PROCEEDINGS OF THE INTERNATIONAL SCIENTIFIC AND PRACTICAL CONFERENCE
Science and Education in Central Asia: Digital Transformation and Sustainable Development
18 December 2026 · online, Zoom · CASJ-2026
SECTION 7. YOUNG RESEARCHERS: STUDENT SECTION (BACHELOR'S AND MASTER'S)

DOI: https://doi.org/10.5281/zenodo.24141910

UDC 004.8

EVIDENCE COMPLETENESS AND ANSWER QUALITY IN GRAPH-ASSISTED RETRIEVAL-AUGMENTED GENERATION

Mukhambetov Bektas, Noken Rakhima, Tarlan Baimyrza, Yskak Zhanibek
Astana IT University, Astana, Kazakhstan

Mukhambetov Bektas — Third-year BSc student in Computer Science, Astana IT University, Astana, Kazakhstan.

Noken Rakhima — Third-year BSc student in Computer Science, Astana IT University, Astana, Kazakhstan.

Tarlan Baimyrza — Third-year BSc student in Computer Science, Astana IT University, Astana, Kazakhstan.

Yskak Zhanibek — Third-year student in Software Engineering, Astana IT University, Astana, Kazakhstan.

Scientific supervisor: Tursynkozha Aisha — PhD, Assistant Professor, School of Artificial Intelligence and Data Science, Astana IT University, Astana, Kazakhstan.

© Mukhambetov Bektas, Noken Rakhima, Tarlan Baimyrza, Yskak Zhanibek, 2026

Abstract: For multi-hop questions it remains unclear whether retrieving more complete evidence actually improves the answers of retrieval-augmented generation. The study compares sparse, dense and entity-mention-graph retrieval with a shared reranker and a local quantized language model on reserved HotpotQA questions, and proposes an evidence-to-answer decomposition that separates the effect of evidence coverage from the effect of evidence utilization. Graph-assisted retrieval kept complete evidence more often than dense retrieval, but its answer gain was not statistically resolved and was explained mainly by how the generator used the evidence rather than by retrieval coverage.

Keywords: retrieval-augmented generation; multi-hop question answering; entity mention graph; evidence coverage; answer quality.

Аннотация: *Для многошаговых вопросов остаётся неясным, улучшает ли более полный поиск подтверждающих фактов ответы генерации, дополненной поиском. В исследовании сравниваются разреженный, плотный и графовый поиск по упоминаниям сущностей с общим переранжировщиком и локальной квантованной языковой моделью на отложенных вопросах HotpotQA, а также предлагается декомпозиция «факты–ответ», разделяющая эффект полноты фактов и эффект их использования. Графовый поиск чаще сохранял полный набор фактов, чем плотный, однако выигрыш в качестве ответа статистически не подтверждён и объясняется в основном тем, как генератор использует факты, а не полнотой поиска.*

Ключевые слова: *генерация, дополненная поиском; многошаговые вопросы; граф упоминаний сущностей; полнота фактов; качество ответа.*

Аңдатпа: *Көп сатылы сұрақтар үшін дәлелдерді толығырақ табу іздеумен толықтырылған генерацияның жауаптарын жақсарта ма деген мәселе әлі де анық емес. Зерттеуде HotpotQA жиынының бөлек қалдырылған сұрақтарында ортақ қайта рангілеушісі және жергілікті кванттелген тілдік моделі бар сирек, тығыз және мәндік атаулар графына негізделген іздеу салыстырылады, сондай-ақ дәлелдердің толықтығы мен*

оларды пайдалану әсерін ажырататын «дәлел–жауап» декомпозициясы ұсынылады. Графтық іздеу тығыз іздеуге қарағанда толық дәлелдерді жиірек сақтады, бірақ жауап сапасындағы жақсару статистикалық түрде расталмады және негізінен іздеу толықтығымен емес, генератордың дәлелдерді пайдалануымен түсіндірілді.

Түйін сөздер: *іздеумен толықтырылған генерация; көп сатылы сұрақтар; мәндік атаулар графы; дәлелдердің толықтығы; жауап сапасы.*

1 Introduction

Retrieval-augmented generation (RAG) combines a retriever that selects passages from an external collection with a language model that writes the answer from them [1]. Universities increasingly use RAG for tutoring and course question answering, often on modest local hardware [2]. For such systems the quality of retrieval and the quality of the answer must be measured separately, because they are produced by different components.

