RAG explained: retrieval-augmented generation, step by step

26 October 2026 · 6 min read · blogs.iictp.org/rag-explained

Retrieval-augmented generation (RAG) gives a language model the documents it needs at the moment it answers, instead of relying only on what it learned in training. It is the standard way to make a model answer from your own material.

TL;DR The short version

- Retrieval-augmented generation gives a model the documents it needs at the moment it answers, so it can answer from your own material.
- Quality is won or lost in chunking and retrieval, and combining keyword and semantic search often beats either alone.
- Measure retrieval and answer quality separately on real questions; RAG reduces unsupported answers but does not remove them.

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

The pipeline

1. **Ingest.** Split documents into chunks and store them with an embedding for each, in a vector index or a search engine.
2. **Retrieve.** For a question, find the chunks most relevant to it.
3. **Augment.** Place those chunks in the prompt with the question.
4. **Generate.** The model answers, ideally citing the chunks it used.

Where quality is won or lost

- **Chunking:** chunks that cut a table or a thought in half retrieve badly.
- **Retrieval:** combining keyword and semantic search often beats either alone.
- **Prompting:** tell the model to answer only from the supplied text and to say when it cannot.
- **Freshness:** re-index when the source changes.

How to know it works

Build a set of real questions with known answers and measure two things separately: did retrieval find the right chunk, and did the model use it correctly. A wrong answer with the right chunk retrieved is a different bug from one where nothing relevant was found.

Limits

RAG reduces unsupported answers but does not remove them. Keep citations visible, restrict what each user can retrieve to what they may see, and test for prompts that try to pull in other people's documents.

Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Lewis et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." arXiv:2005.11401.

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