

When Relation-Free Retrieval Is Not Relation-Aware: A Directionality Stress Test for Linear Graph RAG

Ding Wu

Independent Researcher
dingwu@alumni.upenn.edu

Abstract

Relation-free graph retrieval avoids explicit relation extraction, but removing relation labels can weaken shallow ranking when argument order determines which evidence is correct. We test this boundary condition with the official LinearRAG retrieval-only pipeline. On a controlled set of 3,000 questions over nine asymmetric relation types, LinearRAG recovers the gold passage at rank 20 in 96.4% of cases but ranks it first in only 18.2%. Its top-1 output is a same-entity wrong-relation passage in 1,481 cases and a reversed-relation passage in 780 cases, for 2,261 direction errors (75.4%); a further 108 cases are same-relation wrong-entity errors and are reported separately. A local Qwen2.5 reader shows that shallow retrieval errors mainly propagate as abstention or answer failure. The result is deliberately bounded: it diagnoses first-stage retrieval under asymmetric ambiguity, not the strongest possible reranked GraphRAG system. We recommend relation-sensitive shallow-ranking and reader-propagation checks in addition to aggregate recall.

1 Introduction

Graph-based retrieval-augmented generation (GraphRAG) organizes documents around entities, relations, and paths so that retrieval can move beyond flat lexical or dense nearest-neighbor matching (Edge et al., 2024; Chen et al., 2025; Hu et al., 2024). Explicit relation extraction, however, can be expensive, noisy, and difficult to maintain. LinearRAG removes typed relations and constructs a relation-free graph over text and entities (Zhuang et al., 2025).

This simplification has a natural boundary condition: asymmetric relations, where subject–object order determines the correct evidence. “Alice acquired BetaCorp” and “BetaCorp acquired Alice’s supplier” share entities and relation words, but only the first supports “Who acquired BetaCorp?” A

relation-free retriever can reach the correct neighborhood while still placing a directionally wrong passage first.

We study that failure as an Insights-style diagnostic rather than proposing a new method. The questions are: (1) is gold evidence absent or merely under-ranked; (2) are shallow failures random or relation-specific; (3) does the pattern persist under paraphrase and on a small real-data slice; and (4) do shallow retrieval errors propagate to a reader? We evaluate the official LinearRAG index and retrieval interface, not full answer generation or a production reranking stack.

2 Background

Relation-free GraphRAG. LinearRAG builds a relation-free Tri-Graph and activates entities from a question, aggregating global importance to rank passages without extracted relation triples (Zhuang et al., 2025). This removes a major construction cost, but the retrieval representation does not explicitly encode argument direction.

Asymmetric relations. Relations such as acquired, defeated, replaced, parent of, and part of are not invariant to argument order. A passage can contain the correct entities and relation cue while supporting the opposite answer. Aggregate Recall@20 can therefore hide a shallow-ranking failure that matters when a reader consumes only the first few passages.

GraphRAG design space. Other graph-RAG systems retain more explicit structure, for example relational paths, query-centric graph chunks, or domain-specific graph expansion (Chen et al., 2025; Wu et al., 2025; Mostafa et al., 2024). These systems are not our comparison target; they illustrate the broader construction–ranking trade-off that makes LinearRAG a useful diagnostic case.

Evaluation context. Standard RAG evaluation emphasizes answer quality or relevant-passage recall (Lewis et al., 2020). GraphRAG benchmarks broaden system-level comparisons (Xiang et al., 2025); our study isolates one mechanism-level diagnostic. Qwen2.5 is used only as a local reader, not as an external judge (Qwen Team, 2025).

3 Experimental Setup

3.1 Controlled stress set

The controlled set contains 3,000 questions and 15,000 passages. Each question is paired with five candidates: one *gold passage*, defined as the unique passage that states the queried directed relation with the correct arguments; one reversed-relation hard negative; one same-entity wrong-relation hard negative; one same-relation different-entity hard negative; and one irrelevant passage. Thus all hard negatives are plausible candidates, but only the gold passage entails the directed answer.

