University of Sri Jayerardenepura

Research proposal

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

Group 15

Meet Our Team ↘

Kavinda Maduranga

WEERASINGHA W.G.K.M.
ICT/20/956

Nipuni Nishadini

De Silva K.N.N.C.
ICT/20/826

Dulan Jayawikrama

Jayawickrama D.S.K.
ICT/20/862

Our Supervicers

↘

Main Supervisor

Dr. Chamara Liyanage
Academic Supervisor,
Department of ICT,
University of Sri Jayewardenepura

External Supervisor

Mr. Hiran Wijesingha
Assistant Director IT,
Sri Lanka Tea Board

Content Outline ↘

Introduction & Background

Many organizations with longstanding histories in specific fields rely on physical files to store legal documents for decision-making. This manual process is inefficient and risks losing valuable institutional knowledge when experts retire or leave.

AI Powered System

Development of an agent system for interacting domain specific knowledge.

Hybrid RAG Model

Combined retrieval and augmented generation with knowledge graphs and vector databases.

Web Portal

Includes a user-friendly document management portal for easy interaction.

Domain-specific Glossary

Industry-specific terms will be integrated to improve the accuracy of the AI-generated response.

Motivation To Study

↘

Collaboration with the Sri Lankan Tea Board

Manual Processing

Modernization Need

Knowledge Loss

Preservation of Knowledge

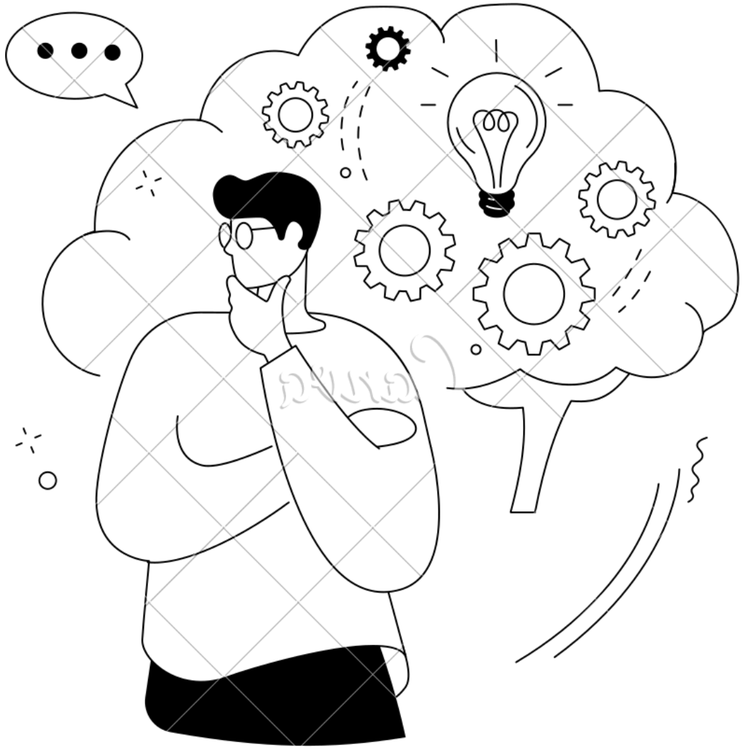
Inefficient Retrieval

Sustainability

The Sri Lanka Tea Board's historical legal documents, containing key decisions and regulations since 1976, are important for decision-making. However, relying on physical files and consultations with retiring experts risks losing critical knowledge. This makes it hard for current staff to access necessary information. The Tea Board has agreed to provide these legal documents, enabling us to continue our research and develop a system that preserves and digitizes this knowledge for future use.

Research Problem

- Organizations with long histories struggle to manage and retrieve critical information from large archives of historical documents, which are essential for decision-making.
- Many still rely on outdated, manual methods, increasing the risk of losing valuable domain expertise as experienced personnel retire.
- This leaves newer employees without the deep knowledge needed for effective decisions. The lack of an efficient knowledge transfer process worsens the issue, creating a knowledge gap.
- An automated system is urgently needed to manage, retrieve, and contextualize this data, ensuring that institutional knowledge is preserved and easily accessible for future decision-making.



