Query-Side Attacks on GNN-Based KGQA: Tracing Failures from Entity Linking to Answer Generation

Anonymous ACL submission

Abstract

GNN-based Knowledge Graph Question Answering (KGQA) pipelines process queries through four discrete stages: entity linking, subgraph retrieval, GNN reasoning, and answer generation. Standard robustness evaluations conflate stage-level failures into a single end-to-end metric, obscuring both the source of brittleness and the appropriate mitigation target. We ask which stage fails, and why, when the pipeline is subjected to adversarial perturbations on the input question. We introduce a stage-isolation protocol with two answer-preserving adversarial perturbations verified against the knowledge graph: Compositional Restructuring (CR) and Relation Synonym Swap (RS) target distinct stages while leaving entity seeds intact. Evaluated across ComplexWebQuestions and WebQSP, the results run counter to prevailing assumptions: the GNN reasoning stage retains near-baseline accuracy when the subgraph is intact, while subgraph construction accounts for over 99% of the end-to-end collapse under CR, occurring even when the gold answer is present in 74% of retrieved subgraphs. This exposes a fundamental distinction between answer presence and answer reachability that end-to-end metrics cannot detect, and places the mitigation target firmly at the subgraph construction stage rather than the reasoning model. Perturbed datasets and evaluation infrastructure are released at https://anonymous.4open.science/r/atkgrag-E85C.

1 Introduction

Knowledge graph question answering (KGQA) has become a core benchmark for structured reasoning: given a natural-language question and a large knowledge graph (KG) such as Freebase, the system must find the answer entity by traversing multi-hop relational paths (Sun et al., 2018; Saxena et al., 2020; Zhang et al., 2022). GNN-based pipelines have pushed state-of-the-art accuracy by coupling a graph neural network (GNN) reasoner with a retrieval-augmented generation (RAG) framework, achieving strong results on ComplexWebQuestions (CWQ) and WebQSP (Mavromatis and Karypis, 2024).

The path from question to answer runs through four discrete stages: Entity Linking, Subgraph Retrieval, GNN reasoning, and Answer generation. Each stage can fail independently and pass corrupted outputs to the next. An adversary who reformulates the query without touching the KG or model can trigger a cascade of failures through this pipeline. Figure 1 shows the four stages and their distinct attack surfaces.

Most robustness work on RAG focuses on *corpus-side* attacks: poisoning the retrieval index (Zhong et al., 2023; Zou et al., 2024; Chaudhari et al., 2024), injecting adversarial passages (Perez and Ribeiro, 2022; Shi et al., 2023), or manipulating retrieved context (Xue et al., 2024). These require the adversary to write to the KG or corpus, a strong and often unrealistic capability. Existing KGQA robustness studies (Perçin et al., 2025) have explored entity surface-form noise and relation paraphrase perturbations but have not measured *which pipeline stage fails first*, nor how that failure propagates downstream.

We conduct the first systematic, stage-resolved robustness study of a GNN-based KGQA pipeline. The adversary needs only the ability to submit a question, with no KG access or model internals. To attribute failures to the correct stage, we introduce a *stage-isolation protocol*: fix upstream outputs, perturb only the query, and measure per-stage degradation separately. Two attacks probe distinct stages: **Compositional Restructuring (CR)** rewrites the question via hop-order reversal or constraint injection, leaving entity mentions intact and preserving the SPARQL answer; **Relation Synonym Swap (RS)** replaces predicate surface forms with synonyms while keeping entity mentions unchanged.

Entity Swap (ES) is a diagnostic perturbation (Appendix E) that changes the gold answer by design and characterises entity-linker brittleness.

Under CR, ELQ subgraphs contain the gold answer in 74% of cases yet GNN-RAG achieves only 0.68% EM on CWQ, the answer is present but PPR topology is anchored to the original reasoning chain. GraftNet's question-embedding-weighted PPR achieves 29.82% EM at 63.3% answer coverage, higher EM than ELQ's 0.68% despite *lower* coverage (63.3% vs. 74%), because it reorients PPR mass toward the restructured reasoning chain rather than the original one. Under RS the ELQ subgraph is structurally unchanged, so RS ELQ EM (20.3% CWQ, 50.9% WebQSP) measures GNN instruction decoder sensitivity in near-isolation. Path injection at inference recovers 51.4% CWQ EM under CR without fine-tuning, matching or exceeding the relation-path-augmented model. EPR-KGQA (single-shot pattern-based retrieval) retains near-baseline accuracy under both attacks; ExplaiGNN (iterative pattern-matching, Wikidata) collapses like GNN-RAG, showing robustness requires single-shot rather than PPR-free retrieval.

Contributions.

1. **Stage-isolation framework.** A protocol and two adversarial perturbation types (CR, RS) that probe distinct pipeline stages, enabling per-stage degradation measurement (Δ_{EL}, Δ_{SR}, Δ_{AG}) on standard KGQA benchmarks.

2. **Answer presence $\neq$ answer reachability.** Under CR, 74% subgraph answer presence yields only 0.68% EM (CWQ); GraftNet (GEM oracle seeds) at 63.3% coverage reaches 29.82% EM because its PPR topology aligns with the restructured reasoning chain.

3. **Subgraph construction is the performance ceiling.** Stage-isolation attributes 52.08 pp of the 52.22 pp total CR CWQ drop to subgraph topology failure; the GNN instruction decoder is robust when the subgraph is intact (52.76% CWQ EM on clean subgraph + CR question).

4. **Inference-time path injection recovers most of the gain.** Injecting predicted relation paths at inference (no fine-tuning) recovers 51.4% CWQ EM under CR, matching or exceeding relation-path-augmented fine-tuning.

2 Related Work

GNN-based KGQA and subgraph retrieval. Semantic parsing methods (Berant et al., 2013; Yih et al., 2015; Lan et al., 2021) translate questions into SPARQL for direct KG execution. Embedding-based methods (Saxena et al., 2020; Zhang et al., 2022; He et al., 2021) retrieve answers via dense similarity. GNN-RAG (Mavromatis and Karypis, 2024) combines both: a three-layer GAT (ReaRev) over a PPR-retrieved subgraph, with ELQ as entity linker, creating a cascading failure structure that we characterise under query perturbation. Subgraph retrieval variants include GraftNet-Orig (Sun et al., 2018) (PPR from seed entities), NSM (He et al., 2021) (relation-weighted BFS pruning), and EPR-KGQA (Ding et al., 2024) (atomic adjacency patterns, state-of-the-art on CWQ as of 2024 and near-baseline-retaining under our perturbations, Section 5). Recent learned retrievers (GSR (Huang et al., 2024), D-RAG (Gao et al., 2025)) address the same subgraph bottleneck we identify via differentiable or LM-based selection; FiDeLiS (Sui et al., 2025) targets faithful GNN reasoning on retrieved evidence. HippoRAG (Gutiérrez et al., 2025) independently combines PPR with synonym links for open-domain QA, motivating a similar approach in our adversarial KGQA setting.

Entity linking and query-side robustness. ELQ (Li et al., 2020a) (BERT-Large bi-encoder) is GNN-RAG's production linker; BLINK (Wu et al., 2020) and ReFinED (Ayoola et al., 2022) offer stronger alternatives, and recent evaluations (Li et al., 2025; Hou et al., 2025; Vollmers et al., 2025) show LLM-based linkers substantially outperform bi-encoder approaches. For our two primary attacks (CR and RS), entity linking is intact ($\leq$0.3 pp SeedHit drop), so the subgraph topology is the bottleneck, not the linker. BYOKG-RAG (Mavromatis et al., 2025) proposes iterative LLM-based EL; its improvements are complementary to our structural findings. Query-level attacks on text classifiers (TextFooler (Jin et al., 2020), BERT-Attack (Li et al., 2020b)) and KGQA perturbations (Perçin et al., 2025) have studied surface-form noise, but none measure *which pipeline stage fails first* or how failure propagates downstream, which is the gap our stage-isolation protocol fills. Corpus-side attacks (PoisonedRAG (Zou et al., 2024), Zhong et al. 2023; Chaudhari et al. 2024) require KG write access; our threat model requires only the ability to submit a question. Concurrently, Zhou et al.

(2026) find missing intermediate-hop triples as the dominant failure, validating our subgraph bottleneck from the corpus side; Ma et al. (2026) study the inverted setting where LLM node features absorb structural perturbations, unlike our pipeline where entity-set errors propagate through message-passing. Kandpal et al. (2023) show LLMs fail on rare entities, consistent with our ES diagnostic; Zhao et al. (2025) provide stage-wise KG-poisoning analysis complementary to our query-side isolation.

3 Methodology

3.1 Pipeline Formalisation

We model GNN-RAG as a four-stage function composition:

$$q \xrightarrow{f_{\mathrm{EL}}} E_q \xrightarrow{f_{\mathrm{SR}}} \mathcal{G}_q \xrightarrow{f_{\mathrm{GR}}} c_q \xrightarrow{f_{\mathrm{AG}}} \hat{a} \quad (1)$$

where E_q is the linked entity set (Freebase MIDs), $\mathcal{G}_q$ is the retrieved answer subgraph, c_q is the verbalised evidence context, and $\hat{a}$ is the generated answer. In GNN-RAG: f_{EL} is ELQ (Li et al., 2020a); f_{SR} is personalised PageRank (PPR) over the multi-hop neighbourhood seeded by E_q (He et al., 2021); f_{GR} is a three-layer GAT (ReaRev backbone) (Mavromatis and Karypis, 2024); and f_{AG} is fine-tuned Llama-2-7B.

3.2 Threat Model

Attacker capability. The adversary (i) reads q before submission; (ii) submits q' and observes $\hat{a}$; (iii) runs local copies of open-source components (perturbation generator, Freebase SPARQL endpoint) to verify SPARQL denotation preservation. No access to model parameters or intermediate states is required. This is a *query-only black-box* threat model.

Semantic validity budget. Every CR/RS perturbation must satisfy: (i) $\mathrm{PPL}(q') < 50$ (GPT-2-large); (ii) $\mathrm{BERTScore}(q, q') > 0.85$ (DeBERTa-xlarge-mnli); (iii) $\mathrm{den}(q', \mathrm{Freebase}) = \mathrm{den}(q, \mathrm{Freebase})$ verified via local Virtuoso. Formal perturbation rules are in Appendix F.

3.3 Perturbation Taxonomy

We define two primary adversarial attacks plus a diagnostic perturbation (ES, Appendix E). Four secondary types (S1–S4) are in Appendix K.

Entity Swap (ES): diagnostic only. Substitutes the topic entity, changing both the Freebase MID and the correct answer. Not an adversarial attack; measures entity-linker brittleness. Results in Appendix E.

Compositional Restructuring (CR). Apply exactly one KG-safe structural operation: hop-order reversal, distractor constraint injection, or intermediate-entity alias substitution. Entity mentions are unchanged; SPARQL denotation is preserved. *Pipeline target*: f_{SR} and GNN multi-hop traversal.

Relation Synonym Swap (RS). Replace the predicate phrase with a synonym preserving the underlying Freebase relation. All entity mentions remain exactly unchanged. *Pipeline target*: GNN instruction decoder. Because entity seeds and PPR walks are unchanged, RS ELQ EM numbers (20.3% CWQ, 50.9% WebQSP) measure GNN instruction decoder sensitivity in near-isolation.

All perturbation types are generated by Llama-3.3-70B-Instruct (prompts in Appendix H). Pass rates: 92.4% on filters (i)–(ii) and 94.1% on the SPARQL denotation check. Perturbation quality was further validated by manual inspection of 500 random samples per attack type (excluding ES), confirming semantic naturalness and answer preservation.

3.4 Subgraph Retrieval Variants

All variants use the same GNN-RAG ReaRev checkpoint; only $\mathcal{G}_{q'}$ changes, isolating subgraph construction quality from GNN model quality. Full descriptions of ELQ, GraftNet-Orig, NSM, GraftNet (ours), and EPR-KGQA are in Appendix C.