The difficulty is greatest for multi-hop questions, which combine facts from several documents. HotpotQA formalised this task and annotated the supporting sentences of each question [3], and iterative retrieval was proposed to find the second, bridge document [4]. However, complete evidence does not guarantee a correct answer: language models use long contexts unevenly [5] and lose accuracy when relevant passages are mixed with noise [6]. Graph-assisted retrieval is a popular way to improve coverage, but its benefit is usually reported only as an end-to-end answer score, so a developer cannot tell whether to improve retrieval or the use of the retrieved context.

The aim of this study is to determine how a lightweight entity-mention graph changes evidence completeness and answer quality of a fixed local generator under a matched context budget, and to quantify how much of the answer difference is explained by coverage and how much by utilization. The object is the RAG pipeline for multi-hop questions; the subject is the relationship between retrieved evidence and answer correctness. Two hypotheses are tested: *H1*, graph-assisted retrieval increases the share of questions with complete evidence compared with dense retrieval; *H2*, this gain translates only partially into answer quality, and most differences are explained by evidence utilization.

The novelty lies in the evaluation design and in the proposed evidence-to-answer decomposition (EAD), which splits the answer difference between two retrieval methods into a coverage effect and utilization effects, and splits the gap to oracle evidence into missing evidence and context dilution.

2 Literature review

Graph-based RAG adds explicit structure between pieces of text. GraphRAG builds an entity graph with a language model and summarises its communities [7]. HippoRAG runs personalized PageRank over an extracted knowledge graph to improve multi-hop retrieval [8], and KG2RAG uses graph links to expand and organise retrieved chunks [9]. Document GraphRAG builds document-structure graphs and reports its largest gains on multi-hop questions, including HotpotQA [10], while a query-aware multi-path approach selects among one-hop, multi-hop and importance-based subgraphs [11]. Most of these methods require a language model to extract entities and relations, which is costly and introduces errors.

Recent analyses question the general benefit of graphs. A comprehensive benchmark finds that graph RAG often underperforms standard RAG and identifies the tasks where structure helps [12]. Zarrinkia et al. show that a leading graph RAG system often retrieves the answer yet reasons incorrectly [13], and evaluation frameworks such as RAGAS score context relevance and answer faithfulness as separate dimensions [14]. Table 1 summarises the comparison.

Table 1 – Comparison of graph-assisted RAG studies

Approach	Graph construction	Measures evidence coverage	Links coverage to answers
GraphRAG	LLM entity graph	No	No
HippoRAG	LLM triples, PageRank	Retrieval recall	No
Document GraphRAG	document structure	Retrieval metrics	No
RAGAS	not applicable	Context relevance	Separate scores
This study	exact title mentions	Complete supporting sentences	Yes, EAD

Table 1 shows that existing studies either do not check whether complete evidence reached the generator or report retrieval and answer scores without linking them. This study fills that gap with a transparent attribution of answer differences to coverage or utilization.

3 Methodology

3.1 Research design and data

The study follows the six-stage design in Figure 1.

