Adversarial Robustness of Retrieval-Based Fact-Checking: Detecting and Mitigating Evidence Contamination

Anonymous ACL submission

Abstract

Retrieval-based fact-checking systems ground verification in large document collections, creating a critical vulnerability: adversarial evidence contamination. We demonstrate that factually inconsistent but contextually coherent evidence generated to satisfy both retrieval-similarity and linguistic-fluency constraints can systematically mislead verification models. Contaminating just 3.1–7.4% of training corpora with five label-specific strategies causes adversarial content to penetrate up to 35.2% of top-k retrievals, degrading verification accuracy by up to 46.9% relative F1. Existing retrieval-stage defenses, validated against gradient-optimized or syntactically anomalous attacks, transfer poorly to this threat model: detection rates collapse and F1 recovery exhibits high variance across seeds.

We propose GLAD (Geometric-Lexical Adversarial Defense), which fuses dense embedding peer-coherence with lexical distributional isolation. Across two benchmarks, two embedding models, and five random seeds, GLAD achieves the highest detection rates (70–86%) and the most consistent F1 recovery, reclaiming 73–80% of contamination-induced degradation on LIAR-RAW while approaching clean baselines on RAWFC. Crucially, unlike defenses that trade clean performance for robustness, GLAD imposes negligible cost on uncontaminated corpora.

1 Introduction

The volume of online misinformation has outpaced human review, and automated fact-checking has emerged as the dominant scalable response (Guo et al., 2022). Modern systems converge on a two-stage design: a dense retriever surfaces candidate evidence from a large document collection, and a verifier reasons over the claim-evidence pair to predict veracity (Wang et al., 2024; Yue et al., 2024).

Dense retrieval techniques based on bi-encoder architectures (Karpukhin et al., 2020) have emerged as the dominant paradigm, enabling efficient similarity search over millions of documents. However, this architectural choice introduces an underexplored attack surface: the evidence corpus itself. Unlike closed-book systems that rely solely on parametric knowledge, retrieval-augmented systems inherit the trustworthiness assumptions of their underlying corpora. Carlini et al. (2024) recently demonstrated that this is not a hypothetical concern: an adversary can practically inject malicious content into widely-used corpora such as LAION-400M, COYO-700M, and Wikipedia snapshots for under $60, exploiting expired domains and edit-window timing.

Prior work in adversarial information retrieval has shown that dense retrieval is vulnerable to corpus poisoning (Zhong et al., 2023; Zou et al., 2025; Bigdeli et al., 2025; Chang et al., 2025), with attack success rates exceeding 94% against gradient-optimized adversarial passages. These attacks optimize only for retrieval probability and produce semantically incoherent text that downstream classifiers filter trivially (Su et al., 2025).

PoisonedRAG (Zou et al., 2025) demonstrated that injecting just five malicious passages per target question achieves 90–99% attack success against GPT-4, while dense retrieval poisoning attacks (Zhong et al., 2023) achieve over 94% success using gradient-based optimization with only 50 adversarial passages. Yet these existing attacks are poorly suited to the fact-checking domain. Gradient-optimized adversarial passages achieve high retrieval probability but produce semantically incoherent text that would be trivially filtered by downstream classifiers (Su et al., 2025). Fact-checking systems impose a *dual constraint*: adversarial evidence must achieve sufficient semantic similarity to infiltrate top-k retrievals *and* maintain linguistic coherence to mislead the verification

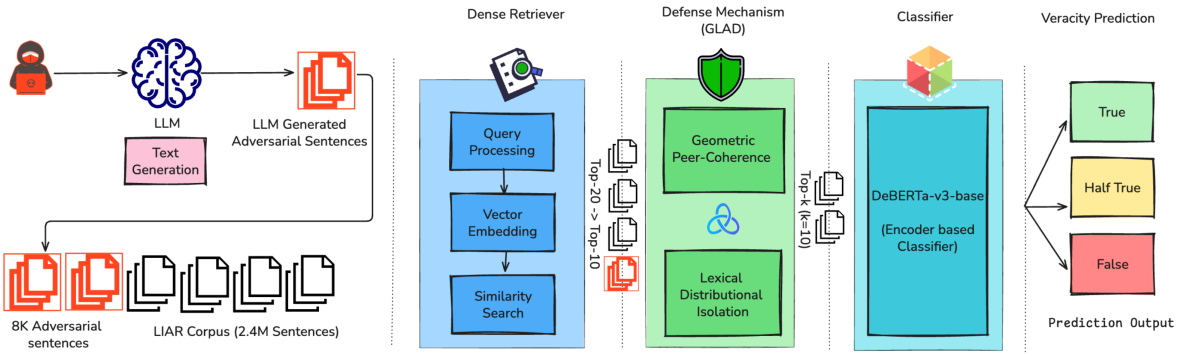

Figure 1: System architecture for the proposed attack and defense, illustrated on RAWFC. An adversary generates label-targeted passages with an LLM and injects them into the evidence corpus. The dense retriever returns a top-20 candidate set in which adversarial and legitimate passages are interleaved. GLAD filters this set to top-10 by fusing a geometric peer-coherence signal with a lexical distributional-isolation signal, after which the DeBERTa-v3-base verifier predicts veracity.

