Hybrid Context Retrieval Augmented Generation Pipeline: LLM-Augmented Knowledge Graphs and Vector Database for Accreditation Reporting Assistance

Candace Edwards

Advisor: Dr. Mahdi Belcaid , Information and Computer Sciences Department

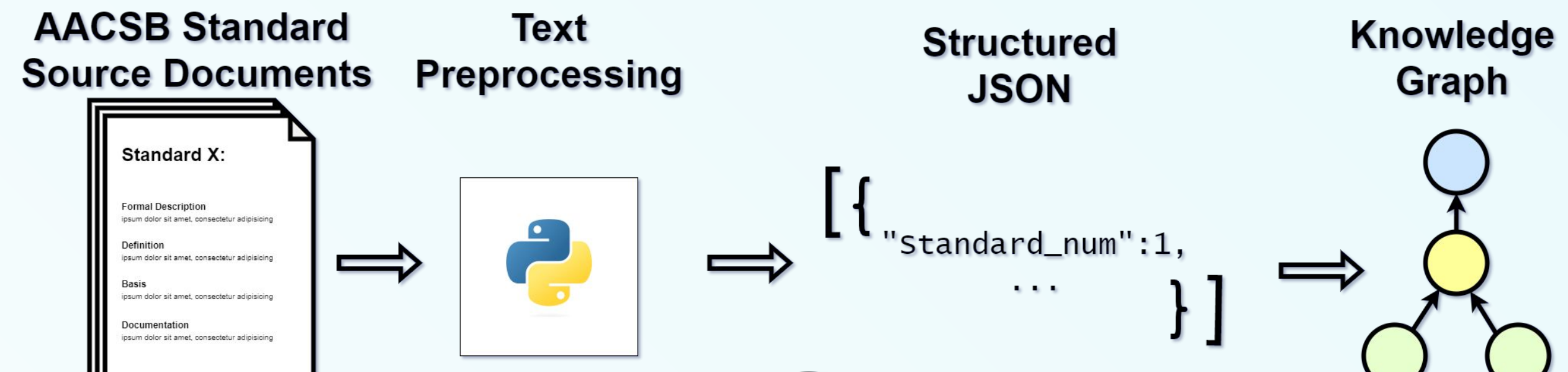
MS Plan B / ICS 699, Spring 2024

INFORMATION & COMPUTER SCIENCES

UNIVERSITY of HAWAIʻI at MĀNOA

Abstract

In higher education, accreditation is a quality assurance process, where an institution demonstrates a commitment to delivering high quality programs and services to their students. For business schools nationally and internationally the Association to Advance Collegiate Schools of Business (AACSB) accreditation is the gold standard. For a business school to receive and subsequently maintain accreditation, the school must undertake a rigorous, time consuming reporting and peer review process, to demonstrate alignment with the AACSB Standards. For this project we create a hybrid context retrieval augmented generation pipeline that can assist in the documentation alignment and reporting process necessary for accreditation. We implement both a vector database and knowledge graph, as knowledge stores containing both institutional data and AACSB Standard data. The output of the pipeline can be used by institution stakeholders to build their accreditation report, dually grounded by the context from the knowledge stores. To develop our knowledge graphs we utilized both a manual construction process as well as an 'LLM Augmented Knowledge Graph' approach. We evaluated the pipeline using the RAGAs framework and observed optimal performance on answer relevancy and answer correctness metrics.



AACSB Standard Source Documents
Standard X:
Formal Description
ipsum dolor sit amet, consectetur adipisicing
Definition
ipsum dolor sit amet, consectetur adipisicing
Basis
ipsum dolor sit amet, consectetur adipisicing
Documentation
ipsum dolor sit amet, consectetur adipisicing
Text Preprocessing
Structured JSON
[{"standard_num":1,
...
}]
Knowledge Graph

