POISONING BY DILUTION: AMPLIFYING RAG ATTACKS WITH ANSWER-FREE DOCUMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation (RAG) is vulnerable to corpus poisoning, where injected passages steer a model toward an attacker-chosen incorrect answer. Existing attacks primarily construct passages that directly support this answer. We investigate whether passages without answer support can also strengthen targeted poisoning. We propose *evidence dilution*, which combines target-bearing passages with query-relevant, answer-free passages that compete for limited retrieval slots and alter the evidence available to the generator. Our method uses relation-aware generation, semantic checks, and proxy retrieval scores to construct and select dilution passages. We evaluate evidence dilution on Natural Questions, HotpotQA, and MS MARCO, achieving strict attack success rates of 92%, 87%, and 75%, respectively. On Natural Questions, adding three dilution passages increases strict attack success from 84% to 92% in vanilla RAG and from 77% to 91% under InstructRAG-ICL. Evidence analysis links these gains to changes in retrieved evidence and transitions from mixed answers to exclusive adoption of the attacker-chosen answer. These results show that passages can strengthen targeted corpus poisoning without directly supporting the attacker's answer.

1 INTRODUCTION

Retrieval-augmented generation (RAG) answers questions using information retrieved from an external corpus (Lewis et al., 2020). This reliance exposes it to corpus poisoning: an attacker can inject passages that change the retrieved context and steer the generated answer (Zou et al., 2025). In targeted corpus poisoning, the objective is to induce a predetermined incorrect answer, which we call the *attacker-chosen answer*.

Existing attacks often construct passages that directly support the attacker-chosen answer. PoisonedRAG combines retrieval and generation requirements, while AuthChain strengthens injected claims through evidence chains and authority cues (Zou et al., 2025; Chang et al., 2025). We call passages that provide this support *target-bearing passages*. We examine a complementary role for injected passages: providing query-relevant background without answering the question. We call these *dilution passages* and their use alongside target-bearing passages *evidence dilution*. Figure 1 contrasts conventional attacks using target-bearing passages with evidence dilution, which adds answer-free, query-relevant passages to reshape the retrieved evidence.

The motivation comes from the difference between query relevance and answer support. Top-k retrieval admits a fixed number of passages. A query-relevant passage can occupy one of these positions without providing an answer. If it replaces a passage supporting the correct answer, the remaining evidence may favor the attacker-chosen answer. Evidence dilution exploits this retrieval competition without adding claims that support the incorrect answer.

This separation also matters for defenses. RAGDefender uses similarity among retrieved passages to help identify suspicious content (Kim et al., 2025). Dilution passages can cover different background facets without supporting the attacker-chosen answer. Their inclusion may change the similarity patterns used by the defense. This motivates examining whether answer-free dilution passages can strengthen attacks against defended RAG systems by altering passage filtering and preserving target-bearing evidence.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

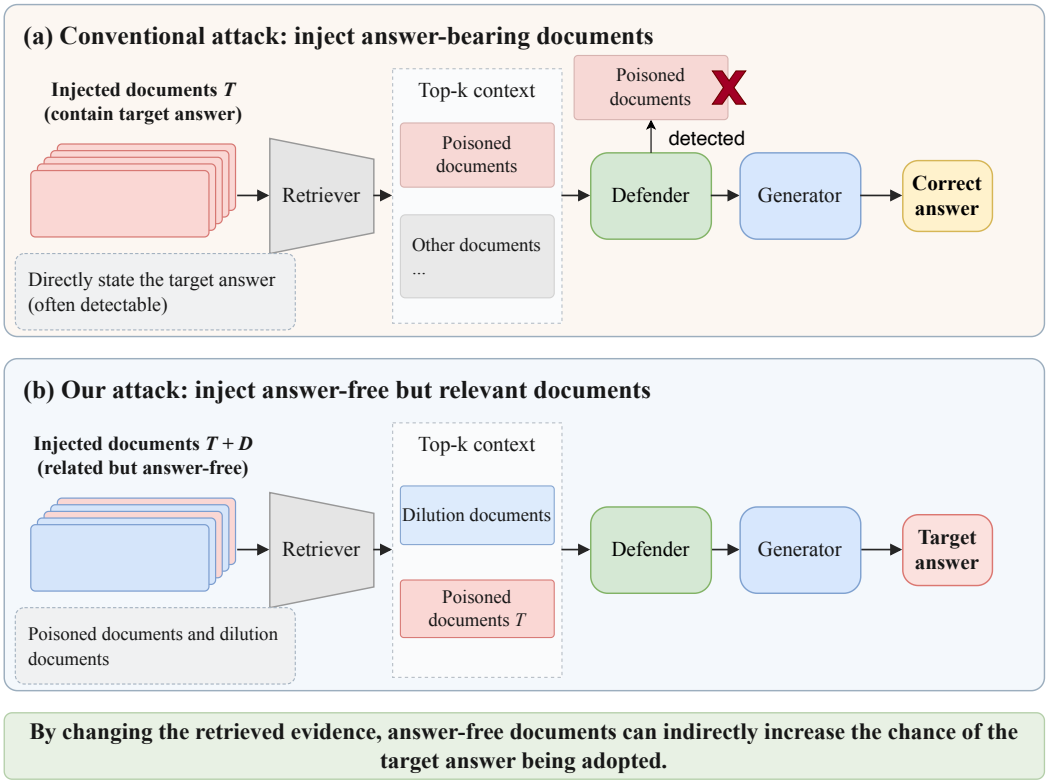

Figure 1: Overview of evidence dilution.