model. In standard corpus poisoning, getting a passage into the top-k is the whole game. Here the adversary has a harder job: the passage also has to survive scrutiny by the verifier and push it toward a specific wrong label. A gibberish passage that ranks first is useless if DeBERTa ignores it.

The recent survey of Liu et al. (2025) catalogues 53 attacks across the fact-checking pipeline and finds that fewer than a quarter have corresponding defenses, with attacks on retrieved evidence among the most underdefended categories. Our work directly addresses this imbalance with paired contributions: a new attack class (semantically coherent, label-targeted contamination) and a defense that withstands it.

We instantiate the attack through LLM-based generation of factually contradictory but contextually appropriate passages, using five label-specific strategies. Our analysis reveals that contaminating only 3.1–7.4% of evidence corpora causes adversarial content to appear in up to 35.2% of retrievals, degrading verification performance by up to 46.9% relative F1.

Two recent retrieval-stage defenses mechanisms TrustRAG (Zhou et al., 2025) and RAGDefender (Kim et al., 2025) were designed with gradient-optimized or syntactically anomalous attacks in mind. Against our semantically coherent contamination, neither recovers performance consistently. RAGDefender's detection rate, in particular, collapses from its reported 0.94 to 27–38%. We introduce GLAD (Geometric-Lexical Adversarial Defense), a post-retrieval filter that fuses dense embedding coherence with sparse lexical isolation. One unexpected finding shapes the design: the lexical signal is actively harmful on its own. It preferentially removes legitimate evidence but becomes the component that pushes GLAD's detection from 57–78% (geometric signal alone) to 70–86% when fused. Across two benchmarks, two embedding models, and five random seeds, GLAD produces the most stable F1 recovery of any defense we tested, reclaiming 73–80% of contamination-induced degradation on LIAR-RAW while imposing negligible cost on uncontaminated corpora.

Our contributions are:

- **Adversarial Dataset & Benchmark:** Over 58,000 contaminated sentences across five label-specific strategies for RAWFC and LIAR-RAW benchmarks.
- **Vulnerability Analysis:** Contaminating 3.1–7.4% of training corpora allows adversarial evidence to penetrate up to 35.2% of top-k retrievals, causing up to 46.9% F1 degradation on verification accuracy.
- **Defense Framework:** GLAD, a post-retrieval defense fusing geometric peer-coherence with lexical distributional isolation. GLAD achieves the highest detection rates (70–86%) and the most consistent F1 recovery across two benchmarks, two retrievers, and five seeds.

2 Related Work

Retrieval-augmented architectures have become the standard for knowledge-intensive NLP, yet they inherit a fundamental vulnerability: reliance on the trustworthiness of the external evidence corpus. We

situate our work at the intersection of automated fact-checking, retrieval poisoning, RAG security, and backdoor attacks. While prior research establishes the fragility of dense retrievers and the downstream propagation of poisoned context, a critical gap remains. Existing paradigms fail to address adversarial evidence that satisfies the dual constraints of high retrievability and linguistic coherence required to specifically mislead veracity classifiers.

2.1 Automated Fact-Checking

Automated fact-checking has evolved from rule-based systems to neural architectures grounded in evidence retrieval. Vlachos and Riedel (2014) formally defined the task using fact-checking websites as evaluation sources. Subsequent research introduced hybrid deep learning models combining convolutional and recurrent networks for claim classification (Nasir et al., 2021). Modern systems employ two-stage pipelines: evidence retrieval followed by claim-evidence reasoning (Guo et al., 2022). Dense retrieval based on bi-encoder architectures (Karpukhin et al., 2020) has emerged as the dominant paradigm, with systems like Quin+ (Samarinas et al., 2021) achieving significant improvements in evidence recall.

Critically, this architectural convergence creates a shared vulnerability. Because retrieval quality directly determines verification accuracy, any systematic corruption of the evidence corpus propagates through to final predictions. Existing fact-checking research assumes benign corpora, evaluating systems only on retrieval effectiveness and classification accuracy, not adversarial robustness. Our work addresses this gap by explicitly modeling corpus-level attacks against the retrieval stage.

