Dr. Babasaheb Ambedkar Technological University, Lonere

Prompt Engineering RAG Assistant

Using Google Gemini and Streamlit

B.Tech in Computer Engineering

Department of Computer Engineering
Dr. Babasaheb Ambedkar Technological University
Lonere, Raigad (MS)

Presented by:

- **Himani Thakur** (PRN: 23030331245060)
- **Siddhi Kavitake** (PRN: 24030331245508)

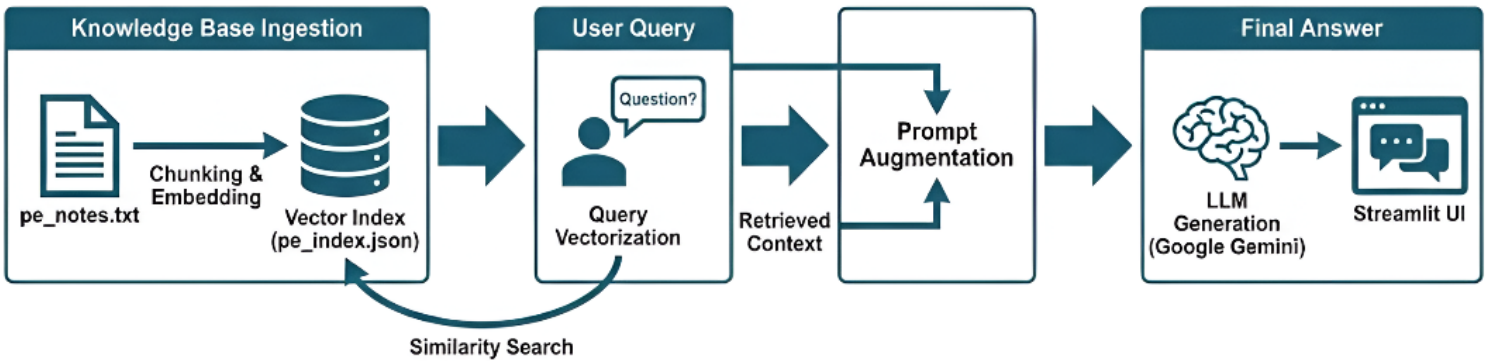
Introduction

- **Project Overview**
  - A chatbot designed to act as a university-level tutor.
  - Specifically tailored for **Prompt Engineering** concepts.
  - Helps students prepare for exams using a custom knowledge base.
- **Core Technology**
  - Built using **Retrieval-Augmented Generation (RAG)**.
  - Powered by **Google Gemini** models.
  - Interface developed in **Streamlit**.

Understanding RAG

- **Definition**
  - RAG stands for Retrieval-Augmented Generation.
  - It allows LLMs to reference an authoritative knowledge base outside of their static training data before generating a response.
- **The Problem it Solves**
  - Standard LLMs may "hallucinate" or provide outdated information.
  - RAG ensures answers are grounded in specific, domain-relevant documents.

Workflow Visualization



Project Implementation Details

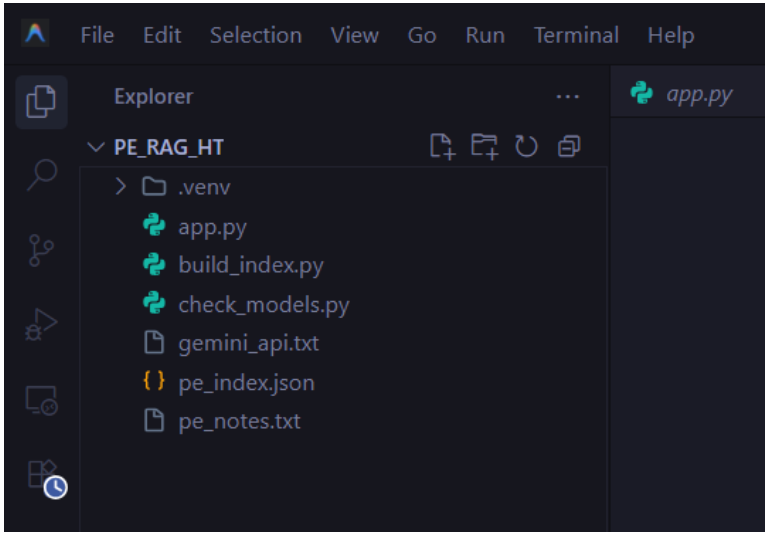
Models Used:

- **Embeddings:** `models/text-embedding-004`
- **Generation:** `models/gemini-2.0-flash`

Process:

1. **Prepare:** Notes are stored in `pe_notes.txt`.
2. **Index:** `build_index.py` vectorises notes into `pe_index.json`.
3. **Chat:** `app.py` retrieves context and generates answers via Streamlit.

