Hybrid Retrieval-Augmented Generation System Using Knowledge Graphs and Vector Databases for Domain-Specific Information Extraction with Glossary-Aided Responses

By

De Silva K.N.N.C.

Jayawickrama D.S.K.

Weerasingha W.G.K.M.

BICT Honours Degree **2024**

Hybrid Retrieval-Augmented Generation System Using Knowledge Graphs and Vector Databases for Domain-Specific Information Extraction with Glossary-Aided Responses

Authors	De Silva K.N.N.C. Software Technology Department of Information and Communication Technology Faculty of Technology University of Sri Jayewardenepura ICT/20/826 ict20826@fot.sjp.ac.lk
	Jayawickrama D.S.K. Software Technology Department of Information and Communication Technology Faculty of Technology University of Sri Jayewardenepura ICT/20/862 ict20862@fot.sjp.ac.lk
	Weerasingha W.G.K.M. Software Technology Department of Information and Communication Technology Faculty of Technology University of Sri Jayewardenepura

	ICT/20/956 ict20956@fot.sjp.ac.lk
Main Supervisor	Dr. Chamara Liyanage Academic Supervisor, Department of Information and Communication Technology, University of Sri Jayewardenepura
External Supervisor	Mr. Indika Hiran Wijesinghe Assistant Director IT, Sri Lanka Tea Board

Declaration

We certify that this proposal has not incorporated, without acknowledgment, any material previously submitted for a degree or diploma at any university. To the best of our knowledge and belief, it does not contain any material previously published or written by another person, except where due reference is made in the text.

Acknowledgement

We would like to express our deepest gratitude to all those who supported us throughout this research project. First and foremost. We extend our sincere thanks to our academic supervisor, Dr. Chamara Liyanage, for his invaluable guidance, insightful suggestions, and constant encouragement. His expertise and dedication were instrumental in shaping this research project and ensuring its successful completion.

We are also grateful to the Department of Information and Communication Technology, University of Sri Jayewardenepura for providing the resources and support needed to conduct this research. Special thanks to the research committee for approving and guiding this project.

A heartfelt thank you to the Sri Lanka Tea Board for granting access to their extensive archives and providing necessary data that formed the foundation of this study. We would like to particularly acknowledge our external supervisor Mr. Indika Hiran Wijesinghe, Assistant Director IT from Sri Lanka Tea Board, whose collaborations and support were crucial in aligning this research with industry needs.

Finally, we would like to acknowledge the assistance of our peers and colleagues who offered feedback and shared their expertise during critical stages of this project.

Table of Contents

List of Figures

List of Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
CNN	Convolutional Neural Network
FastAPI	A Python-based framework for building APIs
HybridRAG	A system that combines vector-based retrieval and graph-based retrieval to enhance information extraction and generation
KG	Knowledge Graph
LangChain	A framework for managing LLM interactions and workflows
LLM	Large Language Model
Milvus	A vector database management system
Neo4j	A graph database management system
NLP	Natural Language Processing
RAG	Retrieval-Augmented Generation
React	A JavaScript library for building user interfaces
SpringBoot	A Java-based framework for creating backend applications

1. Introduction and Background

1.1 Overview of Proposed Study

The proposed research aims to develop an AI-powered agent system that integrates domain-specific knowledge through a Retrieval-Augmented Generation (RAG) model, combined with a knowledge graph and vector database which referred to as "Hybrid RAG." Many organizations, especially those with longstanding histories in specific fields, rely on traditional physical files to store legal documents, which serve as key references for decision-making. However, this manual approach can be inefficient and leads to the loss of critical institutional knowledge when experts retire or leave the organization. The new system seeks to address this by extracting and organizing information from historical documents to provide users with accurate, contextually relevant insights. By simulating interactions with a domain expert, the system will streamline the retrieval and interpretation of complex historical data, thus supporting more informed decision-making processes.

The proposed Hybrid RAG system will combine the capabilities of Large Language Models (LLMs) with a knowledge graph and vector database with integrating domain specific glossary to generate accurate responses to user queries. Textual data will be transformed into vector embeddings and stored in a graph database, enabling efficient information retrieval while clarifying relationships between different data points. This approach ensures prompt access to past decisions, facilitating deeper insights into future organizational strategies. The project also includes the development of a web-based document management portal and an intuitive user interface, allowing authorized personnel to maintain the knowledge graph and engage with the system. Additionally, this integration of domain-specific knowledge will improve response precision, making the system an essential tool for enhancing both data management and decision-making processes across organizations.

