Sponser Information：

- Wu-chang Feng (wuchang@pdx.edu)
- Ameeta Agrawal (ameeta@pdx.edu)

Goal：

The goal of this project is to develop a Retrieval-Augmented Generative AI system (RAG) that supports conversations about local social services. The system will retrieve accurate, up-to-date information from external databases and use large language models (LLMs) to generate Natural language responses for users.

- Backend API Development: The primary task is to develop a REST API backend capable of handling user queries about social services. Use LLM to generate natural language responses based on user queries and information in the database.
- Sponsors need this agent to be available on multiple messaging platformsSponsors such as WhatsApp, Facebook Messenger, especially need to be available on Slack(multiple clients)
- Support comprehensive online social service data sources.（vector database）
- Need to support conversation, not just one question and one answer, need to support historical conversation records, and have the ability to associate context.

  Eventually we may need to use Docker to package all environments and components and upload them to the cloud, such as Google Cloud. Not Sure for rest api or clients？

previous group completed：

The previous team implemented an automatic gmail responder, built a vector database using PSU resources, used the RAG architecture to retrieve relevant content from the database through user queries and used LLM to generate replies. We will obtain the code of the program and continue to develop this project based on it.

Issues:

1. Only supports question and answer, Http get endpoint for everthing This leads to a question-and-answer situation and length limit for the URL. I'm not sure I understand it correctly（ 9-10 minutes of video）This time we are expected to use not only get, but also post
2. Gmail client only
3. PSU resources only

Components：

- Backend API
- Vector Database
- Langchain
- multiple one-off clients for messaging platforms

Website address of social service resources specially mentioned by the sponsor：

https://rosecityresource.streetroots.org/
https://centralcityconcern.org/ ##The sponsor specifically mentioned that the information on this website is more difficult to integrate and extract than the former

Concepts, tools, and frameworks may need ：

Retrieval Augmented Generative：
RAG is an AI framework for retrieving facts from an external knowledge base to ground large language models (LLMs) on the most accurate, up-to-date information and to give users insight into LLMs' generative process
https://research.ibm.com/blog/retrieval-augmented-generation-RAG

Vector Database：
A vector database indexes and stores vector embeddings for fast retrieval and similarity search, with capabilities like CRUD operations, metadata filtering, horizontal scaling, and serverless.
https://www.pinecone.io/learn/vector-database/

Rest API：
https://www.redhat.com/en/topics/api/what-is-a-rest-api

Langchain：
LangChain is an open source framework that lets software developers working with artificial intelligence (AI) and its machine learning subset combine large language models with other external components to develop LLM-powered applications.
https://www.techtarget.com/searchenterpriseai/definition/LangChain

Large language Model：
LLM such as Chatgpt, Gemini

Docker:
Docker's ability to package applications and their dependencies means that developers do not need to manually configure the operating system, language runtime, or dependent libraries
https://www.docker.com/