Security and Privacy of Machine Learning, 2025 Critique G6: Security and Privacy Risks in RAG – (1) On the Vulnerability of Applying Retrieval-Augmented Generation within Knowledge-Intensive Application Domains (2) SafeRAG: Benchmarking Security in Retrieval-Augmented Generation of LLM

Shih-Yu Lai
National Taiwan University
Taipei, Taiwan
akinesia112@gmail.com

Critique: (1) On the Vulnerability of Applying Retrieval-Augmented Generation within Knowledge-Intensive Application Domains

Summary

This paper investigates how Retrieval-Augmented Generation (RAG) systems fail under *data injection* and related manipulations in knowledge-intensive domains (e.g., medicine, law). The authors formalize attacker capabilities at multiple points in the RAG pipeline (index, retrieve, filter, generate), and empirically show that carefully crafted passages can be (i) persistently retrieved, (ii) survive filtering, and (iii) steer or suppress answers by the generator. The work complements prior *corpus poisoning* attacks on retrieval corpora [1], [2], and is conceptually adjacent to recent *blocker/jamming* threats that induce refusals [3]. In contrast with contemporaneous directions on certified defenses to retrieval corruption [4], the paper emphasizes *stealthy*, broadly applicable perturbations that remain top-ranked across query variants, highlighting the brittleness of dense retrievers and rerankers in real deployments. Overall, the study argues that RAG's early-stage choices (what gets retrieved) dominate downstream safety and correctness, especially in high-stakes domains.

Strengths

- **Clear threat modeling across pipeline stages.** The decomposition by attack surface (KB poisoning, retrieval-time mixing, post-filter injection) gives practitioners a concrete map of where integrity can fail.
- **Stealth and transferability.** Emphasis on attacks that maintain high embedding similarity makes the findings relevant beyond any single retriever, aligning with prior poisoning literature [1].
- **High-stakes domain framing.** Positioning results in knowledge-intensive settings surfaces real risks (omitted caveats, contradictory evidence), surpassing toy QA benchmarks.
- **Bridges two lines of work.** The paper sits between poisoning attacks [1], [2] and refusal/jamming threats [3], giving a fuller picture of *both* misleading and silencing failures.

Weaknesses

- **Defense coverage is light.** While the study exposes failures, it does not systematically evaluate *retrieval-side defenses* (e.g., lexical+dense ensembles, provenance/trust scoring, activation-profile outliering) or contrast them against the proposed attacks. Readers seeking practical recipes must look to defense-oriented work [4].
- **Assumption of KB mutability.** Many enterprise RAG systems gate corpus updates (curation, provenance). The threat model would be stronger with *no-insertion* settings (pure query redirection) or poisoning under strict ingestion validators.
- **Evaluator dependence and construct validity.** If success is measured with a single LLM evaluator or narrow MC-style probes, results can reflect evaluator idiosyncrasies rather than ground truth. More human auditing or cross-evaluator checks would improve external validity.
- **Limited multimodality and structure.** Many knowledge-intensive stacks fuse tables, images, or graphs. Text-only attacks are illuminating but leave open

whether similar vulnerabilities hold for heterogeneous retrievers and cross-modal consistency checks.

Potential Improvements / Extensions

- **Retrieval hardening ablations.** Quantify how much each simple mitigation helps: (i) BM25+dense ensembles with disagreement penalties, (ii) source-provenance trust priors, (iii) cross-query stability (retrieve on paraphrases; downweight unstable hits), (iv) contradiction graphs to demote outliers.
- **Attack budget, stealth, and cost trade-offs.** Map success to attacker effort: number of poisoned docs, lexical overlap constraints, age/provenance restrictions, and required ranking positions. Calibrate against certified-robust baselines [4].
- **From retrieval to answer consensus.** Evaluate isolate-then-aggregate schemes (per-context answering + consensus) to limit single-document influence and compare against end-to-end baselines targeted by poisoning [1], [2].
- **Multimodal and structured RAG.** Extend to tabular+text or image+text pipelines with cross-modal contradiction detection and schema-aware filters.

Questions for the Authors

1) **Transferability:** How do attacks crafted on one encoder (e.g., E5/BGE) transfer to others and to hybrid rerankers? Are there families of perturbations that generalize across encoders?
2) **Stability testing:** If the system paraphrases each query (style/grammar shifts) and aggregates results, how much does retrieval stability improve or degrade attack success?
3) **Curation constraints:** Under strict ingestion (whitelists, per-source quotas, freshness limits), what is the smallest feasible poisoning budget to maintain top-k presence?
4) **Defense interactions:** Do blocker/jamming-style contexts [3] combine with poisoning to amplify refusal while keeping high retrieval scores?
5) **Human harm assessment:** For medical/legal use cases, did you measure omission/commission harms with expert annotators, beyond multiple-choice correctness?

Critique: (2) SafeRAG: Benchmarking Security in Retrieval-Augmented Generation of Large Language Model

I. Summary

The paper proposes **SafeRAG**, a benchmark and evaluation framework for security in retrieval-augmented generation (RAG) systems. The authors argue that existing attack formulations (noise, conflict, toxicity, DoS) often fail to bypass modern RAG components, thus underestimating risk. SafeRAG introduces four *harder-to-filter* attacks: *silver noise* (partially correct distractors), *inter-context conflict* (minimally perturbed contradictions), *soft ad* (implicit toxicity via promotional content), and *white DoS* (refusal disguised as safety warnings). A Chinese news–based dataset with human curation is built; metrics cover both retrieval safety (RA) and generation safety (F1 variants, AFR/ASR). Experiments across multiple retrievers/filters/generators show substantial performance degradation and interesting model-specific vulnerabilities. The work situates itself relative to prior RAG poisoning and blocker attacks [2], [3], [5] and to general RAG robustness/benchmarks [6], [7].