1.2 Project Scope

A modern, web-based portal will be developed to manage the organization's repository of historical documents. This portal will allow authorized users to upload, view, and validate documents, ensuring that the knowledge base remains comprehensive, up-to-date, and accessible to all personnel. This feature is designed to streamline document management processes in organizations that previously relied on manual handling of paper-based archives.

1.2.1 Development of Document Management Portal

The project will create an interactive modern web-based portal to manage the Sri Lanka Tea Board's vast repository of historical documents. This portal will enable authorized users to upload, view, and validate documents, ensuring that the knowledge base remains current and comprehensive.

1.2.2 Construction of Knowledge Graph and Vector database

The most significant aspect of this project involves the construction of a detailed knowledge graph derived from the textual content of documents spanning several decades. The text will be divided into smaller, manageable chunks, which will form nodes within the knowledge graph. Relationships between these nodes will be mapped to accurately represent the interconnected nature of the information, facilitating effective retrieval and understanding of the context. Additionally, to enhance the retrieval process, a vector database will be integrated. Each text chunk will be transformed into vector embeddings, capturing the semantic meaning of the content. These embeddings will be stored in the vector database, allowing for efficient similarity-based searches. The combination of the knowledge graph's structured relationships and the vector database's semantic search capabilities will enable more precise and context-aware information retrieval.

1.2.3 Integration of Advanced Technologies

The RAG system will integrate advanced technologies such as Large Language Models (LLMs), LangChain, Milvus and Neo4j database to ensure robust information retrieval. User prompts will be transformed into vector embeddings and matched against the knowledge graph, retrieving the most relevant information. This approach allows the system to simulate expert-level responses based on historical data and domain knowledge.

1.2.4 User Interface Design

A user-friendly, intuitive interface will be developed, allowing personnel across the organization to interact with the system seamlessly. Through this interface, users will be able to pose inquiries in natural language, with the system retrieving and presenting information drawn from the RAG System and enhanced by LLM-generated responses. The interface will be designed to accommodate users with varying levels of technical expertise.

1.2.5 Enhancement with Domain-Specific Glossary

The system will incorporate domain-specific glossary to improve the accuracy of information retrieval and interpretation. For organizations with specialized terminologies, a custom glossary can be added to ensure that user queries are accurately understood, and contextually relevant responses are provided. This feature will enable the system to adapt to the unique requirements of each organization.

1.2.6 Scalability and Maintenance Considerations

The system architecture will be designed for scalability, ensuring that it can accommodate an increasing volume of documents and user queries over time. It will also prioritize maintainability, allowing for straightforward updates, bug fixes, and the integration of new features as required. This ensures the system remains functional and adaptable as the organization's needs evolve.

1.2.7 Collaboration and Ethical Compliance

The project will ensure compliance with ethical standards regarding the handling of proprietary and sensitive information. Close collaboration with participating organizations will be essential to ensure that all ethical and legal requirements are met. Necessary approvals and clearances will be obtained to guarantee adherence to protocols established by both the organization and the research team.

1.3 Motivation for the study

The motivation for this study arises from the need to modernize the Sri Lanka Tea Board's approach to managing and retrieving historical data from legal documents. These documents, which encapsulate important decisions and regulatory acts since 1976, are essential for informed decision-making within the tea industry. Currently, they depend on traditional methods, such as physical file management and consultations with several experienced personnel, poses significant risks. As the workforce changes and experienced personnel retire, much of this knowledge risks being lost or becoming inaccessible.

Additionally, the complexity of the tea industry and its regulatory environment makes it challenging for current staff to navigate and retrieve relevant information from historical documents. By developing this system, the study aims to bridge this knowledge gap and ensure that critical information remains accessible to future generations.

1.4 Rationale and Justification

The rationale for this study is based on the critical need for organizations with extensive histories to manage their vast repositories of historical documents more efficiently. Many organizations across various sectors rely on traditional methods for storing and retrieving vital records, which often contain decades of institutional knowledge and decision-making processes. As the volume of documents grows and key personnel retire or leave, the ability to access and utilize this information becomes increasingly difficult. This creates a gap in institutional memory, leading to inefficiencies in decision-making, delays, and the potential loss of crucial historical insights.

