Poster: DMI-RAG: Dual Mutual Information Optimization for Robust Graph-based Retrieval-Augmented Generation

Zeming Fei*, Hongming Fei†, Prosanta Gope‡, Xiaoyang Wang§, Biplab Sikdar†, Ying Zhang*
*University of Technology Sydney †National University of Singapore
‡University of Sheffield §University of New South Wales

***Abstract*—Graph-based Retrieval-Augmented Generation (GraphRAG) is increasingly deployed in security-critical settings yet remains acutely vulnerable to knowledge graph (KG) poisoning: adversaries who corrupt even a small fraction of KG triples can silently hijack LLM outputs without triggering conventional intrusion detection. Existing defenses are *retrieval-centric* and leave the underlying model structurally naive. We present DMI-RAG, the first defence to explicitly decompose KG poisoning into orthogonal semantic and structural failure modes, hardening the LLM itself via *dual mutual information* (DMI) optimization—combining Direct Preference Optimization (DPO) for semantic robustness with Graph Contrastive Learning (GCL) for structural robustness. Three co-evolving agents establish a progressive adversarial hardening curriculum. On HotpotQA under 30% graph noise, DMI-RAG achieves a Robustness Drop Rate of 13.9%—a 2.1× improvement over the best baseline (29.9%)—while gaining +9.6% absolute Exact Match; gains are consistent across three datasets and three adaptive adversary scenarios.**

1. Introduction

GraphRAG systems underpin legal discovery, clinical decision support, and financial intelligence [1]—domains where a single incorrect output has material consequences. Unlike text-level attacks detectable by anomaly monitors, KG poisoning operates at the *data layer*: an adversary with write access to the KG backend silently corrupts the graph retrieved at inference time, leaving no trace. PoisonedRAG [2] demonstrates that as few as 5–10 injected triples per query suffice to flip LLM answers. Three orthogonal attack primitives cover the threat landscape: *edge insertion* (spurious triples polluting retrieved subgraphs), *edge deletion* (severing multi-hop chains), and *relation modification* (corrupted predicate semantics invisible to structural monitors). A defense addressing only one primitive leaves the other two fully exploitable—yet no prior work addresses all three jointly.

Existing defenses are *retrieval-centric* [6], [8]: they filter before the LLM, leaving the model structurally naive. While ATM [9] applies adversarial fine-tuning to RAG and can be adapted to GraphRAG settings, it treats retrieved subgraphs as flat text, leaving the model blind to graph topology. **DMI-RAG** is the first defense to explicitly decompose KG poisoning into orthogonal semantic and structural failure modes, addressing each with a dedicated MI objective that text-centric methods cannot provide.

2. Threat Model & Problem Formulation

Threat Model. We model a *white-box-KG / black-box-LLM* adversary with write access to a bounded fraction ϵ of KG edges but no access to LLM weights or the retrieval mechanism. The adversary selects $\Delta\mathcal{E}$ with $|\Delta\mathcal{E}|/|\mathcal{E}| \leq \epsilon$ to maximize the probability of eliciting target incorrect answers, evaluated at $\epsilon \in \{0.1, 0.2, 0.3\}$ with equal fractions of each attack primitive; even at ϵ=0.1, standard GraphRAG systems suffer over 20% performance degradation. Adaptive adversary experiments (§4) additionally grant full knowledge of the defense.

Problem Formulation. Given KG $\mathcal{G}=(\mathcal{V}, \mathcal{E}, \mathcal{R})$ and query q, retrieval yields subgraph $\mathcal{G}_{\text{sub}}$. Under noise model $\mathcal{N}_\epsilon$, the LLM observes corrupted $\tilde{\mathcal{G}}_{\text{sub}}$. We seek θ^* minimizing:

$$\theta^* = \arg\min_\theta \ \mathbb{E}_{\mathcal{N}_\epsilon}\Big[\mathcal{L}\Big(\mathcal{M}_\theta(q, \tilde{\mathcal{G}}), a^*\Big) - \mathcal{L}(\mathcal{M}_\theta(q, \mathcal{G}), a^*)\Big]. \tag{1}$$

This gap decomposes into two orthogonal failure modes: (i) *semantic corruption*—wrong relation semantics mislead the LLM (addressed by DPO); and (ii) *structural corruption*—broken topology derails multi-hop chains (addressed by GCL).

