Technical Notes

Secure RAG System Integrated with Copilot Studio

System Overview

This solution integrates Microsoft Copilot Studio with a FastAPI-based Retrieval-Augmented Generation (RAG) pipeline. It ensures secure, role-based access to documents (e.g., PDFs in Google Drive) and delivers LLM-generated answers only from authorised content with citations.

Layer	Tools / Libraries
Interface	Microsoft Copilot Studio + MCP HTTP API
Backend Server	FastAPI, Pydantic, CORS Middleware
Document Loader	LangChain's PyPDFLoader
Vector Store	FAISS
Embeddings	Ollama
LLM for QA	Ollama (local) – Phi3:mini
Access Control	Email Filter in FastAPI

Working Process:

1. **User Input via Copilot Agent**
   A Copilot Studio agent sends the user's query and identity via HTTP POST to a FastAPI endpoint.
2. **RBAC Filtering**
   The backend maps the user's identity to a list of accessible documents and only loads those for processing.
3. **Document Loading & Embedding**
   - PDFs are chunked and parsed with metadata.
   - Embeddings are generated using a local embedding model.
   - Vectors are stored or searched via FAISS.
4. **Contextual Retrieval**
   A top-k similarity search fetches only relevant chunks based on the prompt.
5. **LLM Answer Generation**
   Retrieved content is passed to an LLM (via RetrievalQA chain) to answer accurately using only the filtered context.

6. **Source-Aware Response**
   The final output includes:
   - LLM-generated answer
   - List of filenames + page numbers used as context
7. **Response Sent Back to Copilot**
   The answer is sent as JSON to Copilot Studio for rendering in the UI.

Security & Access Control

- Users are only shown answers grounded in **documents they have access to**.
- Fine-grained RBAC can be extended using email-based filtering.
- Hallucinations are minimised with strict context grounding and minimal answer fallbacks.

Future Enhancements Scope

- SharePoint API integration for dynamic file access control.
- Cloud Hosting