Project Design Document

Graph-Aware Language Intelligence System for Research Synthesis

Remigiusz Sek
Kacper Gutowski
Dawid Koterwas
Andrii Norets

30 June 2025

Mentors: Esther Ruano Hortoneda

Contents

1 Introduction **3**
- 1.1 Problem statement 3
- 1.2 Project objective 3

2 Key Features **3**

3 Low-Level Design **3**

4 Technologies **4**
- 4.1 Programming languages 4
- 4.2 Frameworks and libraries 5
- 4.3 Tools and services 5

5 What is your solution? Is it software as a service or a product? **5**

6 What is the industry or domain under which your solution falls? **6**

7 Who is your target audience? **6**

8 Who are your competitors? **6**

9 What is your added value? **7**

1 Introduction

1.1 Problem statement

Writing the related work and references sections for research papers within the field of computer science is a notoriously time-consuming and challenging task for researchers. This process demands not only summarizing individual computer science publications but also understanding and articulating the complex and nuanced relationships specific to this discipline. While current Large Language Models (LLMs) are capable of generating text, they often fail to capture the deep relational structures inherent in CS literature, resulting in literature reviews that may lack coherence and depth.

1.2 Project objective

Our project focuses on the domain of computer science research by taking advantage of an existing semantic graph as an intermediate resource. This semantic graph structurally captures the relationships between a main (citing) paper, the works it references, and the core concepts they encompass. The primary objective is to deliver a practical tool, in the form of a web-based application embedding the semantic graph and model, that will directly improve researchers' workflow by helping them generate high-quality, coherent related work and references sections for CS papers. This tool will simplify and accelerate the complex process of synthesizing literature, making it easier for users to produce accurate and insightful content.

To achieve this, we will investigate various methods for leveraging the existing semantic graph to guide language model outputs. After evaluating different approaches, we will select the most effective technique and integrate it within the tool to maximize its impact on the quality and relevance of generated text.

2 Key Features

- **Feature 1:** Context-aware References generation
- **Feature 2:** Automated Related Work drafting
- **Feature 3:** Semantic graph integration
- **Feature 4:** Interactive web-based platform
- **Feature 5:** Efficient and reliable literature synthesis

3 Low-Level Design

To be able to use the images in the document, you would need to save the diagrams as image files (e.g., 'high$_l$*evel.png*', '*evaluation*$_p$*ipeline.png*', '*internal*$_s$*ystem.png*')*andplacetheminthesa*