3. DMI-RAG Framework

3.1. Dual MI Objective

DMI-RAG jointly maximizes tractable lower bounds on two MI channels:

$$\mathcal{J}(\theta) = \underbrace{I_{\text{sem}}(A; \mathcal{G} \mid Q)}_{\text{DPO (semantic)}} + \lambda \underbrace{I_{\text{struct}}(\mathcal{G}; \mathcal{G}^+ \mid Q)}_{\text{GCL (structural)}}. \tag{2}$$

DPO [3] maximizes the semantic term via pairwise preference data contrasting clean-graph against corrupted-graph responses. InfoNCE maximizes the structural term by contrasting clean/augmented subgraph pairs against adversarial negatives. Both terms share the LLM backbone parameters θ while maintaining separate projection heads for the structural channel, ensuring co-training without interference. KL

TABLE 1. HOTPOTQA RESULTS UNDER 30% KG NOISE. **BOLD**: BEST; UNDERLINE: SECOND BEST. †: SIGNIFICANT VS. BEST BASELINE ($p<0.05$).

Method	Clean EM	Noisy EM	RDR (%)
Vanilla SFT	38.2	19.5	49.0
Self-RAG [5]	39.8	22.3	44.0
RetRobust [6]	41.2	27.4	33.5
RobustRAG [8]	41.7	28.9	30.7
RAAT [7]	42.0	28.1	33.1
ATM [9]	43.1	30.2	29.9
DMI-RAG	**46.2**	**39.8**†	**13.9**
vs. ATM	*+3.1*	*+9.6*	2.1× *lower*

regularization $\beta\,\mathrm{KL}[\pi_{\mathrm{main}}\|\pi_{\mathrm{ref}}]$ stabilizes joint optimization by preventing one objective from dominating.

3.2. Multi-Agent Co-Evolution

Three specialized agents interact iteratively. **Positive Generator** π_{pos}: semantic-preserving augmentations (edge dropout, feature masking) as GCL anchors. **Negative Generator** π_{neg}: DPO-trained to craft maximally confusing yet structurally plausible corruptions; a GCL coherence penalty prevents trivially disconnected negatives. **Main Generator** π_{main}: fine-tuned under $\mathcal{L}_{\mathrm{task}}+\lambda_1\mathcal{L}_{\mathrm{DPO}}+\lambda_2\mathcal{L}_{\mathrm{GCL}}+\beta\,\mathrm{KL}[\pi_{\mathrm{main}}\|\pi_{\mathrm{ref}}]$. As π_{main} hardens it demands stronger corruptions from π_{neg}, yielding a self-improving curriculum. Convergence requires 3–4 iterations (≈13 h/iter on 16×A100, 230 K extra parameters, 3.5× training overhead vs. vanilla fine-tuning, with negligible inference cost (+38% latency, 120 queries/s on a single A100).

4. Evaluation

Setup. We evaluate on HotpotQA [4], MusiqueQA, and MultihopQA (2–6 hops), injecting KG noise into Wikidata5M at $\epsilon \in \{0.1, 0.2, 0.3\}$. All models use Qwen-7B; results averaged over five seeds ($p<0.05$, paired t-test). We report Exact Match (EM) and $\mathrm{RDR}=1-\mathrm{EM}_{\mathrm{noise}}/\mathrm{EM}_{\mathrm{clean}}$ (lower is better): RDR of 30% means a system answering 40 questions correctly on clean graphs answers only 28 under attack.