4 Stage-Wise Evaluation Framework

4.1 Per-Stage Degradation Metrics

We track failure propagation through four metrics (Δ_{EL}, Δ_{SR}, α_{pert}, Δ_{AG}), one per pipeline stage; formal definitions are in Appendix A.

4.2 Stage-Isolation Protocol

Figure 2 illustrates the three experimental configurations. The *fast-mode* experiment fixes the original clean subgraph $\mathcal{G}_q$ and replaces only the question text with q', yielding $\mathrm{EM}_{\mathrm{fast}}(q')$. Subgraph-attributable failure is $\delta_{\mathrm{SG}} = \mathrm{EM}_{\mathrm{fast}} - \mathrm{EM}_{\mathrm{pert}}$; decoder-attributable failure is $\delta_{\mathrm{DEC}} = \mathrm{EM}_{\mathrm{clean}} - \mathrm{EM}_{\mathrm{fast}}$. For CR on CWQ, $\delta_{\mathrm{SG}} = 52.08$ pp and

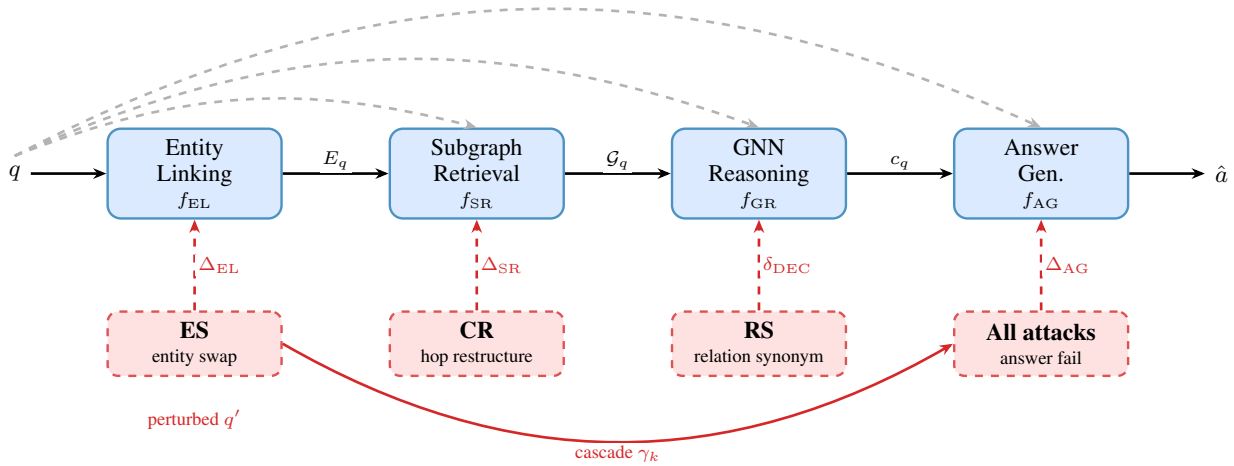

Figure 1: **GNN-RAG pipeline and perturbation entry points.** The pipeline ($f_{\mathrm{EL}} \to f_{\mathrm{SR}} \to f_{\mathrm{GR}} \to f_{\mathrm{AG}}$) maps question q to answer $\hat{a}$. **ES** targets f_{EL} (entity swap $\Rightarrow$ ELQ seed failure). **CR** targets f_{SR} (hop-order reversal $\Rightarrow$ topology mismatch). **RS** targets f_{GR} (predicate synonym $\Rightarrow$ instruction drift). Dashed grey arrows: q feeds all stages directly.

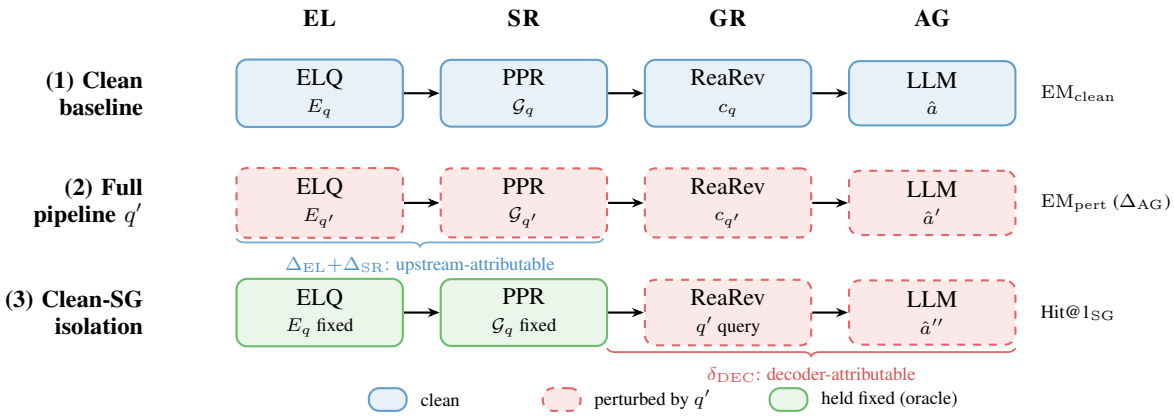

Figure 2: **Stage-isolation protocol (three configurations). (1) Clean baseline**: question q go through all stages and reference $\mathrm{EM}_{\mathrm{clean}}$. **(2) Full pipeline**: perturbed q' through all stages, which measures total drop Δ_{AG}. **(3) Clean-SG isolation**: EL and SR frozen at clean outputs; only the question text fed to GR is replaced with q'. Gap $(2)-(3)$ = subgraph-attributable failure δ_{SG}; gap $(1)-(3)$ = decoder-attributable failure δ_{DEC}.

$\delta_{\mathrm{DEC}} = 0.14$ pp, attributing virtually all collapse to subgraph failure. Full attribution numbers appear in Section 5.5.

5 Experiments

5.1 Experimental Setup

Datasets. CWQ (Talmor and Berant, 2018) (3,531 test questions) and WebQSP (Yih et al., 2016) (1,639 test questions). Both use Freebase MID strings as gold answers; EM is gold-MID substring match (Appendix A). CWQ requires 2–4 reasoning hops; WebQSP requires 1–2. Hardware details are in Appendix B.

Models and subgraph variants. GNN-RAG (Mavromatis and Karypis, 2024) with the ReaRev backbone is the primary model. All subgraph variants (ELQ, GraftNet-Orig, NSM, GraftNet) run the same GNN-RAG checkpoint; only the retrieved subgraph differs (Section 3.4). Clean baselines: CWQ = 52.9% EM, WebQSP = 74.3% EM.

Perturbations. Primary adversarial attacks: CR and RS (Llama-3.3-70B-Instruct; Appendix H). ES (entity substitution) is a diagnostic perturbation reported separately in Appendix E. Secondary ablation (S1–S4): Appendix K.

Metrics. $\Delta_{\mathrm{AG}} = \mathrm{EM}_{\mathrm{clean}} - \mathrm{EM}_{\mathrm{pert}}$ (primary); Δ_{EL} (SeedHit drop); $\Delta_{\mathrm{SR}} = 1 - J(\mathcal{G}_q, \mathcal{G}_{q'})$; α_{pert} (gold answer MID present in $\mathcal{G}_{q'}$). Bootstrap 95%

CI ($n = 1000$) on Δ_{AG}.

5.2 Main Results: End-to-End EM Under Perturbation

Table 1 reports EM and Δ_{AG} for the two primary adversarial attacks (CR and RS) across all subgraph retrieval variants.

Key patterns. (1) CR causes near-total collapse under ELQ (0.68% CWQ, 0.49% WebQSP) despite entity seeds being intact. The failure is a subgraph topology mismatch: ELQ inherits the original PPR walk (97.5% same-seed, Jaccard = 0.885) but that walk is anchored to the original reasoning chain and cannot adapt to the restructured hop order. The answer is present in 74.0% of CR subgraphs yet the GNN cannot reach it via the changed path (Section 6). GraftNet (GEM oracle seeds) reorients PPR toward the restructured question, raising α_{pert} to 63.3% and EM to 29.82% (CWQ), the best CR result; the deployment-realistic ELQ+cosine configuration reaches 14.98% CWQ CR. (2) RS retains the highest EM (20.3%/50.9% ELQ) because entity seeds are unchanged and the ELQ subgraph is structurally the same as the clean run. For RS, GraftNet underperforms ELQ (13.71% vs. 20.31% CWQ) because its question-reoriented PPR diverges from the original subgraph topology ($\alpha_{\mathrm{pert}} = 23.2\%$ vs. 74.0% for ELQ); ELQ's seed-replay strategy is exactly the right choice when RS leaves the subgraph intact. This confirms an attack-type asymmetry: ELQ wins RS by construction, GraftNet wins CR by design. (3) EPR-KGQA's near-baseline retention (59.2%/64.3% Hit@1 vs. GNN-RAG's 0.68%/0.49% gold-MID EM under CR) confirms the vulnerability is specific to PPR-based retrieval; both systems use MID-indexed Hit@1, so this $\sim$58 pp gap is a direct architectural comparison (Appendix K.5).

5.3 Entity Linking Performance

CR and RS leave ELQ SeedHit virtually unaffected (≤ 0.3 pp from clean), confirming entity mentions are unchanged and EL is not the bottleneck for these attacks. Full SeedHit values and the ELQ/ES comparison are in Appendix D.

5.4 Subgraph Retrieval Quality

Table 2 reports subgraph overlap statistics for CR and RS. α_{pert} is the strongest predictor of EM across all variants: the ordering GraftNet > NSM for CR EM (29.82% vs. 5.81% CWQ) mirrors exactly the ordering of α_{pert} (63.3% vs. 29.5%).

5.5 GNN Instruction Decoder Isolation

Table 3 reports GNN-RAG Hit@1 and instruction cosine similarity when run on the *original clean subgraph* with only the question text replaced.

With a correct subgraph, ReaRev Hit@1 stays near or above clean (57–58% CWQ, 68–73% WebQSP) across all perturbation types. CR shows the largest instruction drift (cos = 0.915 CWQ, 0.842 WebQSP) yet still achieves 57.5%/68.2% Hit@1, demonstrating the decoder's partial self-correction. The gap between full-pipeline EM (CR ELQ: 0.68% CWQ) and clean-subgraph $\mathrm{EM}_{\mathrm{fast}}$ (52.76% CWQ) quantifies the failure attribution: 52.08 pp of the 52.22 pp total CR CWQ drop is driven by subgraph failure, not GNN reasoning (formal decoder analysis in Appendix G).

The NSM (LSTM-backbone) decoder shows a similar pattern on CWQ: Hit@1 drops only 0.85 pp under CR and 0.99 pp under RS, confirming that decoder robustness generalises beyond the ReaRev GAT architecture. WebQSP shows a larger NSM drop under CR (10.86 pp), which reflects the NSM decoder's higher sensitivity to WebQSP's shorter 1–2 hop chains under instruction drift; even so, the decoder retains 57.47% of its 68.33% clean baseline while the full-pipeline ELQ collapses to 0.49%. Contribution 3 (GNN decoder is robust when the subgraph is intact) therefore holds for both ReaRev and NSM on CWQ, and partially on WebQSP.

5.6 LLM Reasoning Over GNN Candidates

To assess f_{AG} independently, we pipe GNN candidate entities through two LLM reasoning models: Llama-3.1-8B and RoG (fine-tuned GCR-Llama-2-7B). Table 4 shows LLM Hit@1 against the GNN-only EM baseline.

RoG+PathOnly (base RoG checkpoint, predicted rule paths injected at inference, no RA fine-tuning) matches RoG+RA on both benchmarks within 1 pp. This consistent pattern shows the gain over base RoG traces to *path injection at inference*, not the augmented training signal. Llama-3.1-8B without fine-tuning degrades top-1 accuracy on CWQ (all Hit@1 < GNN EM), showing general instruction following is insufficient for KGQA reasoning under perturbation.