Objectives & Outcomes ↘

Main objective

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.

- Web-based document management portal
- LLMs integration for context-aware responses
- Structured knowledge base from historical document
- Domain-specific glossary
- User-friendly interface
- Systems future needs (Scalability and adaptability)

Literature Review ↘

Hybrid RAG systems that combine knowledge graphs and vector-based retrieval offer enhanced information extraction from unstructured documents. These systems improve performance by leveraging both structured and unstructured data. Our research advances this by integrating a domain-specific dictionary, ensuring more accurate and contextually relevant retrieval for diverse document types.

Challenge	Difficulty in managing unstructured documents (e.g., scanned PDFs).
Solution	Hybrid RAG systems enhance information extraction
Method	Combination of knowledge graphs and vector-based retrieval.
Advantage	Improved performance for structured and unstructured data.
Innovation	Hybrid RAG systems enhance information extraction using domain-specific glossary
Objective	Address existing gaps and enhance contextually relevant responses.

Methodology ↘

1 Data Collection and Preprocessing

- Document Management Portal
- Text Extraction

2 Knowledge Graph and Vector Store Creation

- Vector store
- Knowledge Graph Construction
- Automation

3 Hybrid Context Retrieval

- Dual Context Extraction (Semantic Search and Graph Traversal)
- Hybrid Retrieval

4 Integration with LLM and Glossary

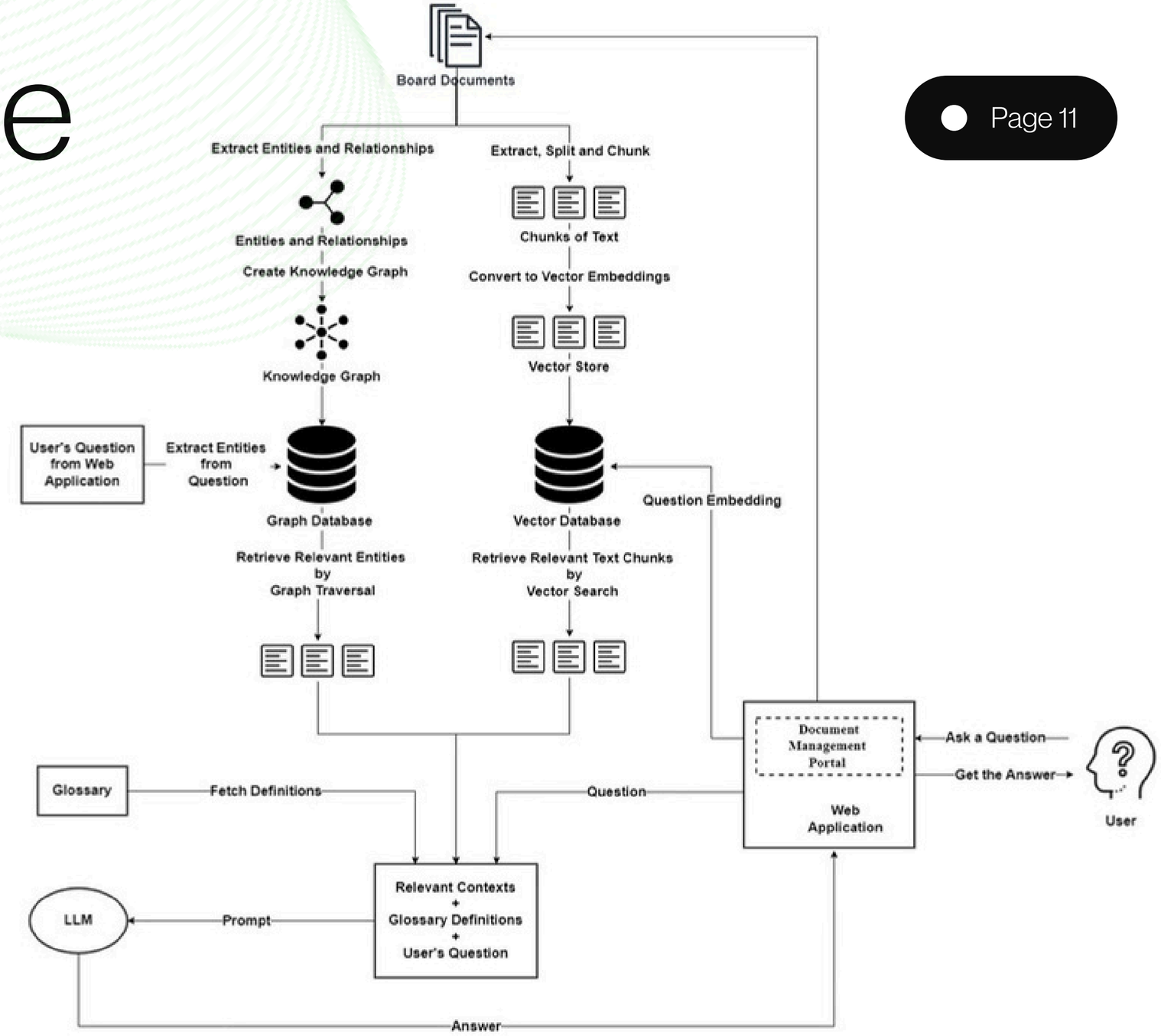
- LLM Integration
- Glossary Integration

