AI/ML Engineer Assignment

Role: AI/ML Engineer
Submission Deadline: July 06, 2025
Objective: Evaluate your practical skills in building a context-aware chatbot using vector databases and LLMs, with an emphasis on minimizing hallucinations and providing traceable sources.

Project Brief:
Develop a prototype chatbot that:

- Uses a vector database to store and retrieve embeddings from three provided documents.
- Integrates an LLM (e.g., OpenAI, Cohere, or open-source LLM) to generate human-like answers to user queries.
- Ensures that each paragraph in the response cites references to the source documents.
- Displays a list of all references at the end of each response.
- Prioritizes factual accuracy and aims to minimize hallucinations.

Requirements:

- Use any programming language/framework you're comfortable with (Python preferred).
- You may use any vector DB (e.g., Pinecone, FAISS, Chroma, or an open-source alternative).
- Demonstrate clear embedding, indexing, retrieval, and prompt integration logic.
- Your code should handle queries end-to-end (from input to reference-backed answer).
- Include a simple CLI, notebook, or web interface to test the chatbot.
- Provide clear setup instructions and environment requirements.

Deliverables:

1. Source code in a GitHub repository (or zip file).
2. Readme with setup instructions and usage guide.
3. Short note (200–300 words) explaining your design choices, libraries used, and how you addressed hallucinations.

Evaluation Criteria:

- Correctness and completeness of your implementation.
- Clarity of references in generated answers.
- Code quality and documentation.
- Simplicity and reproducibility.
- Creativity in minimizing hallucinations.

Note:
The three documents will be provided separately. Feel free to use additional open-source libraries for embeddings and LLM orchestration.

Best Wishes!