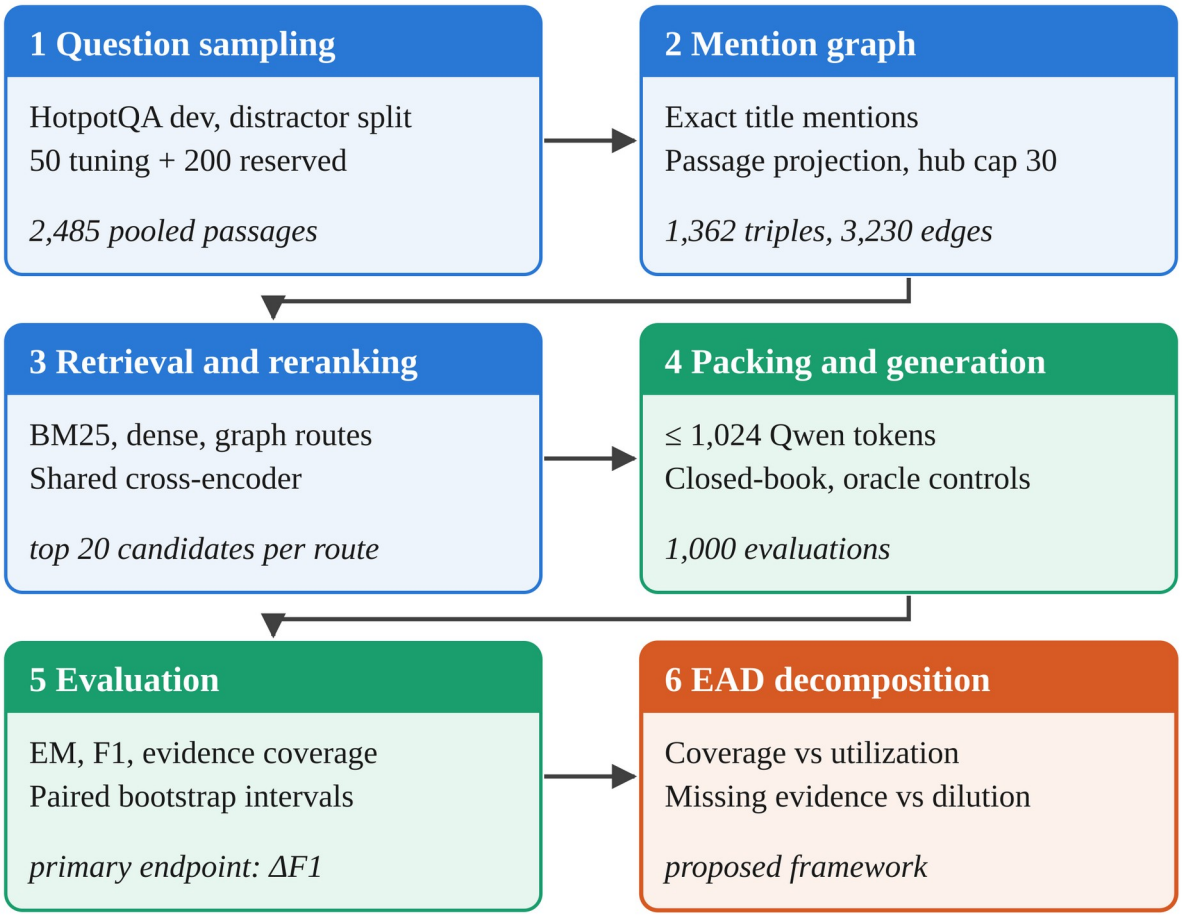


Figure 1 – Research design

Figure 1 shows that stages 1–3 prepare the questions, the graph and the ranked candidates, and stages 4–6 generate, evaluate and decompose the answers. A deterministic hash ordering with seed 20261029 selects 250 records from the HotpotQA development distractor release: 50 for tuning and 200 reserved for evaluation (159 bridge and 41 comparison questions). All their passages form one shared corpus of 2,485 documents. Questions, answers and supporting-fact labels are never used for indexing or graph construction.

Titles of at least four characters are matched at word boundaries; each match yields a mention triple with sentence provenance. Passages are linked if they mention the same entity or if one mentions the title of the other, and entities mentioned in more than 30 passages are ignored. The graph has 1,362 triples and 3,230 edges, and 1,226 documents (49.34%) remain isolated.

3.2 Retrieval and generation pipeline

Figure 2 shows the evaluated pipeline.

