User Manual & System Instructions

System Infrastructure

- **Model:** Qwen2.5-7B-Instruct (Full Precision)

System Requirements

- **Hardware:** Requires a GPU with 16GB-20GB of VRAM to support the full 7B model at high precision.

Operating Steps

1. Start the server and upload numbered PDFs to the `/ingest` endpoint.
2. Submit specific queries (e.g., 'What is the rate in Contract 5?') to ensure the system applies correct filters.
3. **Run System Evaluation:** Navigate to the **Evaluation** tab in the Gradio UI to test the pipeline's factual accuracy.
4. **Set Parameters:** Use the slider to select the number of test questions you want the system to generate.
5. **Execute Test:** Click **Run Evaluation Test** and wait for the LLM to process the documents, retrieve answers, and grade itself.

LLM-as-a-Judge Evaluation Pipeline

The system includes an automated, self-evaluating pipeline to guarantee RAG accuracy without requiring manual human testing:

- **Synthetic Q&A Generation:** The system randomly samples a single chunk from your ingested documents and forces the LLM to generate a highly specific question. It explicitly includes the source filename in the prompt (e.g., *"According to Contract_8.pdf..."*) to prevent context hallucination.
- **RAG Retrieval & Generation:** The synthetic question is passed back into the standard RAG pipeline. The system retrieves the best context and generates an answer, completely blind to the synthetic ground truth.
- **Automated Judging:** The LLM acts as an impartial judge. It compares the newly generated **RAG Answer** against the **Ground Truth**.
    - It awards a `[1]` if the RAG answer lies, contradicts the truth, or misses critical context.
    - It awards a `[2]` if the RAG answer is factually accurate and contains the same (or better) core information.
- **Performance Metrics:** The UI outputs an **Overall Preference Score** (the percentage of questions the pipeline passed) alongside a detailed, step-by-step breakdown of every question, ground truth, RAG answer, and the judge's exact justification.