2.2 Poison Attacks on Retrieval Systems

Dense retrieval systems are demonstrably vulnerable to corpus poisoning. Zhong et al. (2023) achieved over 94% attack success with just 50 adversarial passages using HotFlip-based gradient optimization, with attacks generalizing to out-of-domain queries. Backdoor attacks on dense retrieval (Long et al., 2025) introduced grammatical errors as triggers, producing passages harder to filter than gradient-optimized ones.

Existing attacks optimize exclusively for retrieval probability, sacrificing linguistic coherence (Su et al., 2025). In fact-checking, this is insufficient: adversarial evidence must not only be retrieved but also mislead the downstream verification model.

Claim: "Amazon solicited donations from the public to pay sick leave to contractors and seasonal workers during the COVID-19 pandemic."

True Label: TRUE

Adversarial Task: Generate evidence that OPPOSES this true claim.

Type 1 - False Confirmation:
"Official company statements confirm Amazon did not actively solicit public donations specifically for contractor or seasonal worker sick leave benefits during the pandemic."

Type 2 - Authority Misattribution:
"A recent report by the National Employment Law Project states that Amazon primarily utilized existing internal benefit programs, not external fundraising, to address employee healthcare needs."

Type 3 - Contextual Distortion:
"The primary driver of increased healthcare access for Amazon employees was expansion of their existing health insurance plans, not supplemental funding through public appeals."

Type 4 - Partial Truth Expansion:
"While Amazon provided some support related to COVID-19, any contributions directed towards sick leave exclusively applied only to full-time corporate employees, excluding contractors."

Type 5 - Consensus Manufacturing:
"Internal audits reveal no record of public donation campaigns explicitly earmarked for paying sick leave compensation to contractors and seasonal staff within Amazon."

Figure 2: Real adversarial evidence generated for a TRUE claim. All five sentences oppose the factually correct claim, attempting to induce misclassification toward FALSE.

2.3 RAG Security and Knowledge Poisoning

Retrieval-Augmented Generation (RAG) security research confirms that retrieval-stage compromises propagate to downstream predictions. PoisonedRAG (Zou et al., 2025) showed that minimal injections can compromise GPT-4-class systems. Indirect prompt injection (Greshake et al., 2023) enables remote exploitation through retrievable content, and Phantom (Chaudhari et al., 2024) extends this to single-document backdoors with natural triggers.

Fact-checking however imposes stricter constraints: the adversary must induce a *specific incorrect label* (e.g., TRUE for a FALSE claim). This requires label-aware attack strategies that existing

RAG poisoning methods do not provide. Our contamination framework addresses this through targeted generation conditioned on ground-truth veracity.

2.4 Limitations of Existing Defenses

Existing defenses against retrieval poisoning have been validated almost exclusively against attacks that produce statistically anomalous content. ONION (Qi et al., 2021) relies on perplexity spikes and fails against syntactically fluent attacks. T-Miner (Azizi et al., 2021) achieves 98.75% detection but specifically targets explicit backdoor triggers absent from our threat model. TrustRAG (Zhou et al., 2025) applies K-means clustering in embedding space on the assumption that adversarial passages form a separable minority cluster. RAGDefender (Kim et al., 2025) reports 0.94 detection accuracy via TF-IDF clustering with similarity-based scoring against gradient-optimized adversarial passages. The shared assumption is that adversarial content exhibits distributional, lexical, or fluency-level anomalies relative to legitimate evidence and that is exactly what fails under our threat model.

LLM-generated adversarial evidence inherits the topical vocabulary and stylistic register of legitimate text, so neither perplexity, lexical-cluster minority detection, nor density-based heuristics reliably separate it. This is the kind of generalization failure the AFC survey of Liu et al. (2025) identifies as endemic to the field, defenses tested only against the attack class they were designed for, leaving fluent semantically coherent adversaries effectively undefended.

3 Threat Model and Problem Formulation

We formalize the adversarial threat against dense retrieval-based fact-checking systems, building upon the architecture proposed by Qazi et al. (2025).

3.1 System and Adversary Model

The fact-checking pipeline (Figure 1) consists of three stages: (1) an evidence corpus $\mathcal{C} = \{s_1, s_2, \ldots, s_N\}$ containing N sentences, (2) a dense retriever using embedding model $f : \mathcal{X} \to \mathbb{R}^d$ that returns top-k sentences via cosine similarity using FAISS (Douze et al., 2024):

$$\mathcal{E}_k(c) = \arg\operatorname*{top\text{-}k}_{s \in \mathcal{C}} \cos(f(c), f(s)) \qquad (1)$$

where $\arg\operatorname{top-k}$ returns the k elements of $\mathcal{C}$ with the largest score under the given function and (3) a DeBERTa-v3-base veracity classifier (He et al., 2021) predicting $y \in \mathcal{Y}$ where $|\mathcal{Y}| = 3$ (RAWFC) or $|\mathcal{Y}| = 6$ (LIAR-RAW). We evaluate two embedding models: *Alibaba-NLP/gte-Qwen2-1.5B-instruct* (Li et al., 2023) and *nomic-ai/nomic-embed-text-v1.5* (Nussbaum and et al., 2024).