The justification for this research lies in its potential to revolutionize the way organizations manage and leverage their historical data. By developing a Hybrid RAG System integrated with a knowledge graph and vector database, this study proposes a modern, scalable solution for addressing the limitations of traditional document management methods. The system will enable organizations to preserve institutional knowledge, make it more accessible, and use it effectively to inform current and future decisions. This technology offers a robust, context-aware approach to retrieving and understanding historical data, allowing organizations to improve operational efficiency, maintain continuity, and make informed strategic decisions based on comprehensive, readily available information.

1.5 Research Problem

Organizations with extensive histories face significant challenges in managing and retrieving critical information from their large archives of historical documents. These documents, often containing vital decisions and institutional knowledge, are essential for informed decision-making. However, many organizations still rely on outdated, manual methods to store and access this information. As experienced personnel retire, there is a growing risk of losing valuable domain expertise, leaving newer employees without the deep knowledge necessary for effective decision-making. The lack of an efficient knowledge transfer process further exacerbates this issue.

This situation highlights the urgent need for an advanced, automated system capable of managing, retrieving, and contextualizing historical data. By developing such a system, organizations can preserve their institutional knowledge, provide easier access to critical information, and ensure continuity in decision-making processes. This solution is particularly important as organizations adapt to modern challenges while still relying on decades-old data and insights.

1.6 Research Objectives, Hypotheses, and Questions

1.6.1 Main Objective

The primary objective of this research is to develop an advanced Hybrid RAG System integrated to improve organizations' ability to manage, retrieve, and utilize historical documents. This system will aid decision-making by providing accurate, contextually relevant information based on decades of documented institutional actions, decisions, and domain-specific knowledge.

1.6.2 Specific Objectives

- To design and implement a web-based document management portal that allows for the efficient upload, viewing, and deletion of documents by authorized personnel.
- To extract knowledge from historical documents into a knowledge base, enabling better organization and retrieval of information
- To integrate a Large Language Model (LLM) with the knowledge base to generate precise and contextually relevant responses to user queries
- To develop a user-friendly interface for querying the system, ensuring that users can easily interact with the RAG system and access the necessary information.
- To incorporate a domain-specific dictionary into the system, enhancing the accuracy of the LLM-generated responses by ensuring correct interpretation of industry-specific terminology.

1.6.3 Hypotheses

- The implementation of a Hybrid RAG system will significantly improve the efficiency and accuracy of information retrieval compared to traditional manual document management methods.
- The use of a domain-specific glossary within the system will enhance the contextual relevance and accuracy of LLM-generated responses, leading to better decision-making across different organizations.

1.6.4 Research Questions

- How can a Hybrid RAG system be effectively utilized to manage and retrieve information from historical documents within an organization?
- What impact will the implementation of this system have on the efficiency and accuracy of information retrieval compared to existing manual methods?
- How does the integration of a domain-specific dictionary influence the performance of the LLM in generating accurate and contextually relevant responses?
- What are the potential challenges in implementing such a system, and how can they be mitigated to ensure successful adoption and long-term sustainability?

1.7 The Conceptual, Theoretical, and Practical Significance of the Research

This research is conceptually significant as it explores the application of Hybrid RAG models in domain-specific knowledge management, contributing to a deeper understanding of how AI-driven systems can be tailored to manage institutional knowledge across diverse sectors. Theoretically, it advances the field of information retrieval by demonstrating the effectiveness of integrating knowledge graphs with vector-based searches and Large Language Models (LLMs), thus offering a novel approach to accessing and interpreting complex historical data. Practically, the research provides a robust solution to the widespread challenge faced by organizations in efficiently managing and retrieving valuable institutional knowledge. This system will aid decision-making processes, improve operational efficiency, and ensure that organizations can continue to leverage their historical data for strategic planning and governance.

1.8 Expected Outcomes/Anticipated Results

The proposed research is expected to result in the development of a robust and efficient Hybrid RAG system, integrated with a knowledge graph and vector database, that will significantly enhance the organization’s ability to manage and retrieve information from its extensive archive of historical legal documents.

