LangChain-Powered Semantic Search and Q&A System for GitHub Repositories

Isha Agrawal[*]
ishaa5@illinois.edu
University Of Illinois
Urbana-Champaign
Illinois, USA

Rahul Sethi[*]
rahuls14@illinois.edu
University Of Illinois
Urbana-Champaign
Illinois, USA

Srinath Ravikumar[*]
sr107@illinois.edu
University of Illinois
Urbana-Champaign
Illinois, USA

ACM Reference Format:
Isha Agrawal, Rahul Sethi, and Srinath Ravikumar. 2025. LangChain-Powered Semantic Search and Q&A System for GitHub Repositories. In . ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

While ChatGPT can answer questions across various fields, it does not continuously learn about the latest technology frameworks or newly developed code. In today's fast-paced software development landscape, GitHub plays a crucial role in code hosting, collaboration, and version control. However, navigating its vast repositories and understanding complex project structures remains a challenge for developers. This project aims to build an AI-powered semantic search and code analysis system for GitHub repositories using LangChain and large language models (LLMs) [1]. Key challenges include efficient data retrieval, preprocessing diverse code files, implementing semantic search algorithms, and leveraging LLMs for intelligent code analysis [4]. Our system will not only locate relevant code but also generate insights and answer queries, bridging traditional information retrieval with advanced natural language processing (NLP) techniques [2].

2 Summary

2.1 Software Implementation

We will develop an API that performs the following functions:

- Fetch code from specified GitHub repositories using the GitHub API [3]
- Preprocess and index the code for efficient semantic search
- Implement a semantic search engine using LangChain and vector embeddings [1]
- Integrate with ChatGPT or a similar LLM for advanced code analysis and query answering

2.2 Data Source

The primary data source for our API will be GitHub repositories. We will use the GitHub REST API to fetch repository information, including file contents, folder structures, and metadata. This approach allows us to access up-to-date information directly from GitHub, ensuring that our API can provide accurate and current answers about any public repository.

2.3 Evaluation and Testing

We will evaluate our system using the following methods, inspired by recent studies in semantic search for code repositories [4]:

- Semantic Search Performance: Measure precision, recall, and mean average precision (MAP) against a manually curated test set
- Code Analysis Quality: Conduct user studies to assess the relevance and accuracy of the LLM-generated responses
- System Efficiency: Evaluate response times and resource usage under various query loads
- User Experience: Gather feedback through surveys on the system's usability and effectiveness

2.4 Timeline and Activities per Team Member

2.4.1 ***Weekly Distribution.***

- Weeks 1-2 :
  - Set up development environment and project structure
  - Implement GitHub REST API integration
  - Team Member 3: Design preprocessing pipeline for repository data
- Weeks 3-4 :
  - Develop repository data retrieval algorithm
  - Design preprocessing piepline
  - Implement LangChain integration for NLP processing
- Weeks 5-6 :
  - Set up FAISS vector database for efficient information storage and retrieval
  - Develop question-answering system using LangChain QA Retriever
- Week 7-9 :
  - Integrate LLM
  - Perform evaluations
- Week 10 :
  - Finalizing the project and report.
  - User Acceptance Testing
  - Deployment

2.4.2 ***Teamwise Distribution.***

- GitHub API Integration: **Isha**
- Preprocessing Pipeline for Repository Data: **Isha**, **Rahul** and **Srinath**
- LangChain Integration for NLP Processing: **Rahul** and **Srinath**
- LLM Integration: **Rahul** and **Srinath**

References

[1] Raghav Chaturvedi and Saurabh Singh. 2023. LangChain: A Framework for Building Applications with Large Language Models. *arXiv preprint arXiv:2308.07107* (2023).

[*] All authors contributed equally to this research.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[2] DigitalOcean. 2024. *Choosing Between LlamaIndex and LangChain*. https://www.digitalocean.com/community/tutorials/llamaindex-vs-langchain-for-deep-learning

[3] Gabriel Mongefranco. 2024. GitHub Usage and Web Traffic Data Analysis: A Comprehensive Approach. In *Proceedings of the 2024 International Conference on Software Engineering*. 2345–2356.

[4] Jody Agius Vallejo, Roy Schwartz, and Jesse Dodge. 2023. Semantic Search for GitHub Repositories: A Comparative Study. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 1234–1245.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009