Proposal: GPT-Driven Custom Chat System with Pinecone RAG and Trainer Feedback Integration

---

Project Overview:
This proposal outlines the development of a custom GPT-driven chat system, accessible via both a web app and WhatsApp. The system will offer advanced features like real-time trainer intervention, uncertainty handling, on-the-fly learning, and Retrieval-Augmented Generation (RAG) using Pinecone and text-embedding-ada-002. It will be deployed locally to ensure data privacy, with external updates controlled securely.

Key Features:

1. **Real-Time Trainer Intervention:**
   - Trainers can monitor live conversations, override responses manually, and provide real-time instructions to the model.
   - Language modifications (e.g., changing "sir" to "hey there") can be applied based on trainer feedback for more contextually appropriate conversations.
2. **Uncertainty Handling and Trainer Control:**
   - The model will alert the trainer when uncertain about a response, allowing trainer intervention to guide the conversation.
   - Trainers can pause the model at any point to take manual control.
   - The system will learn from trainer input through a real-time feedback loop to enhance future interactions.
3. **System Prompt Training and Feedback Storage:**
   - The GPT model will follow a predefined system prompt for conversational style, tone, and guidelines.
   - Trainer feedback during live sessions will be appended to the system prompt and stored in MongoDB to continually improve conversation flow.
4. **Retrieval-Augmented Generation (RAG) with Pinecone:**
   - Pinecone will be used for knowledge storage and retrieval, enabling the GPT model to respond with contextually relevant information from a vectorized knowledge base.
   - The *text-embedding-ada-002* model from OpenAI will generate embeddings for efficient knowledge retrieval.
5. **Local Deployment with Controlled Updates:**
   - The system will be deployed locally to ensure data security and privacy.
   - External updates will be managed without sharing data externally, ensuring full control over information flow.
6. **On-the-Fly Learning:**

    - The system will support real-time learning, updating response strategies instantly based on trainer input.
    - Conditional logic and new rules will be applied immediately, allowing the system to evolve in real-time.
7. **Accessibility through Web App and WhatsApp:**
    - The system will be accessible via both a web-based interface and WhatsApp, offering seamless user interaction.
    - Full WhatsApp integration will allow users to engage with the GPT model through a widely used platform.

Technical Stack:

- **Backend:** Python with Streamlit for chat interactions and trainer interventions.
- **Frontend:** Next.js for a user-friendly web app interface.
- **Model:** Hugging Face GPT model for conversation handling.
- **RAG System:** Pinecone with text-embedding-ada-002 for vectorized knowledge retrieval.
- **Database:** MongoDB to store system prompts and real-time trainer feedback.
- **Integration:** Full WhatsApp integration for chat interactions.
- **Deployment:** Local deployment with controlled external updates.

Milestones and Timeline (27 Days):

- **Milestone 1 (Days 1-10):**
    - Backend setup using Python and Streamlit for chat and trainer interventions.
    - Frontend setup using Next.js for the web app interface.
    - Initial GPT model integration for basic chat functionality.
    - WhatsApp integration for chat access.
    - MongoDB setup to store trainer feedback and system prompts.
    - Pinecone setup for RAG with text-embedding-ada-002.
    - Initial internal testing.
- **Milestone 2 (Days 11-20):**
    - Develop real-time trainer intervention and dynamic language modification features.
    - Implement uncertainty handling for trainer input when the system is unsure.
    - System prompts for GPT's conversational style and tone.
    - MongoDB integration for storing real-time feedback and system prompt updates.
    - Pinecone integration for knowledge-backed responses using text-embedding-ada-002.

- **Milestone 3 (Days 21-27):**
  - Finalize on-the-fly learning features for real-time strategy adjustments.
  - Complete Pinecone RAG system for knowledge-backed responses.
  - Finalize local deployment ensuring data privacy.
  - End-to-end system testing for web app, WhatsApp integration, and trainer interventions.
  - User interface refinements and final documentation preparation.