Backend AI Engineer Project

Project

Your goal is to design and develop a Python-based REST API using FastAPI for executing a langchain agent that answers a question from the user using a web scraping tool. The system has to have the following requirements:

- An endpoint that receives a user query and sends back a response.
- In the endpoint, trigger the langchain agent that uses the web search tool and return its response.
- You can use any web search tool of your choice that you find in the langchain documentation.
- You MUST use the ollama langchain extension in order to run an LLM locally for your agent. (see ollama langchain docs. https://python.langchain.com/docs/integrations/llms/ollama/).

In addition to the code, you MUST include the following:

- A video recording of explaining everything below by walking through the code, diagram, data model, and 2-3 challenges/solutions you faced and resolved in this exercise
- A high-level system diagram showing the different services in your system and how they interact with each other
- Brief documentation on the endpoint of your API
- High-level answers to the following questions:
    - How would you implement authentication?
    - If you had to scale this system up to serve 1000s of requests per hour, how would you do it?
    - How would you implement logging in your API?
    - How would you test this API?

Try to fulfill as many of these requirements as possible before the deadline. You may use any offline or online resources to complete this exercise, but please cite your sources if you borrowed or generated code (you may provide a link or name the tool you used). We will be available to answer if you have questions while working on the project.

Submission

Please submit the all relevant code required to run your project (Github repo's public link). We're specifically looking for:

- A video recording of explaining everything below by walking through the code, diagram, data model, and 2-3 challenges/solutions you faced and resolved in this exercise
- The system diagram and data model describing your system
- All code required to run your project
- A README or similar on (a) how to run your code locally, (b) any considerations and tradeoffs you want us to know about, (c) the answers to the questions above; (c) your sources
- Documentation on the inputs and outputs of your API