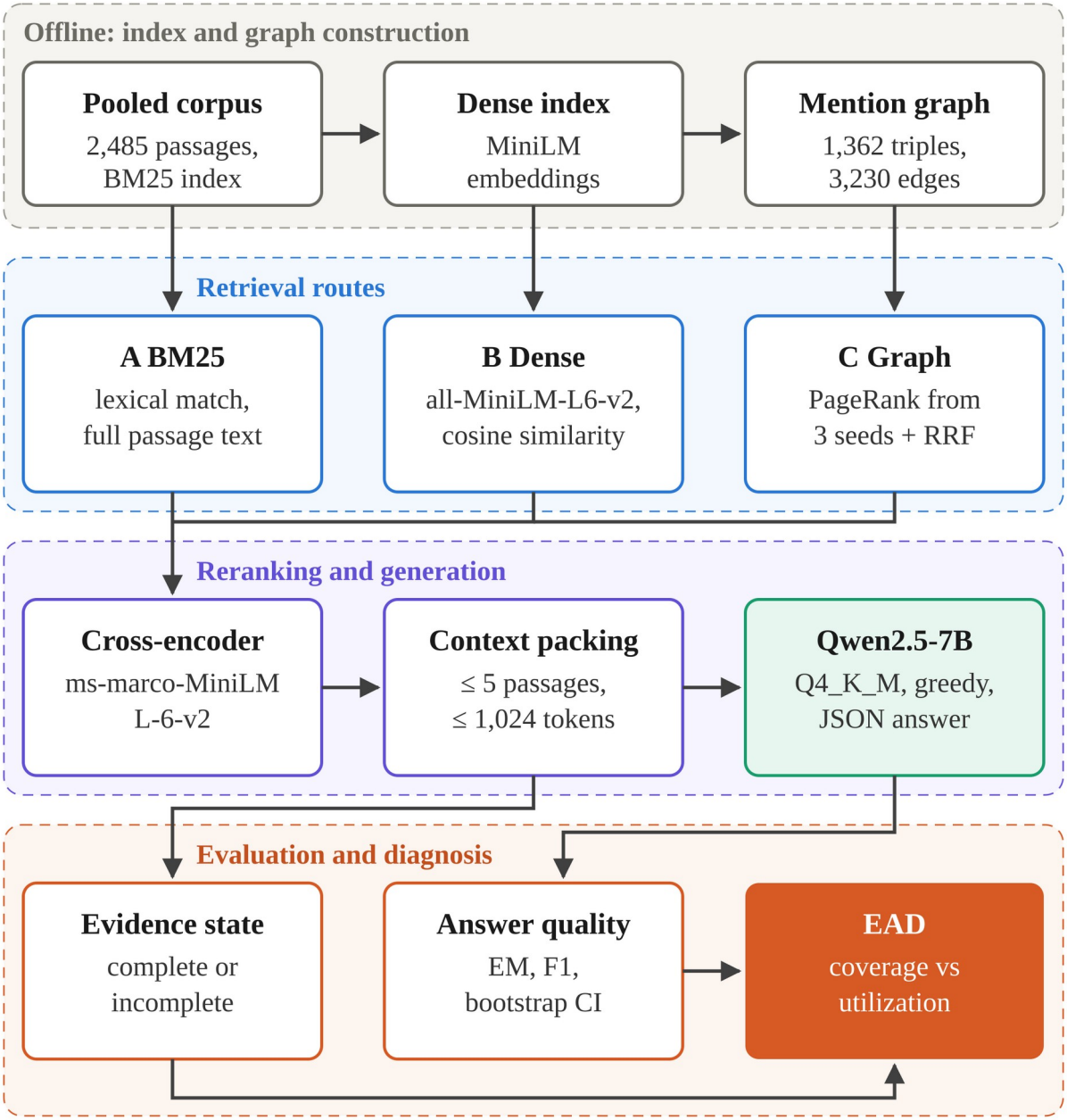


Figure 2 – Architecture of the evaluated pipeline and of the proposed diagnosis

Figure 2 shows three retrieval routes over one corpus that feed one reranker, packer and generator. BM25 uses Lucene settings k1 = 1.2 and b = 0.75 [15]. Dense retrieval uses normalised all-MiniLM-L6-v2 embeddings. The graph route takes the three best dense passages as seeds of personalized PageRank:

where is the vector of passage scores after iterations, is the restart vector with equal mass on the seeds, is the row-normalised adjacency matrix and is the restart probability. Dense and graph rankings are fused by weighted reciprocal rank fusion [16]:

where and are the ranks of passage in the two lists and is the graph weight selected on the tuning questions only.

