CS-E4780 Scalable Systems and Data Management Course Project Enhancing LLM Inference with GraphRAG

Submission deadline: 28.11.2025

Responsible teacher: Prof. Bo Zhao
Staff: Dr. Tuo Shi, Cong Yu, Zhongxuan Xie, Jiaxin Guo, Mustapha Abdullahi, Zheyue Tan
cs-e4780@aalto.fi

1 Background

Recent advances in retrieval-augmented generation (RAG) have shown the effectiveness of combining large language models (LLMs) with external knowledge sources to improve factuality and reduce hallucinations. Traditional RAG systems rely primarily on text-based retrieval over unstructured corpora, where documents are chunked and indexed for semantic similarity search. While this paradigm works well for many knowledge-intensive tasks, it struggles with capturing the rich relational structure that underlies complex domains such as scientific literature, enterprise knowledge bases, or multi-hop reasoning problems. In these settings, relevant information often depends not just on local text passages but also on how entities, concepts, and events are connected across documents.

To address this gap, GraphRAG extends the RAG framework by integrating graph representations of knowledge into the retrieval process[2]. Instead of treating text chunks as isolated units, GraphRAG encodes entities and their relations into a knowledge graph, enabling retrieval along semantic paths that reflect contextual and structural dependencies. This graph-enhanced retrieval allows LLMs to perform reasoning over multi-hop relationships, capture higher-order semantics, and provide more grounded and coherent answers[3]. By bridging unstructured text with structured graph knowledge, GraphRAG has emerged as a promising paradigm for tasks requiring deep contextualization, such as question answering, dialogue systems, and domain-specific information extraction.

2 Data Set: Nobel Laureates

We use a dataset of Nobel laureates and their mentorship relationships, which after some preprocessing yields a nested JSON structure encoding parent-child links between scholars and their mentors. Each entry contains only minimal information: a name, a type label ("laureate" if they won a Nobel Prize, or "scholar" otherwise), and, when applicable, the category and year of the award. The coverage is limited to four prize categories—Physics, Chemistry, Medicine, and Economics—with all Nobel awards between 1901 and 2021 represented. Notably, the mentorship network extends far beyond Nobel-era laureates, tracing scholarly lineages back to the 16th century, including figures such as Galileo and Newton, which makes it possible to explore long-term historical patterns of knowledge transmission.

To enrich this primary dataset, we integrate information from the official Nobel Prize API. The API provides structured, high-quality metadata about laureates, including biographical details, birth and death dates, institutional affiliations, and prize motivations. Once this is done, the resulting knowledge graph in Kuzu looks something like this Figure 2. This enrichment greatly expands the scope of analysis, enabling queries such as: How many Physics laureates were born in the United States but later affiliated with the University of Cambridge? Who were the mentors of female Chemistry laureates? Which laureates were mentored by other laureates outside of U.S. institutions? The resulting graph is significantly more informative, combining deep historical mentorship data with authoritative Nobel Prize records, and offers a powerful resource for studying both individual trajectories and the evolution of scientific communities.

3 Project Requirements and Problem Definition

This course project is designed to help students become familiar with building a graph-based Retrieval-Augmented Generation (RAG) system using the Nobel Laureates dataset with Kuzu. The starting codebase is [1]. Please follow the README to run the scripts. As a first step, students are expected to run and examine the provided demo workflow to gain an understanding of the different modules and their interactions. The project is organized into the following tasks:

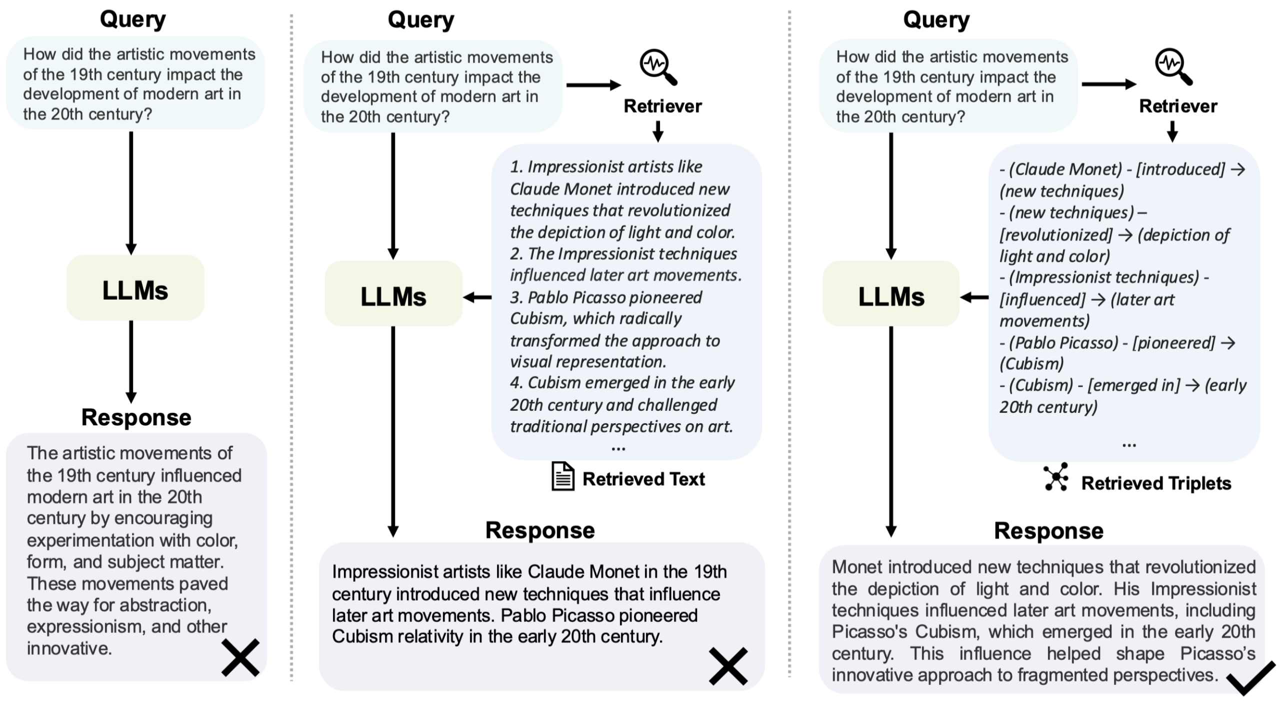
Task 1. Text2Cypher Improvement Students will extend the Text2Cypher component with the following enhancements:

- Dynamically select few-shot exemplars based on similarity to the input question.
- Implement a self-refinement loop: generate → validate (syntax check using a dry-run `EXPLAIN`) → repair if validation fails.
- Add a rule-based post-processor (e.g., enforce lowercase comparisons, ensure proper property projection).

Learning outcome: Gain hands-on experience with prompt engineering and iterative query generation.

Task 2. Caching & Performance Students will focus on improving system efficiency through:

- Adding an LRU cache to support pruning and caching of Text2Cypher results (keyed by question hash and schema hash).
- Benchmarking pipeline latency at the granularity of individual stages.
- Producing a flamegraph or simple timing breakdown to visualize performance.



How did the artistic movements of the 19th century impact the development of modern art in the 20th century?

The artistic movements of the 19th century influenced modern art in the 20th century by encouraging experimentation with color, form, and subject matter. These movements paved the way for abstraction, expressionism, and other innovative.

1. Impressionist artists like Claude Monet introduced new techniques that revolutionized the depiction of light and color.
2. The Impressionist techniques influenced later art movements.
3. Pablo Picasso pioneered Cubism, which radically transformed the approach to visual representation.
4. Cubism emerged in the early 20th century and challenged traditional perspectives on art.
...

Impressionist artists like Claude Monet in the 19th century introduced new techniques that influence later art movements. Pablo Picasso pioneered Cubism relativity in the early 20th century.

