L3 acts after harmful content enters the retrieved context. Detection methods inspect the passages or model behavior: *RevPRAG* (Tan et al., 2025) uses activations, *AV Filter* (Choudhary et al., 2026) uses attention variance, and *FilterRAG/ML-FilterRAG* (Edemacu et al., 2025) learn corpus-level or statistical filters. *RAGShield* (Patil, 2026) checks numerical claims across sources and time,

while *InstructDetector* (Wen et al., 2025b) identifies embedded instructions. *RobustRAG* (Xiang et al., 2024) instead isolates passages and certifies the answer under bounded corruption.

Another way is to constrain interaction among retrieved documents. Isolation methods include *SDAG* (Dekel et al., 2026), which uses sparse document attention, and *MMA-RAGT* (Singh et al., 2026), which extends stateful trust isolation to multimodal agentic RAG. Robust-generation methods discount misleading evidence in different ways: *Discern-and-Answer* (Hong et al., 2024) removes it, *InstructRAG* (Wei et al., 2025) synthesizes denoising rationales, *Astute RAG* (Wang et al., 2025b) reconciles internal and retrieved knowledge, and *RbFT* (Tu et al., 2025) fine-tunes against misleading or counterfactual context. In addition, *ATM* (Zhu et al., 2024) and *INFO-RAG* (Xu et al., 2024) train adversarial or unsupervised refinement, while *Credibility-Aware Generation* (Pan et al., 2024), *Fact-Centric Alignment* (Hao et al., 2025), and *RPO* (Yan et al., 2025) align generation at the source, fact, or preference level. The evaluation-primary *ABDD* study (*LLMs Trust Humans More*) (Li et al., 2025a) is cross-listed for its CDEQ mitigation.

A parallel L3 line controls disclosure after retrieval. *SD-RAG* (Masoud et al., 2026) sanitizes context, *PAD* (Wang et al., 2026a) protects decoding, and *PPRAG* (Zhang et al., 2026c) combines privacy-aware retrieval with context transformation, while *SEAL-Tag* (Xie et al., 2026) aggregates evidence tags across query sequences to limit linkability and mosaic disclosure, giving T2 coverage. L3 is broad, but many methods require model access, extra inference, or detectable abnormal signals.

4.4 Access Control, Privacy, and Confidentiality

L4 controls who can access retrieved knowledge and what can be inferred from it. Authorization and membership defenses include *AC-RAG* (Chen et al., 2025a), which integrates fine-grained authorization, and *Retrieval Pivot* (Thornton, 2026), which enforces authorization at each graph-traversal hop; *Retrieval Pivot* is defense-primary, with its S4 leakage audit as a secondary role. In addition, *Mirabel* (Choi et al., 2025b) and *Ensemble Privacy Defense* (Fu et al., 2026) suppress membership signals.

Privacy mechanisms limit disclosure from text, retrieval, or generation. *DPVoteRAG* (Koga

¹Two unrelated works are both named *RAGuard*; we use *RAGuard-Defense* and *RAGuard-Bench* for disambiguation.

et al., 2024), *RAG with Differential Privacy* (Grislain, 2025), *LPRAG* (He et al., 2025), *VAGUE-Gate* (Hemmat et al., 2025), and *InvisibleInk* (Vinod et al., 2025) apply differential privacy at token, entity, retrieval, or generation level. *SAGE* (Zeng et al., 2025b) replaces private corpora with synthetic alternatives, *DP-SynRAG* (Mori et al., 2026) creates a reusable differentially private corpus under a one-time budget (T2), and *AURA* (Wang et al., 2026d) adulterates proprietary graphs so extracted copies lose utility.

Backend methods protect retrieval through architecture or cryptography. For instance, *RemoteRAG* (Cheng et al., 2025b) secures cloud retrieval and *ppRAG* (Ye et al., 2025a) uses distance-preserving encryption. Additionally, *FedE4RAG* (Mao et al., 2025), *C-FedRAG* (Addison et al., 2024), and *Privacy-Aware RAG* (Zhou et al., 2025b) use federated or isolated retrieval, while *SAG* (Zhou et al., 2025c) adds formal security and *SecureRAG* (Bassit and Boddeti, 2025) provides an end-to-end protected architecture. These controls commonly trade utility, latency, or retrieval quality for protection.

4.5 Summary and Observations

Among 56 defense-primary papers, L1: 4, L2: 12, L3: 21, and L4: 19. The stage mismatch is largest upstream: 42 attacks begin at S1, but only four defenses begin at L1. The target mismatch is also sharp: only DP-SynRAG and SEAL-Tag are T2 (2/56), while the other 54 defenses are T1. Finally, the signal mismatch persists because fluent, retrievable S1 payloads weaken the anomaly and separability assumptions common in L2/L3. Section 6 will further revisit these three mismatches.

5 Secure-RAG Evaluation Studies

Within the scope of Section 2.3, *benchmark studies* provide reusable resources, whereas *systematic evaluation studies* characterize secure-RAG failures. Table 1 maps 15 evaluation-primary works and attack/defense cross-lists (excluded from the count) to SLOT.

5.1 Benchmark Studies

Integrity (O1). The largest cluster tests whether adversarial or misleading inputs compromise answer integrity. *Rag ’n Roll* (De Stefano et al., 2024) evaluates indirect prompt manipulation in end-to-end applications rather than isolated components.

Paper	Objective			Target		Surface			
	O1	O2	O3	T1	T2	S1	S2	S3	S4
<i>Benchmark Studies</i>									
<i>Rag ’n Roll</i> (2024)	✓	–	–	✓	–	✓	✓	✓	–
<i>BIPIA</i> (2025)	✓	◦	–	✓	–	–	–	✓	–
<i>Generative-Search Robustness</i> (2024a)	✓	–	–	✓	–	–	–	–	–
<i>ReEval</i> (2024b)	✓	–	–	✓	–	–	–	✓	–
<i>SafeRAG</i> (2025)	✓	✓	–	✓	–	✓	✓	✓	–
<i>RSB</i> (2025b)	✓	✓	–	✓	–	✓	◦	◦	–
<i>PoisonArena</i> (2026a)	✓	–	–	✓	◦	✓	✓	–	–
<i>RAGuard-Bench</i> (2025a)	✓	–	–	✓	–	◦	–	–	–
<i>OpenRAG-Soc</i> (2026)	✓	–	–	✓	–	✓	◦	✓	–
<i>MPIB</i> (2026)	✓	–	–	✓	–	–	–	✓	–
<i>S-RAG</i> (2025d) ^{primary:A}	–	–	✓	✓	–	–	–	–	✓
<i>SMA</i> (2025) ^{primary:A}	–	–	✓	✓	–	–	–	–	✓
<i>Privacy Protection in RAG</i> (2026c) ^{primary:D}	–	–	✓	✓	–	–	–	–	✓
<i>KE-Bench</i> (2026)	–	–	✓	–	✓	–	◦	–	✓
<i>MedPriv-Bench</i> (2026)	–	–	✓	✓	–	–	–	–	✓
<i>SEAL-Tag</i> (2026) ^{primary:D}	◦	–	✓	◦	✓	–	–	–	✓
<i>Systematic Evaluation Studies</i>									
<i>Instruction-Following Retrievers</i> (2025)	✓	–	–	✓	–	–	✓	–	–
<i>RAG LLMs are Not Safer</i> (2025)	✓	–	–	✓	–	–	◦	–	–
<i>ABDD</i> (2025a)	✓	–	–	✓	–	–	–	–	–
<i>Good&Bad</i> (2024) ^{primary:A}	–	–	✓	✓	–	–	◦	◦	✓
<i>Beyond Text</i> (2025c) ^{primary:A}	–	–	✓	✓	–	–	◦	◦	✓
<i>mRAG Privacy</i> (2026) ^{primary:A}	–	–	✓	✓	–	–	–	–	✓

Table 1: Coverage map of secure-RAG evaluation studies. Gray cross-lists show their primary class (A: attack; D: defense); unmarked works are evaluation-primary. **Objective:** O1 integrity, O2 availability, O3 confidentiality; **Target:** T1 fixed/known, T2 aggregate/cross-query; **Surface:** S1 corpus/indexing, S2 retrieval, S3 context/prompt, S4 extraction. ✓/◦/– denote central, partial, and not primary coverage, respectively.

Generative-Search Robustness (Hu et al., 2024a) black-box tests generative search systems with adversarial factoid questions, while *ReEval* (Yu et al., 2024b) replaces answers in coherent evidence and adds relevant distractors for transferable stress tests. *PoisonArena* (Chen et al., 2026a) models multi-adversary competition, compares defenses, and provides partial T2 coverage through transfer to held-out equivalent queries. *RAGuard-Bench* (Zeng et al., 2025a) uses naturally occurring web misinformation rather than synthetic poison; *OpenRAG-Soc* (Guo and Wei, 2026) jointly tests injection, poisoning, and defenses on web/social corpora; and *MPIB* (Lee et al., 2026) compares indirect-injection harms and defenses across tasks.

