Employing LLMs in Higher Education

Gabriel Abreu, Timothy Hickey, Jessica Liebowitz

Department of Computer Science, Brandeis University, Waltham, MA

Introduction

Inspired by the needs of the *Brandeis Office of Investment Management*, this work aims to implement LLMs in academic administrative offices. This raises three key challenges:

- **Privacy**: Institutional data cannot be exposed to external APIs.
- **Resources**: Limited computational power restricts models to < 7B parameters.
- **Efficiency:** Large context windows lead to long inference times.

Goal: Build a **local, privacy-preserving system** that remains cost-efficient while enabling analytical tasks such as information retrieval and document search.

Methods

1. **VectorRAG** (Baseline): A traditional RAG pipeline retrieves document chunks from a **vector database** using semantic similarity. Chunks are embedded, indexed, and passed with the query to the LLM for context-grounded responses.

2. **GraphRAG** (Proposed): Extends RAG by introducing a **knowledge graph (KG)** representation using *networkx* and *spaCy*.

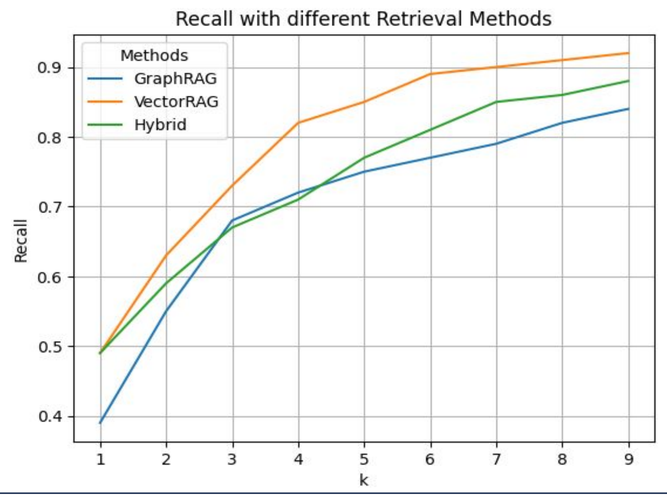
3. **Evaluation Strategy**: We replace expensive LLM-based evaluation with a **relevance-based metric** derived from labeled datasets like *Natural Questions (NQ)*.

4. **Metric: Recall**, we measure how many relevant answer sentences were retrieved for each query. A high Recall indicates effective retrieval, even if some retrieved chunks are less relevant.

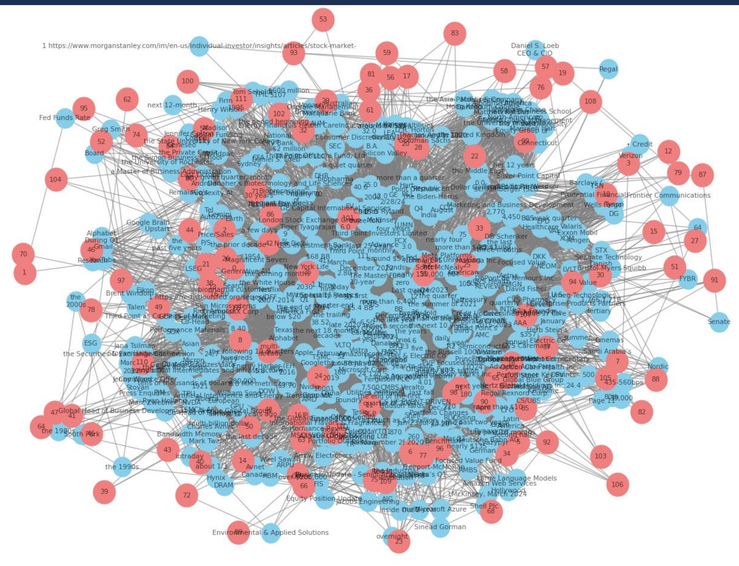
Results

- **VectorRAG** achieves the highest recall at all levels.
- **GraphRAG** preprocessing time: **1 hour** vs. **2 hours** for VectorRAG.

While VectorRAG yields better recall, GraphRAG remains more efficient and interpretable.



Graph Visualization



Conclusions

GraphRAG demonstrates that retrieval frameworks can be optimized for **efficiency, transparency, and low hardware requirements**.

Although VectorRAG excels in **recall**, GraphRAG's **faster preprocessing** and **semantic structure** make it ideal for real-time or r**apidly changing datasets**.

Insights

- **Efficiency trade-off:** GraphRAG achieves **3× faster** setup without costly embeddings.
- **Practical deployment:** can run locally without cloud APIs or high-end clusters.
- **Ideal for privacy-sensitive domains:** education, finance, administration.

Retrieval can thrive even in **resource-limited environments,** showing it does not always require massive infrastructure.

References

1. Sarmah, B., Hall, B., Rao, R., Patel, S., Pasquali, S., & Mehta, D. (2024). *HybridRAG: Integrating Knowledge Graphs and Vector Retrieval Augmented Generation for Efficient Information Extraction*.

2. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. Neural Information Processing Systems; Curran Associates, Inc

3. Yepes, A. J., You, Y., Milczek, J., Laverde, S., & Li, R. (2024, March 16). *Financial Report Chunking for Effective Retrieval Augmented Generation*.

4. Yepes, A. J., You, Y., Milczek, J., Laverde, S., & Li, R. (2024, March 16). *Financial Report Chunking for Effective Retrieval Augmented Generation*.

Paper