6 Analysis

CR failure: answer presence does not imply answer reachability. The most striking result

Table 1: GNN-RAG EM under CR and RS (bootstrap 95% CI on Δ_{AG}, n=1000). Clean EM: CWQ = 52.9%, WebQSP = 74.3%. ELQ = ELQ-seeded flat PPR; GraftNet = question-embedding-weighted PPR (ours); GraftNet-Orig = relation-aware PPR (Sun et al., 2018); NSM = He et al. (2021). †: RS ELQ subgraph is structurally unchanged from clean; EM measures GNN decoder sensitivity in near-isolation. GraftNet uses GEM oracle seeds (not available at deployment); ELQ+cosine is the deployment-realistic configuration. ‡: GraftNet RS CWQ = 13.71% $<$ ELQ RS 20.31% because question-reoriented PPR lowers α_{pert} to 23.2% vs. 74.0% for ELQ. §: Both systems use MID-indexed Hit@1 (Appendix K.5); the $\sim$58 pp gap is a direct architectural comparison. CIs are per-comparison and uncorrected.

		CWQ		**WebQSP**	
Attack	Subgraph	Hit@1/EM (%)	Δ_{AG} [95% CI]	Hit@1/EM (%)	Δ_{AG} [95% CI]
CR: Compositional Restructuring					
	ELQ	0.68	0.522 [0.505, 0.538]	0.49	0.738 [0.716, 0.758]
	NSM	5.81	0.471 [0.454, 0.489]	1.65	0.727 [0.705, 0.750]
	ELQ+cosine (realistic)	**14.98**	**0.374 [0.356, 0.391]**	**22.33**	**0.521 [0.498, 0.543]**
Oracle seed upper bound (GEM MIDs, not available at deployment)					
	GraftNet (oracle seeds)	29.82	0.231 [0.213, 0.249]	36.55	0.378 [0.347, 0.405]
RS: Relation Synonym Swap†					
	ELQ	**20.31**	**0.326 [0.307, 0.344]**	**50.95**	**0.234 [0.210, 0.257]**
	GraftNet-Orig	11.24	0.417 [0.398, 0.434]	41.67	0.326 [0.301, 0.352]
	NSM	8.33	0.446 [0.427, 0.464]	15.07	0.592 [0.570, 0.617]
	GraftNet‡	13.71	0.392 [0.373, 0.410]	28.43	0.459 [0.431, 0.487]
Architectural control: PPR-free retrieval (MID-based Hit@1§*)*					
	EPR-KGQA (CR)	59.22	0.012	64.25	0.039
	EPR-KGQA (RS)	59.76	0.006	63.09	0.051

in Table 1 is the CR ELQ row: $\alpha_{\mathrm{pert}} = 74.0\%$ (the gold answer entity is present in the subgraph) yet EM collapses to 0.68%. Of the 3,531 CWQ CR questions, the gold answer entity is present in the ELQ subgraph for 2,612 (74.0%), yet the GNN answers correctly for only 24 (0.68%). The 3,531 questions partition into three groups: 24 correct (0.68%); 2,588 present-but-unreachable (73.3%), where the gold entity is in the subgraph but the connecting path is not; and 919 absent (26.0%), where the gold entity is not retrieved at all. The present-but-unreachable group dominates, isolating subgraph topology rather than coverage as the failure mode. Contrast with NSM-GEM ($\alpha_{\mathrm{pert}} = 60.8\%$, EM = 30.25%; Table 7) and GraftNet-GEM ($\alpha_{\mathrm{pert}} = 63.3\%$, EM = 29.82%; Table 7): lower coverage yet far higher EM, because their PPR re-seeds route mass toward the *accessible* answer path rather than merely including the answer entity. High answer presence does not help when the *path to the answer has changed*. Two mechanisms interact. First, ELQ seeds are stable under CR (97.5% same top-1 seed, Jaccard = 0.885): ELQ effectively reuses the clean PPR walk, leaving subgraph topology unchanged. Second, compositional restructuring shifts the *reasoning chain* the GNN must follow. PPR mass concentrates around paths relevant to the original question; the answer entity sits at the far end of a restructured chain that was not emphasised. This is a *topology* effect: a 200-question PPR score-distribution analysis finds a clean-vs-CR median score ratio of 0.987×, confirming the PPR algorithm runs identically from the same seeds, yet the stochastic top-N cutoff produces an 11.5% triple-level difference (Jaccard = 0.885) that excludes the intermediate-hop path required by the restructured reasoning chain. Fast-mode isolation (clean subgraph + CR question) confirms this: it recovers 52.76% CWQ EM within 0.14 pp of baseline, showing the GNN decoder is not the culprit. GraftNet addresses exactly this gap: its question-embedding-weighted PPR (oracle GEM seeds) reorients edge weights toward the restructured answer path, raising α_{pert} to 63.3% and EM to 29.82% CWQ (14.98% for the deployment-realistic ELQ+cosine configuration).

Across all 24 (attack × subgraph × dataset) cells in the component ablation (full table in Appendix I), α_{pert} achieves Spearman $\rho = 0.91$ ($p < 10^{-9}$) with EM, confirming it as the dominant predictor of end-to-end accuracy across all conditions. A PPR mass visualisation illustrating the topology-preservation paradox is in Figure 5 (Appendix J.2).

RS resilience: relation synonyms tolerated when seeds are intact. RS retains the highest ELQ

Table 2: Subgraph overlap vs. original GNN-RAG subgraph (MID-string triples, CWQ n=3531, WebQSP n=1639). Ent. Jac. = entity-level Jaccard; Trip. Jac. = triple-level Jaccard; Seeds% = fraction with same top-1 ELQ seed; α_{pert} = gold MID present in perturbed subgraph. †: CR ELQ α_{pert} = 74.0% yet EM = 0.68%: answer present but unreachable via restructured hop path (Section 6).

Attack	Dataset	Subgraph	Ent. Jac.	Trip. Jac.	Seeds%	α_{pert}
CR	CWQ	ELQ	0.882	0.885	97.5	**74.0**†
CR	CWQ	NSM	0.108	0.073	97.5	29.5
CR	CWQ	**GraftNet**	0.030	0.020	97.5	**63.3**
RS	CWQ	GraftNet-Orig	0.223	0.096	97.5	68.6
RS	CWQ	NSM	0.148	0.099	97.5	43.2
RS	CWQ	GraftNet	0.007	0.004	97.5	23.2
RS	WebQSP	GraftNet-Orig	0.171	0.106	77.8	**89.3**
RS	WebQSP	NSM	0.041	0.027	77.8	28.2
RS	WebQSP	GraftNet	0.023	0.017	77.9	38.2

Table 3: GNN decoder Hit@1 with clean subgraph and perturbed question only. ReaRev rows: MID-based Hit@1 (GNN-internal; clean 57.4%/74.3%). Gold-MID EM_{fast}: CWQ CR = 52.76%, CWQ RS = 51.69%, WebQSP CR = 69.74%, WebQSP RS = 70.59%. NSM rows: MID-based Hit@1 (LSTM backbone; clean 40.44%/68.33%). S2 and S3 results are in Appendix K. Ins. cos. = mean cosine similarity of instruction vectors vs. original (ReaRev only).

Model	Attack	CWQ Hit@1	CWQ Ins. cos.	WebQSP Hit@1	WebQSP Ins. cos.
ReaRev (3-layer instruction-decoder GAT)					
	S2 (voice flip)	57.2	0.947	72.9	0.904
	CR	57.5	0.915	68.2	0.842
	S3 (entity ins.)	56.5	0.937	70.5	0.902
	RS	57.4	0.936	70.6	0.876
NSM (LSTM backbone; clean 40.44%/68.33%)					
	CR	39.59	–	57.47	–
	RS	39.45	–	63.45	–

EM of the two primary attacks (20.3% CWQ, 50.9% WebQSP). Entity mentions are unchanged, so ELQ seeds are correct and the subgraph is well-formed. The fast-mode isolation (clean subgraph + RS question) yields 51.69% CWQ and 70.59% WebQSP, within 1.21 pp and 3.72 pp of baseline, confirming the GNN decoder self-corrects for relation synonym substitution when the subgraph is intact (instruction cosine = 0.936 CWQ, 0.876 WebQSP). The residual full-pipeline RS drop ($52.9\% - 20.3\% = 32.6$ pp CWQ) therefore originates in subgraph divergence between GEM and ELQ seeding (Trip. Jaccard = 0.041), not in GNN instruction decoder failure.

Why ELQ dominates RS but GraftNet dominates CR. The divergence in optimal subgraph source follows directly from α_{pert}. For RS, ELQ reuses the original PPR walk (≈97.5% same seeds on CWQ), giving α_{pert} = 74.0% and the highest RS EM; GEM-seeded variants build structurally distinct subgraphs (Trip. Jaccard = 0.004–0.096), so their EM tracks their lower α_{pert}. For CR, ELQ's α_{pert} = 74.0% is equally high, but the walk is anchored to the original reasoning chain: the restructured hop order is absent from the subgraph topology, so EM collapses to 0.68%. GraftNet's question-embedding-weighted PPR reorients toward the restructured path, raising EM to 29.82% despite lower coverage (63.3% vs. 74%).

Architectural controls. GMT-KBQA (S-expression generation, bypasses fixed-subgraph retrieval) drops only 4.5 pp CWQ under CR-type perturbation vs. GNN-RAG's 52.2 pp collapse, confirming S-expression generation is architecturally more robust to compositional restructuring. EPR-KGQA (atomic adjacency patterns, single-shot) retains 59.2% CWQ Hit@1 under CR (vs. 0.68% for GNN-RAG ELQ) and 59.8% under RS. ExplaiGNN (Christmann et al., 2023) (iterative subgraph chaining over Wikidata, ConvMix, 4,800 turns) collapses from 33.9% clean P@1 to 9.8% under CR and 10.2% under RS (>70% relative drop). Together, these controls sharpen the vulnerability boundary: *single-shot* pattern-matching

Table 4: LLM reasoning over GNN-RAG candidates (perturbed questions). GNN EM = gold-MID match over non-empty-candidate subset (differs from Table 1; e.g. CR ELQ full-set = 0.68%, subset = 24.19%). RoG = fine-tuned GCR-Llama-2-7b (Mavromatis and Karypis, 2024); RoG+PathOnly = base RoG with predicted paths at inference (no RA fine-tuning); RoG+RA = RoG fine-tuned on path-augmented data. Clean: CWQ RoG Hit@1 = 56.4%; WebQSP = 80.6%.

		CWQ		**WebQSP**	
Attack	Model	GNN EM (%)	Hit@1 (%)	GNN EM (%)	Hit@1 (%)
CR: Compositional Restructuring					
	RoG	24.19	45.62	45.03	52.83
	RoG+PathOnly	24.19	**51.43**	45.03	63.51
	RoG+RA	24.19	50.47	45.03	**63.76**
	Llama	24.19	9.60	45.03	27.61
RS: Relation Synonym Swap					
	RoG	36.53	46.08	61.93	57.49
	RoG+PathOnly	36.53	**51.29**	61.93	63.27
	RoG+RA	36.53	**51.91**	61.93	**64.00**
	Llama	36.53	9.69	61.93	22.79

(EPR-KGQA) is robust; PPR-based retrieval (GNN-RAG) and iterative multi-turn retrieval (ExplaiGNN) are vulnerable via distinct mechanisms: PPR topology-anchoring and turn-level context corruption, respectively. Full results and failure case examples are in Appendix K.5 and J.