Integrity and availability jointly (O1, O2). Avail-

ability rarely appears as a first-class target; it is mostly covered as a single task within broader poisoning suites. *BIPIA* (Yi et al., 2025) benchmarks indirect prompt injection across application tasks, attack goals, and injection positions, and evaluates boundary-aware and explicit-reminder defenses in black- and white-box settings. *SafeRAG* (Liang et al., 2025) injects Noise, Conflict, Toxicity, and DoS at each pipeline stage and exposes 14 component-level evaluations. *RAG Security Bench (RSB)* (Zhang et al., 2025b) evaluates 13 poisoning types, spanning targeted, DoS, and trigger-DoS attacks, against defenses across sequential, branching, conditional, looping, multi-turn, multimodal, and agentic RAG.

Confidentiality (O3). A growing line evaluates leakage from the retrieval corpus, a property specific to RAG and distinct from generic prompt leakage. *S-RAG* (Zeng et al., 2025d) audits, under black-box access, whether a user’s personal data has been ingested into the index. *SMA* (Sun et al., 2025) introduces source-aware membership auditing to semi-black-box and multimodal retrieval. *Privacy Protection in RAG* (Zhang et al., 2026c) introduces a knowledge-graph-based, fine-grained privacy processor and evaluates extraction and personal identification alongside answer utility. *KE-Bench* (Qi et al., 2026) standardizes knowledge-extraction evaluation across retrievers, generators, and datasets, with RandText, DGEA, CopyBreak, and IKEA baselines for aggregate recovery of initially unknown datastore content (T2). *MedPriv-Bench* (Guan et al., 2026) jointly evaluates medical contextual leakage and clinical utility across three privacy-preserving methods. *SEAL-Tag* (Xie et al., 2026) combines a PII-Evidence Table with probabilistic-circuit routing and evaluates direct, linkability, adaptive-extraction, and PET-Spoofing attacks alongside utility and latency.

5.2 Systematic Evaluation Studies

Beyond benchmark construction, systematic studies clarify what secure-RAG should measure. On integrity, *Instruction-Following Retrievers* (BehnamGhader et al., 2025) measures how malicious instructions exploit instruction-following retrievers at S2 to surface attacker-preferred information, while *RAG LLMs are Not Safer* (An et al., 2025) compares RAG and non-RAG safety, showing that retrieved context can induce unsafe outputs even when the model and documents are individually safe, and evaluates retriever-aware red teaming.

ABDD (Li et al., 2025a) introduces the Authority Bias Detection Dataset to measure bias under conflicting user-provided and retrieved evidence, and proposes the Conflict Detection Enhanced Query (CDEQ) mitigation. It is evaluation-primary because measurement is central, with CDEQ as a secondary defense contribution.

On confidentiality, *Good&Bad* (Zeng et al., 2024) shows that retrieval introduces a corpus-leakage channel but can mitigate certain parametric-memorization risks. *Beyond Text* (Zhang et al., 2025c) demonstrates private-information extraction across vision- and speech-language RAG. *mRAG Privacy* (Al-Lawati and Wang, 2026) evaluates image membership inference and image-caption retrieval attacks in vision-centric mRAG. Together, they show where leakage appears and how it changes across modalities and assumptions.

5.3 Limitations of Current Evaluation

Coverage is uneven along the axes of our threat model, and the gaps below cluster around the insights we revisit in Section 6.

Target axis: T1 dominance. Of the 15 evaluation-primary works, KE-Bench is the only study centered on T2; PoisonArena provides partial T2 coverage through held-out query variants. The other 14 use fixed instances or query sets as their primary setting. T2 evaluation must aggregate stance shift, refusal or degradation, or recovered content over a query distribution or interaction sequence.

Surface axis: multi-surface coverage without a unified harness. *SafeRAG* (Liang et al., 2025), *OpenRAG-Soc* (Guo and Wei, 2026), and *Rag ’n Roll* (De Stefano et al., 2024) already exercise multiple attack surfaces or pipeline stages. There is no unified, defense-aware, pluggable cross-surface harness that exposes common interfaces for S1–S4 and L1–L4, leaving cross-surface interactions and failure propagation difficult to evaluate.

Adaptive evaluation is rarely reported. Few defenses (notably RobustRAG (Xiang et al., 2024) and RAGShield (Patil, 2026)) are evaluated against attackers aware of the defense. Across studies, headline attack success rate (ASR), distribution-level stance shift, and certified accuracy are not directly rankable because they evaluate different targets under different attacker budgets and assumptions. Appendix D details these conditions.

No standardized defense harness. The two defense-primary works cross-listed here (Privacy Protection in RAG (Zhang et al., 2026c), SEAL-

Tag (Xie et al., 2026)) center on the defenses they propose, while benchmarks that compare multiple defenses (BIPIA (Yi et al., 2025), RSB (Zhang et al., 2025b), PoisonArena (Chen et al., 2026a), OpenRAG-Soc (Guo and Wei, 2026), MPIB (Lee et al., 2026), KE-Bench (Qi et al., 2026), and MedPriv-Bench (Guan et al., 2026)) evaluate them under benchmark-specific threat protocols. No shared harness exists for comparing arbitrary defenses against a standardized attack suite, leaving each defense paper to set its own conditions and making cross-defense comparison unreliable. Our data release contains the screening process, selected-paper metadata, and machine-readable SLOT labels; Appendix D provides a protocol specification, not executable evaluation software.

Confidentiality metrics are incommensurable across studies and modalities. The O3 cluster now spans reusable privacy audits, leakage benchmarks, and attack-defense evaluations, but each study reports leakage in incompatible units (membership AUC, exact-match extraction, PII recall, etc.), preventing cross-study severity comparison. The gap widens across modalities, where no shared interface-level leakage rate exists.

6 Insights and Future Directions

The SLOT map exposes three mismatches and motivates six directions: close the target, stage, and signal gaps (I1–I3); build surface-complete defenses (I4); strengthen confidentiality controls (I5); and evaluate multimodal and agentic RAG (I6).

I1: Study more realistic targets. T2 accounts for 31/64 attack-primary works, but only 2/56 defense-primary works and 1/15 evaluation-primary studies. Therefore, defenses and evaluations should audit reliability across a query distribution or interaction sequence (T2), aggregating stance shift (O1), refusal or degradation (O2), or recovered content and query cost (O3).

I2: Close the stage mismatch. Although 42 attacks begin at S1, only 4 defenses begin at L1. Thus, corpus admission, provenance, monitoring, and remediation require more attention.

I3: Close the signal mismatch. Recent S1 attacks produce fluent, retrievable, corpus-fitting payloads, while many L2/L3 defenses still look for semantic anomalies. Defenses should therefore build on provenance, cross-query behavior, or causal influence rather than surface abnormality.

I4: Build surface-complete defenses. A surface-complete stack places an explicit control against every S1–S4 surface and tests each control against an adaptive attacker who knows all non-secret defense details. A shared harness should fix surface-specific budgets and report a compact SLOT card covering S/L/O/T, attacker access, victim setup, utility, latency, and uncertainty; Appendix D provides the template.

I5: Strengthen confidentiality controls. Most confidentiality defenses remain T1 and rely on differential privacy, encryption, or architectural change. T2 evaluations should additionally report recovered corpus fraction and query cost alongside policy violations, utility, and latency.

I6: Evaluate multimodal and agentic RAG. Multimodal leakage and complex poisoning already span vision, speech, multi-turn, and agentic RAG (Zhang et al., 2025c; Al-Lawati and Wang, 2026; Zhang et al., 2025b). *AgentPoison*, *Memory-Graft*, and *MINJA* further show that poisoned memory can affect later queries or actions across sessions (Chen et al., 2024; Srivastava and He, 2025; Dong et al., 2025). Evaluation should therefore test persistence across modalities, turns, and sessions, including whether a payload triggers an action.

7 Conclusion

We surveyed secure RAG within an external-knowledge boundary that separates RAG-introduced or RAG-amplified risks from inherent LLM flaws. We conducted a systematic literature search and curated 135 works. We release the screening process, selected-paper metadata, and machine-readable SLOT labels. The synthesis clarifies which failures fall within secure RAG and reveals three mismatches: a target mismatch (31/64 T2 attacks versus 2/56 T2 defenses and 1/15 T2-centered evaluation studies), a stage mismatch (42 primary-S1 attacks versus 4 primary-L1 defenses), and a signal mismatch (fluent, retrievable, corpus-fitting S1 attacks versus L2/L3 defenses that often assume anomalous or separable poison). We discussed surface-complete defense stacks, comparable evaluation, practical confidentiality controls, and persistence-aware multimodal and agentic testing as directions for closing these gaps.

Limitations

First, secure RAG is a fast-moving area, and our review is a snapshot through April 30, 2026. Later work may change the observed distribution. Second, LLM-assisted screening and expert selection can introduce errors, while English-language and full-text availability constraints may underrepresent work from some communities. Finally, our tables and observations provide a qualitative synthesis rather than a precise meta-analysis, and should be read as highlighting structural patterns rather than exact field-level measurements.

