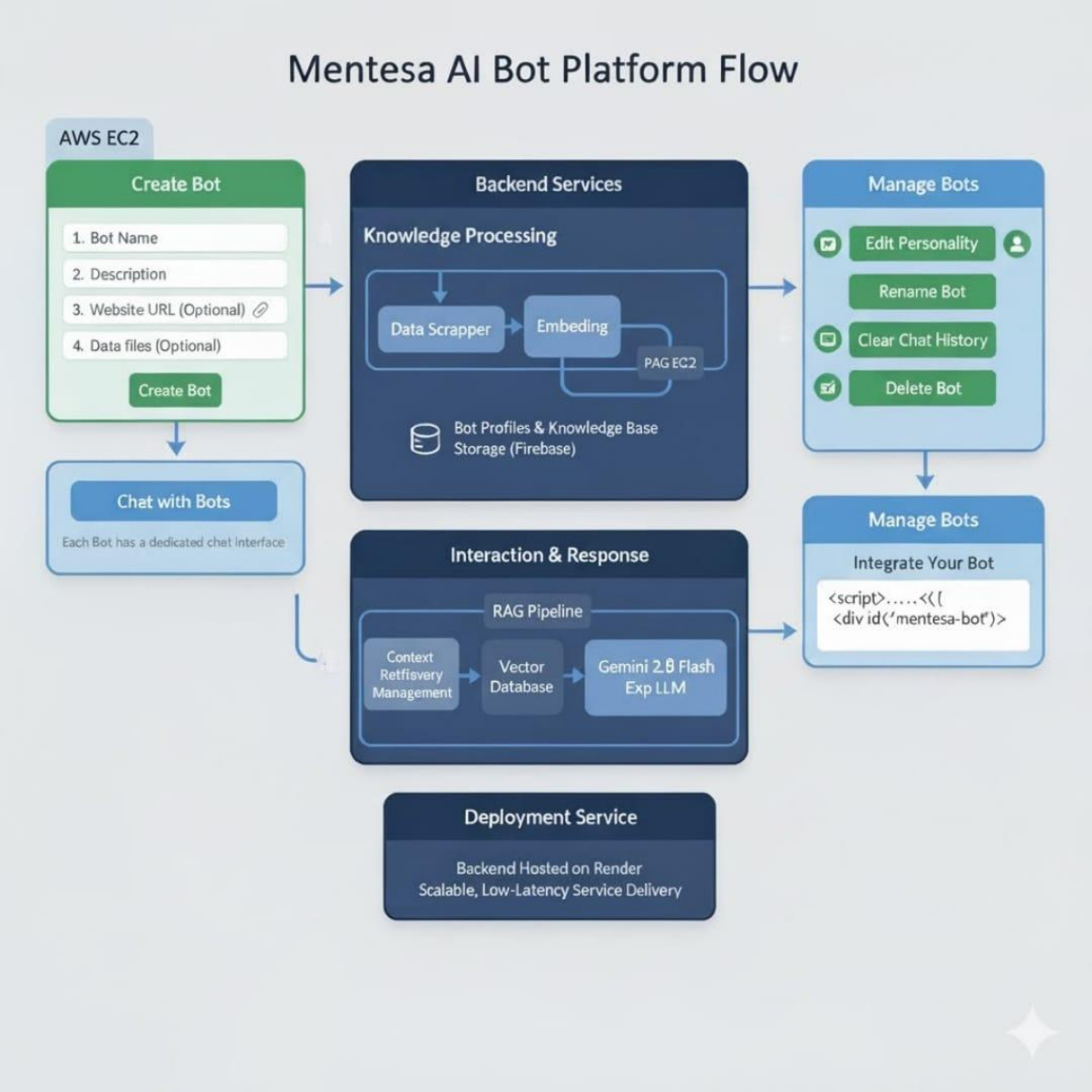
Technical Summary

Mentesa is an AI-powered bot-building platform designed to let users create highly personalized, knowledge-driven chatbots without needing coding or machine-learning expertise. The platform combines document processing, vector search, Retrieval-Augmented Generation (RAG), and large language models like Gemini flash 2.0 exp to deliver intelligent, context-aware bots that can be deployed anywhere. It supports uploading files, extracting knowledge, storing bot configurations, chatting with custom AI agents, and integrating them into websites or applications.

At the core of the platform is a seamless flow: users create a bot by entering its name, description, and uploading reference documents. These documents are processed into embeddings using vector databases such as Chroma or FAISS. A RAG pipeline connects these embeddings with the Gemini LLM to produce context-rich responses. All bot profiles, metadata, and conversation memory are maintained in Firebase, ensuring smooth management and scalability. Users can further edit, rename, delete, or integrate their bots using a simple HTML script. The backend is hosted on Render or AWS EC2, allowing low-latency performance even at scale.

Key Features & Components

- **Bot Creation:** Users provide bot details (name, description, website link, reference files) and instantly generate a personalized AI bot.
- **Knowledge Processing:**
    - Extracts text from PDFs, DOCX, websites, and uploaded files
    - Generates embeddings with Sentence Transformers
    - Stores indexed chunks in FAISS or Chroma
- **RAG-Based Intelligence:**
    - User query → vector search → relevant document chunks
    - Gemini 2.0-Pro / Flash-Exp LLM generates accurate, grounded responses
- **Bot Management:**
    - Edit personality
    - Rename bots
    - Clear chat history
    - Delete or integrate bots
- **Memory Handling:** Stores chat history to provide continuity and personalized answers.



The Mentesa platform is built on a modern and flexible technology stack that ensures performance, scalability, and seamless integration. On the frontend, it uses Streamlit to provide an intuitive and responsive user interface. The backend is powered by Python, with FastAPI or Flask serving as the primary frameworks for handling APIs and bot interactions. For storage, the system relies on Firebase Firestore, which maintains all metadata, bot profiles, embedding references, and conversation memory. The platform uses FAISS or Chroma DB as its vector store, enabling fast similarity search and efficient retrieval of embedded document chunks. Gemini 2.0-flash-exp serves as the core LLM engines responsible for generating intelligent, context-aware bot responses. Document processing is handled using tools like PyPDF2, pdfplumber, and textract, ensuring accurate extraction from PDFs and other file types. On the deployment and DevOps side, Mentesa leverages Render and AWS EC2 for hosting, with GitHub or GitLab supporting version control and CI/CD workflows.

To run the system smoothly, the recommended hardware includes at least 16 GB of RAM and a quad-core CPU for efficient document processing and embedding generation. Although Mentesa can operate on CPU-only systems, an NVIDIA GPU with at least 8 GB VRAM (such

as an RTX 3060 or above) is beneficial for faster embedding computation and heavy inference tasks. A 256 GB SSD ensures sufficient storage for embeddings, logs, and temporary files, while a stable internet connection of 50 Mbps or higher helps maintain low-latency communication between the client, backend, and cloud services.