The set spans acquired, founded, defeated, replaced, caused, parent, CEO, located, and part-of. Half the questions use relation-specific templates and half are cached paraphrases. Archived experiment records preserve the template/paraphrase split and aggregate outcomes, but not the exact paraphrase-generation prompt. We therefore treat the paraphrase comparison as a secondary robustness check rather than a fully reproducible generation procedure.

3.2 Real-data slice

We also evaluate 300 public examples from **2Wiki-MultiHopQA** (2Wiki) over 3,323 chunks. A gold passage is a chunk whose title belongs to the benchmark supporting-fact titles for the query. The slice is strongly imbalanced: CEO 118, parent 94, located 84, and founded 4. It is used only as a qualitative transfer check; we make no relation-level claim for founded.

3.3 Retrievers and reader

We evaluate a local clone of the official LinearRAG pipeline at commit 5da5262, calling `index()` and `retrieve()` but not `qa()`. Diagnostic baselines separate entity overlap, lexical cues, dense similarity, and direction-sensitive argument order. Entity-only ranks by overlapping extracted entities. BM25 uses tokenized passages. Dense retrieval uses cosine similarity from all-MiniLM-L6-v2. Hybrid combines per-query min-max normalized

BM25 and dense scores with $\alpha \in \{0.25, 0.5, 0.75\}$. Relation-aware lexical adds fixed relation-cue, expected-argument, and reversed-template indicators to BM25. These baselines are diagnostic controls, not proposed replacements.

Reader propagation uses a local Qwen2.5-7B-Instruct model. It receives only the supplied context and is instructed to answer “unknown” when unsupported. The 300-example stress reader comparison covers LinearRAG top-1, top-5, top-20, Hybrid top-5, and a gold-passage oracle. A 50-example real-slice reader check is retained only as an artifact-level sanity check because long real contexts caused frequent truncation. No external API is used. Retrieval metrics are R@1, R@3, R@5, R@20, MRR, mean gold rank, wrong-relation error rate, reversed-relation error rate, and entity-trap rate; reader metrics include exact match, contains match, abstention, wrong/reversed answer rates, latency, and context length.

4 Results

4.1 Gold evidence is usually under-ranked

Table 1 shows the main boundary condition. LinearRAG reaches R@20 of 0.964, but R@1 is 0.182 and R@5 is 0.434. The gold passage is therefore commonly available to a deep retriever but absent from the shallow context a reader is likely to receive. Hybrid $\alpha = .25$ reaches R@1 of 0.575 and R@20 of 0.988, indicating that lexical and dense query evidence can repair part of the shallow ordering.

4.2 Top-1 failures are relation traps

The top-1 taxonomy contains 1,481 same-entity wrong-relation outputs and 780 reversed-relation outputs. These sum to **2,261 direction errors** (75.4%). The remaining 108 same-relation wrong-entity outputs measure entity discrimination and are reported separately. Only 85 top-1 outputs are irrelevant, while 546 are gold. The dominant failure is therefore relation ambiguity rather than random drift.

4.3 Paraphrase and real-data checks

Templated and paraphrased questions show similar shallow behavior in the archived aggregate results, although the missing exact paraphrase prompt limits reproducibility of this comparison. Relation-level variation is more informative: defeated and

Method	R@1	R@3	R@5	R@20	MRR	Rev. err	Wrong-rel. err	Entity trap
Official LinearRAG	.182	.297	.434	.964	.317	.260	.494	.756
Entity-only proxy	.142	.330	.469	.794	.292	.098	.659	.766
BM25	.436	.638	.706	.910	.557	.270	.223	.535
Dense MiniLM	.467	.664	.748	.950	.593	.319	.080	.402
Hybrid $\alpha = .25$	.575	.764	.849	.988	.693	.329	.061	.395
Hybrid $\alpha = .50$	.571	.746	.841	.990	.683	.241	.160	.410
Hybrid $\alpha = .75$	.515	.698	.802	.979	.633	.237	.210	.464
Relation-aware lexical	.465	.694	.791	.970	.601	.254	.219	.515

Table 1: Controlled direction-stress retrieval results on 3,000 questions. LinearRAG often reaches gold by rank 20 but under-ranks it shallowly.