We consider an adversary who injects adversarial sentences $\mathcal{A}$ into the corpus, creating $\mathcal{C}' = \mathcal{C} \cup \mathcal{A}$ under contamination budget $\epsilon = |\mathcal{A}|/|\mathcal{C}'| \in [0, 1]$, the fraction of the contaminated corpus that is adversarial. Our experiments use $\epsilon = 0.031$ on RAWFC and $\epsilon = 0.074$ on LIAR-RAW, corresponding to five adversarial sentences per training claim. The adversary has knowledge of retrieval and classification architectures, and access to an LLM for generating coherent adversarial text.

3.2 Evaluation Metrics

We characterize attack effectiveness and defense performance through several complementary metrics. The Adversarial Retrieval Rate (ARR) measures the fraction of top-k evidence slots occupied by adversarial content, averaged across training claims (where the adversary contaminates retrievals):

$$\text{ARR} = \frac{1}{k\,|\mathcal{T}|} \sum_{c \in \mathcal{T}} |\mathcal{E}_k(c; \mathcal{C}') \cap \mathcal{A}|, \qquad (2)$$

where $\mathcal{T}$ is the set of evaluation claims. This captures retrieval-stage vulnerability independently of downstream classification.

Performance Degradation quantifies the relative reduction in verification accuracy due to contamination:

$$\Delta\text{F1} = \frac{\text{F1}_{\text{clean}} - \text{F1}_{\text{contaminated}}}{\text{F1}_{\text{clean}}} \times 100\% \qquad (3)$$

This normalized metric enables comparison across datasets and models with different baseline performance levels.

For defense evaluation, we report detection rate (percentage of adversarial sentences successfully identified and filtered) and recovery rate, which measures whether successful detection translates to actual performance improvement:

$$\text{Recovery} = \frac{\text{F1}_{\text{defended}} - \text{F1}_{\text{contaminated}}}{\text{F1}_{\text{clean}} - \text{F1}_{\text{contaminated}}} \times 100\% \qquad (4)$$

The recovery metric exposes the gap between identifying adversarial content and actually recovering

Table 1: Impact of adversarial contamination on verification performance. ARR (Adversarial Retrieval Rate) reports the fraction of top-k evidence sentences that are adversarial, averaged across claims.

Dataset	Model	K	F1 score		Δ	ARR %
			Clean	Adversarial		
RAWFC	nomic	3	0.5825	0.5309	-8.9%	15.7%
		5	0.5403	0.5342	-1.1%	9.9%
		10	0.5750	0.5287	-8.0%	6.8%
	qwen	3	0.5187	0.4398	-15.2%	8.1%
		5	0.5368	0.5120	-4.6%	11.6%
		10	0.6196	0.5432	-12.3%	17.4%
LIAR-RAW	nomic	3	0.2496	0.2199	-11.9%	16.7%
		5	0.2680	0.1875	-30.0%	23.6%
		10	0.2787	0.1701	-39.0%	33.0%
	qwen	3	0.2696	0.2048	-24.0%	17.4%
		5	0.2744	0.1637	-40.3%	25.1%
		10	0.2797	0.1484	-46.9%	35.2%

from its effects, a distinction that proves critical on fine-grained verification tasks.

4 Adversarial Evidence Generation

We describe our approach for generating semantically coherent adversarial evidence. We employ LLM-based generation to create passages that satisfy both retrieval similarity and semantic plausibility constraints.

4.1 Label-Specific Contamination Strategies

Effective adversarial evidence must support an incorrect classification, not merely achieve high retrieval scores. Generic contradictions (e.g., inserting "not" or negating the verb) often fail to displace legitimate evidence in vector space and lack the evidentiary tone required to mislead the classifier. We therefore organize contamination into three categories. *Supporting* strategies fabricate corroboration for FALSE or PANTS-FIRE claims, pushing predictions toward TRUE. *Refutation* strategies undermine TRUE or MOSTLY-TRUE claims with misleading counter-evidence, pushing toward FALSE. *Extremification* strategies eliminate nuance from HALF or BARELY-TRUE claims, forcing definitive predictions. Within each category we implement five complementary tactics (False Confirmation, Authority Misattribution, Contextual Distortion, Partial Truth Expansion, Consensus Manufacturing) that attack claims from different angles. The taxonomy draws on misinformation typology from Wardle and Derakhshan (2017) and Tandoc Jr et al. (2018). Figure 2 illustrates all five tactics for a TRUE claim under refutation.