Ethical Considerations

This work is a literature survey and introduces no new attacks, defenses, or experiments. All included material is previously published and cited to the original sources. This work introduces no new exploit code or undisclosed vulnerabilities. Although machine assistance supported screening, experts made the final inclusion and SLOT-tagged decisions, and venue prestige was not a selection criterion. We recognize the dual-use risk of organizing attack methods in one place; we mitigate it by emphasizing threat assumptions, crossed boundaries, and corresponding controls. The released screening process, selected-paper metadata, and machine-readable SLOT labels are intended to support reproducible analysis and stronger protection research.

Acknowledgments

This work was partially supported by the National Key R&D Program of China under Grant No. 2023YFF0725100; NSFC under Grant No. U22B2060; Guangdong Basic and Applied Basic Research Foundation under Grant No. 2026A1515010227; PolyU TDG under Grant No. TDG25-28/TBP/11-IICA; PolyU Internal Research Fund (No. BDZ3); PolyU LTC Project (Grant TDLEG25-28/IICA/P/05, No. 48EM); NSFC/RGC Joint Research Scheme under Grant No. N_PolyU5179/25; Hong Kong RGC under Grant No. PolyU25600624; ITF under Grant Nos. ITS/052/23MX and PRP/009/22FX; industrial sponsors; and RMGS under Grant Nos. P0045948, P0048183, P0048191, P0046453, and P0060272.

References

- Parker Addison, Minh-Tuan H Nguyen, Tomislav Medan, Jinali Shah, Mohammad T Manzari, Brendan McElrone, Laksh Lalwani, Aboli More, Smita Sharma, Holger R Roth, et al. 2024. C-fedrag: A confidential federated retrieval-augmented generation system. *arXiv preprint arXiv:2412.13163*.
- Ali Al-Lawati and Suhan Wang. 2026. [Do multimodal rag systems leak data? a comprehensive evaluation of membership inference and image caption retrieval attacks](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 9139–9154.
- Lukas Ammann, Sara Ott, Christoph R Landolt, and Marco P Lehmann. 2025. Securing rag: A risk assessment and mitigation framework. In *2025 IEEE Swiss Conference on Data Science (SDS)*, pages 127–134. IEEE.
- Bang An, Shiyue Zhang, and Mark Dredze. 2025. Rag llms are not safer: A safety analysis of retrieval-augmented generation for large language models. 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 5444–5474.
- Tessa E Andersen, Ayanna Marie Avalos, Gaby G Dagher, Min Long, et al. 2025. D-rag: A privacy-preserving framework for decentralized rag using blockchain. In *CS & IT Conference Proceedings*, volume 15. CS & IT Conference Proceedings.
- Maya Anderson, Guy Amit, and Abigail Goldsteen. 2025. Is my data in your retrieval database? membership inference attacks against retrieval augmented generation. In *International Conference on Information Systems Security and Privacy*, volume 2, pages 474–485. Science and Technology Publications, Lda.
- Atousa Arzanipour, Rouzbeh Behnia, Reza Ebrahimi, and Kaushik Dutta. 2025. Rag security and privacy: Formalizing the threat model and attack surface. *arXiv preprint arXiv:2509.20324*.
- 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.
- Amina Bassit and Vishnu Boddeti. 2025. [SecureRAG: End-to-end secure retrieval-augmented generation](#). In *The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance*.
- Parishad BehnamGhader, Nicholas Meade, and Siva Reddy. 2025. Exploiting instruction-following retrievers for malicious information retrieval. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12962–12980.

- Matan Ben-Tov and Mahmood Sharif. 2025. [Gaslighting the retrieval: Exploring vulnerabilities in dense embedding-based search](#). In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*.
- Andreea-Elena Bodea, Stephen Meisenbacher, Alexandra Klymenko, and Florian Matthes. 2026. Sok: Privacy risks and mitigations in retrieval-augmented generation systems. *arXiv preprint arXiv:2601.03979*.
- Alberto Castagnaro, Umberto Salviati, Mauro Conti, Luca Pajola, and Simeone Pizzi. 2025. The hidden threat in plain text: Attacking rag data loaders. In *Proceedings of the 18th ACM Workshop on Artificial Intelligence and Security*, pages 170–181.
- Zhiyuan Chang, Mingyang Li, Xiaojun Jia, Junjie Wang, Yuekai Huang, Ziyou Jiang, Yang Liu, and Qing Wang. 2025. [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.
- Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwal, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. 2024. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems*.
- 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 Trans. AI Secur. Priv.* Just Accepted.
- Bingxiang Chen, John Tackman, Manu Setälä, Timo Pöranen, and Zheyang Zhang. 2025a. [Integrating access control with retrieval-augmented generation: A proof of concept for managing sensitive patient profiles](#). In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, SAC ’25, page 915–919, New York, NY, USA. Association for Computing Machinery.
- Liujia Chen, Xiaofang Yang, Yuanzhuo Lu, Jinghao Zhang, Xin Sun, Qiang Liu, Shu Wu, Jing Dong, and Liang Wang. 2026a. Uncovering competing poisoning attacks in retrieval-augmented generation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 414–424.
- Tailun Chen, Yu He, Yan Wang, Shuo Shao, Haolun Zheng, Zhihao Liu, Jinfeng Li, Zhizhen Qin, Yuefeng Chen, Zhixuan Chu, Zhan Qin, and Kui Ren. 2026b. [Mirage: Misleading retrieval-augmented generation via black-box and query-agnostic poisoning attacks](#). *Preprint*, arXiv:2512.08289.
- Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. 2024. [Agentpoison: Red-teaming llm agents via poisoning memory or knowledge bases](#). In *Advances in Neural Information Processing Systems*, volume 37.
- Zhuo Chen, Yuyang Gong, Jiawei Liu, Miaokun Chen, Haotian Liu, Qikai Cheng, Fan Zhang, Wei Lu, and Xiaozhong Liu. 2025b. [Flippedrag: Black-box opinion manipulation adversarial attacks to retrieval-augmented generation models](#). In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, CCS ’25, page 4109–4123. ACM.
- Mingyue Cheng, Yucong Luo, Jie Ouyang, Qi Liu, Huijie Liu, Li Li, Shuo Yu, Bohou Zhang, Jiawei Cao, Jie Ma, et al. 2025a. A survey on knowledge-oriented retrieval-augmented generation. *arXiv preprint arXiv:2503.10677*.
- 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. *arXiv preprint arXiv:2405.13401*.
- Yihang Cheng, Lan Zhang, Junyang Wang, Mu Yuan, and Yunhao Yao. 2025b. Remoterag: A privacy-preserving llm cloud rag service. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3820–3837.
- Zirui Cheng, Jikai Sun, Anjun Gao, Yueyang Quan, Zhuqing Liu, Xiaohua Hu, and Minghong Fang. 2025c. [Secure retrieval-augmented generation against poisoning attacks](#). In *2025 IEEE International Conference on Big Data (BigData)*.
- 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.
- Chanwoo Choi, Jinsoo Kim, Sukmin Cho, Soyeong Jeong, and Buru Chang. 2025a. [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*.
- Yujin Choi, Youngjoo Park, Junyoung Byun, Jaewook Lee, and Jinseong Park. 2025b. Safeguarding privacy of retrieval data against membership inference attacks: Is this query too close to home? In *EMNLP (Findings)*, pages 8241–8258.