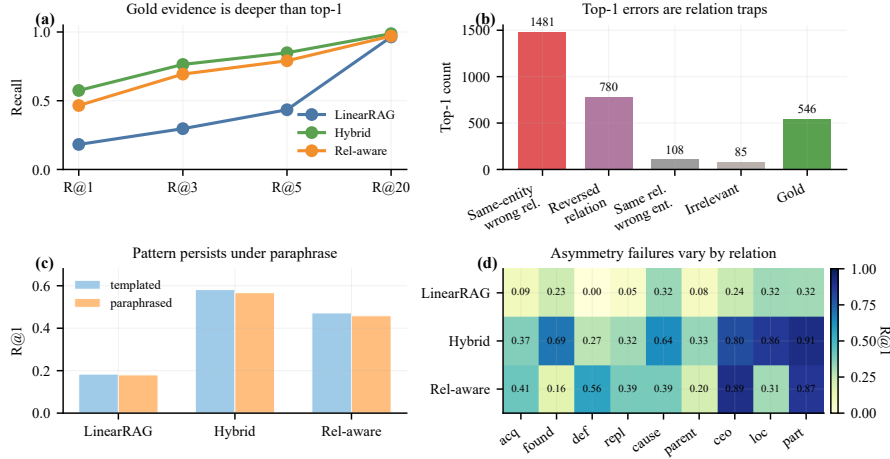


Figure 1: Directionality diagnostics. (a) LinearRAG reaches gold by rank 20 but under-ranks it shallowly. (b) Top-1 outcomes concentrate in same-entity wrong-relation and reversed-relation passages; the 108 same-relation wrong-entity cases are not direction errors. (c) Templated and paraphrased questions show similar shallow behavior in the archived aggregate results. (d) Relation types vary sharply, with defeated and acquired especially brittle.

Method	R@1	R@3	R@5	R@20	MRR
LinearRAG	.470	.587	.680	.750	.553
Entity-only	.197	.300	.333	.403	.258
BM25	.727	.847	.887	.947	.800
Dense	.347	.423	.467	.557	.400
Hybrid $\alpha = .75$	.743	.857	.897	.950	.811
Rel.-aware	.723	.857	.880	.947	.799

Table 2: Real 2Wiki relation slice (300 examples). This imbalanced slice is a qualitative check only.

acquired are brittle, whereas CEO and located retain stronger lexical anchors.

On the 300-example 2Wiki slice, LinearRAG reaches R@1 of 0.470 and R@20 of 0.750. BM25 and relation-aware lexical reach R@1 of 0.727 and 0.723. This result should not be interpreted as a broad superiority claim for BM25. The slice is small, relation-imbalanced, and lexically favorable: query and support text often preserve explicit entity and relation strings. It is a boundary check showing that literal overlap can be unusually strong in this construction.

4.4 Shallow errors propagate to the reader

With LinearRAG top-1 context, the reader obtains a contains-match rate of 0.227 and abstains in 0.760 of cases. Top-5 improves only modestly to 0.290. LinearRAG top-20 reaches 0.550 but uses roughly 425 context tokens, compared with 101 for top-5. Hybrid top-5 reaches 0.497 at a similar context length, while the gold oracle reaches 0.953. Wrong-relation answer rates remain low; failures are mainly abstention or an unsupported answer.

This single-reader experiment is a propagation diagnostic, not evidence that all readers behave identically. A second reader was not available in the surviving experiment package, so model-specific behavior remains a limitation.