Figure 2. Processing

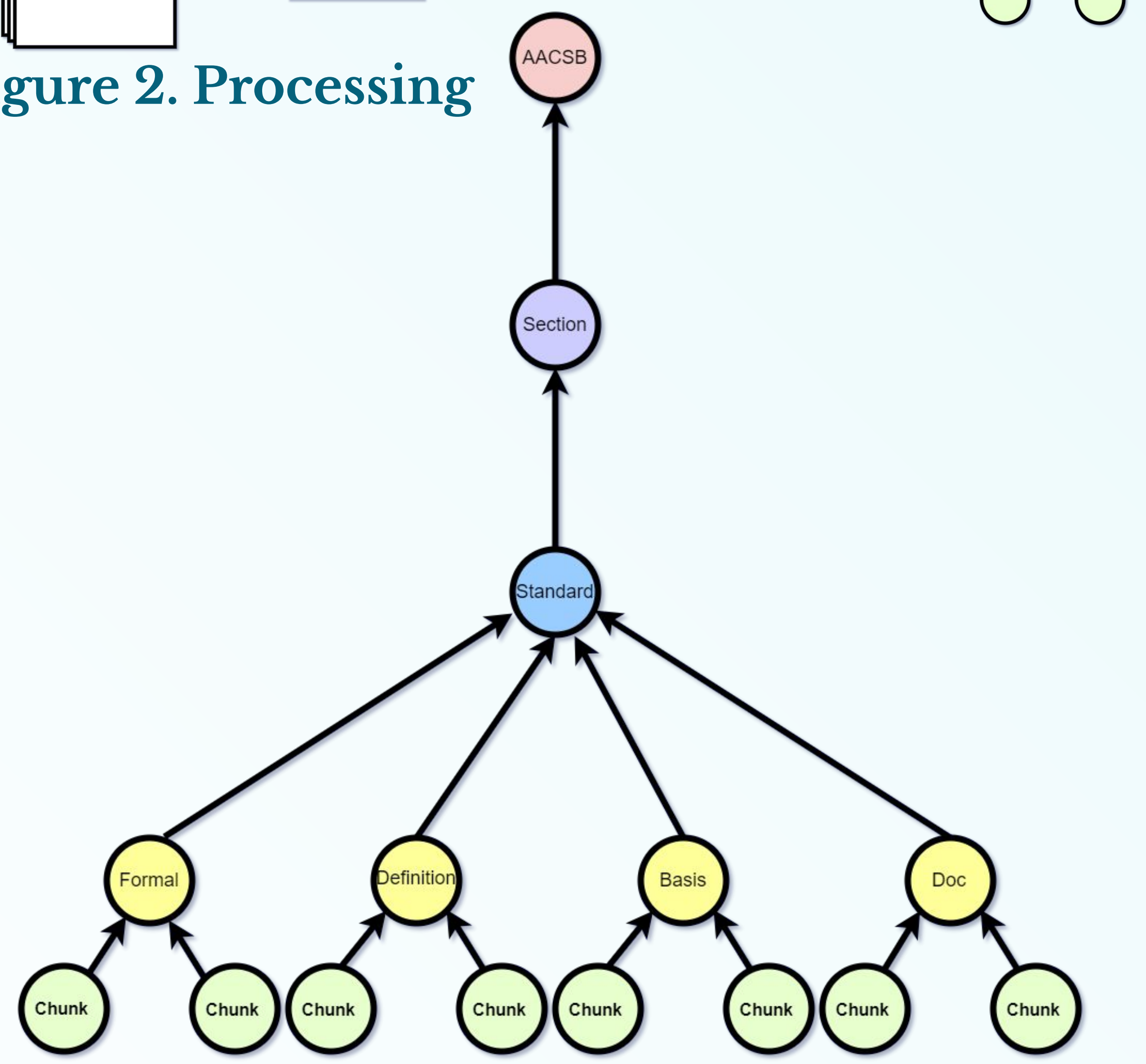


Figure 3. Structure

Figure 1 (Right): Hybrid Context RAG Sequence. User initiates query. Query is used to generate multiple related queries which are then passed into the embedding model to create the query vectors. Vector index returns the top k most similar embeddings based on cosine similarity to query rectors. Knowledge graph returns information based on the Cypher query version of the original query. In the generated, the hybrid contexts are combined with the original query and a response to user query is synthesized and returned.

Figure 2 (Above): Manual processing of AACSB unstructured data into JSON and subsequent knowledge graph.

Figure 3 (Above): Hierarchical graph structure of AACSB Standards.

Methodology

1. Establishing the Knowledge Base
   - Manually construct knowledge graph from AACSB data (Figure 2), maintain given information hierarchy (Figure 3).
   - Unstructured institutional documents processed by LLM, extracting node types and relationships. LLM performs: Entity Typing, Entity Resolution, Coreference Resolution, and Relation Extraction.
   - Knowledge graphs are linked based on metadata classification on institutional document relevance to AACSB Standard nodes.

2.. Semantic Layer
   - Vector embeddings generated for each "Chunk" node and stored in the vector index.

3. Hybrid Context Pipeline (Figure 1):
   - Query Optimization : User query is used to generate multiple related queries.
   - Multi-Source Retrieval: Vector index and knowledge return data based on queries.
   - Generator: Using the hybrid context as a grounding knowledge source and the original query, a natural language response to the user query is produced..