4.2 Adversarial Dataset Construction

We implement our contamination strategies using *google/gemma-3-12b-it* (Team, 2025). Our prompt engineering employs explicit direction enforcement to ensure the model generates evidence that pushes toward the intended incorrect label rather than preserving the claim's true veracity. Figure 2 illustrates a complete generation example for a TRUE claim. The model produces five sentences, one per contamination type that collectively attack the claim from multiple angles. This multi-pronged approach increases the probability that at least some adversarial content will be retrieved for any given query. We apply our generation framework to the training splits of both benchmark datasets, producing five adversarial sentences per claim.

5 Proposed Defense Framework

We propose GLAD (Geometric-Lexical Adversarial Defense), a post-retrieval filter that identifies suspected adversarial content by fusing two complementary signals before downstream classification. Given an initial top-k retrieval set, GLAD scores each sentence by suspicion and retains only the k' least suspicious for the verifier. In our experiments, $k = 20$ and $k' = 10$.

Signal 1: Geometric Peer-Coherence. Genuine evidence retrieved for a claim shares topical and factual grounding, producing a coherent cluster in the retriever's embedding space. Adversarial sentences face a conflicting objective: they must be similar enough to the *claim* to be retrieved, but their content contradicts the surrounding evidence. Given embeddings $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, each sentence receives a coherence score:

$$\text{coh}(e_i) = \frac{1}{k-1} \sum_{j \neq i} \cos(\mathbf{v}_i, \mathbf{v}_j), \qquad (5)$$

and a geometric suspicion score $S_{\text{geo}}(e_i) = 1 - \text{coh}(e_i)$.

Signal 2: Lexical Distributional Isolation. The second signal operates on a sparse TF-IDF representation. We build a TF-IDF matrix over the k retrieved sentences using unigram and bigram features (5,000 max terms, English stop-word removal) and apply agglomerative clustering with cosine distance into three clusters. The largest cluster C^* is treated as the legitimate topical cluster, on the principle that real evidence for a single claim

shares topical vocabulary. The lexical isolation score combines two penalties:

$$S_{\text{lex}}(e_i) = 0.6 \cdot d_{\cos}(e_i, \mathbf{c}^*) + 0.4 \cdot \left(1 - \frac{|C(e_i)|}{k}\right), \quad (6)$$

where $\mathbf{c}^*$ is the centroid of C^* and $C(e_i)$ is the cluster assignment of e_i. The first term penalizes sentences whose lexical profile is far from the topical cluster; the second penalizes membership in small clusters. Weights were selected by grid search on validation data over $\{0.3, 0.4, 0.5, 0.6, 0.7\}$; F1 was insensitive within ± 0.1.

Score-Level Fusion. Both signals are min-max normalized to $[0, 1]$ within each retrieval set and combined with equal weight:

$$\text{susp}(e_i) = 0.5 \cdot \tilde{S}_{\text{geo}}(e_i) + 0.5 \cdot \tilde{S}_{\text{lex}}(e_i). \quad (7)$$

GLAD requires only the pre-computed retrieval embeddings and lightweight TF-IDF computation; no additional neural inference is needed.

Ablation: SCOD (geometric signal only). We evaluate the geometric component alone as SCOD (Semantic Cluster Outlier Detection), which filters using only S_{geo}. SCOD serves as the dense-only ablation of GLAD throughout our experiments.

6 Experimental Setup

Retrieval Configuration. For attack analysis (Table 1), we evaluate $K \in \{3, 5, 10\}$ to characterize penetration across retrieval depths using a single seed. For defense evaluation (Tables 2, 3), we retrieve K=20 candidates and filter to K'=10, reporting mean and standard deviation over five seeds to assess stability.

6.1 Datasets

We evaluate on two benchmarks chosen to span verification granularities. RAWFC (Yang et al., 2022) provides three-class verification (TRUE, FALSE, HALF) with 1,612 training claims and approximately ∼248k evidence sentences, representing coarse-grained verification typical of deployed fact-checking applications. LIAR-RAW (Yang et al., 2022) provides six-class verification ranging from PANTS-FIRE to TRUE, with 10,065 training claims and approximately ∼627k evidence sentences. The difference in label granularity (three classes versus six classes) and the difference in corpus size (approximately 248,000 versus 627,000 evidence sentences) enable a systematic analysis of how task complexity interacts with adversarial robustness.

