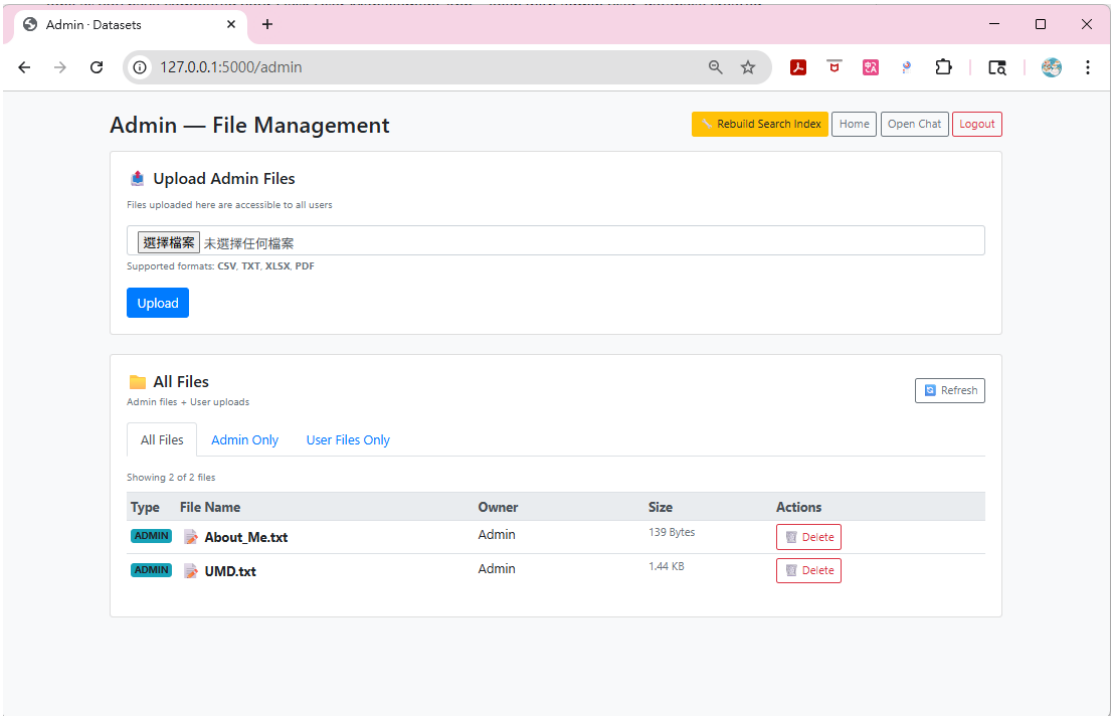
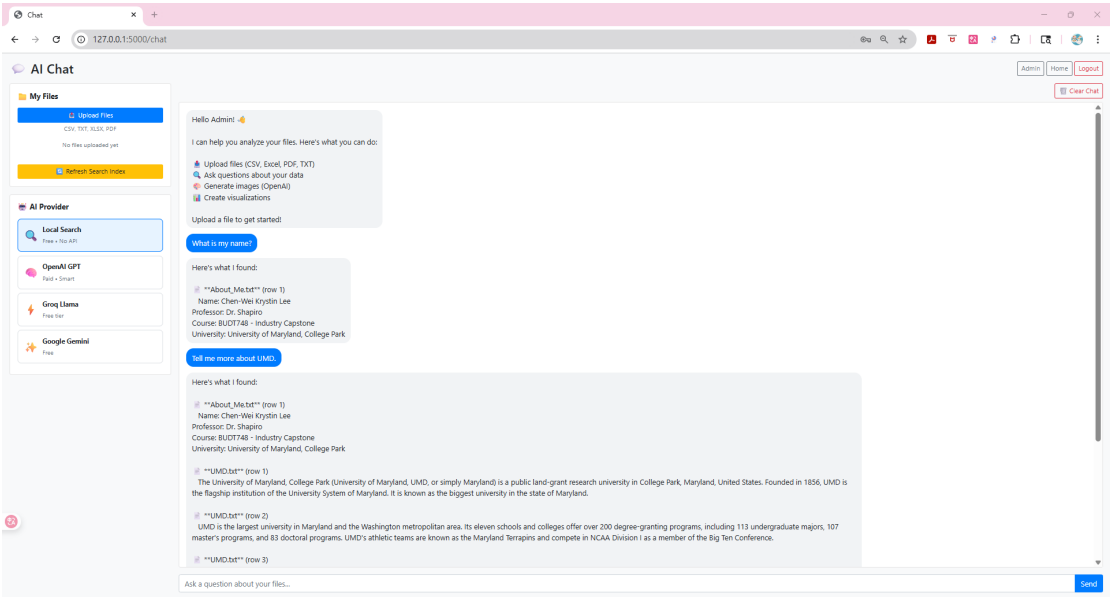
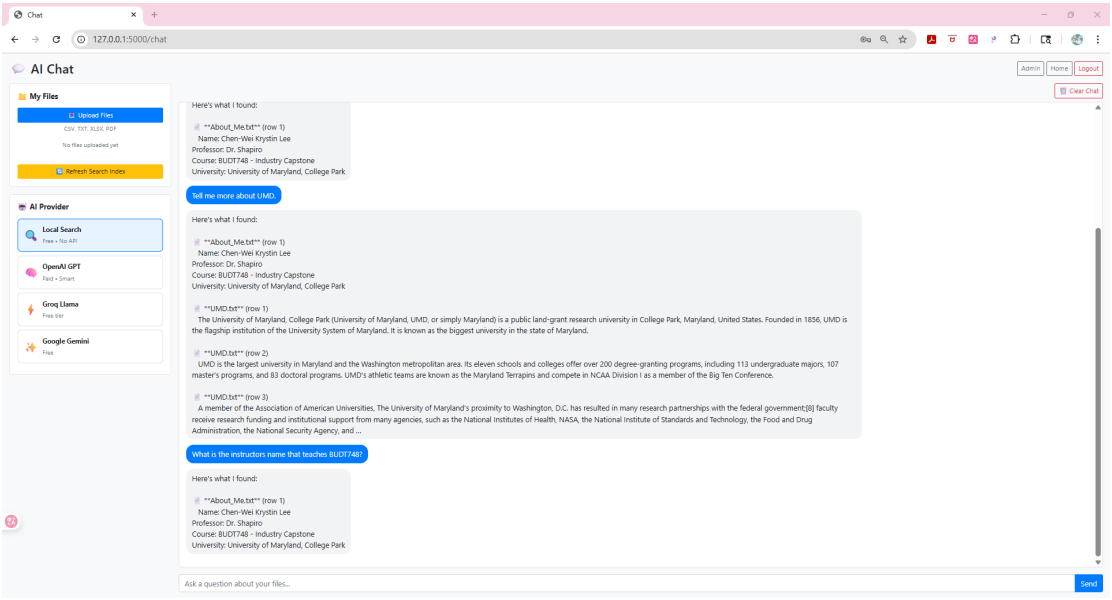
1. Screenshot of Admin File Upload page