Each route passes its top 20 candidates to the same ms-marco-MiniLM-L-6-v2 cross-encoder. Packing admits at most five passages, 256 tokens per passage and 1,024 tokens in total, counted with the generator’s tokenizer. The generator is Qwen2.5-7B-Instruct [17] quantised to Q4_K_M and run with llama.cpp on an RTX 3050 Ti laptop GPU with 4 GB of memory, using greedy decoding and an answer-only JSON format; insufficient evidence should yield “noanswer”.

A closed-book control receives no context, and an oracle control receives all annotated supporting sentences.

3.3 Metrics and proposed decomposition

The primary endpoint is the graph-minus-dense difference in macro answer token F1. Exact match (EM) is also reported, and complete evidence for question is , where is the set of annotated supporting sentences and is the packed context. Paired 95% intervals use 2,000 bootstrap resamples of questions and, separately, of clusters sharing a gold source.

For method , let be the share of questions with complete evidence, the EM among them and the EM among the remaining questions. Then

where the first term is the contribution of complete-evidence questions and the second the contribution of correct answers despite missing evidence. For methods A and B the EM difference is split exactly as

where is the value for B minus the value for A, is the mean completeness and is the mean accuracy lift from complete evidence. The first term is the coverage effect and the other two are utilization effects. The gap to the oracle control is split as

where is the EM with oracle evidence, the first term is the missing-evidence component and the second is the context-dilution component, the loss caused by complete but longer and noisier contexts. The decomposition is descriptive, because evidence states are observed rather than randomised.

4 Results

Table 2 reports all five conditions on the 200 reserved questions; there were 1,000 evaluations and no failed request.

Table 2 – Evidence coverage and answer quality on the reserved questions

Configuration	Complete evidence	EM	F1	Abstain	Context tokens
BM25 + reranking	68.0%	37.0%	47.2%	67	702.0
Dense + reranking	67.0%	34.0%	44.8%	74	700.6
Graph + reranking	71.0%	37.0%	46.9%	69	703.1
Closed book	–	4.5%	6.5%	175	0.0
Oracle evidence	–	49.5%	64.1%	36	149.0

Table 2 shows that the routes used almost identical context lengths, so differences are not caused by the amount of text. The graph route had the highest complete-evidence rate, and its F1 was 2.1 percentage points (pp) higher than that of dense retrieval, with 95% intervals of [−0.1, +4.7] for question resampling and [−0.2, +4.6] for source clusters. Both include zero, so a directional advantage is not established, and BM25 reached a similar F1 (47.2%). Without reranking, coverage was 51.5% for dense and 67.0% for graph retrieval; the common reranker narrowed this 15.5 pp advantage to 4.0 pp, which supports H1 in a weaker form. By question type, graph F1 exceeded dense F1 by 1.6 pp on bridge and 4.1 pp on comparison questions, the latter based on only 41 questions.