4.1. Main Results

DMI-RAG dominates all baselines on every metric. The 2.1× RDR reduction holds across noise levels (RDR = 4.2% vs. ATM's 12.3% at 10% noise; 8.9% vs. 21.4% at 20%) and generalizes across datasets: MusiqueQA (+8.6 EM) and MultihopQA (+8.2 EM) under identical conditions, confirming that dual MI optimization learns generalizable robustness rather than dataset-specific patterns.

4.2. Ablation Study

Removing the co-evolutionary curriculum costs the most (−8.3 EM); removing the positive agent costs −7.4 EM (mode collapse without clean anchors); removing DPO costs −4.0 EM (relation modification attacks succeed); removing GCL costs −2.2 EM (structural attacks succeed); removing KL regularization costs −1.7 EM. The adversarial curriculum yields super-additive gains that neither component achieves alone.

4.3. Adaptive Adversary Analysis

Three adaptive scenarios grant the adversary full knowledge of DMI-RAG. **ScenA** (gradient-aligned against DPO): fails because corruptions effective against the semantic channel simultaneously strengthen the GCL contrastive signal—complementary subspaces maintain detection rate at 100%. **ScenB** (ℓ_2-bounded): absorbed by the fixed-MLP projection head. **ScenC** (joint semantic+structural): jointly optimizing both surfaces forces a trade-off—perturbations defeating the semantic channel are detectable by the structural channel and vice versa, empirically strengthening detection across all seeds ($\text{adv-DR} \geq \text{rand-DR}$).

5. Conclusion

DMI-RAG establishes that KG-poisoning attacks can be countered at the *model level*, achieving 13.9% RDR vs. 29.9% for the best baseline—a 2.1× improvement. By mapping orthogonal attack surfaces to independently optimizable MI objectives, DMI-RAG yields robustness that is both empirically strong and adaptive-adversary-resistant. Future work targets certified robustness bounds under ℓ_0-constrained perturbations, extensions to dynamic KGs, and reduced training overhead via distillation.

References

[1] D. Edge et al., "From local to global: A graph RAG approach to query-focused summarization," *arXiv*:2404.16130, 2024.

[2] W. Zou et al., "PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation," in *Proc. USENIX Security*, 2024.

[3] R. Rafailov et al., "Direct preference optimization: Your language model is secretly a reward model," in *Proc. NeurIPS*, 2023.

[4] Z. Yang et al., "HotpotQA: A dataset for diverse, explainable multi-hop question answering," in *Proc. EMNLP*, 2018.

[5] A. Asai et al., "Self-RAG: Learning to retrieve, generate, and critique," in *Proc. ICLR*, 2024.

[6] O. Yoran et al., "Making retrieval-augmented language models robust to irrelevant context," in *Proc. ICLR*, 2024.

[7] F. Fang et al., "Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training," in *Proc. ACL*, 2024.

[8] C. Xiang et al., "Certifiably robust retrieval-augmented generation against retrieval corruption," in *Proc. ICML*, 2024.

[9] J. Zhu et al., "ATM: Adversarial tuning multi-agent system makes a robust retrieval-augmented generator," in *Proc. EMNLP*, 2024.

DMI-RAG: Dual Mutual Information Optimization System for Robust Graph-Based Retrieval-Augmented Generation