5 Discussion

What the result does and does not show. The study does not show that relation-free GraphRAG is generally ineffective. LinearRAG removes a costly relation-extraction step, and high R@20 shows that it often reaches the correct evidence neighborhood.

The narrower result is that entity-sharing but directionally wrong passages can be ranked above gold under asymmetric ambiguity.

Where reranking fits. A lightweight cross-encoder or LLM reranker could score query–passage pairs after LinearRAG candidate generation and explicitly compare argument roles. Such a stage may preserve the construction efficiency of a relation-free graph while restoring shallow precision. However, reranking cannot recover a gold passage that is absent from the candidate pool, and it adds latency and model dependence. The present experiment isolates the first-stage ranking; it should not be read as the performance ceiling of a full production pipeline.

Diagnostic interpretation. Entity-only, lexical, dense, hybrid, and relation-aware lexical baselines separate three signals: entity overlap, surface relation cues, and direction-sensitive argument order. Their role is mechanistic diagnosis, not a new system claim.

6 Conclusion

Relation-free graph retrieval can be efficient and practical, but it is not automatically relation-aware when argument order matters. In the official LinearRAG retrieval-only setup, gold evidence is often recovered at deeper ranks while same-entity wrong-relation and reversed-relation passages dominate shallow errors. A local reader analysis shows that these errors propagate mainly as answer failure or abstention. Relation-free GraphRAG evaluations should therefore report relation-sensitive shallow-ranking diagnostics and reader propagation, not only aggregate recall.

Limitations

The controlled set deliberately amplifies asymmetric direction errors and does not estimate their prevalence in natural corpora. The real-data check is 2Wiki-only and relation-imbalanced. Archived experiment artifacts preserve the split and aggregate results but not the exact paraphrase-generation prompt or original generation script. We therefore do not claim a complete reproducibility release for the stress set, and the paraphrase result should be treated as secondary. We did not evaluate a cross-encoder or LLM reranker, so the claim concerns the tested first-stage retrieval rather than the strongest possible production pipeline. The reader analysis

uses one local Qwen2.5 model, and multi-reader validation remains future work.

Acknowledgments

We thank the Insights 2026 reviewers for identifying the error-category arithmetic issue and for prompting clearer definitions of gold evidence, 2WikiMultiHopQA, reranking scope, synthetic-data construction, and single-reader limitations.

References

- Boyu Chen, Zirui Guo, Zidan Yang, Yuluo Chen, Junze Chen, Zhenghao Liu, Chuan Shi, and Cheng Yang. 2025. PathRAG: Pruning graph-based retrieval augmented generation with relational paths. arXiv:2502.14902.
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Metropolitansky. 2024. From local to global: A Graph RAG approach to query-focused summarization. arXiv:2404.16130.
- Yuntong Hu, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling, and Liang Zhao. 2024. GRAG: Graph retrieval-augmented generation. arXiv:2405.16506.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*.
- Radeen Mostafa, Mirza Nihal Baig, Mashaekh Tausif Ehsan, and Jakir Hasan. 2024. G-RAG: Knowledge expansion in material science. arXiv:2411.14592.
- Qwen Team. 2025. Qwen2.5 technical report. Technical report.
- Yaxiong Wu, Jianyuan Bo, Yongyue Zhang, Sheng Liang, and Yong Liu. 2025. Query-centric graph retrieval augmented generation. arXiv:2509.21237.
- Zhishang Xiang, Chuanjie Wu, Qinggang Zhang, Shengyuan Chen, Zijin Hong, Xiao Huang, and Jin-song Su. 2025. When to use graphs in RAG: A comprehensive analysis for graph retrieval-augmented generation. arXiv:2506.05690.
- Luyao Zhuang, Shengyuan Chen, Yilin Xiao, Huachi Zhou, Yujing Zhang, Hao Chen, Qinggang Zhang, and Xiao Huang. 2025. LinearRAG: Linear graph retrieval augmented generation on large-scale corpora. arXiv:2510.10114.