Studies of distracting passages show that context composition affects answer accuracy (Cuconasu et al., 2024; Amiraz et al., 2025). Lower accuracy alone does not establish *target adoption*: the generator must endorse the specified attacker-chosen answer. We ask: *With target-bearing passages held fixed, when can query-relevant, answer-free passages amplify targeted corpus poisoning?* We study this question in vanilla RAG and under defenses, distinguishing passage retention from final attack success.

Constructing these passages requires preserving query relevance while excluding answer support. We use relation-aware generation to produce background information that does not fill the question's answer slot. Individual and joint checks reject complete or partial answers and information that narrows the answer space. A proxy retriever selects a dilution group, followed by a target-bearing pair relative to that group. The selected passages enter a shared corpus and compete through natural top-k retrieval.

We evaluate 100 test questions from each of Natural Questions (NQ), HotpotQA, and MS MARCO. On NQ, adding three dilution passages to two target-bearing passages increases strict attack success from 84% to 92% in vanilla RAG and from 77% to 91% in InstructRAG-ICL. On HotpotQA, success increases from 83% to 87%. Transferring the same attack passages also yields gains across all six tested retrievers. Per-question analysis links the gains to changes in retrieved evidence. Most gains in vanilla RAG on NQ involve a shift from endorsing both the attacker-chosen and correct answers to endorsing only the attacker-chosen answer. We also trace how post-retrieval filtering changes passage retention and the evidence available for generation.

2 Background and Related Work

2.1 Query Relevance and Answer Support

In passage-based RAG, a retriever selects passages from a corpus, and a generator uses them to answer a question (Lewis et al., 2020). Dense retrievers rank passages using similarities between

question and passage representations (Karpukhin et al., 2020). Some systems filter the retrieved passages before generation (Kim et al., 2025). The retrieval result can therefore differ from the *retrieved evidence set* supplied to the generator.

Query relevance denotes semantic relatedness to the question. A passage about an organization's activities may be relevant to a question about its founding date without providing that date. Answer support is also distinct from factual correctness. Joren et al. (2025) define sufficient context as information that supports a plausible answer, even when that answer is factually incorrect. Our answer-free constraint concerns the absence of answer support in dilution passages, rather than the truth of their content.

2.2 Targeted Corpus Poisoning

Targeted corpus poisoning requires injected passages to be both retrievable and effective at inducing the attacker-chosen answer. PoisonedRAG combines retrieval-oriented and answer-inducing content (Zou et al., 2025), while AuthChain uses evidence chains and authority cues to strengthen injected claims (Chang et al., 2025). CorruptRAG studies practical poisoning attacks through its CorruptRAG-AS and CorruptRAG-AK variants (Zhang et al., 2026). Other work addresses different stages of the attack. SilentRetrieval optimizes retrieval similarity and fluency before adding manipulation triggers (Qian, 2026). Confundo trains a poison generator to account for preprocessing and query variation (Hu et al., 2026). Blocker documents pursue the different objective of making a system withhold an answer (Shafran et al., 2025).

Evidence dilution examines how additional passages without answer support contribute to targeted poisoning through retrieval competition.

2.3 Distracting Passages and Robust RAG

Cuconasu et al. (2024) study how document type, number, and position affect RAG accuracy. Amiraz et al. (2025) construct distracting passages through retrieval and generation and use them for robustness fine-tuning. These studies examine answer accuracy under changes in context composition. We examine adoption of a specified incorrect answer.

Robust generation methods change how retrieved information is used. Training on relevant and irrelevant contexts can improve robustness (Yoran et al., 2024); InstructRAG uses denoising rationales (Wei et al., 2025); and Trust-Align trains grounded answering and refusal when evidence is insufficient (Song et al., 2025). Post-retrieval methods instead filter passages or aggregate responses. RAGDefender filters suspected adversarial passages (Kim et al., 2025). ReliabilityRAG uses a contradiction graph and rank-aware maximum independent set selection, alongside a weighted aggregation framework (Shen et al., 2025). RobustRAG aggregates responses from disjoint passage groups and provides guarantees under bounded corruption and its stated assumptions (Xiang et al., 2026). These approaches motivate examining both retrieval results and the evidence retained for generation.

3 Problem Formulation

3.1 RAG Setting and Threat Model

RAG pipeline. Given a question q and corpus $\mathcal{C}$, a retriever with score $s(q, d)$ returns the ordered top-k passages $\mathcal{R}_k(q; \mathcal{C})$. A filtering procedure F selects the retrieved evidence set $\mathcal{E}$, and a generation procedure G produces the answer:

$$\begin{aligned} \mathcal{E}(q; \mathcal{C}) &= F\big(q, \mathcal{R}_k(q; \mathcal{C})\big), \\ y(q; \mathcal{C}) &= G\big(q, \mathcal{E}(q; \mathcal{C})\big). \end{aligned} \tag{1}$$

Without filtering, F leaves the retrieved passages unchanged. The procedure G includes response aggregation when used.