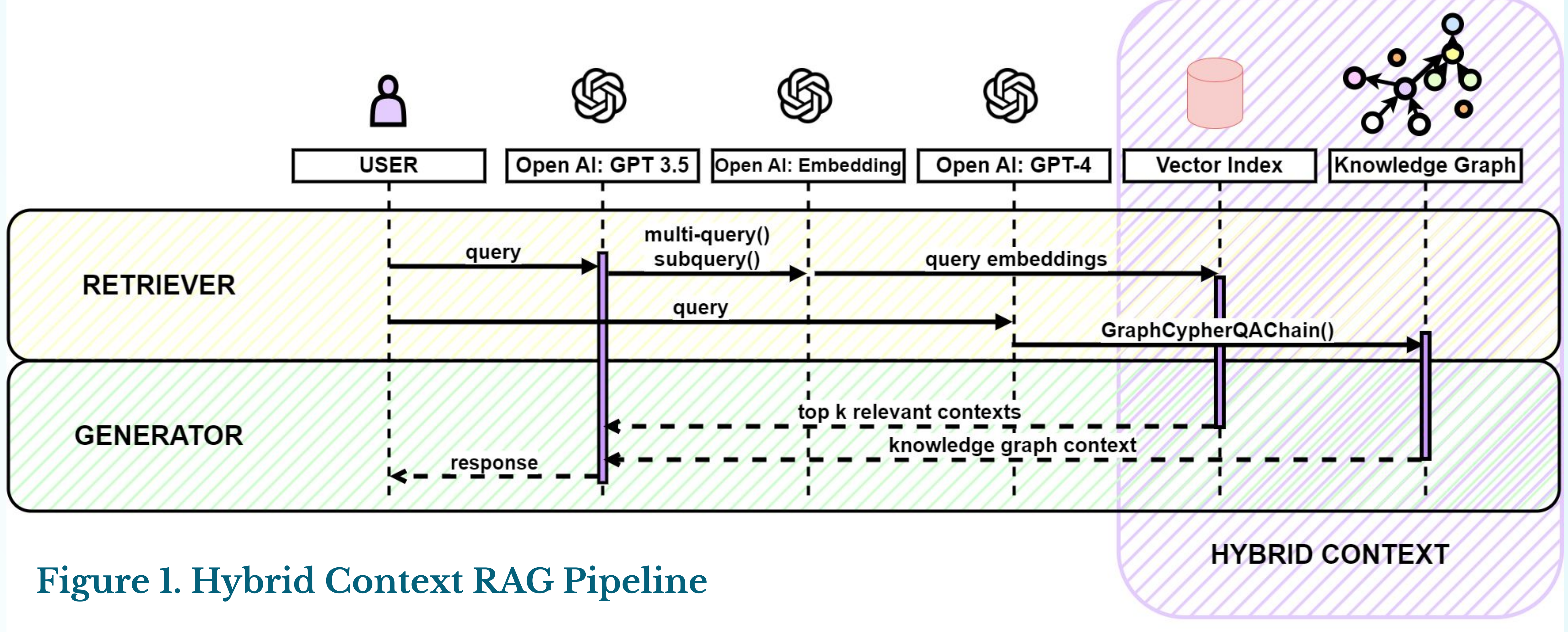


Figure 1. Hybrid Context RAG Pipeline

Results

- The pipeline's performance varied across metrics, with scores ranging from low to fair.
- Performance was observed to be better for AACSB classified queries compared to other categories.
- Expanding the validation query set could enhance the precision of pipeline evaluation, albeit with a time-consuming process due to the manual creation of queries and ground truth information.

Table 1: Advanced RAG Pipeline Metric Summary

Metric	Context Relevance	Faithfulness	Answer Relevancy	Context Recall	Answer Correctness
Mean	0.440	0.252	0.778	0.708	0.787
Median	0.429	0.000	1.000	0.901	1.000
Min	0.000	0.000	0.000	0.000	0.000
Max	0.873	1.000	1.000	0.984	1.000

Table 2: Advanced RAG Pipeline Metric Summary (by category)

Category	Context Relevance	Faithfulness	Answer Relevancy	Context Recall	Answer Correctness
AACSB	0.524	0.522	1.000	0.900	0.813
HYBRID	0.217	0.000	1.000	0.915	1.000
INST	0.481	0.000	0.333	0.300	0.611

Conclusion

Hybrid context pipeline achieved notable performance on AACSB-related queries.

Challenges lie in the non-deterministic nature of LLMs in knowledge graph generation, leading to uncertainty in graph structure and difficulty in generating useful Cypher queries from natural language. For best performance Cypher queries may need to be fine-tined based on the graph schema for each institution.

The process creates a large and complex graph. Future work could involve experimenting with pruning the LLM-constructed knowledge graph to enhance efficiency improving accuracy.