- Sarthak Choudhary, Nils Palumbo, Ashish Hooda, Krishnamurthy Dj Dvijotham, and Somesh Jha. 2026. [Through the stealth lens: Attention-aware defenses against poisoning in RAG](#). In *Forty-third International Conference on Machine Learning*.
- Cody Clop and Yannick Teglia. 2024. Backdoored retrievers for prompt injection attacks on retrieval augmented generation of large language models. *arXiv preprint arXiv:2410.14479*.
- Gianluca De Stefano, Lea Schönherr, and Giancarlo Pellegrino. 2024. Rag and roll: An end-to-end evaluation of indirect prompt manipulations in llm-based application frameworks. *arXiv preprint arXiv:2408.05025*.
- Sagie Dekel, Moshe Tennenholtz, and Oren Kurland. 2026. Addressing corpus knowledge poisoning attacks on rag using sparse attention. *arXiv preprint arXiv:2602.04711*.
- Gelei Deng, Yi Liu, Kailong Wang, Yuekang Li, Tianwei Zhang, and Yang Liu. 2024. [Pandora: Jailbreak gpts by retrieval augmented generation poisoning](#). In *Proceedings of the 2024 Workshop on AI Systems with Confidential Computing*.
- Aritra Dhar, Vasilije Stambolic, and Lukas Cavigelli. 2026. [Rag-pull: Turning retrieval into a code-injection channel via invisible unicode perturbations](#). Preprint, arXiv:2510.11195.
- Shen Dong, Shaochen Xu, Pengfei He, Yige Li, Jiliang Tang, Tianming Liu, Hui Liu, and Zhen Xiang. 2025. [Memory injection attacks on llm agents via query-only interaction](#). In *Advances in Neural Information Processing Systems*, volume 38.
- Kennedy Edemacu, Vinay M Shashidhar, Micheal Tuape, Dan Abudu, Beakcheol Jang, and Jong Wook Kim. 2025. Defending against knowledge poisoning attacks during retrieval-augmented generation. *arXiv preprint arXiv:2508.02835*.
- Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2024. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint arXiv:2407.11005*.
- Haowei Fu, Bo Ni, Han Xu, Kunpeng Liu, Dan Lin, and Tyler Derr. 2026. Ensemble privacy defense for knowledge-intensive llms against membership inference attacks. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2786–2799.
- Aoran Gan, Hao Yu, Kai Zhang, Qi Liu, Wenyu Yan, Zhenya Huang, Shiwei Tong, and Guoping Hu. 2025. Retrieval augmented generation evaluation in the era of large language models: A comprehensive survey. *arXiv preprint arXiv:2504.14891*.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
- Runpeng Geng, Yanting Wang, Ying Chen, and Jinyuan Jia. 2025. Unic-rag: Universal knowledge corruption attacks to retrieval-augmented generation. *arXiv preprint arXiv:2508.18652*.
- 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. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 3807–3826.
- Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. [Not what you’ve signed up for: Compromising real-world llm-integrated applications with indirect prompt injection](#). In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pages 79–90.
- Nicolas Grislain. 2025. Rag with differential privacy. In *2025 IEEE Conference on Artificial Intelligence (CAI)*, pages 847–852. IEEE.
- Shaowei Guan, Yu Zhai, Hin Chi Kwok, Jiawei Du, Xinyu Feng, Jing Li, Harry Qin, and Vivian Hui. 2026. Medpriv-bench: Benchmarking the privacy-utility trade-off of large language models in medical open-end question answering. *arXiv preprint arXiv:2603.14265*.
- Haoze Guo and Ziqi Wei. 2026. Hidden-in-plain-text: A benchmark for social-web indirect prompt injection in rag. *arXiv preprint arXiv:2601.10923*.
- Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. 2024. A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions. *arXiv preprint arXiv:2410.12837*.
- Hyeonjeong Ha, Qiusi Zhan, Jeonghwan Kim, Dimitrios Bralios, Saikrishna Sanniboina, Nanyun Peng, Kai-Wei Chang, Daniel Kang, and Heng Ji. 2026. [MM-PoisonRAG: Disrupting multimodal RAG with local and global knowledge poisoning attacks](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33804–33826, San Diego, California, United States. Association for Computational Linguistics.
- Jia Hao, Chunhong Zhang, Jiarun Liu, Haiyu Zhao, Zhiqiang Zhan, and Zheng Hu. 2025. All that glitters is not gold: Improving robust retrieval-augmented language models with fact-centric preference alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11277–11292.
- Longzhu He, Peng Tang, Yuanhe Zhang, Pengpeng Zhou, and Sen Su. 2025. Mitigating privacy risks

- in retrieval-augmented generation via locally private entity perturbation. *Information Processing & Management*, 62(4):104150.
- Yu He, Yifei Chen, Yiming Li, Shuo Shao, Leyi Qi, Boheng Li, Dacheng Tao, and Zhan Qin. 2026. [External data extraction attacks against retrieval-augmented large language models](#). *IEEE Transactions on Information Forensics and Security*.
- Arshia Hemmat, Matin Moqadas, Ali Mamanpoosh, Amirasoud Rismanchian, and Afsaneh Fatemi. 2025. Vague-gate: Plug-and-play local-privacy shield for retrieval-augmented generation. 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 3715–3730.
- 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.
- Xuming Hu, Xiaochuan Li, Junzhe Chen, Yinghui Li, Yangning Li, Xiaoguang Li, Yasheng Wang, Qun Liu, Lijie Wen, Philip S Yu, et al. 2024a. Evaluating robustness of generative search engine on adversarial factoid questions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10650–10671.
- Zhibo Hu, Chen Wang, Yanfeng Shu, Hye-Young Paik, and Liming Zhu. 2024b. Prompt perturbation in retrieval-augmented generation based large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1119–1130.
- 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.
- Yang Jiao, Xiaodong Wang, and Kai Yang. 2025. 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*, pages 656–667.
- Minseok Kim, Hankook Lee, and Hyungjoon Koo. 2025a. [Rescuing the unpoisoned: Efficient defense against knowledge corruption attacks on rag systems](#). In *Annual Computer Security Applications Conference (ACSAC)*.
- San Kim, Jonghwi Kim, Yejin Jeon, and Gary Geunbae Lee. 2025b. Safeguarding rag pipelines with gmp: A gradient-based masked token probability method for poisoned document detection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24597–24614.
- Tatsuki Koga, Ruihan Wu, Zhiyuan Zhang, and Kamalika Chaudhuri. 2024. Privacy-preserving retrieval-augmented generation with differential privacy. *arXiv preprint arXiv:2412.04697*.
- Junhyeok Lee, Han Jang, and Kyu Sung Choi. 2026. Mpib: A benchmark for medical prompt injection attacks and clinical safety in llms. *arXiv preprint arXiv:2602.06268*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Yuxuan Li, Xinwei Guo, Jiashi Gao, Guanhua Chen, Xiangyu Zhao, Jiaxin Zhang, Quanying Liu, Haiyan Wu, Xin Yao, and Xuetao Wei. 2025a. LLMs trust humans more, that’s a problem! unveiling and mitigating the authority bias in retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28844–28858.
- 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. IEEE.
- Zizhong Li, Haopeng Zhang, and Jiawei Zhang. 2026. Token-level precise attack on rag: Searching for the best alternatives to mislead generation. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 3193–3206.
- Jiacheng Liang, Yuhui Wang, Changjiang Li, Rongyi Zhu, Tanqiu Jiang, Neil Gong, and Ting Wang. 2026. [Graphrag under fire](#). In *2026 IEEE Symposium on Security and Privacy (SP)*.
- Xun Liang, Simin Niu, Zhiyu Li, Sensen Zhang, Hanyu Wang, Feiyu Xiong, Zhaoxin Fan, Bo Tang, Jihao Zhao, Jiawei Yang, et al. 2025. 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.
- Bo Lin, Shangwen Wang, Liqian Chen, and Xiaoguang Mao. 2026. [Exploring the security threats of knowledge base poisoning in retrieval-augmented code generation](#). *IEEE Transactions on Software Engineering*.
- Jiale Liu, Jiahao Zhang, and Suhang Wang. 2026. Exposing privacy risks in graph retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 18073–18093.

