ML Engineer task

Develop a **Question & Answer (Q&A) pipeline** that can provide fluent, human-readable answers to user queries based on the selected dataset.

Use a **Retrieval-Augmented Generation (RAG)** approach as the foundation. The pipeline should include:

- Create a **vector store** of the documents with all the fields necessary
- A **retriever-ranker layer** to identify relevant context
- A **pretrained LLM layer** to generate the final answer

The goal is to simulate a business application that can quickly respond to simple (simulated) business questions.

(For this task, a "business question" is any query that could reasonably be asked of the Wikipedia or other selected dataset.)

Technical Requirements:

- Implement the entire solution in **Python**
- Use **pretrained NLP models** only — do not train new models from scratch
- Do **not** use any **paid API** services
- You may freely use **open-source repositories**, packages, and APIs — we will evaluate your understanding of the tools and your reasoning for using them
- The solution must be **containerised** (e.g., using Docker) and put behind an API to ensure portability and reproducibility
- No need to build a full front end — a **minimal interface** such as Streamlit or equivalent is sufficient for demonstration purposes
- Include a method to **evaluate the quality of the generated answers**
- Provide **2–3 suggestions** on how this solution could be made accessible to end users
-

Expected Outputs

- A **GitHub repository** containing all code from data ingestion to the final Q&A pipeline
- The repo can be public or private — please invite: @IgnaczPeti

Evaluation Criteria

- **Code quality** – 30%
- **Reproducibility of data processing** – 20%
- **Justification and use of pipeline components** – 40%
- **Documentation** – 10%