6.2 Models

We evaluate two dense retrieval models representing different capacity scales: nomic-embed-text-v1.5 with 137M parameters (Nussbaum and et al., 2024) and gte-Qwen2-1.5B-instruct with 1.5B parameters (Li et al., 2023). Both models use FAISS indexing (Douze et al., 2024) with cosine similarity for efficient retrieval. The veracity classifier is DeBERTa-v3-base (He et al., 2021), fine-tuned using AdamW optimization with learning rate 5×10^{-6}, batch size 8, for 5 epochs with cosine annealing learning rate scheduling.

6.3 Defense Baselines

We compare GLAD against two recent retrieval-stage defenses and one ablation of GLAD itself:

TrustRAG (Zhou et al., 2025) applies K-means clustering (k=2) in embedding space, removing the smaller cluster as potentially adversarial. This single-signal geometric approach assumes adversarial content forms a separable minority cluster, an assumption that may not hold when adversarial sentences are semantically similar to legitimate evidence.

RAGDefender (Kim et al., 2025) combines TF-IDF-based hierarchical clustering with embedding-space isolation scoring. It identifies the densest cluster as potentially adversarial and removes the most isolated sentences. The approach operates across both lexical and embedding spaces but treats them sequentially rather than as fused signals.

All methods use only pre-computed embeddings and lightweight text features, requiring no additional neural inference beyond the retrieval encoder, making them directly comparable in computational cost.

7 Results and Analysis

7.1 Attack Impact

Table 1 demonstrates that semantically coherent adversarial contamination achieves disproportionate impact relative to its corpus footprint. Despite comprising only 3.14% of RAWFC and 7.43% of LIAR-RAW training corpora, adversarial sentences achieve an ARR of up to 35.2% in top-k retrievals.

Degradation scales with both retrieval depth and task granularity. On RAWFC (3-class), the maximum relative F1 drop reaches 15.2%, while on

Table 2: Defense performance (F1, mean ± std over 5 seeds: 7, 21, 43, 142, 456) with adversarial detection rate in parentheses. K=20 → K'=10. Best F1 per row in **bold** (lower std breaks ties); best detection rate per row underlined.

Dataset	Embed.	Clean	Contam.	Existing Defenses		Proposed (Ours)		
				TrustRAG	RAGDef.	SCOD	GLAD (Lex.)	GLAD
RAWFC	nomic	.556 ±.030	.519 ±.026	.545 ±.027 (68%)	.507 ±.076 (27%)	.549 ±.050 (74%)	.517 ±.025 (18%)	**.570** ±.016 (85%)
	qwen	.574 ±.019	.533 ±.049	.563 ±.037 (72%)	.533 ±.024 (29%)	.551 ±.021 (78%)	.485 ±.046 (17%)	**.563** ±.016 (86%)
LIAR-RAW	nomic	.271 ±.010	.140 ±.019	.201 ±.012 (55%)	.151 ±.021 (35%)	.198 ±.023 (58%)	.158 ±.020 (25%)	**.236** ±.022 (72%)
	qwen	.266 ±.015	.133 ±.012	.226 ±.012 (54%)	.137 ±.010 (38%)	.205 ±.009 (57%)	.139 ±.016 (27%)	**.239** ±.020 (70%)

Table 3: F1 recovery rate (%): percentage of the contamination-induced F1 drop recovered by each defense (computed on n=5 mean F1 values). Higher is better; 100% = full recovery; negative = defense *worsens* performance. Best per row in **bold**.

Data	Emb.	Trust.	RAGDf.	SCOD	Lex.	GLAD
RAWFC	nomic	70.5	−33.7	79.9	−7.7	**138.8**
	qwen	72.6	0.1	42.7	−116.3	**72.0**
LIAR-RAW	nomic	46.7	8.4	44.6	13.5	**73.4**
	qwen	70.3	3.3	54.2	4.4	**80.1**

LIAR-RAW (6-class) the same attack achieves up to 46.9% relative F1 degradation. Two factors compound on LIAR-RAW: (i) the finer label space means adversarial evidence need only shift predictions by one adjacent label to constitute a successful attack, and (ii) the larger corpus (∼627k vs ∼248k sentences) provides more legitimate competition, making the consistent retrieval penetration of adversarial sentences more significant.

7.2 Defense Effectiveness

Table 2 presents the full ablation across all defense configurations. We draw two main observations.

GLAD is the only method that remains competitive across every dataset–retriever combination. On LIAR-RAW, GLAD achieves the best defended F1 under both embeddings (nomic: .236, qwen: .239) with the cross-seed standard deviation of (.022 and .020) respectively. On RAWFC-qwen, GLAD (.563) and TrustRAG (.563) post identical mean F1; GLAD's lower cross-seed standard deviation (.016 vs .037) breaks the tie in its favour. Critically every other defense underperforms GLAD in at least one configuration: TrustRAG on LIAR-RAW-nomic, SCOD on LIAR-RAW (both retrievers), RAGDefender on RAWFC-nomic, and the lexical signal on every cell.

