Submission Document — AI-Based Expert Call Query Tool

Overview

The AI-Based Expert Call Query Tool enables offline Q&A on expert call transcripts via an easy-to-use Streamlit interface. It ensures data privacy and fast local processing without internet access.

How the Tool Was Built

- **PDF Upload & Parsing**: PDFs are uploaded via the Streamlit app and split into smaller text chunks.
- **Embeddings**: Used `sentence-transformers/all-MiniLM-L6-v2` for semantic embeddings.
- **Vector Store**: ChromaDB stores embeddings for similarity search.
- **LLM**: Mistral 7B model accessed locally via Ollama.
- **Frontend**: Streamlit interface allows users to upload files and ask questions easily.

How Accuracy Was Improved

- **Domain-Specific Embeddings**: Using MiniLM for high-quality document vectorization.
- **Context Limitation**: Only top-4 retrieved documents are passed to the LLM to reduce noise.
- **Preprocessing**: Basic text cleaning to remove headers/footers and keep relevant text.

How Hallucination Was Minimized

- **Source Attribution**: Each answer displays a list of documents used to ground the answer.

- **Prompt Engineering**: Carefully crafted prompts to instruct the model to only answer from the provided context.

- **Offline Execution**: All operations run on the local machine — no external calls — ensuring no hallucination from the internet.