1.8.1 Enhanced Information Retrieval

The RAG system will facilitate the swift and accurate retrieval of contextually relevant information from historical documents, improving decision-making processes within the organization. This will address the current inefficiencies associated with manual document searches and the reliance on institutional memory.

1.8.2 Preservation and Utilization of Institutional Knowledge

By structuring and storing the extracted knowledge in a graph and vector database, the system will preserve valuable institutional knowledge that might otherwise be lost as experienced personnel retire. This will ensure continuity in the organization’s operations and provide new staff with easier access to historical data.

1.8.3 Improved Decision-Making

The system's ability to generate accurate responses based on historical data will enable the organizations to make more informed decisions. This will be particularly valuable in situations where understanding past decisions and regulations is critical to addressing current challenges in the longstanding organizations.

1.8.4 User-Friendly Interface

The development of an intuitive user interface will allow users, regardless of their technical expertise, to easily interact with the system, ensuring widespread adoption and effective use.

1.8.5 Contribution to Academic and Practical Knowledge

The research will contribute to the academic understanding of RAG systems particularly in the context of document management. The practical outcomes will provide a model that can be adapted by other government and industry bodies facing similar challenges in managing large volumes of historical data.

1.8.6 Long-Term Sustainability

By ensuring that the system is both maintainable and scalable, this project will provide a long-term solution for organizations seeking to preserve and utilize their institutional knowledge effectively. The system will support the continuous adaptation and growth of these organizations, enabling them to make well-informed decisions and maintain their competitive advantage in an evolving industry landscape.

2. Literature Review

Organizations across various industries face significant challenges in managing large volumes of legal and operational documents, especially when stored in unstructured formats such as scanned PDFs. Efficient and accurate information extraction is crucial for effective decision-making and regulatory compliance, yet traditional retrieval methods often struggle due to the unstructured nature of the data. Advanced methods like Hybrid Retrieval-Augmented Generation (RAG) systems, which combine both knowledge graph and vector-based approaches, have immerged as promising solutions to overcome these limitations (Sarmah et al., 2024).

Our research proposes a Hybrid RAG system that utilizes both knowledge graphs and vector-based retrieval for enhanced information extraction. This system is designed for organizations managing diverse datasets, enabling more precise and contextually relevant responses specially integrating context relevant dictionary feature when querying structured and unstructured documents. This literature review evaluates prior research on knowledge graph construction, vector-based retrieval, and RAG systems, identifying relevant contributions and gaps our research seeks to address.

Knowledge graphs have become a key tool in domain-specific information extraction, particularly for unstructured data. Zhao et al. (2020) present a method for constructing a domain-specific knowledge graph for technical documents using TextCNN-based topic extraction model. The study illustrated how knowledge graphs can reveal semantic relationships between technical concepts, enhancing retrieval and analysis (Zhao, Pan, & Yang, 2020). Their use of Neo4j for graph storage and visualization underscores the adaptability of knowledge graphs for various domains. However, while Zhao's work focused on technical documents, our research expands the application to a wider variety of documents, particularly those stored as unstructured PDFs, where domain-specific terminology is often present.

Sarmah et al. (2024) took this a step further by proposing a HybridRAG system that integrates VectorRAG and GraphRAG techniques for enhanced retrieval. Their system demonstrated superior performance compared to traditional RAG approaches, particularly in environments that require both structured and unstructured data to be processed. By utilizing knowledge graphs for structured information and vector-based methods for unstructured text, their system achieved higher retrieval accuracy (Sarmah et al., 2024). This hybrid approach is highly relevant to our research, as organizations often deal with document corpora that contain structured metadata as well as extensive unstructured text. By adopting a hybrid system, our research seeks to offer a robust solution for retrieving information from diverse document types.

Moreover, Hu et al. (2024) emphasized the potential of combining LLMs with knowledge graphs for constructing domain-specific knowledge graphs from unstructured text. Their approach involved using LLMs for data annotation, particularly in contexts where labeled data is scarce (Hu et al., 2024). This methodology is critical to our research, as it provides a framework for building knowledge graphs from unstructured, unlabeled documents—a frequent challenge faced by organizations managing large, unstructured datasets.

