Improving Multi-Hop Retrieval for Question Answering via Bipartite Question-Oriented Graphs

MICAH MCCOLLUM, DR. SUSAN GAUCH

UNIVERSITY OF ARKANSAS

CONTACT: MICAHLEE.MCCOLLUM@GMAIL.COM

UNIVERSITY OF ARKANSAS

IARIA

Micah McCollum

- Senior undergraduate in Computer Science at the University of Arkansas (Fayetteville, AR). Graduating Spring 2026.
- Research and Teaching Assistant at University of Arkansas, Summer 2025.
- Broadly interested in information retrieval, data management, and related areas.

How to answer this question?

Who was the first president of the United States?

How to answer this question?

Who was the first president of the United States?

1. Gold Answer: George Washington

How to answer this question?

Was the first president of the United States born on a leap year?

How to answer this question?

Was the first president of the United States born on a leap year?

1. Who is the first president of the United States?
2. What is their birthdate?
3. Is their birth year a leap year?

How to answer this question?

Was the first president of the United States born on a leap year?

1. Who is the first president of the United States?
2. What is their birthdate?
3. Is their birth year a leap year?
4. **Gold Answer: Yes!**

Problem

Many questions are *single-hop*, but sometimes we cannot avoid *multi-hop*

Multi-hop questions require combination of multiple, interdependent pieces of info to answer

We need ***flexibility*** in natural language interactions and ***confidence*** in natural language answers

Solution

Proposal: *Bipartite Question-Oriented Graph* (BiQOG) encoding question-passage relations

QuOTE (Neeser et al., 2025): Generate questions for each passage based on whether passage answers questions

*Why not explicitly model it as a **graph?***

- We already have implicit connections
- Are there questions that can answer different passages than the one they were generated from?

Background

1. RAG (Lewis et al. 2021): Dynamic retrieval of non-parametric knowledge
2. HyDE (Gao et al., 2023): Hypothetical Document Embeddings
3. QuOTE (Neeser et al., 2025): Question-Oriented Text Embeddings

Approach

1. LLM Hypothetical Question Generation
2. LLM-Guided Graph Construction
   a. Pre-connect generated questions with their origin passages
   b. "Does passage answer question?"
   c. Semantic filtering
3. Simple Graph Traversal
   a. Seed similar questions
   b. At each hop, explore a "width" of nodes based on vector similarity and collect neighbors
   c. After number of hops, return collection and do final re-rank