No single-signal method is reliable across both datasets. SCOD is the strongest non-GLAD defense on RAWFC-nomic (.549) but underperforms GLAD on all LIAR-RAW configurations. TrustRAG performs competitively on LIAR-RAW-qwen (.226) but exhibits high variance (std .037) on RAWFC-qwen. RAGDefender, despite its reported effectiveness against gradient-optimized passages, performs poorly against our semantically coherent attack: detection rates fall to 27–38%, and on RAWFC-nomic the defense *actively degrades* performance below the contaminated baseline (.507 vs .519, a recovery of −33.7%), indicating severe over-filtering of legitimate evidence. This confirms that defenses designed for syntactically anomalous adversarial content do not transfer to the linguistically fluent attacks that fact-checking systems face in practice.

7.3 Recovery Analysis

Table 3 quantifies recovery rates computed on mean F1 values across five seeds.

GLAD achieves the highest recovery on LIAR-RAW under both embeddings: 73.4% (nomic) and 80.1% (qwen), substantially reclaiming the 48–50% contamination-induced F1 drop. No method reaches full (100%) recovery on LIAR-RAW, consistent with the 28–30% of adversarial content that remains undetected and continues to influence predictions.

On RAWFC-nomic, GLAD's recovery of 138.8% exceeds 100% because GLAD-defended F1 (.570) marginally exceeds clean (.556) within cross-seed variance (Table 4); $> 100\%$ values should be read as full recovery, not a beneficial filtering side-effect. Across all four settings, GLAD achieves the highest recovery in every cell: RAGDefender and the standalone lexical signal post negative recovery in at least one configuration, while TrustRAG and SCOD remain positive but consistently below GLAD.

Finally, Table 4 confirms that GLAD imposes negligible cost when no attack is present: across

Table 4: Effect of GLAD filtering with and without contamination. **Clean**: no contamination, no filter. **GLAD (contam.)**: contaminated corpus with GLAD applied. **GLAD (clean)**: un-contaminated corpus with GLAD applied (tests whether filtering harms performance when no attack is present). F1 mean ± std over 5 seeds (7, 21, 43, 142, 456).

Dataset	Embed.	Clean	GLAD (contam.)	GLAD (clean)
RAWFC	nomic	.556 ±.030	.570 ±.016	.561 ±.032
	qwen	.574 ±.019	.563 ±.016	.562 ±.035
LIAR-RAW	nomic	.271 ±.010	.236 ±.022	.270 ±.009
	qwen	.266 ±.015	.239 ±.020	.269 ±.017

all four configurations, the absolute F1 difference between unfiltered retrieval and GLAD on uncontaminated corpora is at most .012.

8 Conclusion

We have demonstrated that retrieval-based fact-checking systems are systematically vulnerable to semantically coherent adversarial evidence contamination. By injecting LLM-generated passages that simultaneously achieve high retrieval similarity and maintain linguistic fluency, an adversary can corrupt up to 35.2% of top-k retrievals with a corpus footprint of just 3.1–7.4%, degrading verification accuracy by up to 46.9% relative F1.

We showed that existing defenses, TrustRAG and RAGDefender fail to provide consistent recovery against this threat model. RAGDefender in particular, which performs well against syntactically anomalous gradient-optimized passages, exhibits detection rates of only 27–38% against our fluent LLM-generated attacks. We proposed GLAD (Geometric-Lexical Adversarial Defense), which fuses dense embedding peer-coherence (SCOD) with sparse lexical distributional isolation. Across two benchmarks, two embedding models, and five random seeds, GLAD achieves the highest detection rates (70–86%) and the most consistent F1 recovery, reclaiming 73–80% of contamination-induced degradation on LIAR-RAW. Critically, GLAD imposes negligible cost on clean corpora, making it suitable for deployment regardless of whether an active attack is present. Our ablation demonstrates that neither signal alone suffices: the lexical component is harmful in isolation but genuinely complementary within the fusion, while the dense component provides a strong individual baseline that GLAD reliably improves upon. We release our adversarial dataset of over 58,000 contaminated sentences across five label-specific strategies, providing a benchmark for future defense research on retrieval-augmented fact-checking.

Limitations and Future Directions

All adversarial evidence in our benchmark was generated by a single model (*google/gemma-3-12b-it*). Frontier models with stronger instruction-following capabilities may produce passages with different distributional signatures, potentially closer to legitimate evidence and harder to detect. Whether GLAD's recovery rates generalize across generators remains an open empirical question.