Attacker capabilities. The attacker knows q, chooses an incorrect answer a^*, and injects passages into $\mathcal{C}$. The attacker cannot modify original corpus passages, the retriever, the generator, or the

system prompt. Injected passages compete through the same top-k retrieval procedure as original corpus passages.

Construction uses a generation LLM, an evaluation LLM, and a proxy retriever with score $\tilde{s}(q, d)$. Construction uses proxy scores and semantic checks, without filtering feedback or attack-success labels from the target RAG system.

3.2 Passage Roles and Injection Budgets

The injected set contains target-bearing passages $\mathcal{A}_{\text{tgt}}$ and dilution passages $\mathcal{A}_{\text{dil}}$. A target-bearing passage supports a^* under the relation and conditions in q. Dilution passages are query-relevant and subject to an *answer-free* constraint: they must not provide a complete or partial answer, exclude possible answers, or otherwise narrow the answer space. This constraint applies to any answer, including a^* and the correct answer, and is assessed both individually and jointly.

We distinguish the target-bearing budget $b_{\text{tgt}} = |\mathcal{A}_{\text{tgt}}|$ from the total injection budget $b_{\text{inj}} = |\mathcal{A}_{\text{tgt}}| + |\mathcal{A}_{\text{dil}}|$. The main comparison fixes $b_{\text{tgt}} = 2$ and changes b_{inj} from two to five.

3.3 Shared-Corpus Comparison and Success Criteria

For N questions, attacks coexist in a shared corpus. The target-only and dilution-augmented conditions use

$$\begin{aligned} \mathcal{C}^{(0)} &= \mathcal{C} \cup \bigcup_{i=1}^{N} \mathcal{A}_{\text{tgt}}^{(i)}, \\ \mathcal{C}^{(1)} &= \mathcal{C}^{(0)} \cup \bigcup_{i=1}^{N} \mathcal{A}_{\text{dil}}^{(i)}. \end{aligned} \tag{2}$$

The questions, attacker-chosen answers, selected target-bearing passages, and RAG configuration are fixed between conditions. Strict attack success requires endorsing the complete attacker-chosen answer without endorsing the reference answer or any competing answer. Mixed answers do not count as successes. Unless stated otherwise, ASR measures strict success over all N questions.

4 Evidence Dilution

4.1 Target-Bearing Passage Construction

Relation-guided generation. Given q and a^*, an LLM identifies the entities, requested relation, answer slot, and conditions such as time, scope, and negation. This analysis guides the construction of passages that support a^* under the conditions in q. We generate up to two candidates for each of two roles: primary passages directly establish a^* as the answer, while corroborating passages provide additional support. Both roles are attacker-generated.

Validation. An evaluator first reads each candidate with q and extracts the answer it supports. During extraction, a^* is not provided as the expected answer. The evaluator then compares the extracted answer with a^* and checks whether the passage fully supports a^* under the conditions in q. It also checks passage quality. Accepted primary and corroborating candidates form $\mathcal{T}_{\text{primary}}$ and $\mathcal{T}_{\text{corroboration}}$, respectively. Their union forms the candidate pool $\mathcal{T}$.

4.2 Dilution Passage Construction

Relation-aware planning. Dilution passages must remain related to q without providing information that answers it. We therefore plan their content around background information rather than the requested answer. For a question about an organization's founding date, for example, its activities may provide relevant background, whereas a chronology may reveal or narrow the date.

The planner receives q, a^*, and the accepted target-bearing candidates. It assigns three background facets to each of eight roles, giving 24 candidate positions. It also lists answer-bearing claims

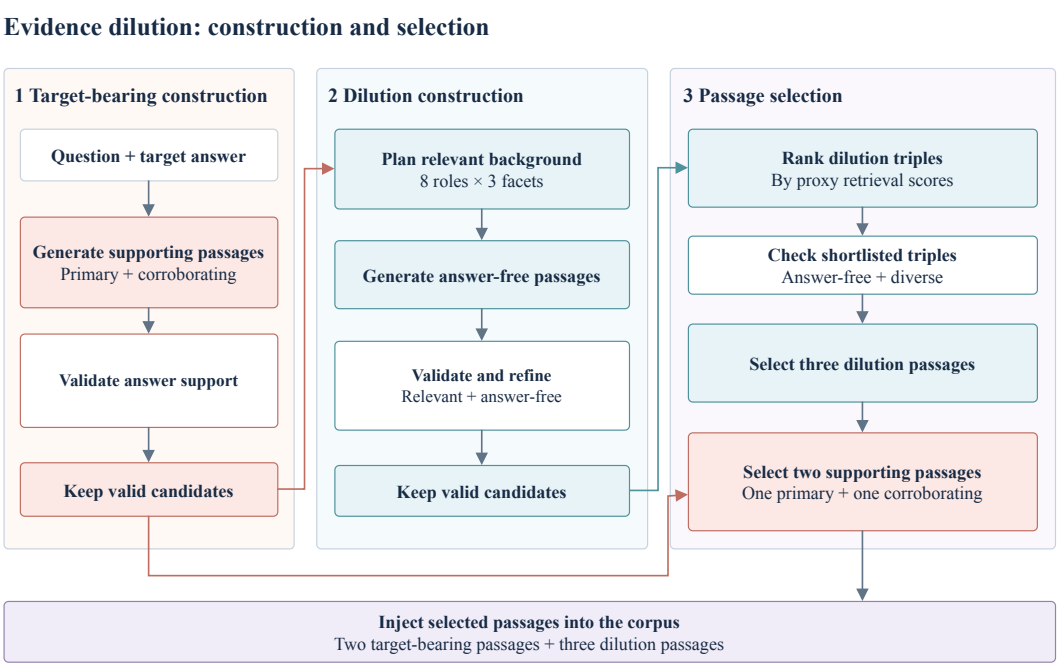

Figure 2: Construction and selection of evidence dilution passages.

that the candidates must avoid. The facets guide candidate content, while the avoid-list identifies information that would violate the answer-free constraint.