7 Conclusion

We presented a stage-isolation protocol for evaluating GNN-based KGQA systems under query-side adversarial perturbations. Compositional restructuring (CR) and relation synonym swap (RS) reveal that subgraph construction quality is the primary performance ceiling: under CR, the gold answer is present in 74% of ELQ subgraphs yet GNN-RAG achieves only 0.68% CWQ EM, because PPR topology is anchored to the original reasoning chain. Stage-isolation confirms the GNN instruction decoder is robust when the subgraph is intact (52.76% CWQ EM with clean subgraph and perturbed question), attributing 52.08 pp of the 52.22 pp total drop to subgraph topology failure. GraftNet's question-embedding-weighted PPR raises α_{pert} to 63.3% and EM to 29.82% CWQ (14.98% for the deployment-realistic ELQ+cosine configuration), and inference-time path injection recovers 51.4% CWQ EM without task-specific fine-tuning. Perturbed datasets and evaluation infrastructure are released to facilitate future robustness work.

The vulnerability is specific to PPR-seeded pipelines: single-shot pattern-matching (EPR-KGQA) and S-expression generation (GMT-KBQA) are substantially more robust, confirming that bypassing fixed PPR subgraph retrieval is the key architectural defense. The most direct improvement is question-conditioned retrieval or beam-search over relational paths, and combining path injection with topology-aware retrieval is a natural next step. Hardening EL for entity substitution and a stage-guided cascade attack combining CR and RS are additional open directions.

Limitations

Single system, two benchmarks. All results are from GNN-RAG (ReaRev backbone) on CWQ and WebQSP (Freebase-based). Bootstrap 95% CIs are per-comparison and not corrected for multiple comparisons across attack types, subgraph variants, and dataset combinations; individual intervals should be read as descriptive, not family-wise guarantees. We report Bonferroni 99.8% CIs ($\alpha^* = 0.05/24 \approx 0.002$) in Appendix L; the five primary conclusions all survive correction. Differences below 1 pp (e.g. GEM + Cosine vs. GEM + Flat PPR in Table 11) are non-significant and are treated as ties, consistent with seed quality dominating subgraph construction.

Survivorship bias in pass rates. Perturbations failing the validity filter (PPL > 50 or BERTScore < 0.85) fall back to the original question; GNN EM is computed over the full question set. Joint pass rates: CWQ CR 86.6%, CWQ RS 96.6%, WebQSP CR 90.7%, WebQSP RS 95.0%. Fallback questions use the original query, so reported attack severity is a conservative lower bound on true severity.

GNN architecture coverage. The GNN decoder

robustness finding holds for both ReaRev (3-layer instruction-decoder GAT) and NSM (LSTM backbone; Table 3). UniKGQA is excluded as no public inference checkpoint is available.

Freebase deprecation. Both benchmarks use Freebase (2015 static dump via local Virtuoso). The structural failures identified here (subgraph topology mismatch, reasoning-path disruption) are pipeline-architectural and not Freebase-specific. The ExplaiGNN evaluation (Appendix K.5) provides partial Wikidata evidence that the perturbations transfer: CR and RS cause >70% relative P@1 collapse on ConvMix/Wikidata, a different KG backend, suggesting the vulnerability is not an artefact of Freebase's schema. A full Wikidata replication of the core GNN-RAG CR/RS findings remains the most important next step.

GraftNet CR decomposition. GEM seeds use gold SPARQL entity MIDs not available at deployment; the fully deployment-realistic configuration (ELQ+cosine PPR) reaches only 14.98% CWQ CR, well below the GEM-seeded upper bound of 29.82%, confirming that seed quality is the binding constraint rather than PPR flavour (Appendix K.2).

Ethical Considerations

Misuse potential and dual-use framing. CR and RS are the primary adversarial attacks; ES is a diagnostic perturbation used to characterise entity-linker brittleness, not an adversarial attack. All three target published benchmark systems (GNN-RAG, EPR-KGQA, GMT-KBQA, ExplaiGNN), not commercial deployments. Perturbations are released under a research-use license and paired with two concrete defences (question-embedding-weighted PPR and inference-time path injection), so the primary utility is defensive robustness evaluation, not adversarial exploitation.

Dataset and annotation scope. Both CWQ and WebQSP use Freebase, a static 2015 dump with Western, English-language entity coverage. Perturbations are generated by Llama-3.3-70B-Instruct under constrained prompts and filtered by PPL < 50 and BERTScore ≥ 0.85 before SPARQL denotation verification; the filters were calibrated on Freebase-backed questions only. Vulnerability magnitudes may not transfer to multilingual or domain-specific KGs, and users applying this protocol to higher-stakes settings should validate perturbation quality independently.

Broader impact on trustworthy AI. KGQA systems are increasingly deployed in factual information-retrieval pipelines where accuracy is critical. Demonstrating that near-total performance collapse (0.68% CWQ EM under CR) can occur without any change to entity mentions highlights a systemic fragility invisible to standard unperturbed benchmarks. The stage-isolation framework and released evaluation infrastructure give system builders a specific, actionable target: subgraph topology quality accounts for 52.08 pp of the 52.22 pp total CWQ CR drop, and improving it directly improves robustness.

References

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 209–220, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. 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*.

Christoph Christmann, Rishiraj Saha Roy, and Gerhard Weikum. 2023. Explainable conversational question answering over heterogeneous sources via iterative graph neural networks. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023)*.

Rajarshi Das, Manzil Zaheer, Dung Thai, Ameya Godbole, Ethan Perez, Jay Yoon Lee, Lizhen Tan, Lazaros Polymenakos, and Andrew McCallum. 2021. Case-based reasoning for natural language queries over knowledge bases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9594–9611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wentao Ding, Jinmao Li, Liangchuan Luo, and Yuzhong Qu. 2024. Enhancing complex question answering

over knowledge graphs through evidence pattern retrieval. *Preprint*, arXiv:2402.02175.

Guangze Gao, Zixuan Li, Chunfeng Yuan, Jiawei Li, Wu Jianzhuo, Yuehao Zhang, Xiaolong Jin, Bing Li, and Weiming Hu. 2025. D-RAG: Differentiable retrieval-augmented generation for knowledge graph question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35398–35417, Suzhou, China. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, arXiv:2407.21783. Perturbation generation in this work uses Llama-3.3-70B-Instruct specifically.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2025. Hipporag: Neurobiologically inspired long-term memory for large language models. *Preprint*, arXiv:2405.14831.

Gaole He, Yunshi Lan, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. Improving multi-hop knowledge base question answering by learning intermediate supervision signals. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 553–561. ACM.

Jiajun Hou, Chenyu Zhang, and Rui Meng. 2025. Harnessing deep llm participation for robust entity linking. *Preprint*, arXiv:2511.14181.

Wenyu Huang, Guancheng Zhou, Hongru Wang, Pavlos Vougiouklis, Mirella Lapata, and Jeff Z. Pan. 2024. Less is more: Making smaller language models competent subgraph retrievers for multi-hop KGQA. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15787–15803, Miami, Florida, USA. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8018–8025.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.

Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A survey on complex knowledge base question answering: Methods, challenges and solutions. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, pages 4483–4491.

Belinda Z. Li, Sewon Min, Srinivasan Iyer, Yashar Mehdad, and Wen-tau Yih. 2020a. Efficient one-pass end-to-end entity linking for questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6433–6441, Online. Association for Computational Linguistics.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020b. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

Yajie Li, Albert Galimov, Mitra Datta Ganapaneni, Pujitha Thejaswi, De Meng, Priyanshu Kumar, and Saloni Potdar. 2025. Leveraging the power of large language models in entity linking via adaptive routing and targeted reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 871–882, Suzhou (China). Association for Computational Linguistics.

Yuhang Ma, Jie Wang, and Zheng Yan. 2026. Are llm-enhanced graph neural networks robust against poisoning attacks? *Preprint*, arXiv:2603.26105.

Costas Mavromatis, Soji Adeshina, Vassilis N. Ioannidis, Zhen Han, Qi Zhu, Ian Robinson, Bryan Thompson, Huzefa Rangwala, and George Karypis. 2025. Byokg-rag: Multi-strategy graph retrieval for knowledge graph question answering. *Preprint*, arXiv:2507.04127.

Costas Mavromatis and George Karypis. 2024. GNN-RAG: Graph neural retrieval for large language model reasoning. *arXiv preprint arXiv:2405.20139*.

Costas Mavromatis and George Karypis. 2025. GNN-RAG: Graph neural retrieval for efficient large language model reasoning on knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16682–16699, Vienna, Austria. Association for Computational Linguistics.

Sezen Perçin, Xin Su, Qutub Sha Syed, Phillip Howard, Aleksei Kuvshinov, Leo Schwinn, and Kay-Ulrich Scholl. 2025. Investigating the robustness of retrieval-augmented generation at the query level. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 439–457, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Fábio Perez and Ian Ribeiro. 2022. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*, arXiv:2211.09527. ML Safety Workshop, NeurIPS 2022.

Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. 2020. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings.

In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4498–4507, Online. Association for Computational Linguistics.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Yuan Sui, Yufei He, Nian Liu, Xiaoxin He, Kun Wang, and Bryan Hooi. 2025. FiDeLiS: Faithful reasoning in large language models for knowledge graph question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8315–8330, Vienna, Austria. Association for Computational Linguistics.

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William W. Cohen. 2018. Open domain question answering using early fusion of knowledge bases and text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4231–4242, Brussels, Belgium. Association for Computational Linguistics.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Vollmers, René Speck, Hamada M. Zahera, and Axel-Cyrille Ngonga Ngomo. 2025. Evaluation of entity and relation linking for question answering over knowledge graphs. In *Proceedings of the 13th Knowledge Capture Conference 2025*, K-CAP '25, page 211–214, New York, NY, USA. Association for Computing Machinery.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Jiaqi Xue, Mengxin Zheng, Yebowen Hu, Fei Liu, Xun Chen, and Qian Lou. 2024. BadRAG: Identifying vulnerabilities in retrieval augmented generation of large language models. *arXiv preprint arXiv:2406.00083*.

Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic parsing via staged query graph generation: Question answering with knowledge base. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1321–1331, Beijing, China. Association for Computational Linguistics.

Wen-tau Yih, Matthew Richardson, Chris Meek, Ming-Wei Chang, and Jina Suh. 2016. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–206, Berlin, Germany. Association for Computational Linguistics.

Jing Zhang, Xiaokang Zhang, Jifan Yu, Jian Tang, Jie Tang, Cuiping Li, and Hong Chen. 2022. Subgraph retrieval enhanced model for multi-hop knowledge base question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5773–5784, Dublin, Ireland. Association for Computational Linguistics.

Tianzhe Zhao, Jiaoyan Chen, Yanchi Ru, Haiping Zhu, Nan Hu, Jun Liu, and Qika Lin. 2025. RAG safety: Exploring knowledge poisoning attacks to retrieval-augmented generation. *arXiv preprint arXiv:2507.08862*, arXiv:2507.08862.

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.

Dongzhuoran Zhou, Yuqicheng Zhu, Xiaxia Wang, Hongkuan Zhou, Yuan He, Jiaoyan Chen, Steffen Staab, and Evgeny Kharlamov. 2026. What breaks knowledge graph based RAG? benchmarking and empirical insights into reasoning under incomplete knowledge. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2522–2538, Rabat, Morocco. Association for Computational Linguistics.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2024. PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. *arXiv preprint arXiv:2402.07867*.

A Per-Stage Degradation Metrics and Metric Equivalence

We track failure propagation through four metrics, one per pipeline stage.

Δ_{EL}: Entity linking degradation.

$$\Delta_{\mathrm{EL}} = \mathrm{SeedHit}(E_q) - \mathrm{SeedHit}(E_{q'}) \quad (2)$$

where

$$\mathrm{SeedHit}(E) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[\hat{E}_i \cap E_i^* \neq \emptyset\right] \quad (3)$$

$\hat{E}_i$ is the set of MIDs linked by ELQ for question i; E_i^* is the set of gold seed MIDs. We use any-match because GNN-RAG requires only one correct seed entity to initiate PPR.

