■ Documentation: Multi-Modal Assistant (Text + Image)

1. Overview

The Multi-Modal Assistant is a Streamlit application that integrates with Ollama to answer questions based on text and images. It runs fully offline (locally) and uses the LLaVA model (Large Language and Vision Assistant) to process both modalities. This tool is useful for understanding images, combining text and image context, and working without cloud dependencies.

2. Features

- Simple web interface built with Streamlit
- Accepts text questions
- Supports image uploads (PNG, JPG, JPEG)
- Queries Ollama LLaVA model locally
- Provides AI-powered responses combining text + image understanding
- Fully offline, no external API key required

3. Requirements

- Python 3.8 or higher
- Streamlit (for the web app)
- Requests (to communicate with Ollama API)
- Ollama installed and running locally (ollama serve)
- LLaVA model pulled into Ollama (ollama pull llava)

4. How It Works

- User enters a text question and optionally uploads an image.
- The app saves the uploaded image temporarily and encodes it as base64.
- The question and image are sent to the Ollama local API endpoint.
- The LLaVA model generates a response based on text and image inputs.
- The answer is displayed in the Streamlit interface and temporary images are deleted.

5. Usage Instructions

- Install Ollama and ensure it is running locally (ollama serve).

- Pull the LLaVA model using: ollama pull llava.
- Launch the app with: streamlit run app.py.
- Open the provided local URL in a browser (usually http://localhost:8501).
- Enter a question and optionally upload an image.
- Click Ask to generate and view the response.

6. Example Workflow

A user uploads an image of a chart and types: 'What does this chart show?'. The app sends the question and image to Ollama. The LLaVA model analyzes the chart and returns an explanation, which is displayed in the app interface.

7. Limitations

- Requires Ollama to be installed and running locally.
- Response speed depends on system performance and image size.
- Image understanding is limited to the LLaVA model's training data.
- Does not retain conversation history (stateless per question).

8. Future Improvements

- Add conversation memory to retain chat history.
- Support multiple images in a single query.
- Allow saving answers as PDF/Markdown reports.
- Provide options to choose different local models (not just LLaVA).
- Enable GPU acceleration for faster image processing.