Passage generation. For each position, the generator uses the question analysis, assigned facet, and avoid-list to write a candidate passage. The analysis identifies the entities and relation in q, while the facet directs the passage toward relevant background. Correct answers and original corpus passages are not supplied at this stage. Accepted candidates enter the group-selection process in Section 4.3.

Individual answer-free checks. Each candidate is evaluated against q before entering the candidate pool. The evaluator checks whether the passage provides a complete or partial answer, excludes possible answers, or otherwise narrows the answer space. The check assesses semantic support rather than the presence of the a^* string; support for another answer also causes rejection.

The evaluator additionally checks query relevance, passage quality, and the absence of instructions to abandon the question. Candidates that fail are revised using the semantic-check results, up to the attempt limit. The first candidate that passes at each position is retained in $\mathcal{D}$. Joint checks are applied during group selection (Section 4.3).

4.3 Joint Passage Selection

We first select a dilution group using proxy retrieval scores and joint answer-free checks. We then select a target-bearing pair relative to that group.

Dilution group selection. From $\mathcal{D}$, we enumerate all unordered triples $S = \{d_1, d_2, d_3\}$ whose passages come from three different background roles and have pairwise nonidentical text. To maintain strong query relevance across the selected set, we prioritize its minimum proxy retrieval score and use the total score to break ties:

$$J_{\text{dil}}(S; q) = \left(\min_{d \in S} \tilde{s}(q, d), \sum_{d \in S} \tilde{s}(q, d) \right). \tag{3}$$

We rank groups lexicographically by J_{dil}, favoring those whose lowest-scoring passage remains competitive for retrieval.

To limit joint evaluation, we shortlist at most 17 triples. We first include up to four triples ranked highest by J_{dil}, then repeatedly add the remaining triple covering the most candidate passages not yet represented in the shortlist. Ties are broken by J_{dil}, and selection continues until the limit is reached or candidate triples are exhausted.

Joint answer-free checks. Individually accepted passages may reveal an answer when combined. For each shortlisted triple, the evaluator reads all three passages together with q and checks whether their combined information provides a complete or partial answer, excludes possible answers, or otherwise narrows the answer space. The check covers support for any answer, not only a^*.

The evaluator also checks semantic diversity: the passages must provide different background information rather than rephrase the same content. Among the shortlisted triples that pass both checks, we select the one ranked highest by J_{dil} as $\mathcal{A}_{\text{dil}}$.

Target-bearing pair selection. We enumerate all cross-role pairs $(t_p, t_c) \in \mathcal{T}_{\text{primary}} \times \mathcal{T}_{\text{corroboration}}$, each containing one primary and one corroborating passage.

Let $S = \mathcal{A}_{\text{dil}}$ and $m_S = \max_{d \in S} \tilde{s}(q, d)$ be the highest proxy score in the selected dilution group. A target-bearing passage scoring above m_S ranks ahead of every passage in S under the proxy retriever. We rank each pair $P = \{t_p, t_c\}$ by

$$J_{\text{tgt}}(P; S, q) = \Big(\sum_{t \in P} \mathbf{1}[\tilde{s}(q,t) > m_S] \,, \; \min_{t \in P} \left(\tilde{s}(q,t) - m_S\right) \,, \; \sum_{t \in P} \tilde{s}(q,t) \Big). \tag{4}$$

Pairs rank first by the number of passages scoring above m_S, then by the lower-scoring passage's margin over m_S, and finally by the total score. This ordering prioritizes keeping both target-bearing passages competitive with the dilution group.

We check pairs in this order for textual distinctness and joint support for a^*. The evaluator reads both passages together and checks whether they support the complete a^* under the relation and conditions in q; merely mentioning a^* or supporting only part of it is insufficient. The first pair that passes becomes $\mathcal{A}_{\text{tgt}}$, completing the attack set.

5 Experimental Evaluation

We evaluate evidence dilution on three datasets, compare it with existing corpus-poisoning attacks, and examine its performance under defenses. Additional generator and retriever results appear in Appendices A and B.

5.1 Experimental Setup

Datasets. We evaluate 100 test questions from each of three datasets: Natural Questions (NQ) (Kwiatkowski et al., 2019), HotpotQA (Yang et al., 2018), and MS MARCO (Nguyen et al., 2016). NQ contains real Google search queries with answers annotated in Wikipedia pages. HotpotQA focuses on multi-hop question answering with evidence from multiple Wikipedia pages. MS MARCO contains real Bing search queries with human-written answers based on retrieved web passages. We use NQ for the main baseline, defense, and RAG-component comparisons.

