Analyzing Open Source Software Ecosystems

Mentors: Dr. Alexey Tregubov, Dr. Jeremy Abramson, Dr. Jim Blythe

Students: Apoorv Dixit, Kai Zheng, Zishen Wei

Motivation

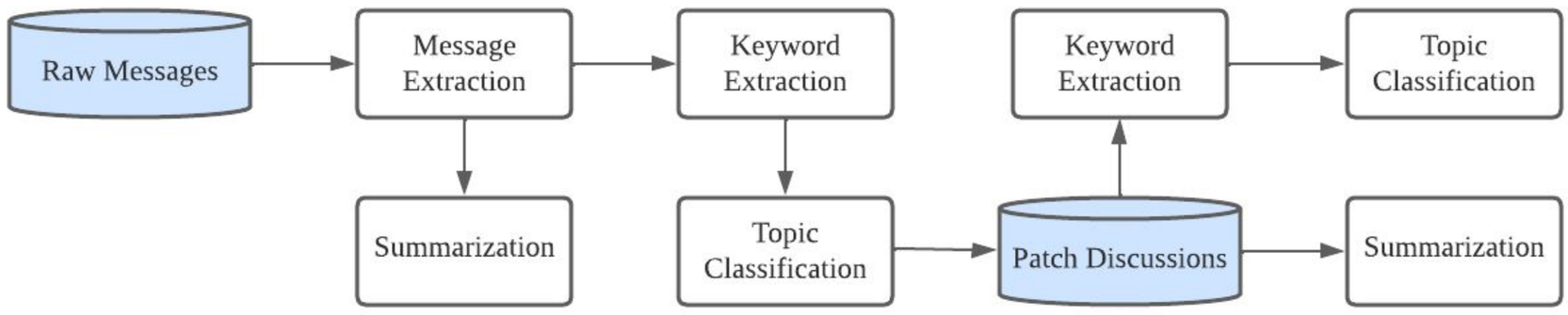
Open source runs a lot of the world's critical software systems, but there is much that's unknown in how maintainers, developers and other parts of the software ecosystem function. This project attempts to analyze code commits of open source software repositories, that includes both source code and patch conversations, to better understand them.

Goals

Gaining a deeper understanding of OSS ecosystems will enable the open source community to identify potential vulnerabilities, and define better development practices. The proposed solution intends to conduct malware analysis and extraction of code information as well as authorship styles.

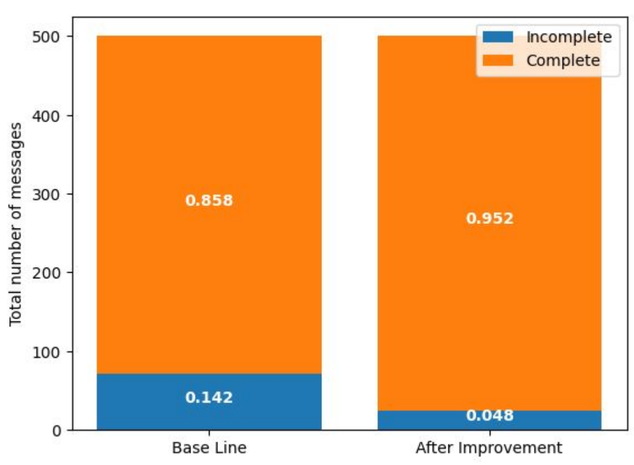
Methods

- Message Extraction on raw messages from *Linux Kernel Mailing List*
- Keyword Extraction, Summarization, and Topic Classification on *individual messages* and *patch discussions*.
- The project aims to leverage the power of large language models (LLM) to accomplish the aforementioned tasks. Various LLMs like Open AI GPT 4, GPT 3.5, Meta Llama 2, and Hugging Face models like KeyBERT have been considered.

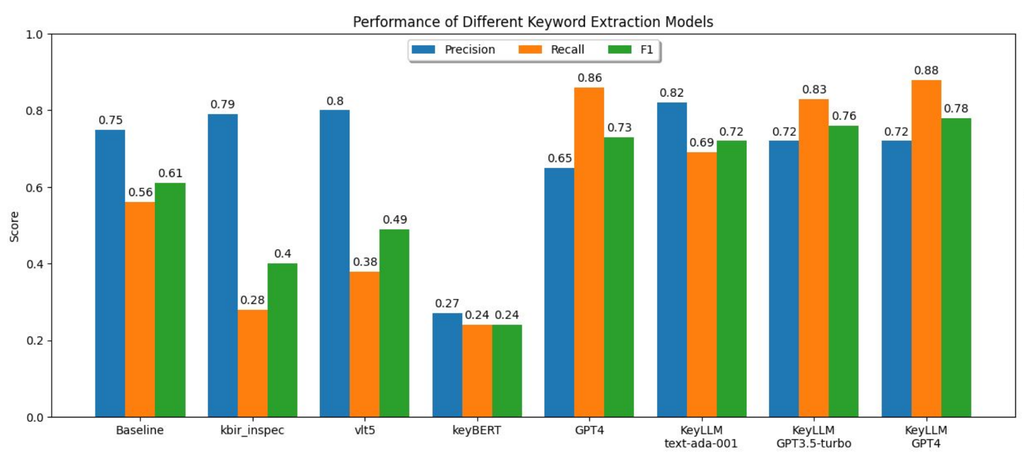


Results

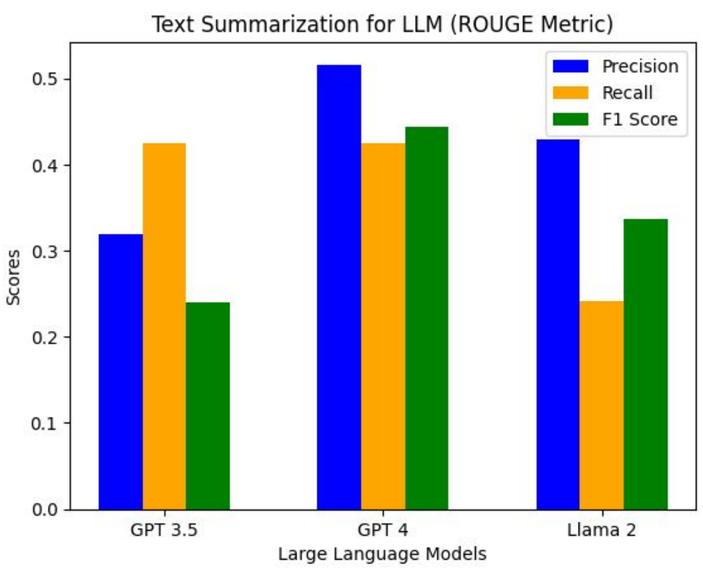
Message Extraction



Keyword Extraction



Summarization



Conclusions

- KeyLLM with GPT4 is the most proficient model at extracting relevant keywords.
- GPT4 generates the most concise summary for messages.

Future Work

- Train and optimize the selected LLM
- Combine messages into different patch conversations
- Perform keyword extraction and summarization on patch conversations