[Start recording]

[Login page]

In the previous demo, we introduced a chatbot that could answer technical queries from uploaded documents.

This is to demonstrate the newer system, which supports better flowchart interpretation and multi-document question-answering using a Retrieval-Augmented Generation pipeline

[Login to the account]

[Username: user | Password: 123]

[Main interface]

On the left panel, we have the chat history with multiple conversations, making it easy to revisit previous queries.

[Visit and scroll through 'Demo Spec Questions' chat]

The main window is where interactions happen — user inputs, model responses, and formatted outputs like step-by-step instructions appear here.

At the top, there's the model selector, letting you switch between local multimodal LLMs for different performance needs.

[Open the model selector and choose the "Llama 3.2 4B (for text) + LLaVA 2.0 7B (for images)" option]

[Create a new chat]

[Click on the spanner icon on the top right to open the right sidebar]

Here, you can tune the model's behavior.

Temperature controls creativity. Lower values make outputs more predictable, while higher values increase variability.

Top-P, to control how much of the probability space the model considers when generating a response.

Max Tokens sets the response length limit.

[Change Hyperparameter tuning to documentation explorer]

The documentation explorer section allows you check the PDFs, flowcharts, or other documents for fast Retrieval

[Keep the right-side bar open]

Let's start by uploading a flowchart. Once uploaded, the system segments it into context chunks, creates embeddings, and indexes them for fast retrieval.

[Attach the Flowchart (1) image from the attachments folder]

[Ask question in chat: "What do you infer from this flowchart?"]

[Wait till reply completes]

We can also ask some follow up Questions to the bot.

[Ask question: "What do I do if 15V is present?"]

[Wait till reply completes]

[Ask question: "What do I do if LED is not lit?"]

[Wait till reply completes]

Now we can attach a different flowchart and ask a few questions from the new one.

[Attach the Flowchart(2) from the attachments folder]

[Ask question: "How much time should I wait from this given flowchart?"]

[Wait till reply completes]

Now let us ask follow up questions from the second flowchart.

[Ask question: "Could you explain what voltages should I look for and when?"]

[Wait till reply completes]

Now let’s upload a technical manual PDF. The system chunks and embeds the document, making it fully query-able.

This enables context retrieval for all future questions.

[Attach the PDF from the attachments folder]

[Ask question: "What do you infer from this attached PDF?"]

[Wait till reply completes]

[End of video]