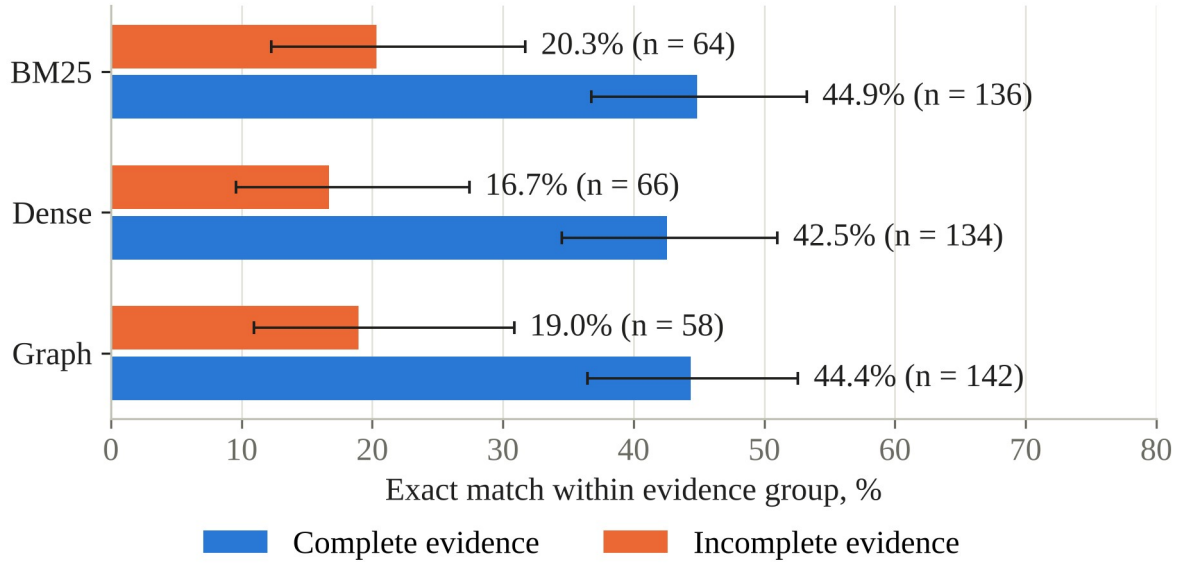


Figure 3 – Exact match conditional on complete or incomplete evidence (95% Wilson intervals)

Figure 3 shows that complete evidence is associated with an EM of 42.5% to 44.9%, against 16.7% to 20.3% with incomplete evidence, a gap of about 25 pp for every route. The intervals within each evidence state overlap strongly across routes, and even complete retrieved evidence gives a lower EM than oracle sentences alone. Table 3 lists the components of the decomposition.

Table 3 – Components of the evidence-to-answer decomposition

Route	c	a	b	EM	Missing evidence, pp	Context dilution, pp
BM25	68.0%	44.9%	20.3%	37.0%	7.9	4.6
Dense	67.0%	42.5%	16.7%	34.0%	8.5	7.0
Graph	71.0%	44.4%	19.0%	37.0%	7.4	5.1