- (Claude Monet) - [introduced] → (new techniques)
- (new techniques) – [revolutionized] → (depiction of light and color)
- (Impressionist techniques) - [influenced] → (later art movements)
- (Pablo Picasso) - [pioneered] → (Cubism)
- (Cubism) - [emerged in] → (early 20th century)
...

Monet introduced new techniques that revolutionized the depiction of light and color. His Impressionist techniques influenced later art movements, including Picasso's Cubism, which emerged in the early 20th century. This influence helped shape Picasso's innovative approach to fragmented perspectives.

Figure 1: Comparision between Direct LLM, RAG, and GraphRAG. Given a user query, direct answering by LLMs may suffer from shallow responses or lack of specificity. RAG addresses this by retrieving relevant textual information, somewhat alleviating the issue. However, due to the text's length and flexible natural language expressions of entity relationships, RAG struggles to emphasize "influence" relations, which is the core of the question. While, GraphRAG methods leverage explicit entity and relationship representations in graph data, enabling precise answers by retrieving relevant structured information.

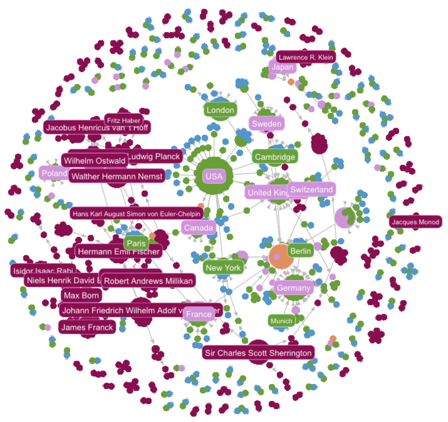


Figure 2: Graph data in Kuzu

Learning outcome: Understand the end-to-end GraphRAG pipeline. Develope the ability to enhance the LLM inference with knowledge graphs. Learn the optimization techniques in GraphRAG systems over real-world datasets.

4 Submission Requirements

Students must implement a system to solve the problems described in Section 3. Each student is required to submit a link to the system's GitHub repository (ensure the repository is set to public visibility) along with a 4 page report detailing the system. For group projects, students must clearly specify the contributions of each member. The detailed requirements for the project report are as follows:

4.1 Project Report Requirements

Template. Use the ACM Proceeding Template: LaTeX[1] or Word [2].
Report Organization. The report should be 4 pages and include the following sections. Students may use their own section titles, but the content should align with the following structure:

(1) **Abstract.**
(2) **Introduction.** Provide a brief background and overview of the project.

[1]https://www.overleaf.com/latex/templates/acm-hypertext-conference-template/pchbkqfnmxgr
[2]https://www.acm.org/binaries/content/assets/publications/word_style/interim-template-style/interim-layout.docx

(3) **System Architecture.** Describe the design choices and motivations in detail.
(4) **Implementation.** Explain how the project was implemented, including the hardware/cloud services, programming tools, and GUI tools used.
(5) **Evaluation.** For **Task 1**, students should report the chosen algorithmic designs and the resulting accuracy improvements. For **Task 2**, students should describe the data structures used, quantify the performance speed-ups achieved, and present the execution times of each stage in the graph RAG pipeline.
(6) **Conclusion.**

References

[1] benyucong. 2025. CS-E4780-project2. https://github.com/benyucong/CS-E4780-project2. GitHub repository.

[2] Haoyu Han, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, Mahantesh Halappanavar, Ryan A Rossi, Subhabrata Mukherjee, Xianfeng Tang, et al. 2024. Retrieval-augmented generation with graphs (graphrag). *arXiv preprint arXiv:2501.00309* (2024).

[3] Cong Yu, Tuo Shi, Matthias Weidlich, and Bo Zhao. 2025. SHARP: Shared State Reduction for Efficient Matching of Sequential Patterns. *arXiv preprint arXiv:2507.04872* (2025).