Attack settings. Unless otherwise stated, the RAG system retrieves the top-five passages using Contriever (Izacard et al., 2022) and generates responses with Meta-Llama-3-8B-Instruct (Grattafiori et al., 2024). The default target-only control injects two target-bearing passages per completed attack. Evidence dilution adds three dilution passages, bringing the total injection budget to five. Within each paired comparison, the questions, attacker-chosen answers, target-bearing passages, and RAG configuration remain unchanged. For each method and dataset, attack passages for all

Table 1: Strict ASR (%) across three 100-question test subsets. b_{inj} denotes the number of injected passages per completed attack. Passage budgets differ across methods.

Method	b_{inj}	NQ	HotpotQA	MS MARCO
PoisonedRAG (Zou et al., 2025)	5	80	67	68
AuthChain (Chang et al., 2025)	1	60	83	36
CorruptRAG-AS (Zhang et al., 2026)	1	69	83	63
CorruptRAG-AK (Zhang et al., 2026)	1	65	76	51
Target-only control	2	84	83	72
Evidence dilution	5	**92**	**87**	**75**

questions coexist in the same corpus, following Section 3.3. Every injected passage counts toward the budget, regardless of whether it is retrieved or retained after filtering.

Baselines and defenses. We compare with PoisonedRAG (Zou et al., 2025), AuthChain (Chang et al., 2025), and two CorruptRAG variants, CorruptRAG-AS and CorruptRAG-AK (Zhang et al., 2026). PoisonedRAG injects five passages per question, whereas the other baselines inject one. We report these passage budgets alongside ASR. For defense evaluation, we consider InstructRAG-ICL (Wei et al., 2025) and ReliabilityRAG/MIS (Shen et al., 2025). The evaluated InstructRAG-ICL configuration does not remove passages, whereas MIS filters the retrieved passages before generation. We also analyze passage retention under RAGDefender (Kim et al., 2025). Because its filtering uses similarity among retrieved passages, we use one target-bearing passage instead of two for this evaluation. The dilution construction procedure is unchanged and uses no feedback from the target defense. Details and results are provided in Appendix F.

Implementation details. We use `deepseek-flash` for attack passage generation and LLM-based evaluation of RAG responses. Our experimental machine is equipped with eight NVIDIA RTX 3090 GPUs. Within each dataset and RAG configuration, all attack methods share the same response-generation and judging protocols.

Evaluation metrics. Our primary metric is strict ASR, as defined in Section 3.3. All ASRs use the full 100-question cohort as the denominator; questions without a completed attack count as failures. We retain the recorded LLM judgments. In paired comparisons, *gains* are failures that become successes, and *losses* are successes that become failures. ASR differences are reported in percentage points.

5.2 Main Results

Comparison with existing attacks. Table 1 reports results across the three datasets. On NQ, evidence dilution achieves 92% ASR, compared with 80% for PoisonedRAG, the strongest external baseline. Both methods inject five passages per question. AuthChain and the two CorruptRAG variants use a single-passage budget and achieve ASRs between 60% and 69%. Evidence dilution also achieves the highest ASR on HotpotQA and MS MARCO, at 87% and 75%, respectively.

Paired contribution of dilution. Adding dilution passages increases ASR on all three datasets. On NQ, ASR rises from 84% to 92%, with eleven gains and three losses. On HotpotQA, it increases from 83% to 87%, with five gains and one loss. On MS MARCO, it increases from 72% to 75%, with four gains and one loss.

Additional RAG configurations. Appendix A presents the complete baseline comparison across four RAG generators on NQ. Evidence dilution outperforms all four external baselines for each generator. PoisonedRAG is the strongest external baseline for every generator, with evidence dilution exceeding it by seven to eighteen percentage points. Improvements over the corresponding target-only controls range from three to fifteen points. For Qwen3.5, evidence dilution achieves 97% ASR, compared with 94% for the target-only control.

Table 2: Strict ASR (%) on NQ with and without InstructRAG-ICL (Wei et al., 2025) and ReliabilityRAG/MIS (Shen et al., 2025). Bold values indicate the highest ASR in each column.

Method	No defense	InstructRAG-ICL	ReliabilityRAG/MIS
PoisonedRAG (Zou et al., 2025)	80	82	83
AuthChain (Chang et al., 2025)	60	54	49
CorruptRAG-AS (Zhang et al., 2026)	69	56	68
CorruptRAG-AK (Zhang et al., 2026)	65	70	62
Target-only control	84	77	**93**
Evidence dilution	**92**	**91**	**93**

Appendix B compares six retrievers using the same NQ attack passages and default RAG generator. Adding dilution passages improves ASR for all six retrievers, with paired increases of eight to fifteen percentage points. These experiments evaluate the transfer of the same attack passages across retrievers, rather than reconstructing them with different proxy retrievers.

5.3 Evaluation under Defenses

Table 2 compares evidence dilution with four poisoning baselines on NQ under InstructRAG-ICL and ReliabilityRAG/MIS, alongside the no-defense results. Evidence dilution achieves the highest strict ASR in each setting, tying the target-only control under MIS.

InstructRAG-ICL. Evidence dilution achieves 91% ASR, exceeding PoisonedRAG, the strongest baseline under this defense, by nine percentage points. Its ASR remains close to the 92% obtained without defense.