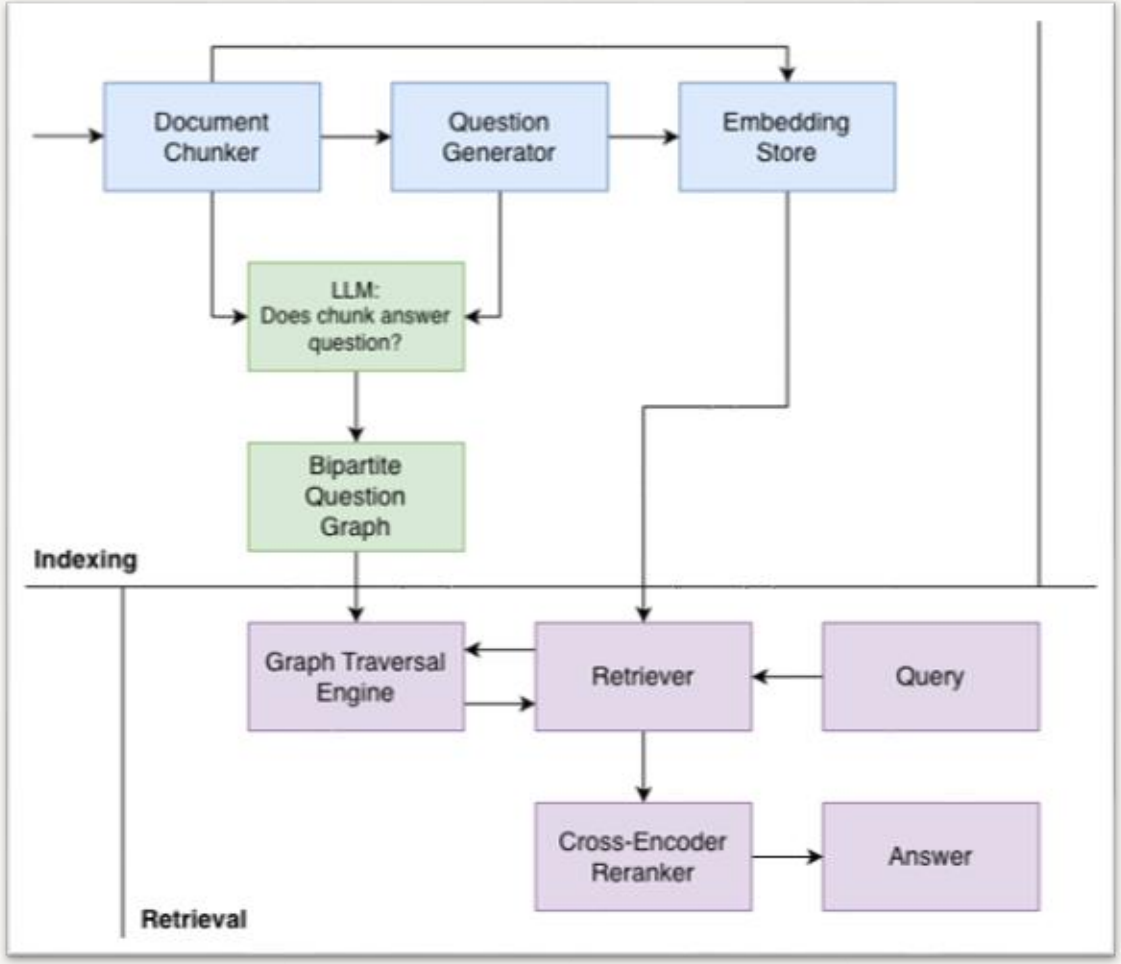


Figure 1. BiQOG High-Level Architecture

Experiment

Dataset: MultiHop-RAG (Tang and Yang, 2023)

- Challenging multi-hop questions from long-form news articles
- *Inference*, *Temporal*, *Comparison*; omit *Null*

Metrics: Average Recall@*k* and Full@*k*

- Recall@*k*: (*# of gold in top-k / # of total gold*)
- Full@*k*: are all gold in the top-*k*?
- Measure coverage: crucial for multi-hop

Baselines:

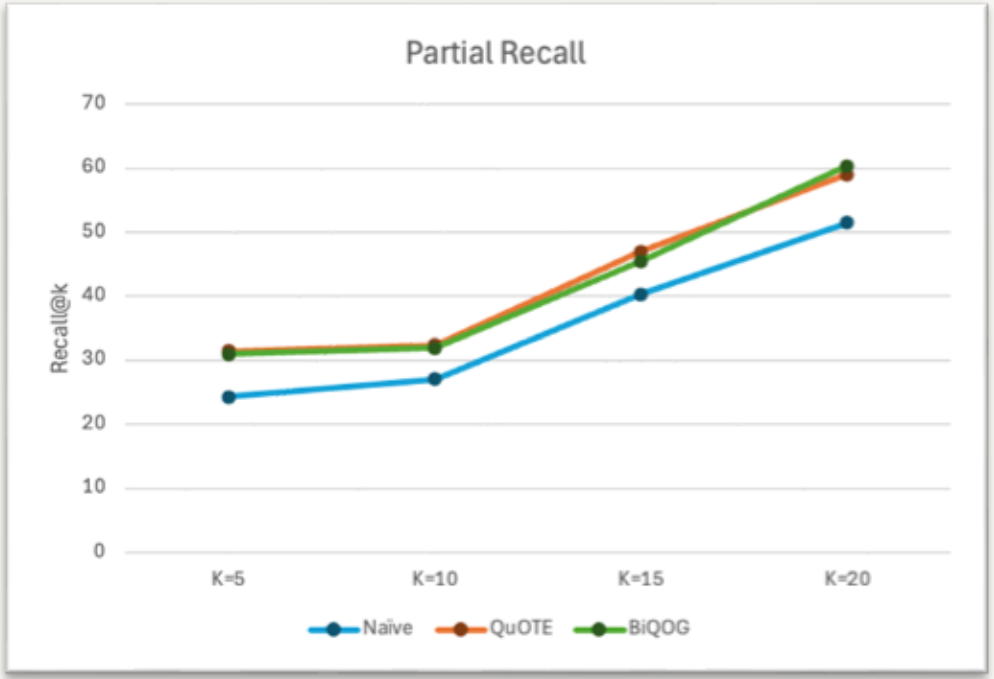
- Naive
- QuOTE

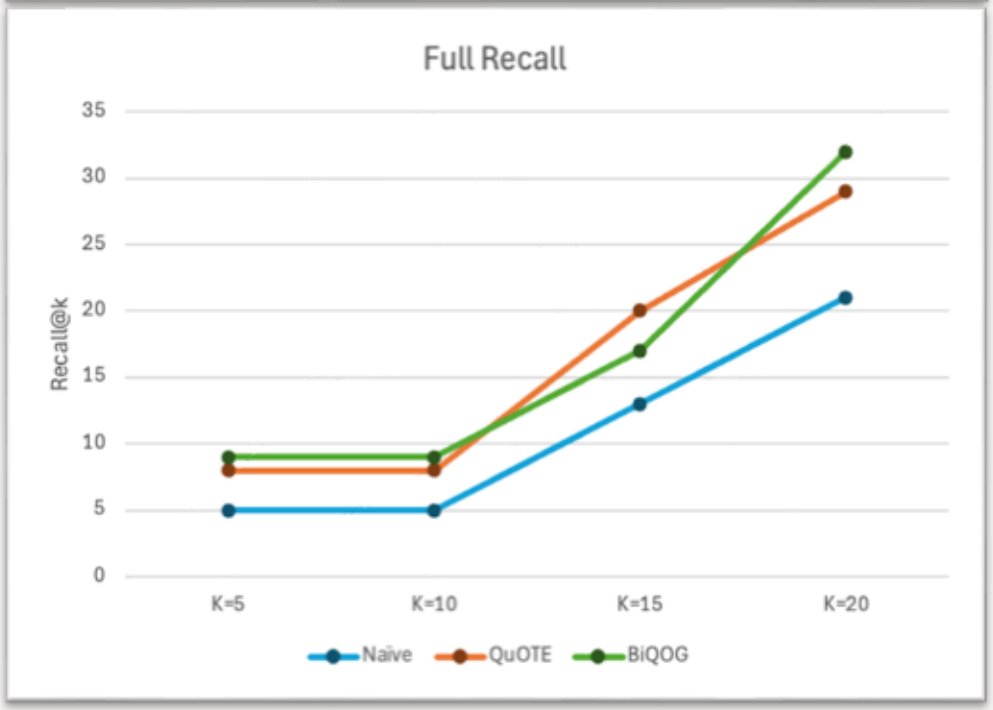
Question:

Does adding an explicit graph structure of questions help multi-hop retrieval?

Hypothesis:

Yes; cross-passage connections may be missing





Results

Approach	Recall			
	Full@5	*Full@20*	*Recall@5*	*Recall@20*
Naive	5.00	21.00	27.08	51.50
QuOTE	8.00	29.00	**32.42**	59.00
BiQOG	**9.00**	**32.00**	31.92	**60.33**

Discussion

- *Does adding explicit graph structure aid retrieval?*
- Graph Structure
  - Sparse, disconnected
  - Limits traversal
  - Few LLM-formed edges
- Improvements:
  - Revisit prompt design
  - Heterogeneous graph
  - Address connectivity