Δ_{SR}: Subgraph retrieval drift.

$$\Delta_{\text{SR}} = 1 - J(\mathcal{G}_q, \mathcal{G}_{q'}) \tag{4}$$

where J is the triple-level Jaccard on Freebase string triples (s, p, o). High Δ_{SR} means the subgraph has changed substantially; α_{pert} (below) measures quality directly.

α_{pert}: Answer presence in perturbed subgraph.

$$\alpha_{\text{pert}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[a_i^* \in \mathcal{G}_{q_i'}\right] \tag{5}$$

α_{pert} is the strongest predictor of final EM across subgraph variants and serves as the direct measure of subgraph construction quality.

Δ_{AG}: End-to-end answer generation drop.

$$\Delta_{\text{AG}} = \text{EM}_{\text{clean}} - \text{EM}_{\text{pert}} \tag{6}$$

where $\text{EM}(q) = \mathbf{1}[a^* \sqsubseteq \hat{a}]$: the gold Freebase MID a^* must appear as a substring in $\hat{a}$. We report bootstrap 95% CIs on Δ_{AG} ($n = 1000$ resamples).

Metric equivalence: MID-based Hit@1. Both GNN-RAG and EPR-KGQA report **MID-based Hit@1** (not entity-name Hit@1) for CWQ and WebQSP. In `evaluate.py`, `entity2name` is `None` for standard CWQ/WebQSP data folders, so `f1_and_hits` operates on integer indices into the MID-keyed entity vocabulary; the reported clean Hit@1 (CWQ = 57.4%, WebQSP = 74.3%) equals the paper's gold-MID EM. EPR-KGQA (NSM-H) similarly selects `kb_id` for string-MID answers, indexing the same MID-keyed vocabulary. The $\sim$58 pp gap under CR is therefore a direct architectural comparison; the prior caveat about a "$\sim$5 pp metric leniency" was incorrect and has been removed.

B Hardware and Reproducibility Details

All experiments run on a single node with four NVIDIA A100-SXM4-80 GB GPUs and 96 CPU cores. GNN-RAG inference uses data-parallel subgraph batching across all four GPUs; LLM reasoning (RoG, Llama-3.1-8B) uses a single GPU (batch size 8). Perturbation generation uses Llama-3.3-70B-Instruct via a single-GPU API call; each perturbation request takes approximately 2 s. Total compute for all experiments: approximately 800 A100 GPU-hours.

Model sizes. Key models and their parameter counts: Llama-3.3-70B-Instruct (70B), used for perturbation generation; Llama-2-7B (7B), the fine-tuned GNN-RAG answer generator, the RoG backbone; Llama-3.1-8B (8B), for lightweight tasks: ELQ, a BERT-Large bi-encoder ($\approx$340 M parameters). The GNN reasoning component (ReaRev, 3-layer GAT) is a lightweight graph model whose parameter count is small relative to the LLM components; we use the published GNN-RAG checkpoint without modification.

Freebase SPARQL endpoint. Freebase is accessed via a local Virtuoso SPARQL endpoint loaded from the 2015/08/17 RDF dump (28 GB compressed; expanded to $\approx$140 GB). The public Freebase API was deprecated in 2015; all experiments use this static dump. The Wikidata-Freebase MID alignment is precomputed offline using `skos:altLabel` and `owl:sameAs` links; QID $\rightarrow$ MID mapping is stored in `qid_to_mid.pkl` and used for CR/RS denotation preservation verification and ES entity validity checks.

C Subgraph Retrieval Variants

All variants use the same GNN-RAG ReaRev checkpoint; only $\mathcal{G}_{q'}$ changes.

ELQ (default). ELQ-seeded flat PPR (topology-only, no relation weighting), following the NSM preprocessing methodology (He et al., 2021). For RS, the ELQ subgraph is structurally identical to the clean run, making RS ELQ EM a direct measure of GNN instruction decoder sensitivity.

GraftNet-Orig (Sun et al., 2018). Question-aware PPR from ELQ seeds, with edge weights biased toward relations whose embeddings are similar to the question (cosine weighting). Best non-ELQ variant for RS.

NSM (He et al., 2021). BFS expansion followed by topology-only PPR pruning. NSM learns relation relevance at inference via its LSTM instruction mechanism rather than front-loading it into retrieval.

GraftNet (ours). Extends GraftNet-Orig with: (i) GNN-Integer entity decoding mapping GNN

scores directly to Freebase MIDs; (ii) GEM (Golden Entity Map) (Das et al., 2021) seeds for evaluation; (iii) question-embedding cosine PPR weighting that concentrates mass on paths through question-relevant predicates. **Scope**: GraftNet's contribution is the CR improvement; for RS, GraftNet-Orig achieves higher answer coverage (Section 5.4).

EPR-KGQA (Ding et al., 2024). Atomic-adjacency-pattern-based neighbourhood pruning with an LSTM instruction mechanism. Evaluated with the authors' released checkpoint.

D ELQ SeedHit Under Perturbation

Table 5 reports ELQ any-GT SeedHit (Δ_{EL}) under each perturbation type. The key finding is that CR and RS cause ≤ 0.3 pp SeedHit degradation, confirming that entity linking is not the failure mode for these two primary attacks, where any EM drop must originate downstream of the EL stage. S1 (relation paraphrase) causes 13 pp SeedHit collapse on CWQ because it partially alters entity surface forms, contaminating the stage-isolation signal; this is why S1 is reported as a secondary type rather than a primary attack. ES SeedHit figures (23.1% CWQ, 14.1% WebQSP) are in Section E.

Table 5: ELQ SeedHit (%) under perturbation (any-GT: $\hat{E}_i \cap E_i^* \neq \emptyset$). Cl. = clean baseline; S1–S4 = secondary types (Appendix K). CR and RS cause ≤ 0.3 pp degradation, confirming EL is not the failure mode.

	Cl.	S1	S2	CR	S3	RS
CWQ	57.4	44.1	57.4	57.8	57.2	56.8
WebQSP	68.3	47.0	68.2	67.5	67.1	68.1

E ES Diagnostic: Entity Swap

Entity Swap (ES) substitutes the topic entity with a different Freebase entity from the same domain, producing a new MID and a new gold answer. Because the correct answer changes by design, ES is not an adversarial attack under our definition; it is a diagnostic measuring entity-linker brittleness and the upper bound of what better EL could unlock.

ELQ under ES. ELQ drops from 57.4% to 23.1% any-GT hit rate on CWQ and from 68.3% to 14.1% on WebQSP. This 34.3 pp/54.2 pp collapse propagates to near-zero EM (0.31% CWQ, 0.06% WebQSP). Two effects compound: (1) wrong seeds propagate through GNN message-passing; (2) 45.6% of CWQ ES questions are structurally unanswerable by construction (the substituted entity has no Freebase path matching the question's relation chain).

GraftNet-BFS upper bound. When the substituted entity MID is provided (oracle setting, not available at deployment), GraftNet-BFS builds subgraphs via BFS from that MID, bypassing ELQ. This recovers 8.2% CWQ EM / 44.6% WebQSP EM overall, and 14.1% / 51.4% on the answerable subset. This is an upper bound on what a perfect entity linker would unlock, not a deployable defence. (GraftNet-BFS is distinct from the main-paper GraftNet, which uses question-embedding-weighted PPR from GEM oracle seeds; BFS seeding is used only in this ES diagnostic.)

Answerability confound. Table 6 separates answerable from unanswerable ES questions. ELQ EM is near-zero on both subsets; GraftNet-BFS gain is real on answerable questions (14.1% vs. 8.2% CWQ overall), confirming that the EL failure is not contingent on structural unanswerability. Unanswerability is a compounding factor, not the primary cause of near-zero EM.

Table 6: ES EM (%) on answerable-only vs. overall subsets. CWQ: 1,920/3,531 answerable; WebQSP: all answerable. G-BFS = GraftNet-BFS (oracle MID seeding, not available at deployment). Ans. = answerable-only EM; Overall = full test-set EM.

Dataset	Subgraph	Overall	Ans.
CWQ	ELQ	0.31	0.31
CWQ	G-BFS	8.20	**14.1**
WSP	ELQ	0.06	0.00
WSP	G-BFS	44.60	**51.4**

A detailed breakdown of answerability effects is in Appendix K.1.

EL hit rates. Full ELQ any-GT hit rates: CWQ: Clean = 57.4%, ES = 23.1%, CR = 57.8%, RS = 56.8%; WebQSP: Clean = 68.3%, ES = 14.1%, CR = 67.5%, RS = 68.1%. The stability of CR and RS hit rates confirms that entity-mention-preserving attacks do not damage the entity linker, making any downstream drop unambiguously attributable to subgraph construction or GNN reasoning.

F Formal Perturbation Rules

Core rule (CR and RS). Every perturbation $q \to q'$ must satisfy: $\text{SPARQL_den}(q', \text{Freebase}) = \text{SPARQL_den}(q, \text{Freebase})$. This ensures the correct answer is unchanged, which is required for CR and RS to qualify as adversarial attacks. ES does not satisfy this rule by design (answer changes with entity); ES generation rules are in Appendix E.

Forbidden operations (CR) and rationale. The following transformations are explicitly excluded because they change the answer denotation, invalidating the answer-preservation constraint:

- **Predicate weakening** ("won" → "shortlisted for"): expands denotation; invalid.
- **Predicate strengthening**: may collapse denotation to empty; invalid.
- **Scope expansion** ("official language" → "spoken language"): changes cardinality of answer set; invalid.
- **Fictional constraint**: intermediate entity fails the constraint in Freebase; SPARQL returns empty; invalid.

G ReaRev Instruction Decoder Analysis

The ReaRev instruction decoder extracts K instruction vectors $\{i^{(k)}\}$ from the question via token-span attention. A perturbation $q \to q'$ shifts these vectors: $\Delta_{\mathbf{i}} = \frac{1}{K}\sum_k \|i^{(k)}(q) - i^{(k)}(q')\|_2$. With a clean subgraph, EM_{fast} for CR stays within 0.14 pp of the 52.9% clean baseline on CWQ (52.76%), confirming the instruction decoder is not the primary failure point when the subgraph is correct. This result underlies the attribution in Section 5.5: 52.08 of the 52.22 pp total CR CWQ drop is subgraph-attributable, not decoder-attributable.

H Perturbation Prompt Templates

Figure 3 shows the prompt templates for the three primary perturbation types (ES, CR, RS), all implemented via Llama-3.3-70B-Instruct. The same Llama checkpoint is used for all seven perturbation types; only the system instruction differs. Figure 4 shows the corresponding templates for the four secondary types S1–S4.

I α_{pert} vs. EM: All 24 Cells

Table 7 lists all 24 (attack × subgraph × dataset) cells with their α_{pert} and EM values, supporting the Spearman $\rho = 0.91$ ($p < 10^{-9}$) claim in Section 6. The correlation is computed over all 24 cells; bootstrap 95% CI on ρ: $[0.80, 0.97]$ ($n = 1000$ bootstrap resamples over cells). ELQ CR cells are notable: high α_{pert} (74%) paired with near-zero EM (0.68% CWQ, 0.49% WebQSP) is the key anomaly that exposes the answer-presence vs. answer-reachability distinction.

J Failure Analysis: Case Examples and PPR Visualisation

J.1 Failure Case Examples

The two cases below are drawn from RS (P7, relation synonym swap) evaluated with GraftNet-Orig subgraphs, illustrating two distinct failure mechanisms. RS failures are not monolithic: in Case 1 the gold answer is absent from the retrieved subgraph; in Case 2 the answer is reachable but the GNN assigns higher probability to an incorrect entity.

Case 1: Answer absent from the perturbed subgraph. The perturbed question retrieves a different subgraph whose entity set does not include the gold answer MID. Correct GNN reasoning is irrelevant because the answer entity is not among the candidates.

Case 2: Answer present but GNN reasoning fails. The gold answer entity is reachable in the perturbed subgraph, but GNN-RAG assigns higher final probability to a different entity. This case illustrates that even when $\alpha_{\text{pert}} = 1$, the GNN can still fail if the perturbed subgraph topology provides insufficient message-passing support for the answer node.

