Title:

Automated Data Query and Retrieval System Using Offline (Free & Open Source) LLMs With CSV, MongoDB, LlamaIndex, and LangChain.

Overview:

This project demonstrates an automated system that allows querying structured CSV data via natural language using open-source, offline Large Language Models (LLMs). It loads CSV data into MongoDB, uses LLMs to dynamically generate queries based on user input, executes them, and returns or saves the results.

Technologies Used:

Python, MongoDB, Pandas, LlamaIndex, LangChain, llama-cpp-python, Offline LLMs (e.g., TinyLlama, Mistral)

Key Components:

CSV Ingestion: Loads CSV into MongoDB as documents.

LLM Query Generator: Accepts natural language, returns MongoDB queries.

Query Executor: Runs the query and fetches results.

Output Handler: Displays results or saves to CSV.

LLM Justification:

Offline LLMs ensure full data privacy, are cost-effective, and run locally without APIs. They suit secure environments and allow full customization.

Challenges Faced:

LLM inference latency on limited hardware → used quantized models

Query accuracy → added templates, manual validation

Data type handling → ensured type-safe conversions

MongoDB schema flexibility → added error-handling for nulls/missing fields

Conclusion:

The system proves LLMs can drive data querying without cloud dependency. It's scalable, secure, and useful in data-sensitive domains. Future work includes a web UI and SQL support.