- Mingrui Liu, Sixiao Zhang, and Cheng Long. 2025a. Mask-based membership inference attacks for retrieval-augmented generation. In *Proceedings of the ACM on Web Conference 2025*, pages 2894–2907.
- Yinuo Liu, Zenghui Yuan, Guiyao Tie, Jiawen Shi, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. 2025b. Poisoned-mrag: Knowledge poisoning attacks to multimodal retrieval augmented generation. *arXiv preprint arXiv:2503.06254*.
- Qianren Mao, Qili Zhang, Hanwen Hao, Zhentao Han, Runhua Xu, Weifeng Jiang, Qi Hu, Zhijun Chen, Tyler Zhou, Bo Li, et al. 2025. Privacy-preserving federated embedding learning for localized retrieval-augmented generation. *arXiv preprint arXiv:2504.19101*.
- Aiman Al Masoud, Marco Arazzi, and Antonino Nocera. 2026. Sd-rag: A prompt-injection-resilient framework for selective disclosure in retrieval-augmented generation. *arXiv preprint arXiv:2601.11199*.
- Junki Mori, Kazuya Kakizaki, Taiki Miyagawa, and Jun Sakuma. 2026. Differentially private synthetic text generation for retrieval-augmented generation (rag). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 1216–1235.
- Ali Naseh, Yuefeng Peng, Anshuman Suri, Harsh Chaudhari, Alina Oprea, and Amir Houmansadr. 2025. Riddle me this! stealthy membership inference for retrieval-augmented generation. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pages 1245–1259.
- Bo Ni, Zheyuan Liu, Leyao Wang, Yongjia Lei, Yuying Zhao, Xueqi Cheng, Qingkai Zeng, Luna Dong, Yinglong Xia, Krishnaram Kenthapadi, et al. 2025. Towards trustworthy retrieval augmented generation for large language models: A survey. *arXiv preprint arXiv:2502.06872*.
- Ruotong Pan, Boxi Cao, Hongyu Lin, Xianpei Han, Jia Zheng, Sirui Wang, Xunliang Cai, and Le Sun. 2024. Not all contexts are equal: Teaching llms credibility-aware generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19844–19863.
- Pankayaraj Pathmanathan, Michael-Andrei Panaitescu-Liess, Cho-Yu Jason Chiang, and Furong Huang. 2025. Ragpart & ragmask: Retrieval-stage defenses against corpus poisoning in retrieval-augmented generation. *arXiv preprint arXiv:2512.24268*.
- KrishnaSaiReddy Patil. 2026. Ragshield: Detecting numerical claim manipulation in government rag systems. *arXiv preprint arXiv:2604.00387*.
- Yuefeng Peng, Junda Wang, Hong Yu, and Amir Houmansadr. 2024. Data extraction attacks in retrieval-augmented generation via backdoors. *arXiv preprint arXiv:2411.01705*.
- Zhenting Qi, Hanlin Zhang, Eric Xing, Sham 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 (ICLR 2025)*.
- Zhisheng Qi, Utkarsh Sahu, Li Ma, Haoyu Han, Ryan Rossi, Franck Dernoncourt, Mahantesh Halappanavar, Nesreen Ahmed, Yushun Dong, Yue Zhao, et al. 2026. [Benchmarking knowledge-extraction attack and defense on retrieval-augmented generation](#). In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*.
- Ayush RoyChowdhury, Mulong Luo, Prateek Sahu, Sarbartha Banerjee, and Mohit Tiwari. 2024. [Confused-pilot: Confused deputy risks in rag-based llms](#). *arXiv preprint arXiv:2408.04870*.
- Jerome H. Saltzer and Michael D. Schroeder. 1975. [The protection of information in computer systems](#). *Proceedings of the IEEE*, 63(9):1278–1308.
- Avital Shafran, Roei Schuster, and Vitaly Shmatikov. 2025. Machine against the {RAG}: Jamming {Retrieval-Augmented} generation with blocker documents. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 3787–3806.
- Yingjia Shang, Yi Liu, Huimin Wang, Furong Li, Wenfang Sun, Wu Chengyu, and Yefeng Zheng. 2026. [Medusa: Cross-modal transferable adversarial attacks on multimodal medical retrieval-augmented generation](#). In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*.
- Chaitanya Sharma. 2025. Retrieval-augmented generation: A comprehensive survey of architectures, enhancements, and robustness frontiers. *arXiv preprint arXiv:2506.00054*.
- Zeyu Shen, Basileal Yoseph Imana, Tong Wu, Chong Xiang, Prateek Mittal, and Aleksandra Korolova. 2025. [ReliabilityRAG: Effective and provably robust defense for RAG-based web-search](#). In *The Thirtieth Annual Conference on Neural Information Processing Systems*.
- Ezzeldin Shereen, Dan Ristea, Shae McFadden, Burak Hasircioglu, Vasilios Mavroudis, and Chris Hicks. 2026. [One pic is all it takes: Poisoning visual document retrieval augmented generation with a single image](#). *Transactions on Machine Learning Research*.
- Shivani Shukla and Himanshu Joshi. 2025. Proof-carrying answers: A systematic protocol for verifiable retrieval-augmented generation with cryptographic provenance. In *2025 Annual Computer Security Applications Conference Workshops (ACSAC Workshops)*, pages 405–413. IEEE.
- 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](#). In *Advances in Neural Information Processing Systems*, volume 38.
- Inderjeet Singh, Vikas Pahuja, Aishvariya Priya Rathina Sabapathy, Chiara Picardi, Amit Giloni, Roman Vainshtein, Andrés Murillo, Hisashi Kojima, Motoyoshi Sekiya, Yuki Unno, et al. 2026. Adversarial intent is a latent variable: Stateful trust inference for securing multimodal agentic rag. *arXiv preprint arXiv:2602.21447*.
- 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.
- Saksham Sahai Srivastava and Haoyu He. 2025. [Memorygraft: Persistent compromise of llm agents via poisoned experience retrieval](#). Preprint, arXiv:2512.16962.
- 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.
- Shixuan Sun, Siyuan Liang, Ruoyu Chen, Jianjie Huang, Jingzhi Li, and Xiaochun Cao. 2025. Sma: Who said that? auditing membership leakage in semi-black-box rag controlling. *arXiv preprint arXiv:2508.09105*.
- 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. *arXiv preprint arXiv:2504.21680*.
- 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. 2024. 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.
- Scott Thornton. 2026. Retrieval pivot attacks in hybrid rag: Measuring and mitigating amplified leakage from vector seeds to graph expansion. *arXiv preprint arXiv:2602.08668*.
- Yiteng Tu, Weihang Su, Yujia Zhou, Yiqun Liu, and Qingyao Ai. 2025. [Rbft: 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*.
- Vishnu Vinod, Krishna Pillutla, and Abhradeep Guha Thakurta. 2025. [Invisibleink: High-utility and low-cost text generation with differential privacy](#). In *Advances in Neural Information Processing Systems*, volume 38.
- Cheng Wang, Yiwei Wang, Yujun Cai, and Bryan Hooi. 2025a. 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.
- Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O Arik. 2025b. 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.
- Haoran Wang, Xiong Xiao Xu, Baixiang Huang, and Kai Shu. 2026a. [Privacy-aware decoding: Mitigating privacy leakage of large language models in retrieval-augmented generation](#). In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*.
- Haowei Wang, Rupeng Zhang, Junjie Wang, Mingyang Li, Yuekai Huang, Dandan Wang, and Qing Wang. 2026b. Joint-gcg: Unified gradient-based poisoning attacks on retrieval-augmented generation systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 35793–35801.
- Haozhen Wang, Haoyue Liu, Jionghao Zhu, Zhichao Wang, Yongxin Guo, and Xiaoying Tang. 2026c. Pidp-attack: Combining prompt injection with database poisoning attacks on retrieval-augmented generation systems. *arXiv preprint arXiv:2603.25164*.
- Weijie Wang, Peizhuo Lv, Yan Wang, Rujie Dai, Guokun Xu, Qiu Jian Lv, Hangcheng Liu, Weiqing Huang, Wei Dong, and Jiaheng Zhang. 2026d. Making theft useless: Adulteration-based protection of proprietary knowledge graphs in graphrag systems. *arXiv preprint arXiv:2601.00274*.
- Yuhao Wang, Wenjie Qu, Yanze Jiang, Lichen Liu, Yue Liu, Shengfang Zhai, Yinpeng Dong, and Jiaheng Zhang. 2025c. [Silent leaks: Implicit knowledge extraction attack on RAG systems through benign queries](#). In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.
- Ziting Wang, Haitao Yuan, Wei Dong, Gao Cong, and Feifei Li. 2026e. [Carrot: A learned cost-constrained retrieval optimization system for rag](#). In *2026 IEEE 42nd International Conference on Data Engineering (ICDE)*, pages 2150–2163.

- Chris M Ward and Josh Harguess. 2025. Adversarial threat vectors and risk mitigation for retrieval-augmented generation systems. In *Assurance and Security for AI-enabled Systems 2025*, volume 13476, pages 80–97. SPIE.
- Zhepei Wei, Wei-Lin Chen, and Yu Meng. 2025. [InstructRAG: Instructing retrieval-augmented generation via self-synthesized rationales](#). In *The Thirteenth International Conference on Learning Representations*.
- Jiayi Wen, Tianxin Chen, Zhirun Zheng, and Cheng Huang. 2025a. A few words can distort graphs: Knowledge poisoning attacks on graph-based retrieval-augmented generation of large language models. *arXiv preprint arXiv:2508.04276*.
- Tongyu Wen, Chenglong Wang, Xiyuan Yang, Haoyu Tang, Yueqi Xie, Lingjuan Lyu, Zhicheng Dou, and Fangzhao Wu. 2025b. [Defending against indirect prompt injection by instruction detection](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19472–19487, Suzhou, China. Association for Computational Linguistics.
- Shangyu Wu, Ying Xiong, Yufei Cui, Haolun Wu, Can Chen, Ye Yuan, Lianming Huang, Xue Liu, Tei-Wei Kuo, Nan Guan, et al. 2024. Retrieval-augmented generation for natural language processing: A survey. *arXiv preprint arXiv:2407.13193*.
- 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. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 16882–16902.
- Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. 2024. Certifiably robust rag against retrieval corruption. *arXiv preprint arXiv:2405.15556*.
- Yilin Xiao, Jin Chen, Qinggang Zhang, Yujing Zhang, Chuang Zhou, Longhao Yang, Lingfei Ren, Xin Yang, and Xiao Huang. 2026. Logicpoison: Logical attacks on graph retrieval-augmented generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5575–5591.
- Jin Xie, Songze Li, and Guang Cheng. 2026. Seal-tag: Self-tag evidence aggregation with probabilistic circuits for pii-safe retrieval-augmented generation. *arXiv preprint arXiv:2603.17292*.
- Shicheng Xu, Liang Pang, Mo Yu, Fandong Meng, Huawei Shen, Xueqi Cheng, and Jie Zhou. 2024. Unsupervised information refinement training of large language models for retrieval-augmented generation. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 133–145.
- Jiaqi Xue, Mengxin Zheng, Yebowen Hu, Fei Liu, Xun Chen, and Qian Lou. 2024. Badrag: Identifying vulnerabilities in retrieval augmented generation of large language models. *arXiv preprint arXiv:2406.00083*.
- 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.
- Shuhua Yang, Jiahao Zhang, Yilong Wang, Dongwon Lee, and Suhan Wang. 2026. Query-efficient agentic graph extraction attacks on graphrag systems. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16017–16041.
- Mengyu Yao, Ziqi Zhang, Ning Luo, Shaofei Li, Yifeng Cai, Xiangqun Chen, Yao Guo, and Ding Li. 2026. Connect the dots: Knowledge graph-guided crawler attack on retrieval-augmented generation systems. *arXiv preprint arXiv:2601.15678*.
- Ruobing Yao, Yifei Zhang, Shuang Song, Neng Gao, and Chenyang Tu. 2025. Ecosaferag: Efficient security through context analysis in retrieval-augmented generation. In *EMNLP (Findings)*, pages 4034–4050.
- Huanyi Ye, Jiale Guo, Ziyao Liu, and Kwok-Yan Lam. 2025a. Efficient privacy-preserving retrieval augmented generation with distance-preserving encryption. In *2025 3rd International Conference on Foundation and Large Language Models (FLLM)*, pages 668–676. IEEE.
- Kai Ye, Liangcai Su, and Chenxiong Qian. 2025b. [Importsnare: Directed 'code manual' hijacking in retrieval-augmented code generation](#). In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security, CCS '25*, page 335–349, New York, NY, USA. Association for Computing Machinery.
- Jingwei Yi, Yueqi Xie, Bin Zhu, Emre Kiciman, Guangzhong Sun, Xing Xie, and Fangzhao Wu. 2025. Benchmarking and defending against indirect prompt injection attacks on large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 1809–1820.
- Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024a. Evaluation of retrieval-augmented generation: A survey. In *CCF Conference on Big Data*, pages 102–120. Springer.
- Xiaodong Yu, Hao Cheng, Xiaodong Liu, Dan Roth, and Jianfeng Gao. 2024b. Reeval: Automatic hallucination evaluation for retrieval-augmented large language models via transferable adversarial attacks. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1333–1351.