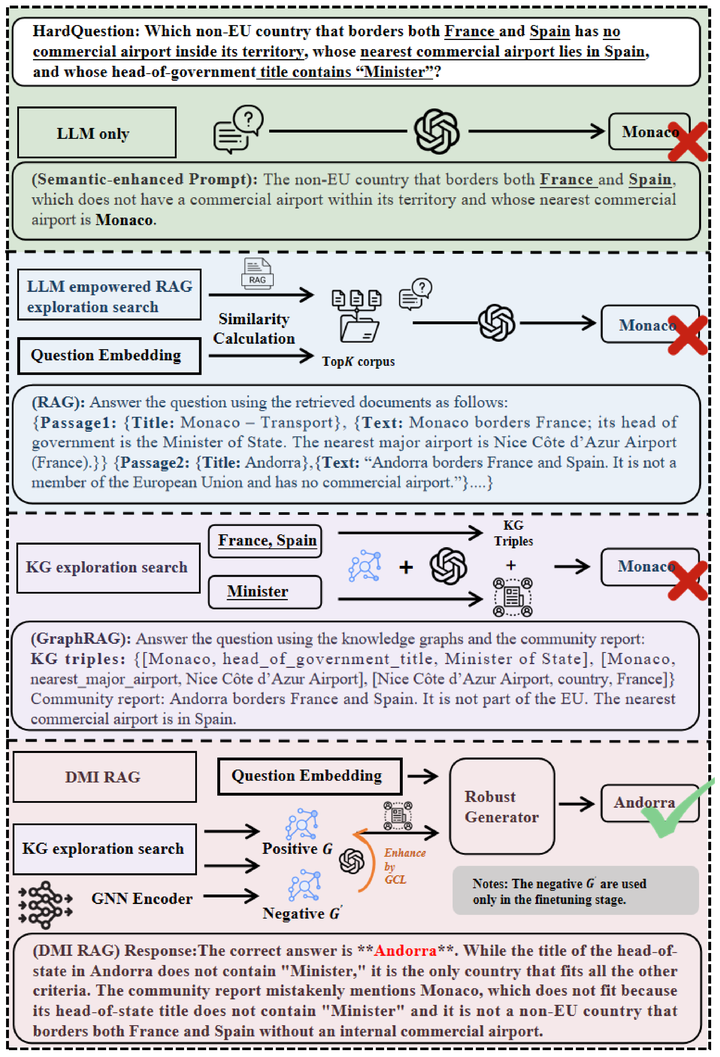
Zeming Fei[1], Hongming Fei[2], Prosanta Gope[3], Xiaoyang Wang[4], Biplab Sikdar[2], Ying Zhang[1]
[1]University of Technology Sydney [2]National University of Singapore
[3]University of Sheffield [4]University of New South Wales