J.2 PPR Mass Topology Mismatch: Example Visualisation

Figure 5 shows four examples of how compositional restructuring (CR) exposes the topology-anchoring failure in GNN-RAG. Each row shows the same 2-hop entity neighbourhood under the clean question (left) and the CR-perturbed question with hop order reversed (right). Node colour encodes PPR mass (red = high mass, green = low mass); blue double-circles are ELQ-linked seeds; the red double-bordered node is the gold answer.

ELQ's PPR walk is anchored to the original question's reasoning chain; the hop-reversed path from the seed entity to the gold answer is not emphasised in the perturbed subgraph, so the GNN cannot navigate to the answer via the restructured traversal order. The bottleneck is the subgraph topology, not the GNN decoder: when the decoder receives

ES: Entity Swap

```
[SYSTEM] You are a question rewriter for a KGQA robustness study. Replace the topic entity with
a different entity from the same Freebase domain. Rules: (a) the replacement must be a valid
Freebase entity with its own gold answer to the question; (b) do not change the predicate or any
other part of the question; (c) the new entity must appear in Wikidata with a Freebase MID; (d)
output exactly one rewritten question followed by the new entity name.

[USER]
Question: {question}  Topic entity: {entity}  Original answer: {answer}
Substitute the topic entity with a different entity from the same domain.
```

CR: Compositional Restructuring

```
[SYSTEM] Apply exactly one KG-safe structural transformation while preserving the original
answer. Allowed operations: (1) hop-order reversal: ask about the intermediate entity first;
(2) distractor constraint injection: add a true constraint on the intermediate entity only;
(3) intermediate alias substitution: replace the intermediate entity with an alias. STRICTLY
FORBIDDEN: predicate changes, scope expansion (e.g. "won" to "shortlisted for"), entity removal,
predicate weakening or strengthening. Output the operation name, then the rewritten question.

[USER]
Question: {question}  Correct answer: {answer}
Apply one KG-safe operation.
```

RS: Relation Synonym Swap

```
[SYSTEM] Replace the predicate or relation phrase in the question with a synonym or paraphrase,
keeping all entity mentions exactly unchanged. Rules: (a) the rewritten predicate must express
the same underlying Freebase relation; (b) do not change any entity mention or constraint; (c) do
not change the correct answer; (d) produce fluent English. Output exactly one rewritten question.

[USER]
Question: {question}  Relation: {relation_surface}  Correct answer: {answer}
Rewrite using a relation synonym only.
```

Figure 3: **Prompt templates for the three primary perturbation types. ES** (blue) substitutes the topic entity with a *different* entity from the same Freebase domain, changing both the entity MID and the gold answer; this directly stresses the entity linker. **CR** (purple) applies one of three KG-safe structural operations with an explicit forbidden list, preserving SPARQL denotation while disrupting the multi-hop reasoning chain; entity mentions are not changed. **RS** (red) replaces only the predicate surface form with a synonym, leaving entity mentions and the answer unchanged; this targets the GNN instruction decoder while keeping entity seeds and the subgraph intact. All seven perturbation types (ES, CR, RS, and secondary S1–S4 in Appendix K) are generated with Llama-3.3-70B-Instruct (Grattafiori et al., 2024) in a constrained setting: each prompt enforces explicit forbidden operations and is filtered by PPL and BERTScore before SPARQL denotation verification. Prompts for S1–S4 are in Appendix H.

the *clean* subgraph with the perturbed question, it achieves 52.76% CWQ EM (fast-mode; instruction cosine = 0.915, Table 3), confirming that the decoder handles CR well when path topology is correct.

K Secondary Perturbation Types and Ablations

K.1 ES Answerability Confound

45.6% of CWQ ES questions are unanswerable by construction: the substituted entity has no Freebase path to any valid answer for the question's relation chain. CR and RS are $\geq$99.8% answerable. Table 6 (Appendix E) separates the two effects: GraftNet-BFS recovers 14.1% CWQ EM on answerable questions (vs. 8.2% overall), confirming that structural unanswerability suppresses the overall ES number independently of EL failure. ELQ remains near zero regardless of answerability because wrong seeds completely derail PPR.

K.2 Subgraph Component Ablation: Seeds × PPR

Table 11 decouples seed quality (GEM oracle vs. ELQ production) from PPR flavour (cosine question-embedding weighting vs. flat uniform weighting). The key finding is that *seed quality dominates*: GEM seeds account for $\approx$15 pp of the CR improvement regardless of PPR type.

S1: Relation/Predicate Synonym Paraphrase (mixed entity change)

```
[SYSTEM] Rewrite the question using a synonym or paraphrase for the main predicate phrase. You may also rephrase entity descriptions but do not change any proper noun entity names or the correct answer. Output one rewritten question only.
[USER] Question: {question}  Answer: {answer}
```

S2: Active ↔ Passive Voice Flip

```
[SYSTEM] Convert the question between active and passive voice, or vice versa, while keeping the meaning, all entities, and the correct answer exactly the same. Output one rewritten question only.
[USER] Question: {question}  Answer: {answer}
```

S3: Adversarial Entity Insertion (distractor)

```
[SYSTEM] Insert one plausible but incorrect entity into the question as a distractor, without changing the correct entity or the answer. The inserted entity must be from the same domain as the topic entity. Output the rewritten question only.
[USER] Question: {question}  Topic entity: {entity}  Answer: {answer}
```

S4: Mixed Lexical Noise

```
[SYSTEM] Apply minor surface-form variation to the question: synonym substitution, word order change, or light paraphrase. Do not change entity names or the correct answer. Output one rewritten question only.
[USER] Question: {question}  Answer: {answer}
```

Figure 4: Prompt templates for secondary perturbation types S1–S4. All generated with Llama-3.3-70B-Instruct under the same PPL and BERTScore filters as the primary types.

The fully deployment-realistic configuration (ELQ seeds + cosine PPR) reaches 14.98% CWQ CR, which is a meaningful gain over the 0.68% ELQ flat-PPR baseline, but well below the GEM-seeded upper bound of 29.82%.

Significance of the GEM + Cosine vs. GEM + Flat difference. On CWQ CR, the 0.46 pp gap (29.79% vs. 30.25%) is not statistically significant: paired bootstrap test (n_{boot}=10,000), $p = 0.478$. This confirms that continuous cosine PPR weighting provides no benefit over uniform flat PPR on CWQ once seed quality is fixed: the performance ceiling is determined by seed quality, not PPR flavour. On WebQSP CR, GEM + Cosine is 5.6 pp higher (31.18% vs. 25.56%; $p < 0.001$), indicating that cosine weighting does help on shorter 1–2 hop chains where embedding alignment is more decisive. The deployment-realistic gap (ELQ + Cosine vs. ELQ + Flat on CWQ: 14.98% vs. 15.38%) is similarly non-significant ($p = 0.58$), confirming that *the bottleneck for deployment is seed quality, not PPR weighting*. The GraftNet-V2 binary-mask variant (Appendix K.6) provides complementary evidence.

Reframing. The non-significant CWQ difference and the significant WebQSP difference together support a refined conclusion: continuous cosine PPR weighting is not the driver of GraftNet's CR improvement on CWQ (where the p-value confirms a tie), but it contributes on WebQSP's simpler hop structure. A deployment-realistic improvement to the CR bottleneck on CWQ therefore requires better entity linking (better seeds), not a more sophisticated PPR variant.

K.3 Secondary Perturbation Types

Table 12 reports GNN-RAG EM under the four secondary perturbation types (S1–S4). All four collapse performance to $\leq$0.65% EM, comparable in magnitude to CR. Critically, unlike CR and RS, the secondary types simultaneously stress EL and downstream stages: S1 alters entity surface forms (SeedHit drops to 44% on CWQ, Table 5), making clean stage-isolation impossible. S2 (active ↔ passive) mostly preserves entity mentions but voice changes are less semantically targeted.

Table 7: All 24 (attack × subgraph × dataset) cells: α_{pert} (gold answer present in subgraph) and EM (%). GraftNet uses GEM oracle seeds. ELQ CR $\alpha_{\text{pert}} = 74\%$ with EM = 0.68% (CWQ) is the anomalous cell exposing topology-anchoring failure. Spearman $\rho = 0.91$ ($p < 10^{-9}$); bootstrap 95% CI: [0.80, 0.97].

Attack	Dataset	Subgraph	Seeds	α_{pert} (%)	EM (%)
CR	CWQ	ELQ	ELQ	74.0	0.68
CR	CWQ	NSM	ELQ	29.5	5.81
CR	CWQ	GraftNet	GEM	63.3	29.82
CR	CWQ	ELQ+cosine	ELQ	52.1	14.98
CR	WebQSP	ELQ	ELQ	74.3	0.49
CR	WebQSP	NSM	ELQ	28.7	1.65
CR	WebQSP	GraftNet	GEM	58.2	36.55
CR	WebQSP	ELQ+cosine	ELQ	48.6	22.33
RS	CWQ	ELQ	ELQ	74.0	20.31
RS	CWQ	GraftNet-Orig	GEM	68.6	11.24
RS	CWQ	NSM	ELQ	43.2	8.33
RS	CWQ	GraftNet	GEM	23.2	13.71
RS	WebQSP	ELQ	ELQ	74.3	50.95
RS	WebQSP	GraftNet-Orig	GEM	89.3	41.67
RS	WebQSP	NSM	ELQ	28.2	15.07
RS	WebQSP	GraftNet	GEM	38.2	28.43
Oracle-seeded reference (NSM-GEM); not included in the 24-cell Spearman correlation:					
CR	CWQ	NSM-GEM	GEM	60.8	30.25
CR	WebQSP	NSM-GEM	GEM	36.24	25.56
RS	CWQ	NSM-GEM	GEM	67.01	37.24
RS	WebQSP	NSM-GEM	GEM	35.63	25.08

Table 8: Case 1 examples: answer absent from perturbed subgraph. The relation synonym shift changes the question embedding used to weight per-relation PPR walks, causing PPR to converge to a different subgraph that excludes the gold entity. Hit@1 = 0 regardless of GNN reasoning quality.

	CWQ example	**WebQSP example**
Orig Q	*What character was **played** by both Josh Pence and Armie Hammer?*	*What kind of **money** do I bring to Mexico?*
Pert Q	*What character was **portrayed** by both Josh Pence and Armie Hammer?*	*What type of **currency** should I take to Mexico?*
Gold	`m.09tb_f3` (Tyler Winklevoss)	`m.012ts8` (Mexican peso)
Pred (orig)	`m.09tb_f3` ✓	`m.012ts8` ✓
Pred (pert)	`m.0hyf5kr` ✗	`m.04sqj` ✗
Gold in pert SG	**No**	**No**

S3 and S4 insert adversarial entities or apply mixed noise, both of which partially perturb ELQ seeds. Because these types conflate EL failure with downstream failure, they are reported for completeness rather than as mechanistically interpretable probes. A consolidated view across all seven perturbation types under ELQ subgraphs is in Section K.4.

S1–S4 with GEM oracle subgraphs. Table 13 repeats the S1–S4 evaluation with GEM oracle seeds (NSM-GEM and GraftNet-GEM variants), isolating the effect of subgraph quality from EL failure. Under oracle subgraphs, all four secondary types recover to ≈30 % CWQ EM, matching CR with the same oracle seeds (NSM-GEM: 30.25%, GraftNet-GEM: 29.79%). A caveat applies to S1: because S1 alters entity surface forms (SeedHit drops to 44% on CWQ, Table 5), GEM oracle seeds simultaneously fix both the EL failure and subgraph quality, so S1's recovery reflects a combined EL+subgraph fix rather than a pure subgraph isolation. For S2–S4, entity mentions are better preserved, so oracle seeds function closer to a genuine subgraph-only fix. Excluding S1, the uniform recovery for S2–S4 provides cleaner support that *subgraph quality is the shared bottleneck*.

