

RAG Explained: Giving AI Your Knowledge Without Retraining

FREE EDUCATIONAL BRIEF · LEARN AI. UNCOVER THE DATA. PROTECT THE TRUTH.

Models know nothing about your price list, your policies, or yesterday's meeting. Retraining a model on your data is expensive and slow — so almost every serious AI product uses retrieval-augmented generation (RAG) instead: find the relevant passages from your documents at question time, hand them to the model, and instruct it to answer only from what it was given.

THE PIPELINE IN FOUR STEPS

» Ingest & chunk

Documents are split into passages small enough to search precisely but large enough to carry meaning.

» Embed & index

Each chunk becomes a numerical vector capturing its meaning, stored in a searchable index.

» Retrieve

A user's question is embedded the same way; the closest chunks by meaning — not just keywords — are pulled.

» Generate with citations

The model answers using those chunks, ideally citing which passage supports each claim.

WHERE RAG BREAKS

Garbage in remains garbage out: outdated documents, duplicated policies, and contradictory versions surface directly in answers. Chunking too coarsely buries the needle; too finely severs context. And retrieval can simply miss — which is why well-built systems say “I don't have that information” instead of improvising. Evaluating a RAG system means testing with real questions and checking the citations, not admiring fluent prose.

WHY THIS MATTERS BEYOND ENGINEERING

RAG is also a governance tool. Because answers trace to source passages, you get auditability: when the system is wrong, you can find the document that misled it. For regulated businesses, that traceability is the difference between deployable and undeployable AI.

INVESTIGATOR'S TAKEAWAYS

- RAG feeds your documents to the model at question time — no retraining.
- Answer quality is capped by document quality; curate before you index.
- Citations are the audit trail; demand them in any RAG product.
- “I don't know” is a feature — improvisation is the bug.
- Test with real user questions, then verify the cited passages.