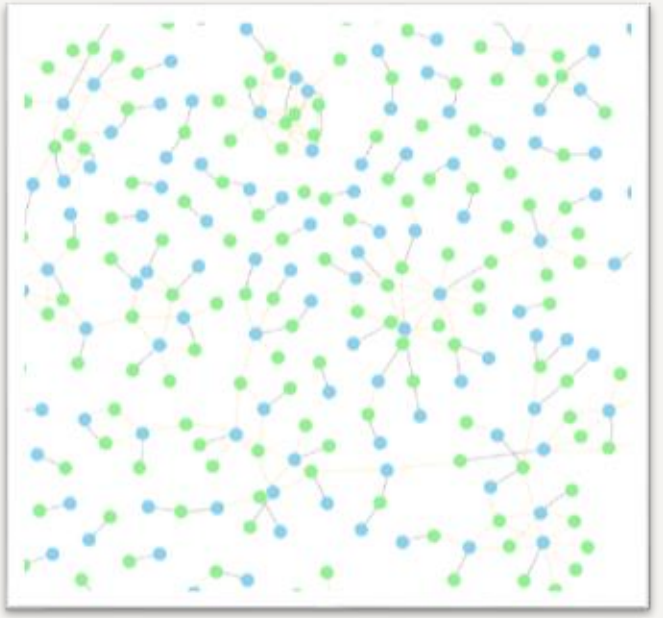


Figure 2. Subgraph visualization (questions nodes in blue)

Statistic	Value
Number of Nodes (|V|)	45382
Number of Edges (|*E*|)	44357
Average Degree	1.95
Number of Original Edges	37507 (84.47%)
Number of LLM Edges	6887 (15.53%)

Limitations

Implementation concerns:

- High LLM usage
  - Token costs
  - Inference/API latency
  - But offline, upfront
- In-memory graph
  - Extremely large workloads may require special solutions

Conclusion

Goal: Increase **flexibility** of and **confidence** in LLM-based applications

- Proposed question-passage bipartite graph to improve multi-hop retrieval
- Results are promising even with potential graph weaknesses
- Shoring up the graph structure could yield further improvements

Future Work:

- Investigate query decomposition: "compositionality gap" (Song and Zheng, 2026)
- Evaluate on larger sample size and other common datasets

References

P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks", *Advances in Neural Information Processing Systems*, vol. 33, 2020.

D. Edge et al., "From local to global: A graph RAG approach to query-focused summarization", arXiv preprint arXiv: 2404.16130, 2025. [Online]. Available: https://arxiv.org/pdf/2404.16130.

L. Gao, X. Ma, J. Lin, and J. Callan, "Precise zero-shot dense retrieval without relevance labels", Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 1762-1777, 2023. doi: 10.18653/v1/2023.acl-long.99.

A. Neeser et al., "QuOTE: Question-oriented text embeddings", arXiv preprint arXiv:2502.10976, 2025. [Online]. Available: https://arxiv.org/pdf/2502.10976.

M. Song and M. Zheng, "A survey of query optimization in large language models", arXiv preprint arXiv: 2412.17558, 2026. [Online]. Available: https://arxiv.org/pdf/2412.17558.

S. Min, V. Zhong, L. Zettlemoyer, and H. Hajishirzi, "Multi-hop reading comprehension through question decomposition and rescoring", Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 6097-6109, 2019. doi: 10.18653/v1/P19-1613.

R. Fu et al., "Decomposing complex questions makes multi-hop QA easier and more interpretable", Findings of the Association for Computational Linguistics, pp. 169-180, 2021. doi: 10.18653/v1/2021.findings-emnlp.17.

J. Li, M. Ren, Y. Gao, and Y. Yang, "Ask to understand: Question generation for multi-hop question answering", Proceedings of the 22nd Chinese National Conference on Computational Linguistics, pp. 569-582, 2023.

Y. Tang and Y. Yang, "MultiHop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries", arXiv preprint arXiv:2401.15391, 2024. [Online]. Available: https://arxiv.org/pdf/2401.15391.