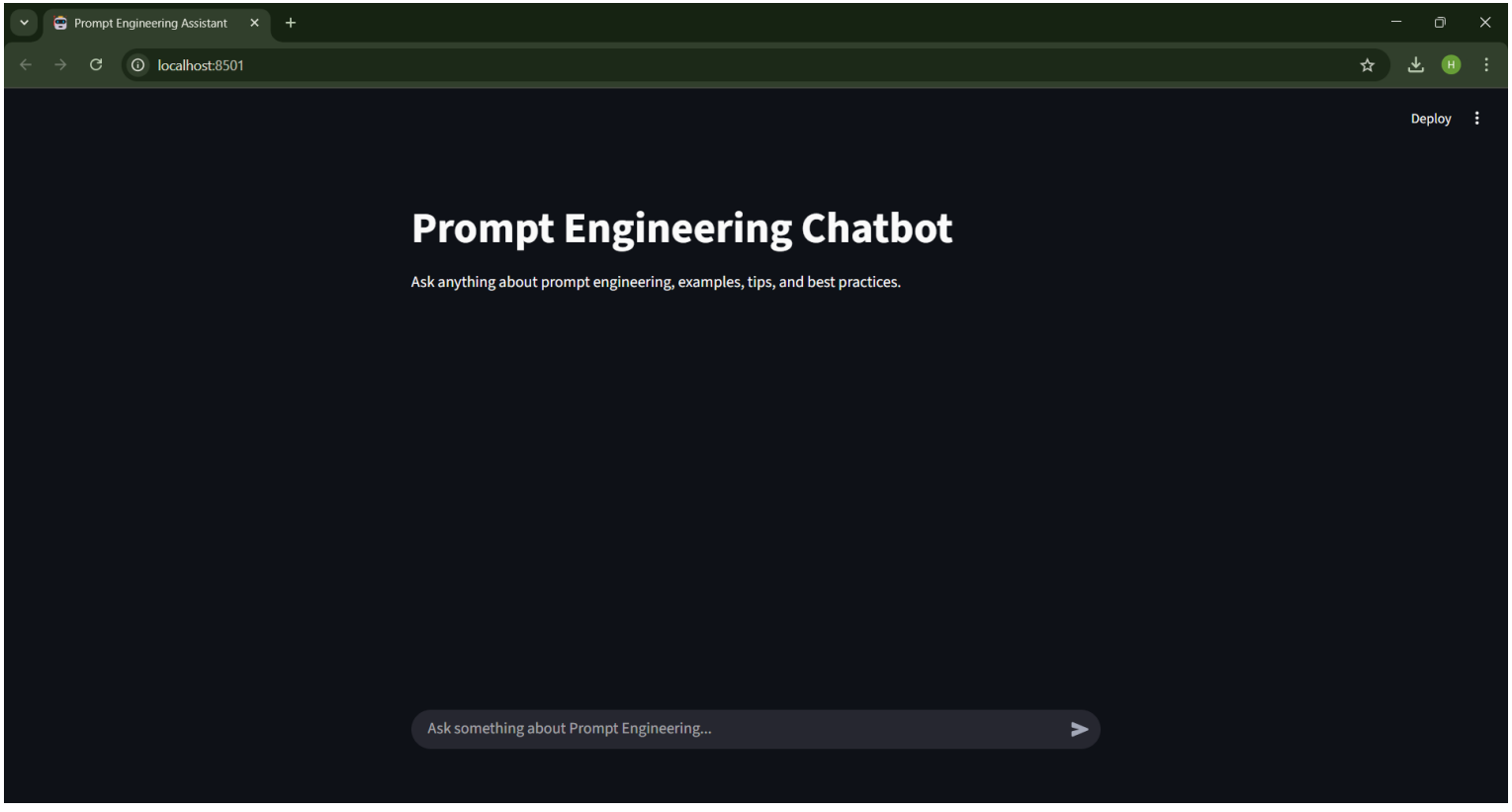
File Structure



Key Files:

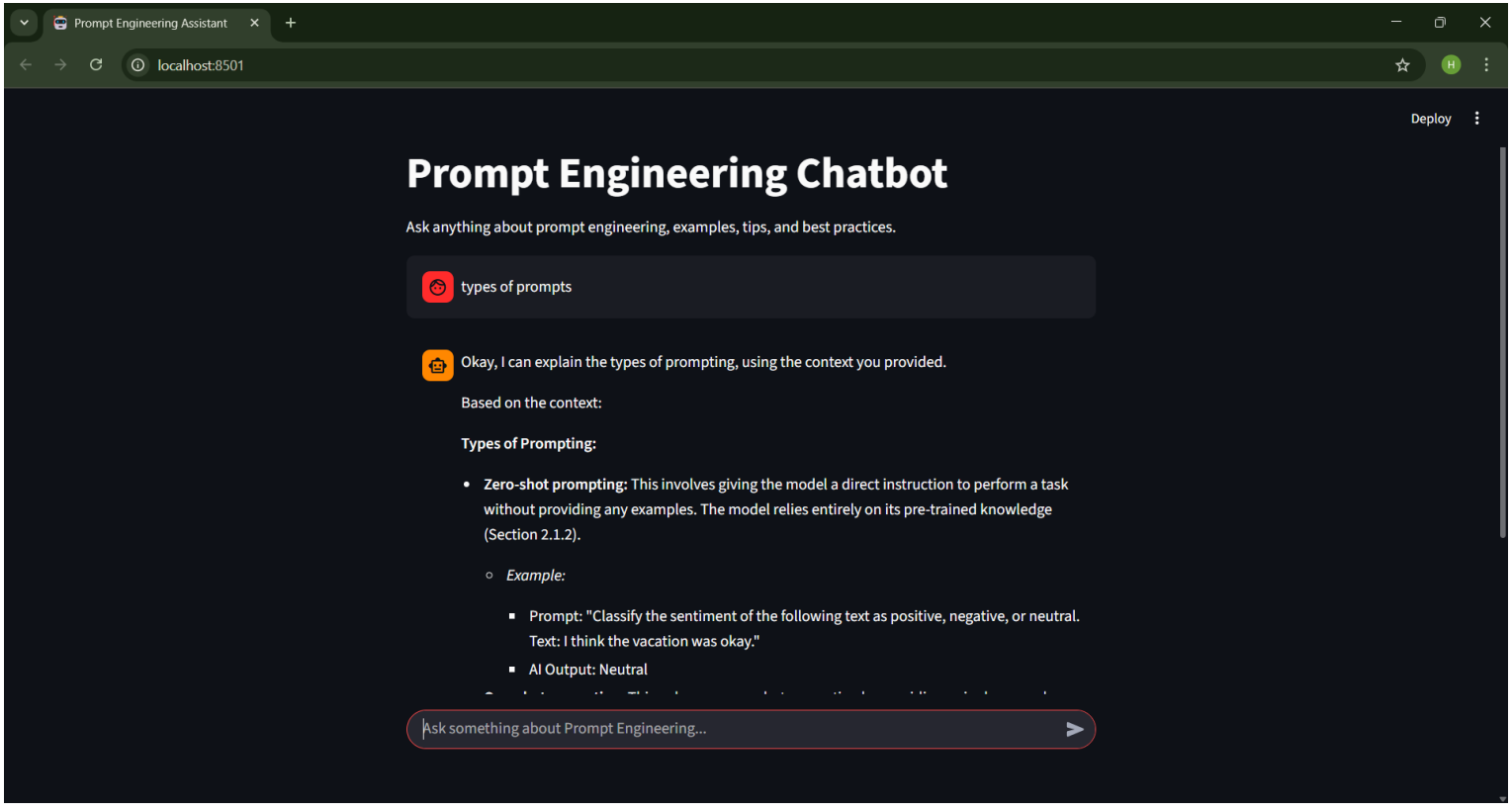
- `app.py`: Main UI & logic.
- `build_index.py`: Creates embeddings.
- `pe_notes.txt`: Knowledge base.
- `pe_index.json`: Vector store.
- `gemini_api.txt`: API Key.

Application Demo: Home Screen



The Streamlit Interface ready for queries.

Application Demo: Chat Response



The chatbot retrieving context and answering a query.

Key Features & Advantages

Features

- Custom Knowledge Base
- Fast responses (Gemini Flash)
- Interactive Web UI

Advantages

- **Accuracy:** Answers are based strictly on provided notes.
- **Relevance:** Tailored specifically for university exams.
- **Extensibility:** Easy to update notes.

Limitations

- **Dependency:** Requires an active Internet connection and valid Google Cloud API Key.
- **Context Window:** Very large knowledge bases may require more complex chunking strategies.
- **Domain Specific:** The chatbot is only as good as the quality of the notes provided in the text file.

Conclusion

- This project demonstrates how RAG can transform a general-purpose LLM into a specialized exam tutor.
- By leveraging **Google Gemini's** advanced embeddings and generation capabilities, we provide students with accurate, context-aware assistance for Prompt Engineering.

Thank You