Our adversary is also non-adaptive: contamination is generated without knowledge of the defense. This is the easier setting. An adversary who knows about GLAD could, for instance, enforce lexical consistency with the retrieved peer set to defeat the isolation signal while maintaining embedding-space proximity. Evaluating robustness under adaptive attack is the most pressing next step.

Finally, our evaluation uses a DeBERTa classifier as the verifier. LLM-based verifiers with long context windows might behave differently in two opposing ways: cross-passage reasoning could make them more robust to individual contaminated sentences, but their instruction-following capability also makes them more susceptible to indirect prompt injection embedded within adversarial passages. Our setup does not capture this tension.

Ethics Statement

This research investigates vulnerabilities in automated fact-checking systems to improve their robustness against adversarial attacks. Our adversarial dataset is released exclusively for defensive research purposes, with clear "ADVERSARIAL" labeling to prevent misuse. We believe responsible disclosure of these vulnerabilities, paired with effective defenses, strengthens the integrity of fact-checking systems deployed to combat misinformation. Our adversarial dataset is released under the CC BY-NC-SA 4.0 license.

References

Ahmadreza Azizi, Ibrahim Asadullah Tahmid, Asim Waheed, Neal Mangaokar, Jiameng Pu, Mobin Javed, Chandan K Reddy, and Bimal Viswanath. 2021. T-Miner: A generative approach to defend against trojan attacks on DNN-based text classification. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2255–2272.

Amin Bigdeli, Negar Arabzadeh, Ebrahim Bagheri, and Charles L. A. Clarke. 2025. Adversarial attacks against neural ranking models via in-context learning. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2025, page 211–220, New York, NY, USA. Association for Computing Machinery.

Nicholas Carlini, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. 2024. Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 407–425.

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.

Harsh Chaudhari, Giorgio Severi, John Abascal, Matthew Jagielski, Christopher A Choquette-Choo, Milad Nasr, Cristina Nita-Rotaru, and Alina Oprea. 2024. Phantom: General trigger attacks on retrieval augmented language generation. *arXiv preprint arXiv:2405.20485*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. *arXiv preprint arXiv:2401.08281*.

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.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *ICLR*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

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

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Fanzhen Liu, Sharif Abuadbba, Kristen Moore, Surya Nepal, Cecile Paris, Jia Wu, Jian Yang, and Quan Z. Sheng. 2025. Adversarial attacks against automated fact-checking: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22968–22990, Suzhou, China. Association for Computational Linguistics.

Quanyu Long, Yue Deng, Leilei Gan, Wenya Wang, and Sinno Jialin Pan. 2025. Backdoor attacks on dense retrieval via public and unintentional triggers. In *Second Conference on Language Modeling*.

Jamal Abdul Nasir, Osama Subhani Khan, and Iraklis Varlamis. 2021. Fake news detection: A hybrid CNN-RNN based deep learning approach. *International Journal of Information Management Data Insights*, 1(1):100007.

Zach Nussbaum and et al. 2024. Nomic embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*.

Alamgir Munir Qazi, John Philip McCrae, and Jamal Nasir. 2025. When retrieval outperforms generation: Dense evidence retrieval for scalable fake news detection. In *Proceedings of the 5th Conference on Language, Data and Knowledge*, pages 255–265, Naples, Italy. Unior Press.

Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2021. ONION: A simple and effective defense against textual backdoor attacks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9558–9566, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chris Samarinas, Wynne Hsu, and Mong Li Lee. 2021. Improving evidence retrieval for automated explainable fact-checking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 84–91, Online. Association for Computational Linguistics.

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

Edson C Tandoc Jr, Zheng Wei Lim, and Richard Ling. 2018. Defining "fake news" a typology of scholarly definitions. *Digital journalism*, 6(2):137–153.

Gemma Team. 2025. Gemma 3.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22.

Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM on Web Conference 2024*, pages 2452–2463.

Claire Wardle and Hossein Derakhshan. 2017. *Information disorder: Toward an interdisciplinary framework for research and policymaking*, volume 27. Council of Europe Strasbourg.

Zhiwei Yang, Jing Ma, Hechang Chen, Hongzhan Lin, Ziyang Luo, and Yi Chang. 2022. A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2608–2621, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zhenrui Yue, Huimin Zeng, Lanyu Shang, Yifan Liu, Yang Zhang, and Dong Wang. 2024. Retrieval augmented fact verification by synthesizing contrastive arguments. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10331–10343, Bangkok, Thailand. Association for Computational Linguistics.

Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. 2023. Poisoning retrieval corpora by injecting adversarial passages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13764–13775, Singapore. Association for Computational Linguistics.

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

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.