5 Answer Generation and Delivery

- LLM Processing
- Frontend Delivery

Architecture Design

↘



Technologies ↘

neo4j

Graph database management system designed to store and query connected data efficiently.

Milvus

Vector database for managing unstructured data, often used in AI and machine learning applications.

LangChain

A framework for building applications that integrate LLMs with external data sources.

Llama 3.1

Large language model developed by Meta for various natural language processing tasks.

React

Popular JavaScript library for building interactive user interfaces, especially for web applications.

spring boot

Java-based framework for creating standalone, production-grade Spring applications with minimal configuration.

FastAPI

Modern, high-performance web framework for building APIs with Python, based on standard Python type hints.

Techniques ↘

1. Semantic Search via Vector Embeddings

Semantic search transforms text and queries into vector embeddings, compares them using similarity metrics, and retrieves relevant results efficiently from Milvus.

2.Knowledge Graph Construction

NER and Relation Extraction identify key entities and relationships in text, creating nodes and edges in Neo4j with the help of the APOC library, building an interconnected graph for structured query context.

3.Hybrid Retrieval Process

The system retrieves context from both semantic search in a vector database and graph traversal in a knowledge graph, combining structured and unstructured data to improve the relevance of LLM-generated answers.

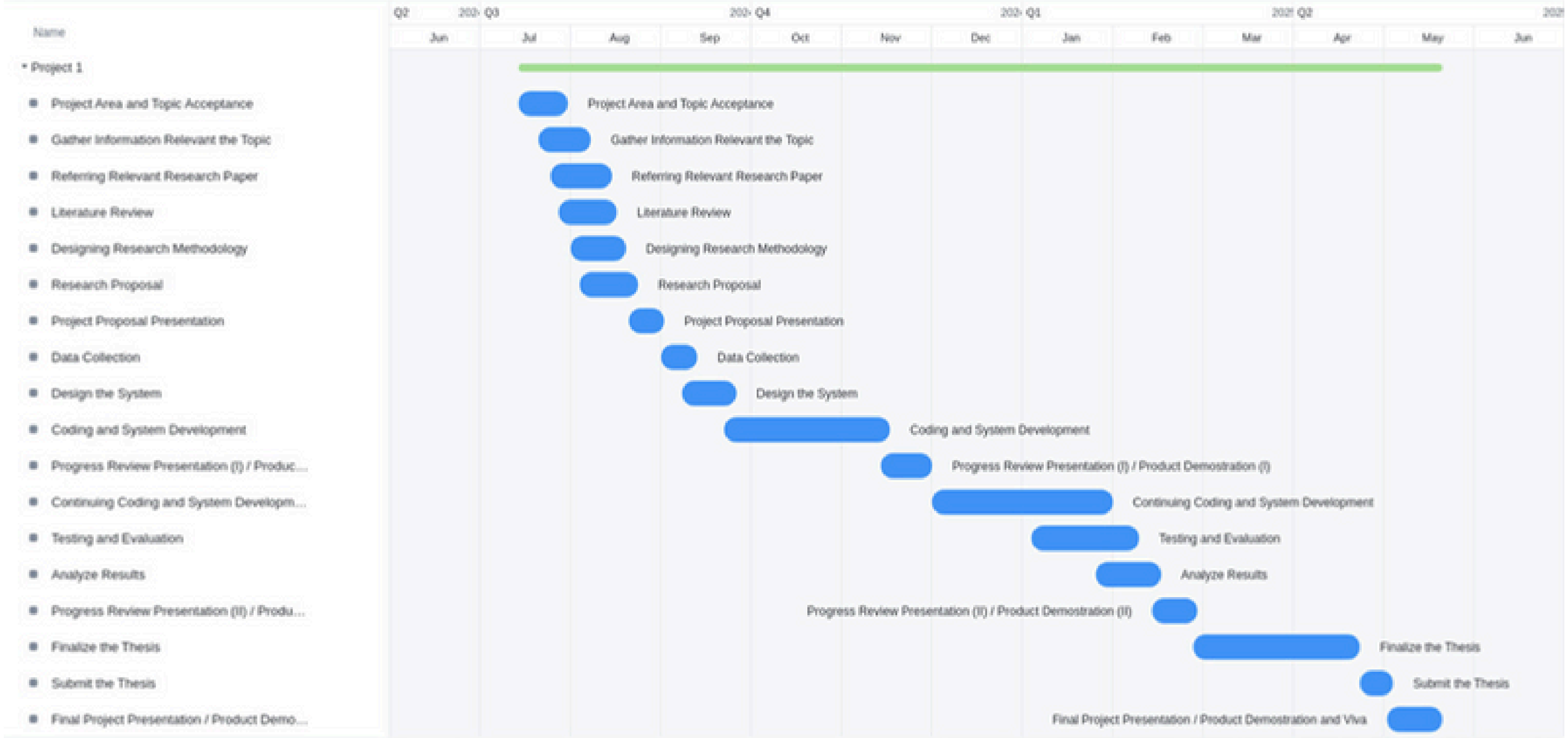
4.Glossary Integration for Domain-Specific Understanding

