A Containment Metric with No Lower Bound: A Null-Control Audit of Retrieval Scoring in GraphRAG

Ahmed Maaloul
ahmed.maaloul@proton.me
https://github.com/ahmedmaaloul/synapse

Abstract

Retrieval-augmented systems built on knowledge graphs (GraphRAG) are frequently evaluated with a *containment* rule: a gold passage counts as retrieved once the context contains the surface form of an entity extracted from it. We show this rule has no lower bound. On a 14-question internal multi-hop benchmark, a *null retriever* that returns nothing but the graph's alphabetical entity vocabulary—no descriptions, no edges, no prose, and no ability to answer any question—scores 100% fact recall under containment and 0% under a prose-grounded rule. The mechanism replicates on HotpotQA (n=50): a graph-only retriever scores 100% Both-gold under containment but 0% under prose-grounding, with 93% of its credited hits coming from entity-name blocks rather than retrieved evidence. Crucially, the inflation is *not specific to one system*: on a second, independently authored GraphRAG (LightRAG), a zero-evidence null retriever again scores 100% under containment and 0% under prose-grounding, establishing the flaw as a property of the metric. Under prose-grounded scoring the graph's advantage is real but budget-neutral—a plain passage baseline matches it while reading 45% of the context. We frame this as a caution for practitioner retrieval-recall evaluations, propose a channel-gated fix that costs prose-carrying systems nothing, and release a reproducible null-control protocol.

1 Introduction

GraphRAG systems [2, 3] extract entities and relations from a corpus, store them as a graph, and retrieve a subgraph as context for a language model. A recurring way to report their retrieval quality, especially in practitioner and open-source evaluations, is a *containment* (string-match) rule: a required fact—typically an entity—counts as retrieved if its surface form appears anywhere in the assembled context.

This paper makes one point precisely. Containment scoring rewards a system for *naming* an entity, not for retrieving *evidence* about it, and a graph retriever names entities by construction. We operationalise the objection with a *null control*: a retriever that ignores the question and returns only the alphabetical list of entity names in the graph. It carries no descriptions, no relations, and no source prose; it cannot answer anything. Under containment it nevertheless attains perfect recall. A metric on which a system carrying zero evidence is indistinguishable from a perfect one has no lower bound, and comparisons made under it are uninterpretable.

Our contributions are:

1. A *null-control* construction and a two-rule (containment vs. prose-grounded) protocol that expose the missing lower bound, with a per-hit channel attribution that localises *why* credit is awarded (Section 3).

2. Evidence that the inflation is *implementation-independent*: it reproduces on two independently authored GraphRAG systems—our own and LightRAG—so it is a property of the metric, not of one codebase (Section 4).
3. A budget-matched reading showing that, once scored on prose, the graph's multi-hop advantage is real but buys no context saving on this data, and a cheap channel-gated fix (Section 4, Section 6).

All numbers are generated by released, seeded scripts and traced to their source report; nothing is hand-copied.

2 Related work

Zeng et al. [7] give an unbiased evaluation framework for GraphRAG, but target *answer*-side bias in LLM-as-judge scoring and question generation; they do not examine retrieval string-containment and use no null control. Their independent finding—that vector-search RAG is a strong low-cost generalist and graph methods win chiefly on global/thematic questions—agrees with our budget-matched reading. An [1] show, in the adjacent domain of long-conversation memory, that verbatim chunks beat LLM-extracted artifacts, naming the mechanism "lossy distillation"; we corroborate the necessity of source text but in document GraphRAG and, distinctly, attack the *measurement* rather than the representation. The systems we audit inject entity names or community summaries into context by construction [2–4], so all are exposed to the metric flaw we characterise. HotpotQA [6] and 2WikiMultihopQA [5] are our external multi-hop testbeds.

3 Method

Systems. We compare, at matched retrieval budgets: a *passage* vector-RAG baseline over the source paragraphs; a *graph-only* retriever (entities, descriptions, and relation lines, no source prose); the full shipped system (graph *and* the source chunks the entities were extracted from); and the *null control* (entity vocabulary only). On LightRAG we map the same four roles onto its `naive`, hybrid graph-only, hybrid full, and a vocabulary-only null over its own extracted entities.

Two scoring rules. A required fact is credited under the *permissive* (containment) rule when its entity surface form appears anywhere in the retrieved context, and under the *strict* rule only when the gold paragraph's own text is returned as prose. We report both; their gap is the *inflation*. For each credited hit we record its *channel*—whether the name arrived in retrieved prose, in a structured entity block, or in a relation line—so the source of credit is explicit.

Effect-size floor. With N questions, one question is worth $1/N$ of a rate, so we refuse to report any gap smaller than one question as a difference. We claim no statistical significance; the null-control result is *constructive* (the existence of a zero-evidence system scoring 100% is a fact about the metric), so small N bounds generality, not the validity of the existence claim.

