Project Proposal

Intelligent Email Parsing and Information Retrieval with RAGs

1. Use Case Description

The goal of this project is to build a **Personal Information Virtual Assistant (VA)** using a Retrieval-Augmented Generation (RAG) approach. The assistant will be designed to retrieve and summarize information from a user's personal emails. This will allow the user to do advanced retrieval form your personal emails.

Target Users:

- **Individuals** who need advanced retrieval and summarizing information based on email contexts
- **Professionals** who want to streamline daily tasks such as email review and information organization.

2. Document Collection & Dataset Creation

To mimic a user's personal document database, the dataset is curated from personal emails. Key steps include:

- Fetching all the emails in Gmail using Google takeout.
- Preprocessing and cleaning email content.
- Storing the processed emails in a structured text format for easy retrieval.
- Generating embeddings and storing them in a vector database.

3. Large Language Models (LLMs) Chosen

For the VA, I will experiment with three state-of-the-art LLMs, chosen based on their compatibility with the RAG pipeline and their ability to handle smaller-scale datasets effectively:

1. **GPT-4**: Renowned for its exceptional natural language understanding and response generation capabilities, making it suitable for complex information retrieval tasks.
2. **Gemini 1.5**: Known for its contextual awareness and efficiency in document retrieval, particularly effective for summarizing and analyzing email content.
3. **Llama 3.1**: A robust model optimized for performance and computational efficiency, making it ideal for lightweight systems.

All models will be accessed and tested using their respective APIs, ensuring seamless integration into the RAG pipeline.

4. Evaluation & Metrics

The VA will be evaluated based on its ability to retrieve accurate information and generate coherent, concise responses. The following metrics will be used for evaluation:

- **Human Evaluation**: The primary evaluation method, focusing on qualitative feedback from users. Responses will be scored on criteria such as relevance, accuracy, coherence, conciseness, and fluency, using a Likert scale (1-5).
- **User Satisfaction**: A qualitative measure based on user feedback collected via structured forms, assessing how effectively the VA meets their needs.
- **Retrieval Accuracy**: Evaluated indirectly through human judgment of how well the retrieved documents align with the query intent.
- **Task Completion Time**: Measured as the time taken for the VA to generate responses, ensuring it operates efficiently for practical use cases.

5. Project Timeline

- **Week 1-2**: Assemble and preprocess the document dataset.
- **Week 3-5**: Implement and test the RAG pipeline using Mistral 7B and Llama 2 7B.
- **Week 5**: Evaluate VA performance on 30+ questions.
- **Week 6**: Optimize the VA by improving retrieval and generation parameters.
- **Week 7**: Finalize the project report and prepare the presentation.