Despite these advancements, several research gaps remain. Existing studies primarily focus on domain-specific implementations, with limited exploration of hybrid systems that apply to more generalized settings. Furthermore, while vector-based and knowledge graph-based RAG systems have been shown to work well individually, few studies have explored the combined power of these techniques in broader, more complex documents (Banerjee et al., 2024; Fang, Meng, & Macdonald, 2024).

Our research aims to address these gaps by developing a hybrid RAG system that integrates knowledge graphs and vector-based retrieval techniques to support organizations across various domains with integrating domain-specific dictionary. By leveraging both structured and unstructured data, we aim to build a system that not only retrieves information efficiently but also provides contextually relevant and accurate responses, tailored to the specific requirements of the user. This hybrid approach ensures greater flexibility and precision in information extraction, advancing decision-making processes for organizations managing large, diverse document repositories (Edwards, 2024; Su et al., 2024) .

3. Methodology

3.1 Proposed Experiments/Investigations and Techniques

The primary objective of this project is to create a HybridRAG system that combines context retrieval from a knowledge graph in a graph database and a vector database, with an added domain-specific glossary to enhance the responses provided by the system. To achieve this, we will conduct a series of experiments focused on both the retrieval and response-generation phases.

3.1.1 Document Management Portal

A web-based portal will be developed to allow authorized personnel to upload, view, and delete board documents. This portal ensures that the RAG remains up-to-date with the latest documents and provides a secure interface for managing the document repository.

3.1.2 Vector Store Setup

The first stage involves building a vector store for the documents provided by the Sri Lanka Tea Board (or other relevant documents for the system's generic aspect). Text extraction will be carried out by breaking documents into fixed-size chunks and converting these chunks into vector embeddings. These embeddings will then be stored in the Milvus vector database, which supports efficient vector similarity search.

3.1.3 Knowledge Graph Creation

In parallel, we will build a knowledge graph using Neo4j. Entities and relationships will be automatically extracted using Named Entity Recognition

(NER) and Relation Extraction models. The APOC (Awesome Procedures on Cypher) library will further automate the generation of relationships between entities. This graph database will provide structured insights for the system to use when responding to user queries.

3.1.4 Context Retrieval for Information Extraction

Once the vector and knowledge graph databases are set up, users will input their queries into the system. For each query, two contexts will be retrieved: one through a semantic search of the vector database and another through graph traversal in the knowledge graph. These two contexts, along with any relevant domain-specific glossary terms, will be provided to the LLM for generating an appropriate answer.

3.1.5 Response Generation and Display

The final stage involves combining the user's query with both contexts and any glossary definitions. The LLM will process this input and generate an answer, which will be displayed in the frontend interface of the web app.

3.2 Justification of Methodology to Realize Objectives

The methodology outlined ensures the project objectives can be effectively met. The hybrid approach of using both a vector database (Milvus) and a graph database (Neo4j) ensures that information retrieval is comprehensive, pulling from both structured (graph) and unstructured (vector) data sources. This improves the quality of the system's responses by providing the LLM with a richer context.

3.2.1 Realization of Main Objectives

The system’s main objective—to allow users to extract meaningful insights from a document-based knowledge base—is achieved by using the semantic search capabilities of the vector database and the structured insights from the graph database. Both techniques complement each other, ensuring the system can handle complex, real-world queries effectively. The domain-specific glossary further strengthens the system’s relevance to the target domain by ensuring precise terminology understanding.

3.2.2 Realization of Specific Objectives

Specific objectives, such as creating an efficient document management portal and allowing seamless integration of new documents into the system, are met through the use of React, SpringBoot, and FastAPI. These technologies enable the development of a user-friendly interface that allows authorized personnel to upload, view, and delete documents. The integration of vector and graph database technologies into this portal ensures that the system remains scalable and robust in terms of both information retrieval and management.

3.3 Other Considerations

3.3.1 Ethical Clearance

Given that the data consists of internal documents from the Sri Lanka Tea Board, ensuring the confidentiality and proper use of these documents is paramount. The project will adhere to any ethical guidelines set forth by both the university and the Sri Lanka Tea Board, including obtaining any necessary approvals or clearances for the use of proprietary or sensitive information.

3.3.2 Collaboration with the Sri Lanka Tea Board

The project is being supervised by the Assistant Director of Information Technology from the Sri Lanka Tea Board, who is acting as an external supervisor. This collaboration provides guidance and ensures that the project aligns with the Board's standards and requirements.