- Linda Zeng, Rithwik Gupta, Divij Motwani, Yi Zhang, and Diji Yang. 2025a. [Worse than zero-shot? a fact-checking dataset for evaluating the robustness of rag against misleading retrievals](#). *Advances in Neural Information Processing Systems*, 38.
- Shenglai Zeng, Jiankun Zhang, Pengfei He, Jie Ren, Tianqi Zheng, Hanqing Lu, Han Xu, Hui Liu, Yue Xing, and Jiliang Tang. 2025b. [Mitigating the privacy issues in retrieval-augmented generation \(RAG\) via pure synthetic data](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24527–24558, Suzhou, China. Association for Computational Linguistics.
- Shenglai Zeng, Jiankun Zhang, Pengfei He, Yue Xing, Yiding Liu, Han Xu, Jie Ren, Shuaiqiang Wang, Dawei Yin, Yi Chang, 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.
- Shenglai Zeng, Jiankun Zhang, Bingheng Li, Yuping Lin, Tianqi Zheng, Dante Everaert, Hanqing Lu, Hui Liu, Yue Xing, Monica Xiao Cheng, et al. 2025c. Towards knowledge checking in retrieval-augmented generation: A representation perspective. 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 2952–2969.
- Zhirui Zeng, Jiamou Liu, Meng-Fen Chiang, Jialing He, and Zijian Zhang. 2025d. S-rag: A novel audit framework for detecting unauthorized use of personal data in rag systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10375–10385.
- Baolei Zhang, Yuxi Chen, Zhuqing Liu, Lihai Nie, Tong Li, Zheli Liu, and Minghong Fang. 2026a. [Practical poisoning attacks against retrieval-augmented generation](#). In *Proceedings of the 31st ACM Symposium on Access Control Models and Technologies*.
- Baolei Zhang, Haoran Xin, Yuxi Chen, Zhuqing Liu, Biao Yi, Tong Li, Lihai Nie, Zheli Liu, and Minghong Fang. 2026b. [Who Taught the Lie? Responsibility Attribution for Poisoned Knowledge in Retrieval-Augmented Generation](#). In *2026 IEEE Symposium on Security and Privacy (SP)*, pages 3396–3415, Los Alamitos, CA, USA. IEEE Computer Society.
- Baolei Zhang, Haoran Xin, Minghong Fang, Zhuqing Liu, Biao Yi, Tong Li, and Zheli Liu. 2025a. Trace-back of poisoning attacks to retrieval-augmented generation. In *Proceedings of the ACM on Web Conference 2025*, pages 2085–2097.
- Baolei Zhang, Haoran Xin, Jiatong Li, Dongzhe Zhang, Minghong Fang, Zhuqing Liu, Lihai Nie, and Zheli Liu. 2025b. Benchmarking poisoning attacks against retrieval-augmented generation. *arXiv preprint arXiv:2505.18543*.
- Jiankun Zhang, Shenglai Zeng, Jie Ren, Tianqi Zheng, Hui Liu, Xianfeng Tang, Hui Liu, and Yi Chang. 2025c. [Beyond text: Unveiling privacy vulnerabilities in multi-modal retrieval-augmented generation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24789–24810, Suzhou, China. Association for Computational Linguistics.
- Quan Zhang, Binqi Zeng, Chijin Zhou, Gwihwan Go, Heyuan Shi, and Yu Jiang. 2024a. Human-imperceptible retrieval poisoning attacks in llm-powered applications. In *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, pages 502–506.
- Yuan Zhang, Jionghan Wu, Rui Li, Tong Zhang, Yujie Song, Chuanyi Li, Shangqi Wang, Hao Shen, Jiao Yin, Jidong Ge, et al. 2026c. [Privacy protection in rag: A novel method and evaluation framework](#). *Information Processing & Management*, 63(3):104505.
- Yucheng Zhang, Qinfeng Li, Tianyu Du, Xuhong Zhang, Xinkui Zhao, Zhengwen Feng, and Jianwei Yin. 2024b. Hijackrag: Hijacking attacks against retrieval-augmented large language models. *arXiv preprint arXiv:2410.22832*.
- Dongfang Zhao. 2024. Frag: Toward federated vector database management for collaborative and secure retrieval-augmented generation. *arXiv preprint arXiv:2410.13272*.
- Jingjie Zheng, Aryo Pradipta Gema, Giwon Hong, Xuanli He, Pasquale Minervini, Youcheng Sun, and Qionghai Xu. 2025. [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.
- 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. *arXiv preprint arXiv:2501.00879*.
- Pengcheng Zhou, Yinglun Feng, and Zhongliang Yang. 2025b. Privacy-aware rag: Secure and isolated knowledge retrieval. *arXiv preprint arXiv:2503.15548*.
- Pengcheng Zhou, Yinglun Feng, and Zhongliang Yang. 2025c. Provably secure retrieval-augmented generation. *arXiv preprint arXiv:2508.01084*.
- Xinyun Zhou, Xinfeng Li, Yinan Peng, Ming Xu, Xuanwang Zhang, Miao Yu, Yidong Wang, Xiaojun Jia, Kun Wang, Qingsong Wen, et al. 2026. Emorag: Evaluating rag robustness to symbolic perturbations. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 2100–2111.

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. *arXiv preprint arXiv:2409.10102*.

Junda Zhu, Lingyong Yan, Haibo Shi, Dawei Yin, and Lei Sha. 2024. Atm: Adversarial tuning multi-agent system makes a robust retrieval-augmented generator. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10902–10919.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2025. PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 3827–3844.

A Search and Screening Protocol

Sources, window, and queries. We searched arXiv, ACL Anthology, OpenReview, Semantic Scholar, and DBLP for work first made public from May 1, 2020 through April 30, 2026, inclusive. Each source query combined at least one RAG term with at least one security term in the title or abstract.

The RAG block was as follows: retrieval augmented OR retrieval-augmented OR RAG OR GraphRAG OR agentic RAG OR multimodal RAG OR graph retrieval augmented generation OR multimodal retrieval augmented generation OR visual document retrieval augmented generation.

The security block was as follows: attack* OR adversarial OR poison* OR backdoor* OR "prompt injection" OR jailbreak* OR hijack* OR malicious OR threat* OR vulnerab* OR compromis* OR manipul* OR corrupt* OR perturb* OR exploit* OR privacy OR private OR confidential* OR leak* OR "data extraction" OR "knowledge extraction" OR "membership inference" OR unauthoriz* OR "access control" OR provenance OR copyright OR watermark* OR steal* OR theft OR protect* OR security OR secure OR defen* OR robust* OR safe* OR trust* OR reliab* OR integrity OR verif* OR audit* OR guardrail* OR resilien* .

Because the DBLP publication-search API does not expose abstracts, both blocks were matched against DBLP titles only.

Title-and-abstract prompt. The first screening round used the following instruction with the returned title and abstract:

Decide whether the external-knowledge path of a generative system is the object of a primary or substantial contribution. Include (I1) attacks on a corpus, index, retriever, ranking, memory, graph, tool content, or retrieved evidence, including indirect prompt injection and extraction; (I2) defenses, secure architectures, access control, selective disclosure, cryptographic protection, or remediation against these risks; (I3) security/privacy benchmarks, red-team studies, surveys, or measurements of behavior under adversarial or untrustworthy retrieved evi-

dence; (I4) methods or evaluations centered on the integrity of false, misleading, conflicting, contaminated, or temporally invalid external evidence; and (I5) query or prompt perturbations designed to manipulate retrieval or a retrieval-backed generator. Exclude ordinary RAG accuracy, retrieval optimization, efficiency, citation quality, benign irrelevant context, generic hallucination reduction, generic LLM safety/privacy, unrelated domain applications, and broad risk discussions without a concrete external-knowledge contribution. Return INCLUDE, EXCLUDE, or UNCERTAIN, the matched criterion, surface, and a brief evidence-based reason.