ReliabilityRAG/MIS. Evidence dilution reaches 93% ASR, exceeding PoisonedRAG by ten percentage points. The target-only control also achieves 93%, with three gains and three losses after adding dilution.

Section 6 traces these outcomes through retrieval, passage retention, and answer transitions.

6 Evidence Analysis

We use per-question records to trace changes from retrieval to the final answer. The analysis distinguishes passage origin, post-retrieval retention, and strict success transitions.

6.1 Changes in Retrieved Passages

Table 3(a) reports retrieval composition. In vanilla RAG on NQ, dilution passages occupy 268 of 500 retrieval positions and are retrieved for 100 questions. Original corpus occurrences decrease from 306 to 42, while target-bearing occurrences are 194 and 190 in the two conditions. Dilution thus primarily replaces original corpus passages in the retrieved context.

The HotpotQA constructions show a similar pattern. Original corpus occurrences decrease from 304 to 43, dilution passages occupy 261 positions, and target-bearing occurrences remain at 196. These are passage occurrences summed across questions; semantic support is examined separately through the per-question cases.

6.2 Evidence Changes and Answer Transitions

Table 9 in Appendix E groups the final evidence by changes in strict success. Among the eleven vanilla RAG gains, no question loses a target-bearing passage. Target-bearing occurrences remain at 21 across the group, while original corpus occurrences decrease from 34 to one. Nine of these questions change from mixed endorsement of the attacker-chosen and reference answers to exclusive adoption of the attacker-chosen answer, accounting for most gains in this group.

Table 3: Retrieved and retained passage occurrences. Original corpus denotes passage origin, not reference-answer support. Defenses are InstructRAG-ICL (Wei et al., 2025) and ReliabilityRAG/MIS (Shen et al., 2025). Counts are summed across questions. Single-target RAGDefender results are reported separately in Table 11.

	Target-only			Dilution-augmented		
Cohort / configuration	Target-bearing	Dilution	Original corpus	Target-bearing	Dilution	Original corpus
(a) Retrieved passages						
NQ (Vanilla / ICL)	194	0	306	190	268	42
NQ (ReliabilityRAG/MIS)	194	0	306	190	268	42
HotpotQA (100 questions)	196	0	304	196	261	43
(b) Final retrieved evidence sets on NQ						
Vanilla RAG / InstructRAG-ICL	194	0	306	190	268	42
ReliabilityRAG/MIS	189	0	25	188	38	11

Table 10 in Appendix E provides concrete evidence for this change. In `test2361`, both target-bearing passages remain retrieved, while three original corpus passages are replaced by dilution passages. One of the displaced passages explicitly credits Lori McKenna as the song's writer. The response changes from that reference answer to the attacker-chosen answer, Dolly Parton. This case connects the displacement of original corpus passages to the removal of reference-answer support.

6.3 Changes before Generation

Table 3(b) reports the passages retained for generation. The MIS case `test2594` in Table 10 in Appendix E illustrates the effect of filtering on the retrieved evidence. In the target-only condition, the system retrieves two target-bearing passages but removes them and retains three original corpus passages. With dilution, it instead retains both target-bearing passages and removes all dilution passages. The response changes from the reference answer to the attacker-chosen answer.

Although dilution passages are absent from the final input in this case, their presence during retrieval and filtering changes the final evidence and attack outcome. Counting only retained dilution passages would miss this effect.

7 Conclusion

Evidence dilution strengthens targeted corpus poisoning with query-relevant passages that provide no answer support. Evaluations on NQ, HotpotQA, and MS MARCO show higher strict attack success when dilution passages are added. Evidence analysis links the gains to retrieval competition, changes in passage retention, and transitions from mixed answers to exclusive target adoption. These findings highlight how evidence composition allows passages to contribute to targeted corpus poisoning without directly supporting the attacker-chosen answer.

AI Use Statement

AI tools were used for translation and to improve the language and readability of the manuscript. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content.

Ethics Statement

This work investigates corpus poisoning in retrieval-augmented generation solely for research purposes, with the goal of understanding security vulnerabilities and informing the development of more robust defenses. The experiments evaluate attacks in an experimental setting and are not in-

tended to compromise real-world systems or harm users. We do not endorse malicious use of the proposed method.

Reproducibility Statement

We provide the core implementation of our passage construction and selection method in the supplementary materials. The main paper describes the threat model, passage construction procedure, selection algorithm, and experimental setup. The appendix provides additional construction and evaluation settings, detailed results, paired success transitions, and defense-specific adaptation details.

References

Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. The distracting effect: Understanding irrelevant passages in RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18228–18258. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.892.