Data and cost. The internal corpus is 14 multi-hop questions over 34 passages. External runs use the HotpotQA dev/distractor split (a seeded sample of 50 questions / 500 paragraphs) and, for the cross-implementation control, the first 15 of that sample. Extraction uses `gpt-4o-mini`; embeddings are local. Total spend across all runs reported here is under $0.50 (estimated).

System	Perm.	Strict	Chars
A. Passage RAG (top-2)	44.0	44.0	871
A′. Passage RAG (top-12, matched)	88.0	88.0	6,154
C. Graph only	**100.0**	**0.0**	2,392
D. Graph + source chunks	100.0	82.0	5,171

Table 1: HotpotQA dev/distractor, n=50. Both-gold recall (%) under the permissive and strict rules, and mean retrieved context (characters). The graph-only system C is inflated by 100 points; 93% of its permissive credit comes from entity blocks, 0% from prose.

LightRAG system	Perm.	Strict	Infl.
L-naive (vector chunks)	100.0	100.0	0.0
L-graph (hybrid, KG only)	**100.0**	**0.0**	**100.0**
L-full (KG + chunks)	100.0	100.0	0.0
L-null (entity vocabulary only)	**100.0**	**0.0**	**100.0**

Table 2: Cross-implementation control on LightRAG v1.5.4, same 15 HotpotQA questions. Both-gold recall (%) and inflation (points). L-graph and L-null draw 100% of their permissive credit from entity blocks.

4 Results

The metric has no lower bound (internal). On the 14-question benchmark the null control scores 100% fact recall under the permissive rule and 0% under the strict rule. It ties the budget-matched passage baseline and outscores six of the seven real systems, while containing, by construction, no evidence about any question.

Replication on HotpotQA (n=50). Table 1 shows the mechanism on external data. The graph-only system (C) reaches 100% Both-gold under containment and 0% under prose-grounding; 93% of its credited gold paragraphs are credited through an entity block and none through retrieved prose. The prose-carrying systems (A, D) show no such gap. Under the strict rule, the graph's advantage is real but *budget-neutral*: a passage baseline matches the shipped system's Both-gold rate at top-5 while reading about 45% of its context (Table 1, A′).

The inflation is implementation-independent (LightRAG). Table 2 runs the identical audit on LightRAG [3]: its own extraction (1,020 entities over 150 paragraphs), its own retrieval, scored by the same code. The result mirrors ours feature for feature. A zero-evidence null retriever (L-null) scores 100% under containment and 0% under prose-grounding; the graph-only mode (L-graph) is inflated by 100 points with all credit from entity blocks; prose-carrying modes are uninflated. Two independently authored systems, serialising their graphs differently, are inflated identically—so the flaw lies in the metric, not the implementation.

5 Threats to validity

Small N. 50 HotpotQA and 14 internal questions; one question is 2.0 and 7.1 points respectively. No significance is claimed; the null-control existence result is constructive, but every *magnitude* is a small-sample estimate. **Retrieval only.** These are retrieval-recall metrics, comparable to published

retrieval numbers but *not* to leaderboard EM/F1; no leaderboard comparison is implied. **Not the flagship metric.** Microsoft's GraphRAG line scores by LLM-judge and HotpotQA's headline is F1/EM; containment is a practitioner retrieval-recall rule. We scope the claim accordingly—though the cross-implementation result shows the flaw is not idiosyncratic. **Extraction non-determinism** and **lexical matching** add noise that is held constant across systems.

6 Conclusion

Containment-style retrieval scoring, common in practitioner GraphRAG evaluations, has no lower bound: a zero-evidence null retriever scores 100% under it and 0% under a prose-grounded rule, on two independent GraphRAG implementations, with the credit traceable to entity names rather than evidence. Under prose-grounded scoring the graph's multi-hop advantage is real but buys no context saving on our data. The fix is cheap and we recommend it: alongside any containment number, report the *channel-gated* rule that credits a fact only when it is materialised as retrieved prose—it removes the null control and costs prose-carrying systems nothing. A second fully-scored public dataset, larger N, and answer-level (EM/F1) metrics remain future work; they widen generality but are not load-bearing for the claim.

Reproducibility. Code, seeds, exact commands, per-run cost, and a provenance table mapping every number to its generating report are released at the repository above under AGPL-3.0.

References

[1] Tao An. CogCanvas: Verbatim-grounded artifact extraction for long LLM conversations. *arXiv preprint arXiv:2601.00821*, 2026.

[2] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

[3] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. LightRAG: Simple and fast retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

[4] Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. *arXiv preprint arXiv:2405.14831*, 2024.

[5] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, 2020.

[6] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[7] Qiming Zeng, Xiao Yan, Hao Luo, Yuhao Lin, Yuxiang Wang, Fangcheng Fu, Bo Du, Quanqing Xu, and Jiawei Jiang. How significant are the real performance gains? an unbiased evaluation framework for GraphRAG. *arXiv preprint arXiv:2506.06331*, 2025.