Full-text prompt. The second round applied the following rule to full-text evidence:

Include only if all three conditions hold: (A) the system uses generation grounded in an external corpus, retrieved documents, graph, multimodal or code store, external memory, tool-returned content, or private inference-time context; (B) security, privacy, or semantic/source integrity of that path is a primary or substantial contribution; and (C) the paper provides a concrete attack, defense, benchmark, empirical measurement, integrity-robustness method, RAG-targeted input attack, or systematic survey/taxonomy. Apply I1–I5 as defined above. Exclude ordinary quality, efficiency, abstention, benign noise, generic LLM safety/privacy, unrelated applications, and non-concrete discussions. Use UNCERTAIN only when the extracted evidence is inadequate or genuinely ambiguous, and flag the record for manual review. Do not reward or penalize venue, citation count, novelty, writing style, or result magnitude. Return the decision, criterion, surface, centrality, paper type, evidence locations, exclusion code, quality flags, confidence, manual-review flag, and a one-sentence rationale.

Both rounds used DeepSeek-V4-Flash-0731.

Consolidation and screening funnel. The source-level merge used lowercased DOI, version-

stripped arXiv ID, source identifiers, and Unicode-normalized exact titles. DBLP records remained screenable from titles alone. The later work-level merge combined DOI and arXiv matches with fuzzy-title, abstract, and manual comparison. Duplicate or superseded versions are represented by one canonical record.

Expert selection and snowballing. Expert review selected a work as *representative* when it added a distinct mechanism, threat setting, control, evaluation resource, modality, domain, or empirical finding needed to cover the eligible landscape. *Technically substantive* required a concrete method, benchmark, or measurement and enough full-text evidence for SLOT coding; venue prestige was not a criterion. Backward and forward citation snowballing then added BIPIA (Yi et al., 2025), MINJA (Dong et al., 2025), and the canonical work by Greshake et al. (Greshake et al., 2023).

Dates and released records. The earliest verified public date determined window eligibility. When metadata supplied only a year, YYYY-01-01 records an explicit year-only fallback rather than a claimed publication day.

The release documents the screening process and contains 555 records, including 135 selected works. It provides canonical identifiers and public paper or code links, screening-stage information and rationales, selected-paper summaries, and primary/secondary SLOT labels.

B SLOT Coding Guide

Primary S/L identifies the first pipeline point at which a paper’s central attack or control acts; a secondary S/L records a substantial mechanism at another stage. Primary O identifies the main measured security objective, while secondary objectives record additional measured effects. T1 fixes the target instance, query, or finite query set and

Stage	Retained
Source retrieval	17,256
Source consolidation	8,262
Title/abstract inclusion	636
Version-level unique works	617
Full-text eligibility	552
Expert selection	132
Snowballing	+3
Final corpus	135

Table 2: Literature-search and screening funnel.

assumes that the attacker knows it before the attack; T2 targets aggregate behavior over unknown future queries or adaptively explored records. Each work is counted once as attack-, defense-, or evaluation-primary; secondary roles are retained and shown in gray when cross-listed.

Boundary cases use the same external-knowledge test. Greshake et al. (Greshake et al., 2023) and BIPIA (Yi et al., 2025) are included because retrieved content carries the attack. JailbreakBench (Chao et al., 2024) is excluded as prompt-only, and RAGBench (Friel et al., 2024) is excluded as ordinary RAG-quality evaluation.

C Scope Comparison with Adjacent Surveys

Table 4 positions our external-knowledge-centered scope relative to adjacent RAG surveys. The dimensions record whether external knowledge is treated as a security object, whether RAG-specific risks are separated from generic LLM risks, and whether the survey connects pipeline stages, attacks, defenses, evaluation, and deployment controls.

D Evaluation Protocol and Reporting Specification

Our proposed six-stage plug-in harness exposes the external-source, ingestion/indexing, retrieval/reranking, context-assembly, generation, and response/remediation stages. Attacks plug into S1–S4 and defenses into L1–L4; each result is reported in the relevant O×T cell. A surface-complete stack must contain an explicit control for every S1–S4 surface and evaluate each control separately against an attacker who knows all non-secret defense details.

The minimum SLOT card reports S/L/O/T, attacker knowledge and access, a surface-specific budget, victim configuration, assumptions, adaptive status, utility, latency, and uncertainty. O1 evaluation should report task success and integrity under attack; O2 should report refusal/failure and resource costs; O3 should report membership or leakage together with recovered fraction and query cost. T1 fixes the target instances or query set before the attack; T2 measures aggregate behavior over unknown future queries or adaptively explored records.

Table 5 illustrates why cross-paper results require protocol-level alignment by comparing four representative studies spanning targeted corpus poi-

Category	Representative work	O/T	Reason for primary placement
S1	PoisonedRAG (Zou et al., 2025)	O1/T1	Injects malicious knowledge before retrieval.
S2	GGPP (Hu et al., 2024b)	O1/T1	Manipulates a known query to retrieve false evidence.
S3	Indirect Injection (Greshake et al., 2023)	O1/T2	Delivers instructions through retrieved external content.
S4	MIA-RAG (Anderson et al., 2025)	O3/T1	Infers membership in a private corpus.
L1	D-RAG (Andersen et al., 2025)	O1/T1	Applies provenance controls to the knowledge substrate.
L2	RAGuard-Defense (Cheng et al., 2025c)	O1/T1	Filters retrieved candidates before generation.
L3	InstructRAG (Wei et al., 2025)	O1/T1	Limits malicious post-retrieval influence.
L4	AC-RAG (Chen et al., 2025a)	O3/T1	Enforces fine-grained authorization for retrieved knowledge.

Table 3: Representative examples for the eight primary Surface/Layer categories. Secondary labels are retained in the released records.

Survey	Primary focus	EK Scope	Risk Bnd.	Pipe. View	A-D-E	Eval.	Deploy.
Gao et al. (2023)	RAG architectures and components	–	–	○	–	○	–
Yu et al. (2024a)	Evaluation metrics and benchmarks	–	–	○	–	✓	–
Wu et al. (2024)	Retrieval modules and applications	–	–	○	–	○	–
Zhou et al. (2024)	Trustworthiness dimensions and benchmark	○	○	○	○	✓	○
Gupta et al. (2024)	RAG landscape and future directions	–	–	○	–	–	○
Ni et al. (2025)	Trustworthy RAG roadmap	○	○	○	○	○	○
Cheng et al. (2025a)	Knowledge-oriented RAG taxonomy	○	–	○	–	○	–
Gan et al. (2025)	Evaluation datasets and meta-analysis	–	–	○	–	✓	–
Ammann et al. (2025)	Risk assessment and mitigation mapping	✓	○	✓	○	○	✓
Sharma (2025)	RAG robustness and benchmarks	–	–	○	–	○	○
Ward and Harguess (2025)	Adversarial threats and risk controls	✓	○	✓	○	–	✓
Arzanipour et al. (2025)	RAG threat model and attack surface	✓	✓	✓	○	–	–
Bodea et al. (2026)	Privacy risks and mitigations	✓	○	✓	○	✓	○
Ours	External-knowledge security in RAG	✓	✓	✓	✓	✓	✓

Table 4: Scope comparison with adjacent RAG-centered surveys. EK Scope: external knowledge as a security object; Risk Bnd.: explicit RAG/generic-LLM boundary; Pipe. View: mapping to knowledge-access stages; A-D-E: attacks, defenses, and evaluation under one threat scope; Eval.: metrics, benchmarks, or protocols; Deploy.: mitigation, recovery, or access control. ✓/○/– denote central, partial, and not-primary coverage.

Work	Threat model, access, and budget	Representative evaluation
PoisonedRAG (Zou et al., 2025) (S1/O1/T1)	Target query and answer known; corpus-write access; five injected passages per query.	NQ, HotpotQA, and MS MARCO; 97% attack success rate (ASR) on NQ with five injected passages per query.
Topic-FlipRAG (Gong et al., 2025) (S1/O1/T2)	Black-box topic/stance manipulation; exact future test queries need not be known.	PROCON and MS MARCO; distribution-level attack success value (ASV) of 0.49 for Qwen2.5 and 0.48 for Llama3.1 on PROCON.
RobustRAG (Xi et al., 2024) (L3/O1/T1)	Defense-aware bounded corruption; certification requires benign evidence to dominate malicious evidence.	RQA-MC, RQA, NQ, and Biography; 69–71% certifiable robust accuracy on RQA-MC and 31–38% on RQA under bounded corruption.
BIPIA (Yi et al., 2025) (S3/L3/O1/T1)	Indirect prompt injection through external content; attack success and clean utility are measured separately.	Five tasks, 250 attacker goals, and 25 LLMs; 626,250 training and 86,250 test prompts, evaluated using ASR and utility metrics.

Table 5: Comparison of representative secure-RAG evaluations under their primary SLOT labels. Secondary labels are omitted for clarity; for BIPIA, S3 and L3 denote its cross-tagged attack and defense roles, respectively.

soning, distribution-level stance manipulation, certified defense, and indirect prompt injection.