3.3.3 Data Security and Privacy

Given the sensitive nature of some documents, data security measures will be implemented to protect the confidentiality of the information. This includes access controls within the document management portal, encryption of stored data, and secure communication protocols between the web app and the backend systems.

3.4 Functional and Non-Functional Requirements

3.4.1 Functional Requirements

Document Management Portal

- The system must provide a secure interface for authorized personnel to upload, view, and delete documents in the knowledge base. Each action should be logged, ensuring traceability.

Knowledge Base Construction

- The system must convert uploaded documents into both vector embeddings (for the vector store) and knowledge graph entities (for the knowledge graph). Automated processes for text chunking, embedding, and entity-relation extraction should be implemented.

Query and Information Retrieval

- Users must be able to submit queries through the web app. The system should retrieve relevant information from both the vector database and the knowledge graph in response to these queries.

Glossary Integration

- The system must detect domain-specific terms in the user's queries and provide definitions from the domain-specific glossary to enhance the context provided to the LLM.

User Interface

- The web app must present the LLM-generated answer to the user in a clear and concise format.

3.4.2 Non-Functional Requirements

Performance

- The system should return answers to user queries with low latency, ensuring a smooth user experience. This includes optimizing both the vector search and graph traversal processes.

Scalability

- The system must be scalable, allowing for the integration of more documents over time without significant performance degradation.

Security

- Only authorized personnel should be able to access the Document Management Portal. All sensitive data should be securely stored and protected from unauthorized access.

Usability

- The web app interface must be intuitive and easy to navigate, ensuring that users with varying levels of technical expertise can interact with the system effectively.

Reliability

- The system must be reliable, with minimal downtime, ensuring consistent availability for users.

Maintainability

- The system should be designed with maintainability in mind, allowing for easy updates, bug fixes, and feature enhancements.

3.5 Specific Use of Technologies and Techniques

3.5.1 Technologies

Neo4j (Graph Database)

- Neo4j is a powerful graph database used for managing and querying the knowledge graph. It will store the structured data extracted from documents in the form of entities (nodes) and relationships (edges). Neo4j allows efficient graph traversal, making it suitable for retrieving contextual information based on the connections between entities.

Milvus (Vector Database)

- Milvus is a high-performance, open-source vector database that specializes in storing vector embeddings. It enables semantic search, where the similarity between vectors (representing text chunks from documents) can be calculated. This helps to retrieve relevant information from unstructured data.

LangChain

- LangChain facilitates the integration of language models (LLMs) with external data sources, such as databases and knowledge graphs. It helps manage the hybrid retrieval process, combining responses from both the graph and vector databases and passing that context to the LLM.

Llama 3.1 (Large Language Model)

- Llama 3.1 is the core model used for generating answers based on the user's query and retrieved contexts. It will analyze the input, apply natural language processing, and generate coherent and accurate answers, especially when enhanced by domain-specific terms from the glossary.

React (Frontend Library)

- React will be used to build the web application's frontend. It provides a dynamic and responsive interface for users to interact with the Document Management Portal, submit queries, and view the system's responses.

SpringBoot and FastAPI (Backend Frameworks)

- These two frameworks will handle the backend logic of the application. SpringBoot will manage the overall system architecture, while FastAPI will handle interactions with the document processing pipeline and databases. This combination ensures a robust and scalable backend system.

3.5.2 Techniques

Semantic Search via Vector Embeddings

- Text extracted from documents is broken down into smaller, fixed-size chunks, which are then converted into vector embeddings. These embeddings represent the semantic meaning of the text and are stored in Milvus. When a user query is entered, it is also converted into a vector embedding, allowing for efficient semantic search through distance metrics like cosine similarity.

Knowledge Graph Construction

- Named Entity Recognition (NER) and Relation Extraction techniques are applied to the document text to identify key entities (nodes) and the relationships between them (edges). The APOC library helps automate the creation of nodes and relationships in the Neo4j database. This process builds a rich, interconnected graph that serves as a structured context for queries.

Hybrid Retrieval Process

- When a query is input by the user, the system retrieves two types of context: one from the vector database through semantic search and another from the knowledge graph through graph traversal. The hybrid approach ensures both structured and unstructured data are considered, enhancing the relevance of the answer generated by the LLM.