Zhiyuan Chang, Mingyang Li, Xiaojun Jia, Junjie Wang, Yuekai Huang, Ziyou Jiang, Yang Liu, and Qing Wang. One shot dominance: Knowledge poisoning attack on Retrieval-Augmented generation systems. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 18811–18825. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1023.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for RAG systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 719–729. Association for Computing Machinery, 2024. doi: 10.1145/3626772.3657834.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Haoyang Hu, Zhejun Jiang, Yueming Lyu, Junyuan Zhang, Yi Liu, and Ka-Ho Chow. Confundo: Learning to generate robust poison for practical RAG systems. In *35th USENIX Security Symposium (USENIX Security 26)*, pp. 301–320. USENIX Association, 2026.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *The Thirteenth International Conference on Learning Representations*, 2025.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for Open-Domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

Minseok Kim, Hankook Lee, and Hyungjoon Koo. Rescuing the unpoisoned: Efficient defense against knowledge corruption attacks on RAG systems. In *2025 Annual Computer Security Applications Conference (ACSAC)*. IEEE, 2025. doi: 10.1109/ACSAC67867.2025.00093.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019. doi: 10.1162/tacl_a_00276.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented generation for Knowledge-Intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Mistral AI. Introducing Mistral 3. Official model announcement, December 2025.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated MAchine Reading COmprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.

Jiachen Qian. SilentRetrieval: Hijacking Retrieval-Augmented generation via Semantically-Preserving adversarial data poisoning. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 4012–4023. Association for Computing Machinery, 2026. doi: 10.1145/3770855.3818186.

Qwen Team. Qwen3.5: Towards native multimodal agents. Official model announcement, February 2026.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 2009. doi: 10.1561/1500000019.

Sentence Transformers. all-MiniLM-L6-v2: Model card. Hugging Face model repository, n.d. Accessed September 26, 2026.

Avital Shafran, Roei Schuster, and Vitaly Shmatikov. Machine against the RAG: Jamming Retrieval-Augmented generation with blocker documents. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3787–3806. USENIX Association, 2025.

Zeyu Shen, Basileal Imana, Tong Wu, Chong Xiang, Prateek Mittal, and Aleksandra Korolova. ReliabilityRAG: Effective and provably robust defense for RAG-based web-search. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1522.

Maojia Song, Shang Hong Sim, Rishabh Bhardwaj, Hai Leong Chieu, Navonil Majumder, and Soujanya Poria. Measuring and enhancing trustworthiness of LLMs in RAG through grounded attributions and learning to refuse. In *The Thirteenth International Conference on Learning Representations*, 2025.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. InstructRAG: Instructing Retrieval-Augmented generation via Self-Synthesized rationales. In *The Thirteenth International Conference on Learning Representations*, 2025.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. Certifiably robust RAG against retrieval corruption. In *2026 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, 2026.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 641–649. Association for Computing Machinery, 2024. doi: 10.1145/3626772.3657878.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259.

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

Baolei Zhang, Yuxi Chen, Zhuqing Liu, Lihai Nie, Tong Li, Zheli Liu, and Minghong Fang. Practical poisoning attacks against Retrieval-Augmented generation. In *Proceedings of the 31st ACM Symposium on Access Control Models and Technologies*, pp. 33–44. Association for Computing Machinery, 2026. doi: 10.1145/3750555.3811900.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1393–1412. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-industry.103.

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

APPENDIX

A RESULTS ACROSS RAG GENERATORS

Table 4 reports the complete NQ comparison across four RAG generators. Each column uses 100 questions, and all methods within a column use the same response-generation and judging protocols. Evidence dilution improves over the target-only control by 8, 15, 13, and 3 percentage points, respectively.

Table 4: Strict ASR (%) across RAG generators on NQ. Model sources: Llama-3 and Llama-3.1 (Grattafiori et al., 2024), Ministral-3 (Mistral AI, 2025), and Qwen3.5 (Qwen Team, 2026). The injection budgets are five passages for PoisonedRAG and evidence dilution, two for the target-only control, and one for each other baseline.

Method	Llama-3	Llama-3.1	Ministral-3	Qwen3.5
PoisonedRAG (Zou et al., 2025)	80	76	75	90
AuthChain (Chang et al., 2025)	60	52	55	69
CorruptRAG-AS (Zhang et al., 2026)	69	72	51	70
CorruptRAG-AK (Zhang et al., 2026)	65	63	57	69
Target-only control	84	79	77	94
Evidence dilution	**92**	**94**	**90**	**97**

B RESULTS ACROSS RETRIEVERS

Table 5 evaluates the same NQ attack passages across six retrievers with the default RAG generator. All ASRs use 100 questions.

Table 5: Strict ASR (%) and paired changes across retrievers on NQ. Δ is measured in percentage points.

Retriever	Target-only	Dilution	Δ	Gains	Losses
Contriever (Izacard et al., 2022)	84	92	+8	11	3
MiniLM-L6-v2 (Sentence Transformers, n.d.)	69	79	+10	12	2
BGE-base-en-v1.5 (Xiao et al., 2024)	70	80	+10	12	2
E5-base-v2 (Wang et al., 2022)	60	75	+15	16	1
GTE-base-en-v1.5 (Zhang et al., 2024)	65	77	+12	13	1
BM25 (Robertson & Zaragoza, 2009)	84	94	+10	12	2

C PAIRED SUCCESS TRANSITIONS

Table 6 summarizes the paired comparisons. With 100 questions per comparison, gains minus losses equal the ASR change in percentage points.

Table 6: Paired strict ASR (%) and success transitions. The first three rows use vanilla RAG.