As these studies evaluate different targets under different attacker access, budgets, assumptions, and metrics, their headline ASR, distribution-level ASV, and certifiable robust accuracy are not directly rankable.

E Full Taxonomy Figures

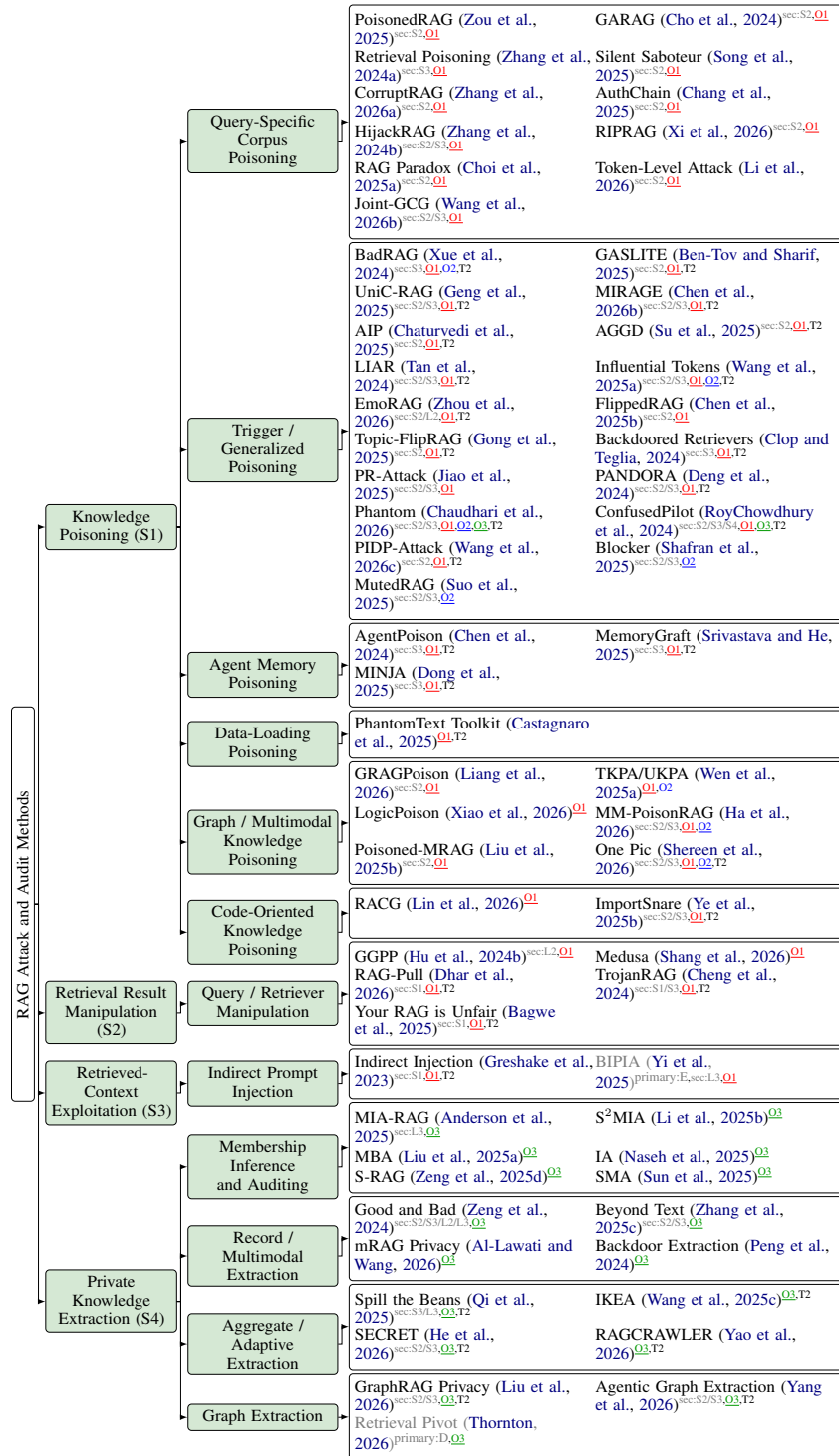


Figure 3: Taxonomy of RAG attack methods by primary attack surface (S1–S4). Each attack-primary method appears once at its first intervention point. The gray BIPIA and Retrieval Pivot cross-lists are evaluation- and defense-primary, respectively, and excluded from attack counts. Gray *sec:* tags record explicit secondary surfaces or defense layers and are excluded from primary counts. Superscripts mark attack objectives: **O1** integrity, **O2** availability, and **O3** confidentiality. T2 denotes aggregate targets over unknown future queries or adaptively explored records, marking 31 of 64 attack-primary works; unmarked works are T1.

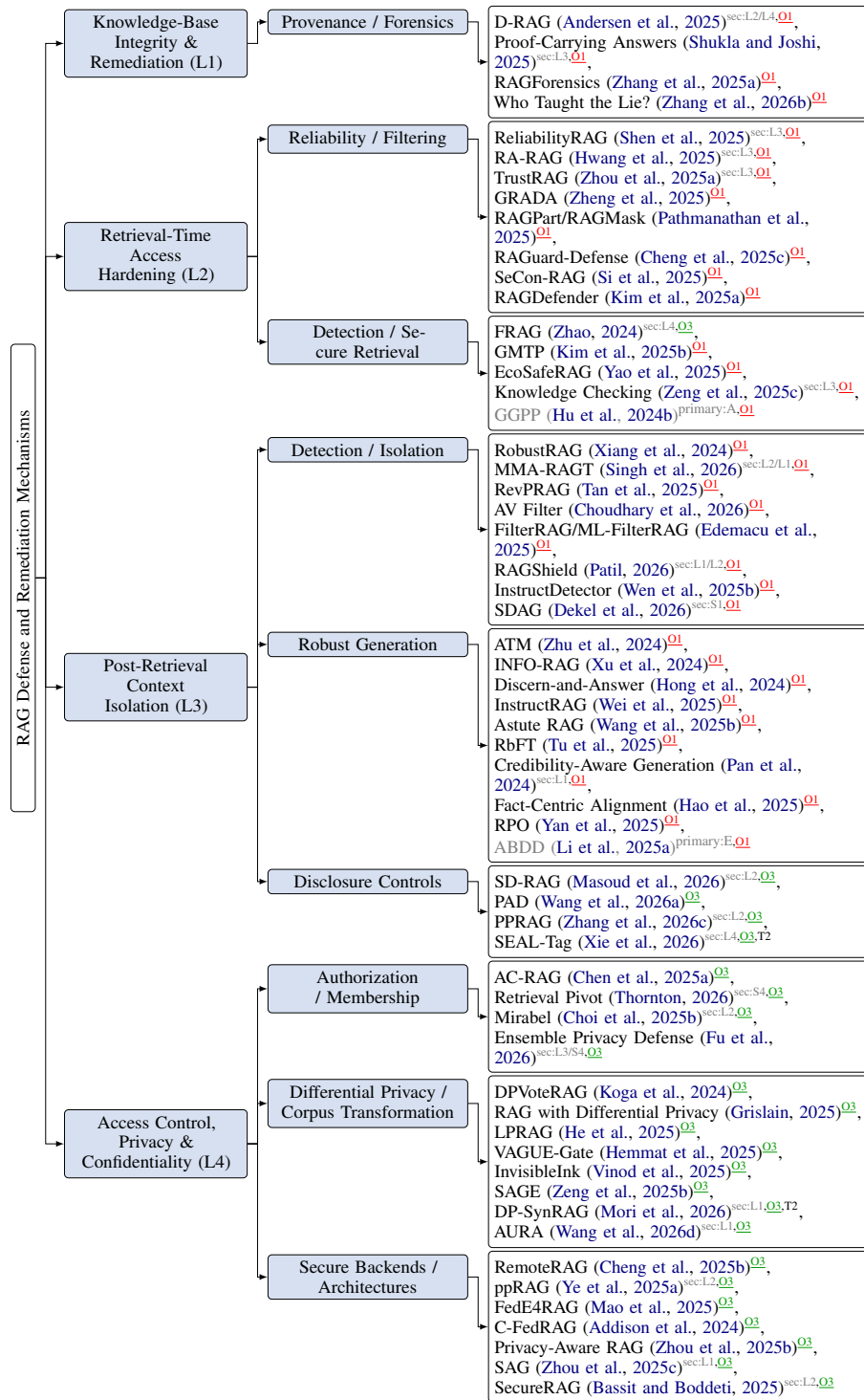


Figure 4: Taxonomy of RAG defense and remediation mechanisms, organized by primary control point (L1–L4). Superscripts mark the security objective each method counters: **O1** integrity, **O2** availability, and **O3** confidentiality. Gray sec: tags record explicit secondary layers or attack surfaces and are excluded from primary counts. The gray GGPP and ABDD cross-lists are attack- and evaluation-primary, respectively. T2 marks 2 of 56 defense-primary works with aggregate protection over unknown future queries or adaptively explored records. Retrieval Pivot is defense-primary, with S4 as a secondary audit role.