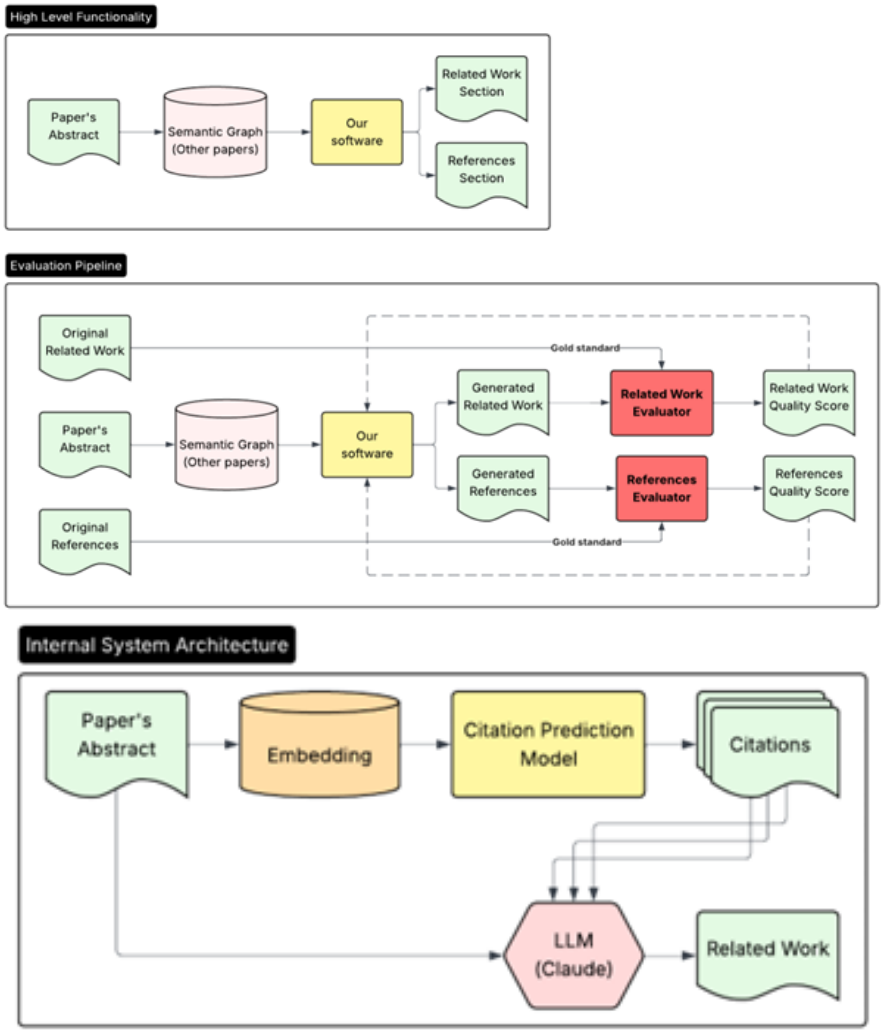

Figure 1: High Level Functionality

4 Technologies

4.1 Programming languages

- **Python:** Python will be used to develop the main application logic, including the data processing pipeline and integration. It enables seamless interaction with core libraries and frameworks necessary for semantic graph analysis.

4.2 Frameworks and libraries

- **PyTorch & PyTorch Geometric:** PyTorch - Machine Learning library for constructing and training Neural Networks; Pytorch Geometric - version of PyTorch specialized for data with graph structure. Both of them will be used to train Graph Neural Networks.
- **Scikit-learn:** Machine Learning library containing many popular tools for classification, regression, clustering, etc. It will be used for creating a baseline.
- **Hugging Face Transformers:** The library to easily use and train various models (BERT, GPT, etc.) from its vast ecosystem.
- **OGBN-ArXiv (Open Graph Benchmark):** Benchmark dataset representing a citation network of computer science papers from arXiv. It provides node features, citation edges, and research area labels. It will be used for training and evaluation.
- **LangChain:** A framework for building language model applications by chaining prompts and integrating external knowledge. It facilitates interaction between embeddings, semantic graph data, and the LLM that receives citations to generate the related work section.

4.3 Tools and services

- **MLflow:** a platform to manage the machine learning lifecycle, including experiment tracking, model versioning, and deployment. For GALIS, MLflow will be integrated to log experiments, track performance metrics, and manage models, supporting reproducibility and analysis of results.
- **Docker:** Platform for developing, shipping, and running applications in containers. In this project, Docker will be used to containerize the standard GNN pipeline, ensuring consistent environments across all development stages.
- **Embeddings and Claude from Amazon Bedrock:** An embedding model will be used to encode the abstract of the input paper into a vector representation. This embedding is passed to our model responsible for retrieving potentially cited papers. The abstracts of these candidate papers, along with the original abstract, are then provided to Claude, which generates the related work and references sections.

5 What is your solution? Is it software as a service or a product?

GALIS falls into the Software as a Product category, as it is delivered and accessed entirely through a web-based application. Our solution leverages a

semantic graph to map out the relationships between academic papers in the CS field and their key concepts. This graph serves as a structured foundation for generating clear and insightful related work and references sections. The user submits the abstract of their academic paper as input, and the system processes it using a semantic graph built from a curated dataset of computer science literature. Based on this analysis, GALIS generates two key outputs: a contextually relevant reference list and a structured draft of the Related Work section.

6 What is the industry or domain under which your solution falls?

GALIS is designed for professionals engaged in scholarly work and scientific development across both academic and industrial computer science research environments. Our approach streamlines and improves the creation of related work and references sections in CS research publications. It leverages a structured semantic graph that captures the relationships and core concepts shared among cited papers. This graph is used internally to guide the generation of a high-quality, contextually relevant related work and references sections, ensuring that citations are coherent and reflect meaningful connections within the existing body of work. The software is intended for academic researchers and R&D professionals in the CS field who require efficient, insightful, and accurate citation management and literature synthesis. The semantic graph plays a central role in supporting the automated generation of related work and references sections, resulting in more logically organized and context-aware citation lists.

7 Who is your target audience?

Researchers working in academic institutions who need to manage and generate high-quality related work and references for their CS publications efficiently.

- Professionals in R&D environments who rely on accurate and context-aware citation practices to support scientific and technical documentation.
- Individuals involved in scholarly work who benefit from structured, coherent, and insightful citation lists to enhance the quality of their research outputs.

8 Who are your competitors?

There are several tools that assist with managing and visualizing research references, such as Connected Papers, Litmaps, and ResearchRabbit. These platforms specialize in helping users explore citation networks and track connections between publications. However, unlike these visualization-focused solutions,

GALIS is designed specifically to automate and enhance the generation of related work and references sections in CS research publications. By leveraging a semantic graph built internally from the relationships among scholarly works, the software ensures related work and references are contextually relevant, logically organized, and tailored for both academic and enterprise R&D professionals. This focus on automating and improving the accuracy of the related work and references sections distinguishes our solution from existing literature mapping tools.

9 What is your added value?

The software adds value by automating and intelligently organizing related work and references sections in research publications. By exploiting a semantic graph to capture relationships between cited works, GALIS ensures references are accurate, logically grouped, and contextually relevant that which standard citation tools do not offer. Our approach saves time, reduces manual errors, and delivers more meaningful, well-structured reference lists and the related work section for both academic and R&D professionals.