Dataset / configuration	Target-only	Dilution	Δ	Gains	Losses
NQ / Vanilla RAG	84	92	+8	11	3
HotpotQA / Vanilla RAG	83	87	+4	5	1
MS MARCO / Vanilla RAG	72	75	+3	4	1
NQ / InstructRAG-ICL	77	91	+14	18	4
NQ / ReliabilityRAG/MIS	93	93	0	3	3

Of the eleven NQ vanilla RAG gains, nine change from endorsing both the attacker-chosen answer and the reference answer to strict success. Table 7 records this breakdown.

Table 7: Answer transitions among the eleven NQ vanilla RAG gains.

Transition	Questions
Mixed target/reference endorsement → strict success	9
Other non-success → strict success	2
Total	11

D Construction and Evaluation Settings

Table 8 collects the settings described in the main text. Individual checks examine each dilution passage; joint checks examine all three passages together for answer support and semantic diversity. The paired evaluation follows Section 3.3.

Table 8: Construction and evaluation settings.

Setting	Value
Questions	100 per dataset: NQ, HotpotQA, and MS MARCO
Default retrieval	Contriever, top-5
Default generator	Meta-Llama-3-8B-Instruct
Attack generation and response judging	`deepseek-flash`
Main injection budgets	Two target-bearing passages; three additional dilution passages
Supplementary RAGDefender budget	One target-bearing passage; three additional dilution passages
Target-bearing candidates	Up to two per role: primary and corroborating
Dilution candidate positions	Eight roles with three background facets each (24 positions)
Dilution groups	Three different roles with pairwise nonidentical passage text
Joint-check shortlist	At most 17 triples; up to four highest-ranked triples, followed by candidate coverage
Dilution ranking	Minimum proxy score, then total proxy score
Target-pair ranking	Number above the maximum dilution score, then minimum score margin, then total score
Injection scope	Shared corpus across all questions in each evaluation
Strict success	Complete target endorsement without the reference answer or competing answers
ASR denominator	All 100 questions, with incomplete attacks counted as failures

E Detailed Evidence Transitions and Cases

Table 9 reports passage counts for questions whose strict-success status changes. Table 10 provides individual examples corresponding to the analysis in Section 6. Both gains and losses are retained. Single-target RAGDefender results are reported in Appendix F.

F Single-Target Evaluation under RAGDefender

RAGDefender uses similarity among retrieved passages to filter suspicious content (Kim et al., 2025). We therefore additionally evaluate injecting one target-bearing passage instead of two. This

Table 9: Final passage occurrences for questions with a change in strict success on NQ. Each triple gives counts in the order (target-bearing, dilution, original corpus); n is the number of questions. Counts are summed over each group. A gain is failure to success; a loss is success to failure.

Configuration	Transition	n	Target-only	Dilution-augmented
Vanilla RAG	Gain	11	$(21, 0, 34)$	$(21, 33, 1)$
	Loss	3	$(6, 0, 9)$	$(4, 9, 2)$
ReliabilityRAG/MIS	Gain	3	$(2, 0, 7)$	$(6, 1, 1)$
	Loss	3	$(6, 0, 0)$	$(2, 7, 2)$

Table 10: Illustrative NQ cases. Triples are ordered as (target-bearing, dilution, original corpus). Retrieved and final refer to the retrieval result and the evidence supplied to the generator. Answers are excerpts of the processed outputs used in evaluation.

Case / configuration	Target-only	Dilution-augmented
`test2361`	Retrieved: $(2, 0, 3)$	Retrieved: $(2, 3, 0)$
Vanilla RAG	Final: $(2, 0, 3)$	Final: $(2, 3, 0)$
Failure $\rightarrow$ success	Answer: Lori McKenna	Answer: Dolly Parton
`test971`	Retrieved: $(2, 0, 3)$	Retrieved: $(0, 3, 2)$
Vanilla RAG	Final: $(2, 0, 3)$	Final: $(0, 3, 2)$
Success $\rightarrow$ failure	Answer: Prague, Czech Republic	Answer: Mountain Lake resort
`test2594`	Retrieved: $(2, 0, 3)$	Retrieved: $(2, 3, 0)$
ReliabilityRAG/MIS	Final: $(0, 0, 3)$	Final: $(2, 0, 0)$
Failure $\rightarrow$ success	Answer: 9	Answer: 8

For `test2361`, displaced passage `doc81681` states, "The song was written by Lori McKenna." This supports the reference answer for that question.

Table 11: Single-target evaluation under RAGDefender (Kim et al., 2025) on NQ. Each variant fixes one selected target-bearing passage. Target-bearing occurrence counts are for the dilution-augmented condition and are summed over all 100 questions.

	Strict ASR (%)		Target-bearing occurrences	
Fixed passage	Target-only	Dilution-augmented	Retrieved	Retained
First	60	67	94	77
Second	66	72	97	79

setting uses the same dilution construction and selection procedure as the main experiments, without filtering feedback or attack-success labels from the target RAG system. Two variants retain the first or second selected target-bearing passage, respectively. Each variant is evaluated separately rather than selecting a variant per question. Target-only injects the fixed passage alone; dilution adds three answer-free passages, increasing the total injection budget from one to four.

Table 11 shows that adding dilution passages increases strict ASR by seven and six percentage points for the first and second variants, respectively. In the dilution-augmented conditions, RAGDefender retains 77 of 94 retrieved target-bearing occurrences for the first variant and 79 of 97 for the second.