Glossary Integration for Domain-Specific Understanding

- The system maintains a domain-specific glossary. When a query contains terms from the glossary, the system fetches their meanings and provides this additional context to the LLM. This ensures that the LLM can handle technical or specialized terms correctly, producing more accurate and informed responses.

LLM Prompt Engineering with Temperature Control

- The LLM is instructed via prompts to only generate an answer based on the provided context. If the context is insufficient, it will inform the user that the question cannot be answered. Additionally, the temperature parameter is set to zero, minimizing the likelihood of hallucinations or irrelevant answers from the LLM.

4. Project Timeline

The successful completion of this research project relies on a well-structured timeline that ensures all key tasks are systematically addressed. The timeline is designed to guide the project from the initial stages of topic selection and literature review through to the final stages of thesis submission and presentation. Each phase of the project is carefully planned to allow adequate time for research, system development, testing, and analysis. This structured approach not only ensures the timely completion of each task but also facilitates continuous progress monitoring, enabling any necessary adjustments to be made to keep the project on track. The following timeline diagram outlines the key milestones and tasks that will be undertaken over the course of the project, providing a clear roadmap to achieving the research objectives.

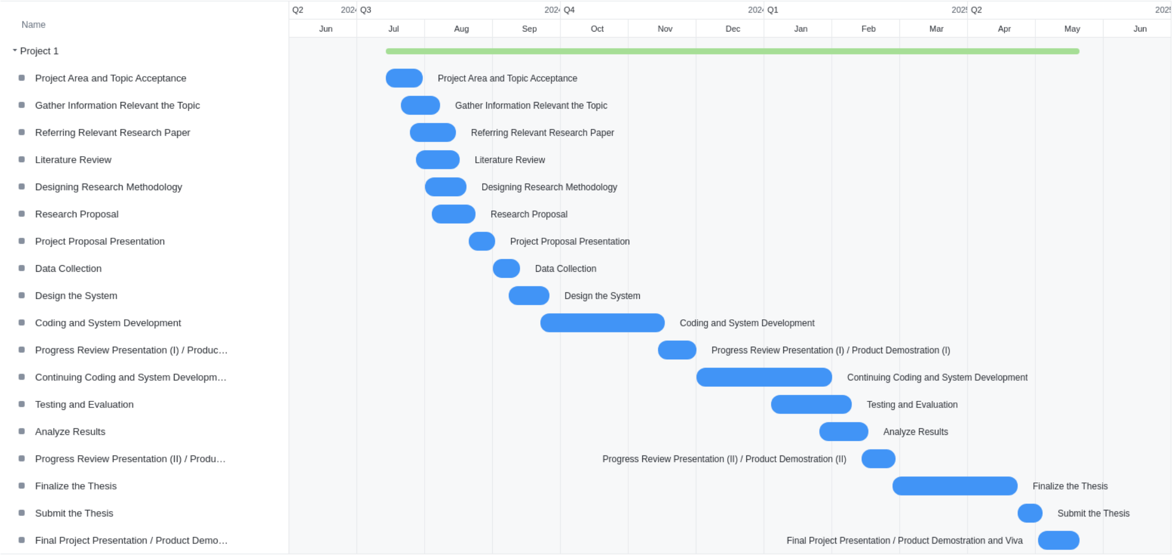

Figure 1: Timeline

5. System Design Overview

The system design for this project is centered around a multi-tier architecture that integrates a web-based frontend, a robust backend, and a graph database. The frontend, developed using React, provides an intuitive interface for users to interact with the system, submit queries, and manage documents. The backend, built with SpringBoot and FastAPI, handles the processing of user requests, manages the knowledge graph stored in Neo4j, and interfaces with LangChain and Llama 3.1 to generate responses. The knowledge graph is the core component, where text chunks from documents are stored as nodes, relationships are established, and vector indexing is applied for similarity searches. This architecture ensures efficient data retrieval, secure document management, and seamless integration of advanced AI techniques for natural language processing.