II. Strengths

- **Attack design novelty.** The four attack classes are carefully specified to *bypass* practical safeguards (retriever, filter, generator), moving beyond obvious or easily-filtered baselines (e.g., naive refusal strings) [3], [5].
- **Pipeline-wide evaluation.** Clear threat model across KB, retrieved, and filtered stages; the same attack families are tested at multiple insertion points—useful for forensics and defense placement.
- **Human-aligned metrics.** The dual perspective (RA for retrieval; F1/AFR for generation) provides interpretable signals and shows high agreement with human judgments.
- **Chinese benchmark contribution.** Valuable coverage of a non-English setting; the news-domain focus keeps facts time-stamped and verifiable.
- **Empirical insights.** Hybrid rerankers are relatively robust to silver noise; compression can *hurt* in conflict but *help* against soft ad/white DoS. Smaller models sometimes appear less trigger-prone than highly safety-aligned LLMs.

III. Weaknesses

- **Metric coupling and invariances.** RA linearly averages recall of gold vs. attack. This assumes equal cost and stationarity across domains; it is insensitive to rank order and to asymmetric harm (e.g., a single white-DoS trigger may be more damaging than several silver-noise hits).
- **Evaluator and construction circularity.** Multiple-choice scoring and evaluator choices (e.g., LLM-based proposition judgments) risk alignment bias and self-consistency effects; some conclusions may change with a different evaluator policy.
- **Domain and modality scope.** Single-modality, Chinese news domain limits generality to enterprise RAG (code/doc QA, legal, clinical). Attacks like soft ad could manifest differently in other domains/languages.
- **Manual attack curation scalability.** Inter-context conflicts and naturalistic soft ads rely on trained annotators; scaling and updating the benchmark will be costly, and drift may outpace revisions.
- **Defense baselines.** While filters/rerankers are covered, certified or *training-time* defenses (robust contrastive retrievers, counterfactual-consistency training, provenance-aware decoding) are not systematically compared [6].

IV. Potential Improvements and Extensions

- **Cost-sensitive RA.** Replace RA with a costed retrieval risk (e.g., weighted by downstream harm or triggerabil-

ity), and incorporate rank-aware measures (nDCG with adversarial penalties).

- **Causal consistency tests.** Add interventions that test whether the generator maintains causal relations under conflict (do-calculus style perturbations) rather than surface agreement.
- **Provenance and trust signals.** Augment contexts with provenance metadata and cryptographic/verifiable attestations; measure whether models learn to *discount* untrusted sources.
- **Adversarial training for retrievers.** Integrate negative mining from silver-noise and conflict variants, akin to robustness training for dense retrieval; report trade-offs.
- **Cross-lingual/multimodal SafeRAG.** Extend to EN and JP (and code/doc), plus image–text RAG; include cross-modal conflicts (image contradicts caption) and advertisement-style toxicity in visuals.
- **Benchmark dynamics.** Release a generator that continually mutates soft ads/white-DoS patterns using black-box optimization to prevent overfitting to static attacks.

V. Questions for the Authors

1) How sensitive are your conclusions to the specific evaluator prompts and to the proposition granularity? Have you tried an abstractive, open-ended scoring alternative without MCQ?
2) Can RA be decomposed to isolate ranking robustness (e.g., Kendall-τ under attack) versus mere inclusion/exclusion?
3) Do hybrid rerankers trade precision for vulnerability under white DoS at higher K? Any evidence of phase transitions as K grows?
4) How would the attack success change if sources carried signed provenance (e.g., content credentials) and the model was trained to condition on it?
5) Could you quantify *collateral* degradation: when defenses block soft ads, do they also suppress legitimate policy or safety text?

References

[1] Z. Zhong, Z. Huang, A. Wettig, and D. Chen, "Poisoning retrieval corpora by injecting adversarial passages," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2023, pp. 13 764–13 775.

[2] W. Zou, R. Geng, B. Wang, and J. Jia, "Poisonedrag: Knowledge poisoning attacks to retrieval-augmented generation of large language models," *arXiv preprint arXiv:2402.07867*, 2024. [Online]. Available: https://arxiv.org/abs/2402.07867

[3] A. Shafran, R. Schuster, and V. Shmatikov, "Machine against the rag: Jamming retrieval-augmented generation with blocker documents," *arXiv preprint arXiv:2406.05870*, 2024. [Online]. Available: https://arxiv.org/abs/2406.05870

[4] C. Xiang, T. Wu, Z. Zhong, D. Wagner, D. Chen, and P. Mittal, "Certifiably robust rag against retrieval corruption," *arXiv preprint arXiv:2407.XXXX*, 2024, preprint; arXiv identifier to be updated if revised.

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

[6] F. Fang, Y. Bai, S. Ni, M. Yang, X. Chen, and R. Xu, "Rag bench: Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training," *arXiv preprint arXiv:2405.20978*, 2024.

[7] Y. Liu, L. Huang, S. Li, S. Chen, H. Zhou, F. Meng, J. Zhou, and X. Sun, "Recall: A benchmark for llms robustness against external counterfactual knowledge," *arXiv preprint arXiv:2309.16125*, 2023.