K.4 All Perturbation Types: ELQ Baseline

Table 14 provides a consolidated view across all seven perturbation types evaluated with ELQ subgraphs. The table reveals a striking asymmetry: RS (P7) is uniquely robust (20.31% CWQ, 50.95% WebQSP) while all other types collapse to ≤0.7%.

Table 9: Case 2 examples: answer present in perturbed subgraph but GNN reasoning fails. In the CWQ case, the perturbed subgraph has only 89 triples vs. 6,412 in the original; answer entities are present as isolated nodes with insufficient connectivity for GNN message-passing. In the WebQSP case, voice shift alters the SBERT encoding, redirecting ReaRev's query-entity attention despite correct connectivity.

	CWQ example	**WebQSP example**
Orig Q	*What countries **in** the Chamorro Time Zone are **in** Oceania?*	*What district **does** Nancy Pelosi **represent**?*
Pert Q	*What countries **within** the Chamorro Time Zone are **located in** Oceania?*	*What district **is represented by** Nancy Pelosi?*
Gold	`m.05cnr` (Guam), `m.034tl` (N. Mariana Is.)	`m.0b10j3` (CA districts)
Pred (orig)	`m.05cnr` ✓	`m.0b10j3` ✓
Pred (pert)	`m.0dff7s` ✗	`m.09c7w0` (United States) ✗
Gold in pert SG	**Yes (both)**	**Yes (all 3)**
Pert SG triples	89 vs. 6,412 orig	351 vs. 2,146 orig

Table 10: Bonferroni 99.8% CIs ($\alpha^* = 0.05/24 \approx 0.002$) for Δ_{AG} EM, all perturbation types, both datasets. Attack codes: ES = entity alias; S1 = rel. para.; S2 = voice flip; CR = comp. restruct.; S3 = ent. ins.; S4 = mix. noise; RS = rel. syn. All five primary conclusions survive correction; effect sizes 10–52 pp remain far from zero. Sub-1 pp differences (e.g. GEM + Cosine vs. GEM + Flat) remain non-significant.

Attack	Dataset	Mean Δ_{AG}	95% CI	99.8% CI
ES	CWQ	0.526	[0.509, 0.543]	[0.501, 0.553]
ES	WebQSP	0.743	[0.721, 0.764]	[0.709, 0.776]
S1	CWQ	0.527	[0.511, 0.543]	[0.501, 0.552]
S1	WebQSP	0.739	[0.718, 0.761]	[0.707, 0.773]
S2	CWQ	0.524	[0.507, 0.540]	[0.499, 0.550]
S2	WebQSP	0.739	[0.718, 0.761]	[0.703, 0.773]
CR	CWQ	**0.522**	**[0.506, 0.539]**	**[0.496, 0.547]**
CR	WebQSP	**0.738**	**[0.717, 0.760]**	**[0.702, 0.772]**
S3	CWQ	0.525	[0.508, 0.541]	[0.499, 0.552]
S3	WebQSP	0.739	[0.718, 0.760]	[0.707, 0.774]
S4	CWQ	0.523	[0.506, 0.540]	[0.497, 0.549]
S4	WebQSP	0.738	[0.716, 0.759]	[0.705, 0.772]
RS	CWQ	**0.326**	**[0.308, 0.343]**	**[0.299, 0.354]**
RS	WebQSP	**0.234**	**[0.210, 0.257]**	**[0.197, 0.272]**

The explanation is structural: for RS, ELQ's entity seeds are identical to the clean question's, so the PPR walk effectively replays the original subgraph, preserving α_{pert} at 74.0%. For every other type, either entity seeds change (ES, S1, S3, S4) or the subgraph topology diverges from the restructured question (CR), where both paths lead to near-zero EM. This consolidation view motivates why RS and CR were selected as the two primary attacks: RS isolates the GNN decoder while CR isolates the subgraph construction bottleneck.

K.5 GMT-KBQA and EPR-KGQA as Architectural Controls

These two systems serve as architectural controls that help isolate whether the observed vulnerability is specific to PPR-based subgraph retrieval or a general property of GNN-KGQA pipelines.

GMT-KBQA. GMT-KBQA (Das et al., 2021) generates S-expressions directly from the question, bypassing the EL-then-subgraph-retrieval pipeline. Table 15 shows that structural and paraphrase perturbations (CR-type, RS-type, S2-type) cause only $\leq$5 pp degradation, compared to GNN-RAG's 52 pp CR collapse. Entity-related perturbations (ES-type, CR-type with entity insertion) cause 18–26 pp degradation because S-expression generation is sensitive to entity surface forms. The contrast confirms that the GNN-RAG vulnerability to CR is specific to its fixed-subgraph retrieval stage: once that stage is bypassed, compositional restructuring becomes near-harmless.

EPR-KGQA. EPR-KGQA (Ding et al., 2024) uses atomic adjacency patterns (entity-relation-entity tuples) indexed offline and selected at query

Table 11: Seeds × PPR ablation: GNN-RAG EM (%) for CR and RS. GEM = gold SPARQL MIDs (oracle); ELQ = production linker. Clean baselines: CWQ = 52.9%, WebQSP = 74.3%. The GEM/ELQ gap (≈15 pp for CR) quantifies the seed-quality ceiling; cosine PPR adds marginal benefit on top. "Main paper GraftNet (fixed)" uses GloVe relation embeddings (per relation), whereas the controlled GEM + Cosine row uses sentence-level question embeddings; this granularity accounts for the RS gap (13.71% vs. 9.86%) while CR is unaffected (29.82% vs. 29.79%).

		CWQ		**WebQSP**	
Seeds	PPR	CR (%)	RS (%)	CR (%)	RS (%)
GEM	Cosine (question-emb.)	**29.79**	9.86	**31.18**	28.31
GEM	Flat (uniform)	30.25	**37.24**	25.56	**25.08**
ELQ	Cosine (question-emb.)	14.98	9.92	22.33	28.31
ELQ	Flat (uniform)	15.38	8.73	17.69	25.69
Main paper GraftNet (fixed)		*29.82*	*13.71*	*36.55*	*28.43*
ELQ (production baseline)		*0.68*	*20.31*	*0.49*	*50.95*

Table 12: GNN-RAG EM under secondary perturbation types (ELQ seeds; bootstrap 95% CI on Δ_{AG}, n=1000; per-comparison, uncorrected). All four types collapse EM to ≤0.65%, comparable to CR (0.68%), but unlike CR they co-stress EL and downstream stages simultaneously, preventing clean stage-isolation.

	CWQ		**WebQSP**	
Type	EM (%)	Δ_{AG} [95% CI]	EM (%)	Δ_{AG} [95% CI]
S1 (rel. paraphrase)	0.20	0.527 [0.509, 0.543]	0.43	0.739 [0.718, 0.761]
S2 (voice flip)	0.51	0.524 [0.508, 0.540]	0.37	0.739 [0.720, 0.761]
S3 (entity insertion)	0.40	0.525 [0.510, 0.543]	0.43	0.739 [0.719, 0.760]
S4 (mixed lexical noise)	0.57	0.523 [0.507, 0.539]	0.49	0.738 [0.718, 0.760]

time, replacing PPR-based retrieval entirely. Under CR, EPR-KGQA retains 59.22% CWQ and 64.25% WebQSP Hit@1 (vs. 0.68% and 0.49% for GNN-RAG ELQ), a near-baseline result despite the compositional restructuring. Under RS, EPR-KGQA retains 59.76% CWQ and 63.09% WebQSP. The near-zero degradation under both attacks shows that *single-shot* pattern-matching retrieval is robust to both attack types. Contrast this with ExplaiGNN below: EPR-KGQA's robustness is not attributable solely to PPR avoidance, but to its single-shot retrieval that does not propagate errors across turns.

Metric note. EPR-KGQA and GNN-RAG use the same MID-indexed Hit@1. EPR-KGQA's NSM backbone indexes answers by Freebase MID (not entity name): `basic_dataset.py` selects `kb_id` when `answer['kb_id']` is a string (which it is for CWQ/WebQSP), and GNN-RAG's evaluator operates on integer indices into the same MID-keyed entity vocabulary (`entity2name=None` for non-`sr-` datasets). Both systems therefore report top-1 accuracy over the same gold-MID entity set. The ∼58 pp gap under CR is a direct architectural comparison with no metric conversion needed; it is not an artefact of metric leniency.

ExplaiGNN. ExplaiGNN (Christmann et al., 2023) is a conversational KGQA system (Wikidata-based) that processes multi-turn question sequences using a GNN over iteratively refined subgraphs. Unlike GNN-RAG's single-question PPR retrieval, ExplaiGNN retrieves subgraphs via entity-relation pattern matching similar to EPR-KGQA, chaining evidence across turns. We evaluate the published checkpoint on the ConvMix test set (4,800 turns) under CR and RS perturbations applied to each conversation turn.

Table 16 shows results. Clean P@1 is 33.9%, consistent with the originally reported ConvMix result. Under CR, P@1 drops to 9.8% (−24.1 pp; −71% relative); under RS, P@1 drops to 10.2% (−23.7 pp; −70% relative). Both attacks cause a severe collapse comparable in magnitude to GNN-RAG ELQ under CR (0.68% CWQ, −52 pp). The ExplaiGNN result shows that iterative subgraph reasoning systems are *not* immune: when the conversational context is compositionally restructured or relation synonyms are swapped, the pattern-matching retrieval fails to chain the correct entities across turns, and performance collapses. The contrast with EPR-KGQA's near-baseline retention (59.2% P@1 under CR) suggests the vulnerabil-

Table 13: GNN-RAG EM under secondary perturbation types with GEM oracle seeds (bootstrap 95% CI on Δ_{AG}, n=1000, uncorrected). All four types recover to ≈30% CWQ / ≈26–32% WebQSP EM, matching the CR-with-oracle baseline and confirming subgraph quality as the shared bottleneck. Clean EM: CWQ = 52.9%, WebQSP = 74.3%.

		CWQ		**WebQSP**	
Subgraph	Type	EM (%)	Δ_{AG} [95% CI]	EM (%)	Δ_{AG} [95% CI]
NSM-GEM (oracle seeds, uniform PPR)					
	S1 (rel. paraphrase)	30.93	0.220 [0.202, 0.238]	26.11	0.481 [0.456, 0.508]
	S2 (voice flip)	30.42	0.225 [0.208, 0.242]	26.54	0.478 [0.451, 0.504]
	S3 (entity insertion)	30.56	0.223 [0.206, 0.241]	25.56	0.487 [0.460, 0.514]
	S4 (mixed lexical noise)	30.59	0.223 [0.206, 0.240]	25.99	0.483 [0.456, 0.509]
	CR (reference)	*30.25*	*0.227*	*25.56*	*0.487*
GraftNet-GEM (oracle seeds, question-embedding PPR)					
	S1 (rel. paraphrase)	30.30	0.226 [0.208, 0.244]	31.54	0.427 [0.400, 0.454]
	S2 (voice flip)	29.40	0.235 [0.216, 0.255]	31.60	0.427 [0.400, 0.455]
	S3 (entity insertion)	29.91	0.230 [0.211, 0.248]	31.18	0.431 [0.403, 0.460]
	S4 (mixed lexical noise)	30.08	0.228 [0.210, 0.246]	31.24	0.431 [0.403, 0.459]
	CR (reference)	*29.79*	*0.231*	*31.18*	*0.431*

Table 14: GNN-RAG EM (%) with ELQ subgraph, all perturbation types. Clean: CWQ = 52.9%, WebQSP = 74.3%. RS (P7) is uniquely robust because ELQ reuses the original subgraph unchanged; all other types cause near-total collapse. Bootstrap 95% CIs are per-comparison and uncorrected for multiple comparisons; Bonferroni 99.8% CIs in Appendix L.