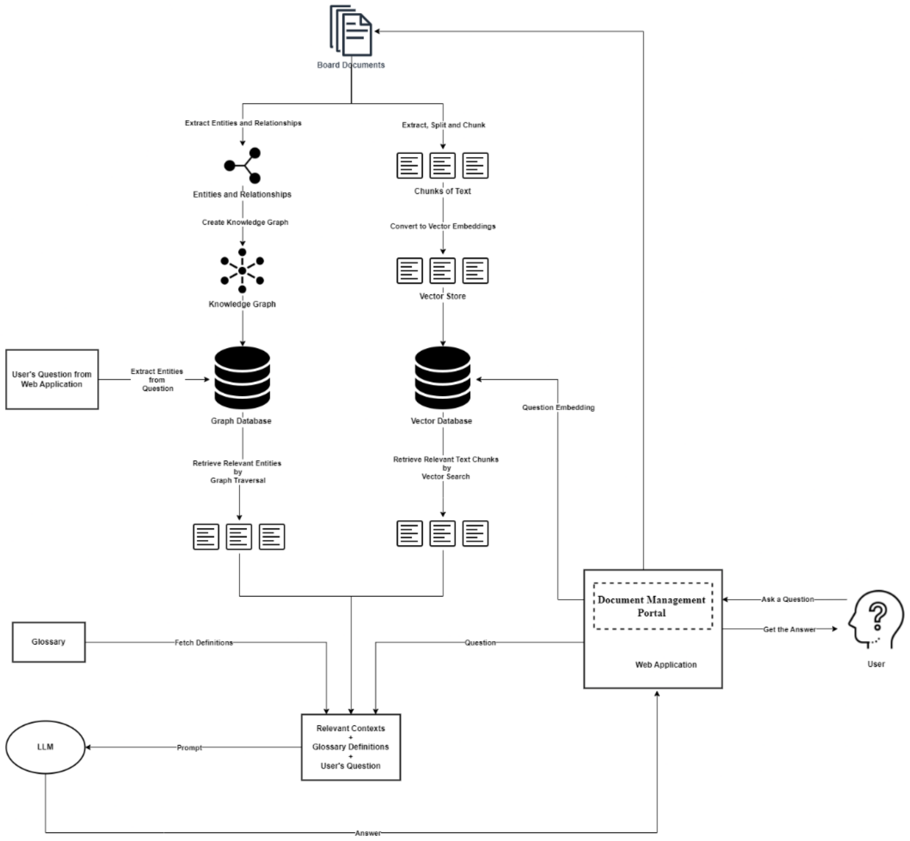
Board Documents
Extract Entities and Relationships
Extract, Split and Chunk
Entities and Relationships
Chunks of Text
Create Knowledge Graph
Convert to Vector Embeddings
Knowledge Graph
Vector Store
User's Question from Web Application
Extract Entities from Question
Graph Database
Vector Database
Question Embedding
Retrieve Relevant Entities by Graph Traversal
Retrieve Relevant Text Chunks by Vector Search
Document Management Portal
Web Application
Ask a Question
Get the Answer
User
Glossary
Fetch Definitions
Question
LLM
Prompt
Relevant Contexts + Glossary Definitions + User's Question
Answer

Figure 2: System Design

6. References

Banerjee, S., Sahoo, A., Layek, S., Dutta, A., Hazra, R., & Mukherjee, A. (2024). Context Matters: Pushing the Boundaries of Open-Ended Answer Generation with Graph-Structured Knowledge Context. *arXiv preprint arXiv:2401.12671v2*.

Edwards, C. (2024). Hybrid Context Retrieval Augmented Generation Pipeline: LLM-Augmented Knowledge Graphs and Vector Database for Accreditation Reporting Assistance. *ICS 699: MS Plan B Capstone Report, University of Hawaii at Manoa*.

Hu, Y., Zou, F., Han, J., Sun, X., & Wang, Y. (2024). LLM-TIKG: Threat intelligence knowledge graph construction utilizing large language model. *Computers & Security*. https://doi.org/10.1016/j.cose.2024.103999

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. *arXiv preprint*. https://arxiv.org/abs/2408.04948v1

Setty, S., Thakkar, H., Lee, A., Chung, E., & Vidra, N. (2024). Improving Retrieval for RAG based Question Answering Models on Financial Documents. *arXiv preprint arXiv:2404.07221v2*.

Zhao, H., Pan, Y., & Yang, F. (2020). Research on Information Extraction of Technical Documents and Construction of Domain Knowledge Graph. *IEEE Access*, 8, 168087-168098. https://doi.org/10.1109/ACCESS.2020.3024070