RAG on a local LLM: private document search that never leaves your building

26 October 2026 · 5 min read · blogs.iictp.org/local-rag-private-document-search

Combine retrieval with a locally hosted model and you can answer questions from sensitive documents without sending them anywhere. Every component has to run on your side, which is where most of the design work lies.

TL;DR The short version

- Retrieval with a locally hosted model lets you answer questions from sensitive documents without sending them anywhere.
- You host every part yourself, including the embedding model and the vector store or search engine.
- Enforce permissions in retrieval rather than leaving them to the model, and test retrieval and answers separately.

Read online, listen to it, or change the text size: https://blogs.iictp.org/local-rag-private-document-search

The parts you host

- An embedding model to turn text into vectors.
- A vector store or a search engine.
- The language model that writes the answer.
- A thin service that ties them together and enforces permissions.

Design points

- **Permissions:** retrieve only documents the asking user may see. Do not leave this to the model.
- **Ingestion:** handle scanned pages, tables and updates.
- **Citations:** show the passages used so answers can be checked.
- **Offline updates:** plan how new models and software reach a machine with no internet.

Test it

Use real questions and measure retrieval and answers separately. See RAG explained for the pipeline.

Related credential: Local LLM Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. IICTP. Local LLM Practitioner credential page. iictp.org/certifications/llmp.

Published by IICTP, an independent technical credential body, on 26 October 2026. The current version of this post is at https://blogs.iictp.org/local-rag-private-document-search. More guides: blogs.iictp.org.