IEEE S&P

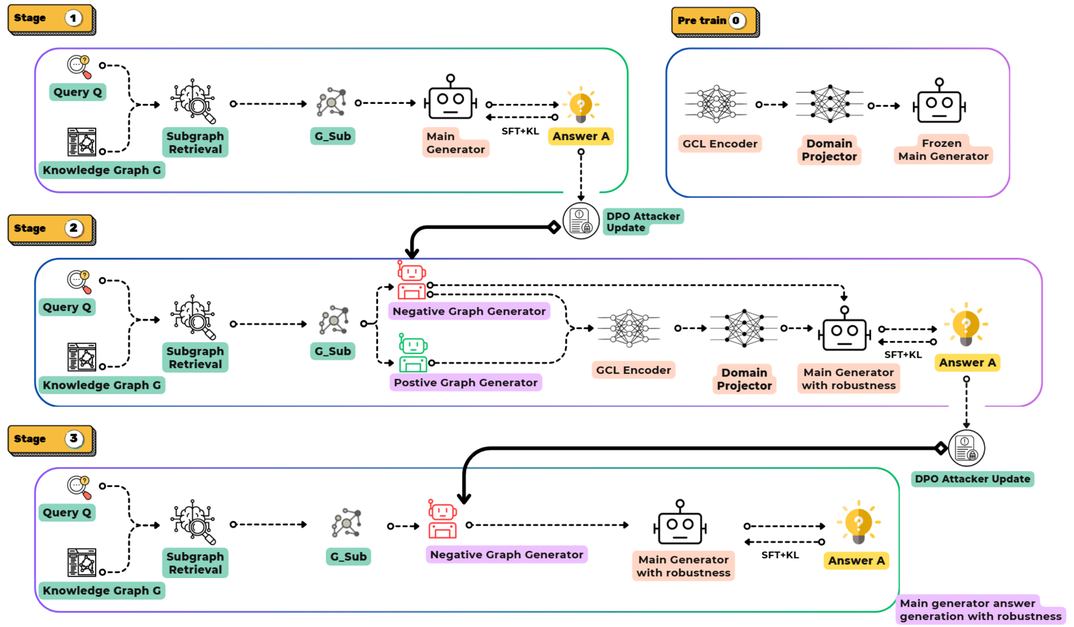
Problem Statement & Motivation

- GraphRAG systems are deployed in security-critical settings but remain acutely vulnerable to knowledge graph (KG) poisoning.
- An adversary with write access can silently hijack LLM outputs without triggering conventional intrusion detection.
- Three principal attack types threaten graph integrity: edge insertion, edge deletion, and relation modification.

Vulnerability Demonstration



Multi-Agent Co-Evolution Architecture



Threat Model

- We model a white-box-KG / black-box-LLM adversary with write access to a bounded fraction ϵ of KG edges but no access to LLM weights.
- We evaluate at noise budgets $\epsilon \in \{0.1, 0.2, 0.3\}$ with equal fractions of edge insertion, deletion, and relation modification. Adaptive adversaries are also considered.

Dual Mutual Information Objective

- DMI-RAG jointly maximizes tractable lower bounds on two mutual information channels:

$$\mathcal{J}(\theta) = I_{sem}(A; \mathcal{G}|Q) + \lambda I_{struct}(\mathcal{G}; \mathcal{G}^{+}|Q)$$

- Semantic robustness is achieved via Direct Preference Optimization (DPO), and structural robustness via Graph Contrastive Learning (GCL).

Agent Co-Evolution & Training Protocol

- **Positive Generator:** Creates semantic-preserving augmentations as GCL anchors.
- **Negative Generator:** DPO-trained to craft maximally confusing yet structurally plausible corruptions.
- **Main Generator:** Learns robust discrimination, progressively hardening against attacks.
- **Iterative Protocol:** Iteration 1 establishes structural awareness via GCL and projections. Iteration 2+ drives progressive adversarial hardening.

Background & Related Works

- Existing defenses (e.g., RetRobust, ATM) are retrieval-centric, filtering before the LLM, leaving the underlying model structurally naive.
- A stealthy attack that preserves surface plausibility can bypass upstream filters, and retrieval guards fail to transfer robustness into parametric knowledge.

Main Results (with F1 Scores)

- On HotpotQA under 30% graph noise, DMI-RAG achieves a Robustness Drop Rate of 13.9% (a **2.1x improvement** over ATM) while maintaining superior semantic similarity.

Method	Clean EM	Noisy EM	Clean F1	Noisy F1	RDR
Vanilla SFT	38.2	19.5	49.7	28.3	49.0%
RetRobust	41.2	27.4	52.8	36.9	33.5%
ATM	43.1	30.2	54.6	39.7	29.9%
DMI-RAG	**46.2**	**39.8**	**57.8**	**49.6**	**13.9%**

Ablation, Scalability & Future Directions

- **Mode Collapse Prevention:** Removing the positive agent causes a severe -7.4 EM drop. It serves as a vital anchor preventing *mode collapse*, ensuring the negative agent produces plausible corruptions rather than trivial noise.
- **Synergy of DMI:** Removing DPO costs -4.0 EM (vulnerable to relation modifications); removing GCL costs -2.2 EM (vulnerable to structural attacks).
- **Scalability:** Our defense generalizes to larger architectures. Qwen-14B achieves 48.7 Clean EM (+2.5) and 42.3 Noisy EM (+2.5), maintaining excellent RDR.
- **Dynamic KGs:** While offline multi-agent training is intensive, future iterations will employ lightweight adapter tuning (e.g., LoRA) to incrementally update structural awareness for evolving KGs without full retraining.