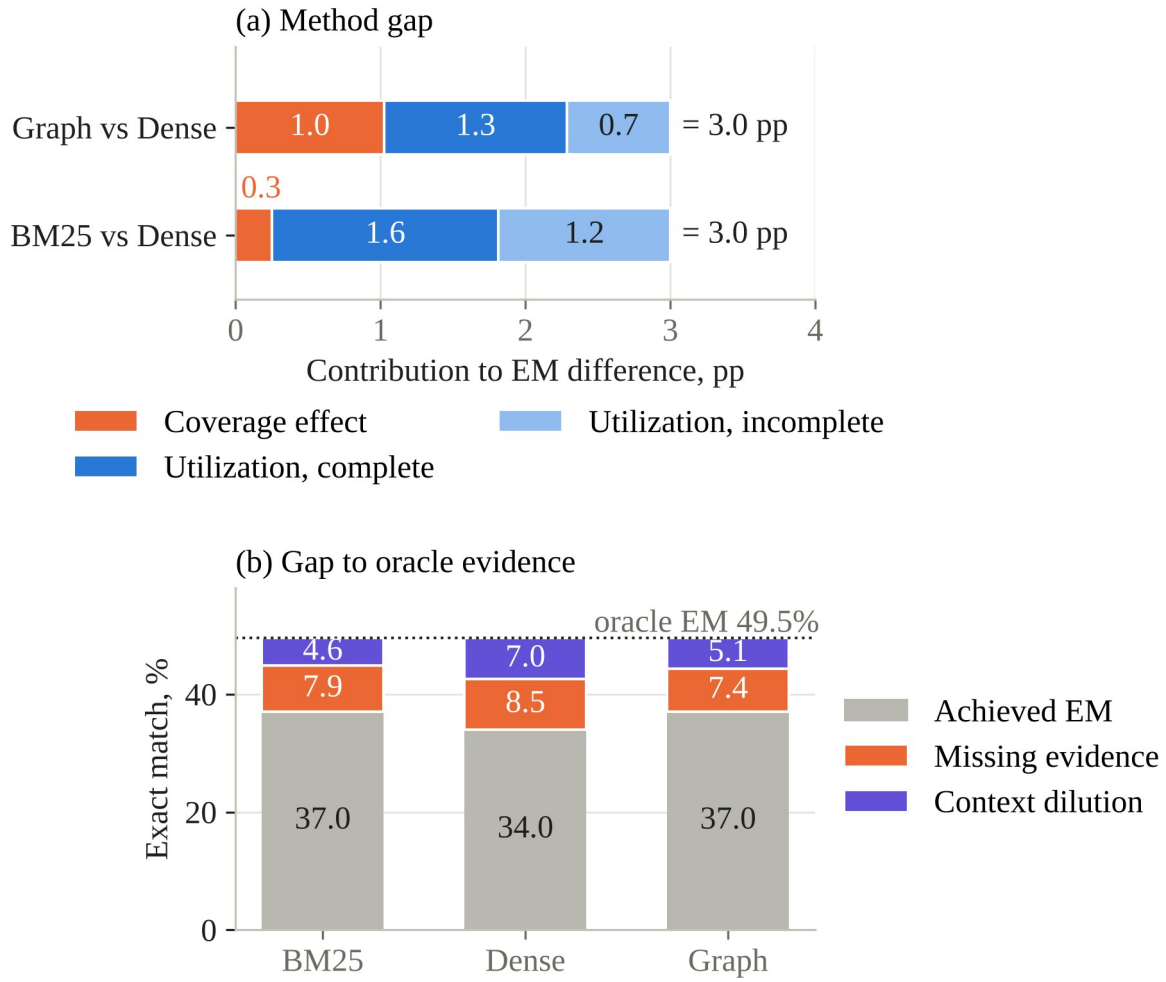


Figure 4 – Evidence-to-answer decomposition: (a) method gap relative to dense retrieval; (b) gap to oracle evidence

Figure 4a shows that of the 3.0 pp EM gain of graph over dense retrieval only 1.0 pp (34%) comes from higher coverage, while 2.0 pp come from higher accuracy within evidence states. BM25 achieves the same gain almost entirely through utilization, with a coverage effect of 0.3 pp. Figure 4b shows that missing evidence explains 7.4 to 8.5 pp of the gap to the oracle, and context dilution explains 4.6 to 7.0 pp, or 37% to 45% of the gap. These results support H2.

5 Discussion

The mention graph increased evidence coverage strongly before reranking and modestly after it, but its answer gain over dense retrieval was small, statistically unresolved and no larger than that of BM25. The decomposition explains why: most differences arise within evidence states rather than from moving questions into the complete state. This agrees with the reasoning bottleneck reported for graph RAG by Zarrinkia et al. and adds a quantitative attribution: in this setting about one third of the gain comes from coverage. The dilution component links the results with long-context studies: the oracle used about 149 tokens, the retrieval routes about 700, and the extra text cost 4.6 to 7.0 pp even when nothing was missing.

The work has practical value for developers of RAG systems that run locally on limited hardware, for example in universities. A mention graph can be built without language-model calls, but it should be compared with a sparse baseline using paired intervals. Recording for each question whether the complete evidence was packed is enough to compute the EAD and decide whether to invest in retrieval or in context selection and compression; when dilution is large, as here, reducing context noise is likely to help more than higher recall.

The study is limited to one corpus of 2,485 passages, one quantised generator and modest retrieval models. The graph is a mention-based projection without alias resolution, annotated supporting sentences may miss valid evidence paths, and EM can penalise valid answer aliases. The

EAD is descriptive and was computed for EM.

6 Conclusion

The aim of the study was achieved. Under the pooled HotpotQA setting, graph reranking retained complete evidence for 71.0% of questions against 67.0% for dense and 68.0% for BM25 reranking, and reached 46.9% F1 against 44.8% and 47.2%; the graph-minus-dense difference of +2.1 pp is not statistically resolved. The proposed decomposition shows that only about one third of the graph-versus-dense gain in EM is due to coverage and that context dilution accounts for 37% to 45% of the gap to oracle evidence. Future work will test alias resolution, other generators, Kazakh and Russian corpora and context selection aimed at the dilution component. As local language models spread in education and public services, separating retrieval from the use of evidence is likely to become a standard part of RAG evaluation.

