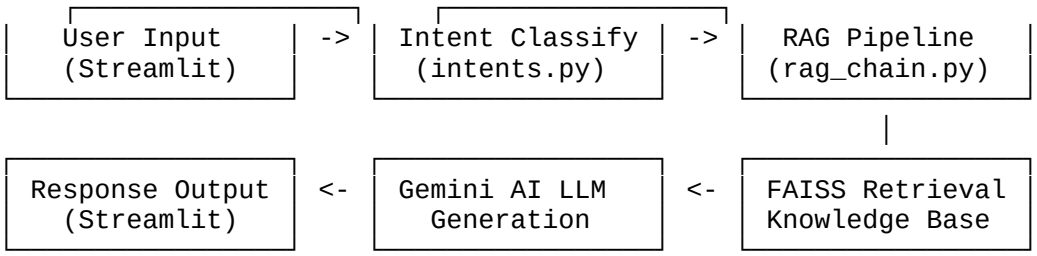
RAG – chat-based IT Support Assistant

Summary

The RAG IT Support Assistant is an intelligent chat-based support system that leverages Retrieval-Augmented Generation (RAG) technology to provide professional IT support solutions. The system combines rule-based intent classification with advanced AI-powered knowledge retrieval to deliver accurate, contextual responses to common technical problems. Built using Google's Gemini 2.0 Flash Lite model (available in the free tier), the solution offers an effective approach to automated technical support.

System Architecture

Architecture Overview



Core Components

1. Frontend Layer (Streamlit Interface)

- Provides interactive web interface with real-time user interaction
- Handles session management and maintains conversation context
- Captures and validates user queries before processing
- Displays formatted responses with proper styling

2. Intent Classification Layer (intents.py)

- Implements rule-based pattern matching using keyword detection
- Maps user queries to six predefined support categories:
    - Network Issue, System Restart, System Settings
    - Browser/Cache, Drivers/Updates, General Support

3. RAG Pipeline Layer (rag_chain.py)

- **Document Processing**: Chunks knowledge base into 700-character segments with 50-character overlap
- **Vector Embeddings**: Converts text to 768-dimensional vectors using Google Text Embedding 004

- **Similarity Search**: Utilizes FAISS for fast, accurate document retrieval (top-4 results)
- **Context Assembly**: Combines retrieved documents with user query for LLM processing

4. AI Generation Layer (Gemini 2.0 Flash Lite)

- Generates professional, contextual responses based on retrieved knowledge
- Maintains consistent tone and formatting across all interactions
- Supports multilingual responses in the same language as user input
- Implements temperature control (0.3) for reliable, consistent outputs

Tools & Technologies

- **Python 3.8+** – Core programming language
- **Streamlit** – Web app framework
- **LangChain** – RAG framework
- **Google Gemini 2.0 Flash Lite** – Fast, cost-effective LLM
- **Google Text Embedding 004** – Semantic embeddings
- **FAISS** – Vector similarity search engine
- **CharacterTextSplitter**, **RetrievalQA**, **PromptTemplate** – Data processing & prompt management

Prompt Engineering Strategy

Design Goal:
Deliver professional, accurate, and actionable IT support answers while ensuring consistency, reliability, and user trust.

Core Prompt:

> *"You are a helpful IT support assistant with access to a knowledge base. Answer respectfully and professionally, avoid slang, give clear step-by-step guidance, format with lists, and politely decline if the answer is not in the provided context. Reply in the user's language when applicable."*

Conclusion

The RAG IT Support Assistant integrates modern AI technologies into a practical, user-friendly support system. Its modular design ensures scalability, while free-tier Google services enable cost-effective operation. Strategic prompt engineering delivers professional, accurate responses, meeting enterprise support standards. By combining rule-based intent classification, semantic search, and advanced language generation, the system offers a robust, adaptable foundation for automated IT support across diverse organizational needs.