Objectives and Scope

Objectives

The primary objective of this dissertation is to systematically evaluate whether an explicit **Knowledge Graph (KG)–based representation** improves the **factual completeness and correctness** of enterprise question answering systems when compared to **vector-only Retrieval-Augmented Generation (RAG)** approaches.

More specifically, the project aims to:

- Model enterprise data using a **structured Knowledge Graph** that captures:
  - Entities such as employees, positions, departments, and projects
  - Explicit **relationships and hierarchies** (e.g., reporting structure, role transitions)
  - **Temporal validity** of relationships (e.g., role held during a time interval)
- Establish a **baseline vector-only RAG pipeline** using dense embeddings and similarity search
- Compare both approaches on **constraint-based and multi-hop queries** that are common in enterprise analytics
- Identify **failure modes of vector-based retrieval** in handling structured constraints and relational dependencies

Scope

The scope of this work is limited to:

- **Synthetic but realistic HR and enterprise datasets**, chosen to ensure reproducibility and controlled experimentation
- Offline evaluation using **deterministic ground truth** derived from the Knowledge Graph
- Focus on **retrieval and grounding quality**, not on LLM fine-tuning or generation quality

The project does **not** aim to:

- Propose new large language models
- Optimize vector databases for scale or latency
- Replace existing RAG systems, but rather to **augment and analyse their limitations.**

Methodology Adopted

Overall Methodology

The methodology follows a **comparative, evaluation-driven approach**, consisting of four main stages:

---

1. Dataset Selection and Preparation

- Three synthetic HR datasets were selected from Kaggle and GitHub
- Datasets include employee records, departments, roles, reporting structure, salary, and employment timelines
- Data was preprocessed into:
  - **Structured CSV data** for graph construction
  - **Textual representations** for vector-based retrieval

Dataset sizes range from **300 to ~3000 rows**, enabling manageable yet realistic experiments.

---

2. Baseline Vector-Only RAG Pipeline

- Textual data is chunked and embedded using a sentence-transformer model
- An **in-memory FAISS vector store** is used for similarity-based retrieval
- Top-*k* chunks are retrieved and treated as context for answering queries
- This setup represents a **standard enterprise RAG baseline**

---

3. Knowledge Graph Construction

- A **Neo4j graph database** is used to construct the Knowledge Graph
- Explicit schema is defined with nodes such as:
  - Person, Position, Department
- Relationships encode:
  - Role transitions (`HELD_POSITION`)
  - Organizational hierarchy (`REPORTS_TO`)
  - Department membership (`WORKS_IN`)
- Temporal attributes (`from`, `to`) are stored on relationships to enable time-aware reasoning

This graph serves as a **ground-truth knowledge representation**.

---

4. Comparative Evaluation

- The same queries are executed against:
  - Vector-only RAG
  - Knowledge Graph–based retrieval
- Queries include:
  - Range constraints (e.g., salary filters)
  - Multi-hop relations (e.g., manager of an employee)
  - Time-bounded role queries
- Outputs are compared against **ground truth derived from the KG**

Progress and Key Findings

Current Progress

- Successfully constructed a **temporal enterprise Knowledge Graph** in Neo4j
- Implemented a **vector-only RAG baseline** using FAISS and open-source embeddings
- Defined and executed representative enterprise queries on both systems
- Established a reproducible evaluation setup using deterministic graph queries as ground truth

---

Key Observations

- Vector-only RAG performs reasonably well for:
  - Simple, single-entity lookup queries
- However, it shows consistent limitations for:
  - Queries requiring **multiple relational hops**
  - Queries involving **numerical or temporal constraints**
  - Queries where relevant information is **distributed across multiple documents**
- Knowledge Graph–based retrieval:
  - Preserves **structural and hierarchical relationships**
  - Ensures **factual completeness** when answering constrained queries
  - Avoids partial or misleading answers caused by top-*k* chunk truncation

These findings confirm that **representation choice significantly impacts answer reliability**, independent of the language model used.