The system includes a domain-specific glossary that provides meanings for specialized terms in user queries, enhancing the LLM's ability to generate accurate and informed responses.

5.LLM Prompt Engineering with Temperature Control

The LLM is prompted to generate answers solely based on the provided context, informing users when context is insufficient, while a temperature setting of zero reduces the risk of hallucinations or irrelevant responses.

Timeline


Name
Project 1
Project Area and Topic Acceptance
Gather Information Relevant the Topic
Referring Relevant Research Paper
Literature Review
Designing Research Methodology
Research Proposal
Project Proposal Presentation
Data Collection
Design the System
Coding and System Development
Progress Review Presentation (I) / Produc...
Continuing Coding and System Developm...
Testing and Evaluation
Analyze Results
Progress Review Presentation (II) / Produ...
Finalize the Thesis
Submit the Thesis
Final Project Presentation / Product Demo...
Q2
Q3
Q4
Q1
Q2
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Project Area and Topic Acceptance
Gather Information Relevant the Topic
Referring Relevant Research Paper
Literature Review
Designing Research Methodology
Research Proposal
Project Proposal Presentation
Data Collection
Design the System
Coding and System Development
Progress Review Presentation (I) / Product Demostration (I)
Continuing Coding and System Development
Testing and Evaluation
Analyze Results
Progress Review Presentation (II) / Product Demostration (II)
Finalize the Thesis
Submit the Thesis
Final Project Presentation / Product Demostration and Viva

Have Questions?

Thank You!