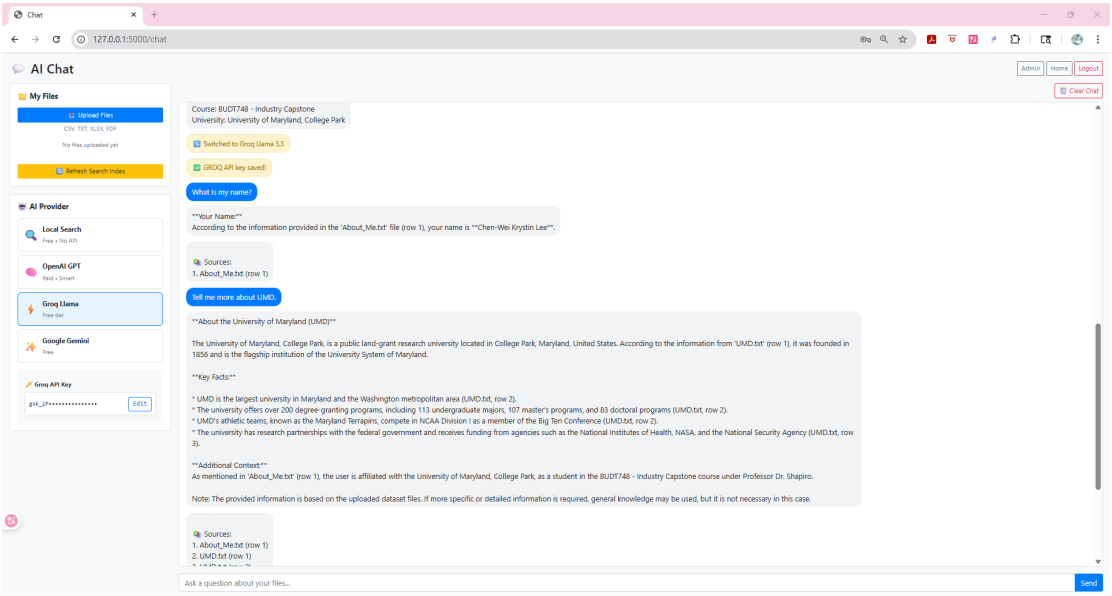


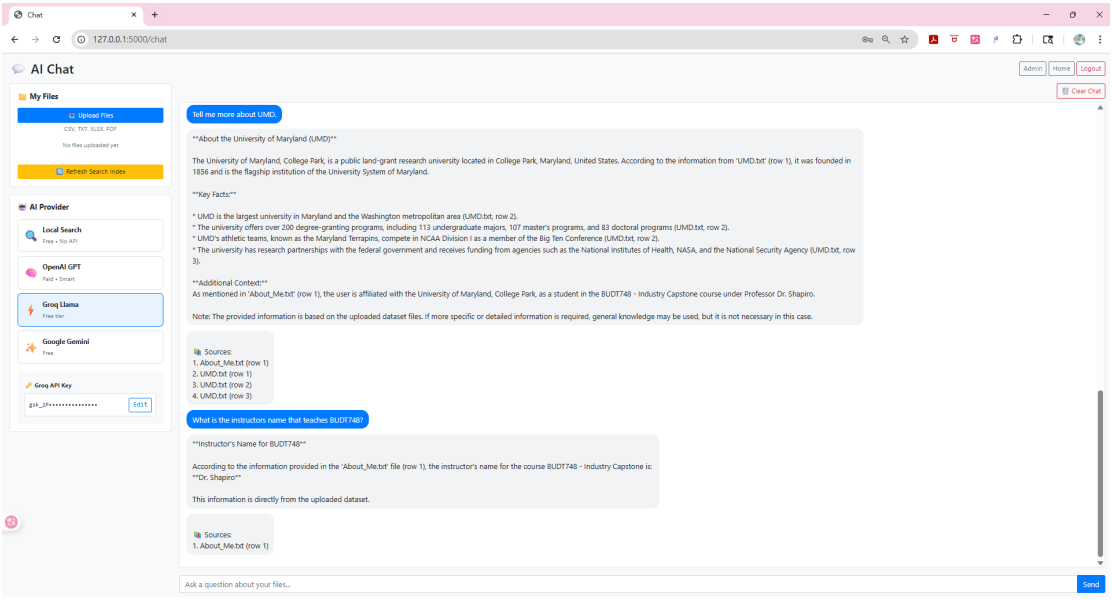
2. Screenshots from Local LLM chatbot showing all 3 questions and answers



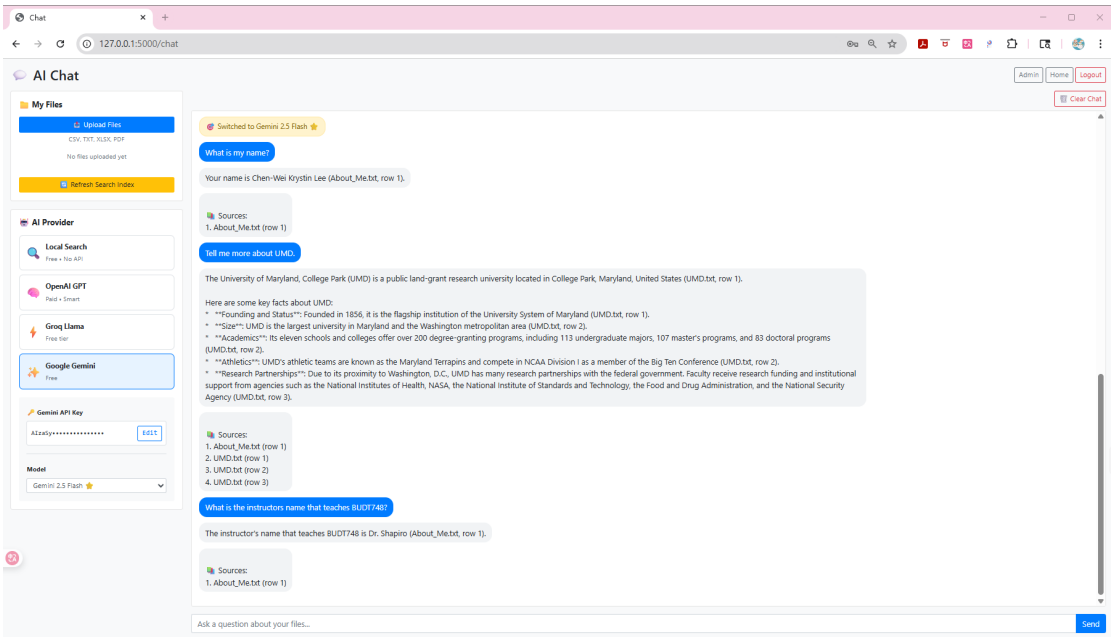


3-1. Screenshots from API chatbot (Groq) showing all 3 questions and answers





3-2. Screenshots from API chatbot (Gemini) showing all 3 questions and answers



4. Brief reflection on differences between Local LLM and API

A Local LLM runs entirely on the user's own hardware, emphasizing data privacy and complete control, with no network latency, making it suitable for handling sensitive information. In contrast, an LLM accessed via an API (like Groq Llama or Google Gemini) executes on the service provider's cloud servers, offering fast deployment, scalability, and easy access to the latest models, but it requires network reliance and sacrifices some degree of data control.

In the chatbot responses, this difference is evident: the answers using Local Search, similar to pure file retrieval or a local model, are more direct and rigid, strictly quoting specific file lines. Conversely, switching to the Groq Llama API or Google Gemini API resulted in answers that were more organized, fluent, and comprehensive, showcasing the model's advantage in information synthesis and narrative style, even though its underlying data still came from the uploaded files.