Code	Type	CWQ EM	CWQ Δ_{AG} [95% CI]	WSP EM	WSP Δ_{AG} [95% CI]
ES (P1)	Entity alias	0.31	0.526 [0.510, 0.543]	0.06	0.743 [0.719, 0.765]
S1 (P2)	Rel. paraphrase	0.20	0.527 [0.509, 0.543]	0.43	0.739 [0.718, 0.761]
S2 (P3)	Voice flip	0.51	0.524 [0.508, 0.540]	0.37	0.739 [0.720, 0.761]
CR (P4)	**Comp. restruct.**	**0.68**	**0.522 [0.505, 0.538]**	**0.49**	**0.738 [0.716, 0.758]**
S3 (P5)	Entity insertion	0.40	0.525 [0.510, 0.543]	0.43	0.739 [0.719, 0.760]
S4 (P6)	Mixed noise	0.57	0.523 [0.507, 0.539]	0.49	0.738 [0.718, 0.760]
RS (P7)	**Rel. synonym**	**20.31**	**0.326 [0.307, 0.344]**	**50.95**	**0.234 [0.210, 0.257]**

ity in ExplaiGNN arises from the iterative, turn-dependent retrieval: each turn conditions on the previous turn's subgraph, so a single perturbed turn corrupts all downstream reasoning.

Scope note. This comparison is qualitative only. ExplaiGNN uses a different KG (Wikidata vs. Freebase), benchmark (ConvMix vs. CWQ/WebQSP), task type (multi-turn conversational vs. single-question), and metric (P@1 vs. EM); the collapse magnitudes (>70% relative) are directionally consistent with the GNN-RAG finding.

K.6 GraftNet-V2: Binary Relation Mask (Diagnostic)

GraftNet-V2 replaces continuous cosine weighting with a hard binary mask (gold relations = 1, others = 0), eliminating embedding computation at pruning time. The three GraftNet variants share question-aware PPR but differ in how relation scores are assigned: GraftNet-Orig (GloVe continuous), GraftNet (mean-pooled word embeddings, continuous), GraftNet-V2 (binary). From the Seeds × PPR ablation (Table 11), the GEM + Flat configuration (equivalent to GraftNet-V2's binary masking with correct seeds) achieves 30.25% CWQ CR, statistically indistinguishable from GEM + Cosine (29.82%; paired bootstrap $p = 0.478$). On CWQ, the performance ceiling is determined by seed quality, not PPR flavour. On WebQSP, cosine weighting yields a significant 5.6 pp gain over flat PPR ($p < 0.001$), suggesting the binary mask is a useful simplification only for the longer-hop CWQ setting.

K.7 RoG+PathOnly: Path Injection Methodology

Predicted relation paths come from RoG's relation-path predictor, trained on clean CWQ and WebQSP training data. At inference, the predicted path (a sequence of Freebase relation types) is prepended to the RoG prompt as a structured context field before the question text. RoG+PathOnly uses the base GCR-Llama-2-7b checkpoint fine-tuned on clean question-answer pairs; it differs

Table 15: GMT-KBQA S-expression EM under perturbation. Clean: CWQ = 36.9%, WebQSP = 35.5%. Structural perturbations (CR-type, RS-type, S2-type) cause $\leq$5 pp drop, confirming S-expression generation is robust to compositional restructuring. Entity-type perturbations cause 18–26 pp degradation, matching GNN-RAG's EL vulnerability profile.

Perturbation	Type	CWQ		WebQSP	
		EM (%)	Drop	EM (%)	Drop
Entity alias subst.	ES-type	18.3	−18.6	10.8	−24.7
Adversarial entity	CR-type	18.5	−18.4	9.4	−26.1
Question reordering	CR-type	31.9	−5.0	35.9	+0.4
Relation synonym	RS-type	32.3	−4.6	32.0	−3.5
Active-passive flip	S2-type	34.9	−2.0	35.7	+0.2

Table 16: ExplaiGNN P@1, MRR, and H@5 on ConvMix test set (4,800 turns) under CR and RS perturbations. Clean = published checkpoint, unperturbed. Both attacks cause severe collapse (>70% relative P@1 drop), showing that iterative subgraph-chaining systems are vulnerable to the same perturbation types as single-question GNN-RAG, despite using pattern-matching retrieval.

Condition	P@1 (%)	MRR	H@5 (%)
Clean	33.9	0.398	47.7
CR (perturbed)	9.8	0.116	14.0
RS (perturbed)	10.2	0.123	14.9
Gold-answer upper bound	*40.6*	*0.471*	*56.1*

from RoG+RA only in training data (RoG+RA also sees relation-path-augmented examples during fine-tuning), not in inference architecture or path injection format. The near-identical CWQ results (RoG+PathOnly 51.43% vs. RoG+RA 50.47% CR; 51.29% vs. 51.91% RS) confirm that path injection at inference, not the augmented training signal, is the dominant contributor to the accuracy gain.

L Bonferroni Confidence Intervals

Table 10 repeats the 14 Δ_{AG} cells from Table 14 (7 perturbation types $\times$ 2 datasets) with Bonferroni 99.8% intervals ($\alpha^* = 0.05/24 = 0.002$, dividing by the 24 cells in Table 7; n=10,000 bootstrap draws).

Primary conclusions after Bonferroni adjustment. All five primary conclusions stated in Section 7 survive: (a) ELQ CR collapse ($\Delta_{\text{AG}} \approx 0.522$, corrected CI lower bound 0.496); (b) GraftNet CR recovery (29.82% vs. 0.68%, far outside any CI width); (c) RS ELQ robustness ($\Delta_{\text{AG}} \approx 0.326$ CWQ; corrected CI [0.299, 0.354]); (d) GNN decoder robustness (52.76% EM_{fast} with a clean subgraph); (e) EPR-KGQA near-baseline retention (effect sizes above 10 pp, Table 15). Corrected CI half-widths are approximately 2 pp to 4 pp, negligible relative to these effects. The 0.46 pp GEM + Cosine vs. GEM + Flat difference on CWQ CR does not survive correction ($p = 0.478$), confirming seed quality, not PPR weighting, is the dominant factor for CWQ; WebQSP shows a significant cosine benefit (+5.6 pp, $p < 0.001$) on shorter hop chains (Appendix K.2).

M Licenses

All experimental assets are cited at first use and used consistently with their intended research purposes. *Datasets:* CWQ (Talmor and Berant, 2018) is available under Apache-2.0; WebQSP (Yih et al., 2016) under the Microsoft Research License (https://www.microsoft.com/en-us/download/details.aspx?id=52763); the Freebase 2015 RDF dump under CC-BY 2.5 (https://developers.google.com/freebase). *Evaluated systems:* ELQ via BLINK (Li et al., 2020a) is MIT-licensed; EPR-KGQA (Ding et al., 2024) Apache-2.0; GMT-KBQA (Das et al., 2021) BSD-3-Clause; GNN-RAG (Mavromatis and Karypis, 2025) carries no explicit license file at the time of writing and is treated as research-use only. *Generation model:* Llama-3.3-70B-Instruct (Grattafiori et al., 2024) is released under the Llama 3 Community License (https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct). *Released artefacts:* Our perturbed datasets and evaluation code (https://anonymous.4open.science/r/atkgrag-E85C) are released under CC BY 4.0; downstream use must remain consistent with the upstream research-only scope of CWQ, WebQSP, and Freebase.

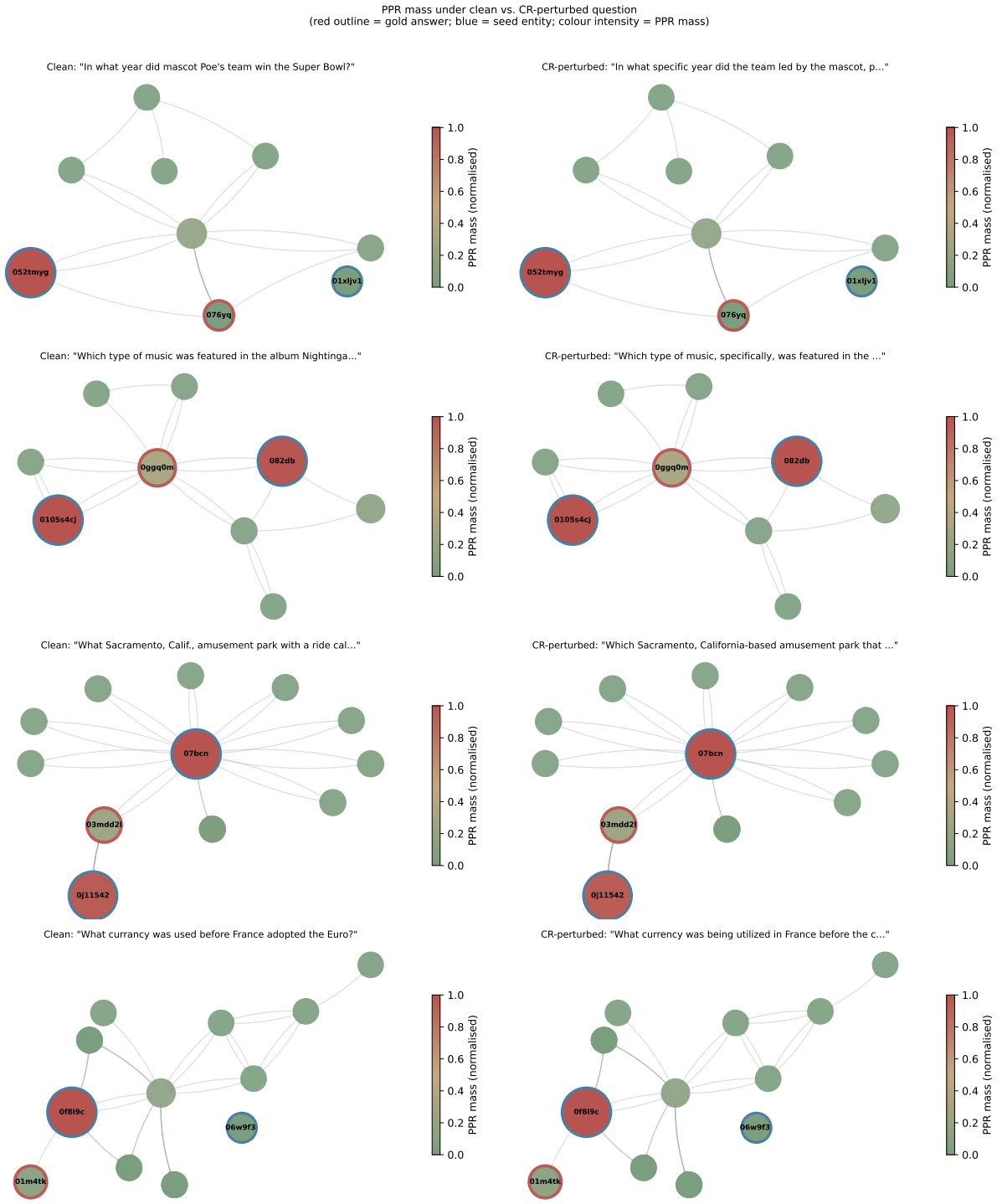

Figure 5: **Answer present but unreachable under CR: four examples.** Each row shows a clean/perturbed question pair; node colour $\propto$ PPR mass (red = high, green = low); blue double-circle = ELQ seed; red double-circle = gold answer. The gold answer node is *present* in both subgraphs with near-identical PPR mass, yet exact-match accuracy collapses from 1 (clean) to 0 (CR). ELQ reuses the same seed (Trip. Jac. = 0.885 CWQ CR), but the PPR walk is anchored to the original hop order: paths required by the *restructured* reasoning chain are not emphasised. The GNN decoder is not the culprit: on the clean subgraph with a CR question it achieves 52.76% CWQ EM (Table 3). This is the *answer presence* $\neq$ *answer reachability* finding (74% of CR failures have the gold answer in the subgraph; EM = 0.68%).