References:

1. Lewis P., Perez E., Piktus A. [et al.] Retrieval-augmented generation for knowledge-intensive NLP tasks // Advances in Neural Information Processing Systems. – 2020. – Vol. 33. – P. 9459–9474. – DOI: https://doi.org/10.48550/arXiv.2005.11401

2. Li Z., Wang Z., Wang W. [et al.] Retrieval-augmented generation for educational application: a systematic survey // Computers and Education: Artificial Intelligence. – 2025. – Vol. 8. – Art. 100417. – DOI: https://doi.org/10.1016/j.caeai.2025.100417

3. Yang Z., Qi P., Zhang S. [et al.] HotpotQA: a dataset for diverse, explainable multi-hop question answering // Proceedings of EMNLP 2018. – 2018. – P. 2369–2380. – DOI: https://doi.org/10.18653/v1/D18-1259

4. Trivedi H., Balasubramanian N., Khot T., Sabharwal A. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions // Proceedings of ACL 2023. – 2023. – P. 10014–10037. – DOI: https://doi.org/10.18653/v1/2023.acl-long.557

5. Liu N.F., Lin K., Hewitt J. [et al.] Lost in the middle: how language models use long contexts // Transactions of the Association for Computational Linguistics. – 2024. – Vol. 12. – P. 157–173. – DOI: https://doi.org/10.1162/tacl_a_00638

6. Chen J., Lin H., Han X., Sun L. Benchmarking large language models in retrieval-augmented generation // Proceedings of the AAAI Conference on Artificial Intelligence. – 2024. – Vol. 38, No. 16. – P. 17754–17762. – DOI: https://doi.org/10.1609/aaai.v38i16.29728

7. Edge D., Trinh H., Cheng N. [et al.] From local to global: a graph RAG approach to query-focused summarization [Electronic resource] // arXiv. – 2024. – arXiv:2404.16130. – DOI: https://doi.org/10.48550/arXiv.2404.16130

8. Jiménez Gutiérrez B., Shu Y., Gu Y. [et al.] HippoRAG: neurobiologically inspired long-term memory for large language models // Advances in Neural Information Processing Systems. – 2024. – Vol. 37. – DOI: https://doi.org/10.48550/arXiv.2405.14831

9. Zhu X., Xie Y., Liu Y. [et al.] Knowledge graph-guided retrieval augmented generation // Proceedings of NAACL 2025. – 2025. – P. 8912–8924. – DOI: https://doi.org/10.18653/v1/2025.naacl-long.449

10. Knollmeyer S., Caymazer O., Grossmann D. Document GraphRAG: knowledge graph enhanced retrieval augmented generation for document question answering within the manufacturing domain // Electronics. – 2025. – Vol. 14, No. 11. – Art. 2102. – DOI: https://doi.org/10.3390/electronics14112102

11. Wei Q., Ning H., Han C., Ding J. A query-aware multi-path knowledge graph fusion approach for enhancing retrieval-augmented generation in large language models // Expert Systems with Applications. – 2026. – Vol. 316. – Art. 131932. – DOI: https://doi.org/10.1016/j.eswa.2026.131932

12. Xiang Z., Wu C., Zhang Q. [et al.] When to use graphs in RAG: a comprehensive analysis for graph retrieval-augmented generation // International Conference on Learning Representations (ICLR 2026). – 2026. – DOI: https://doi.org/10.48550/arXiv.2506.05690

13. Zarrinkia Y., Srinivasan V., Thomo A. The reasoning bottleneck in Graph-RAG: structured prompting and context compression for multi-hop QA [Electronic resource] // arXiv. – 2026. – arXiv:2603.14045. – DOI: https://doi.org/10.48550/arXiv.2603.14045

14. Es S., James J., Espinosa Anke L., Schockaert S. RAGAs: automated evaluation of retrieval augmented generation // Proceedings of EACL 2024: System Demonstrations. – 2024. – P. 150–158. – DOI: https://doi.org/10.18653/v1/2024.eacl-demo.16

15. Apache Lucene. BM25Similarity [Electronic resource]. – Version 10.5.2. – 2026. – URL: https://lucene.apache.org/core/10_5_2/core/org/apache/lucene/search/similarities/BM25Similarity.html (accessed: 29.10.2026).

16. Cormack G.V., Clarke C.L.A., Büttcher S. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods // Proceedings of the 32nd International ACM SIGIR Conference. – 2009. – P. 758–759. – DOI: https://doi.org/10.1145/1571941.1572114

17. Qwen Team. Qwen2.5 technical report [Electronic resource] // arXiv. – 2025. – arXiv:2412.15115. – DOI: https://doi.org/10.48550/arXiv.2412.15115