

RAG For Beginners



Why LLMs Need Help

Core idea: LLMs don't always know, so they sometimes guess.

- Depend on pretraining → no access to latest info
- Cannot access private/company data
- May hallucinate (answer confidently but incorrectly)
- Not designed as factual databases

Analogy: A very smart person with a fixed memory — but can't open new books unless you give them.

What is RAG?

RAG = Retrieval-Augmented Generation

A method that lets AI search your documents first, then answer using that information.

- **Retrieve**: Find the relevant text
- **Augment**: Give that text to the LLM
- **Generate**: LLM writes an accurate answer based on evidence

Value: More accurate, grounded, and up-to-date responses.

How **RAG** Solves LLM Limitations

- Reduces hallucinations
- Uses your private documents (PDFs, policies, product info)
- No model retraining needed
- Answers stay consistent and explainable
- Works for customer support, HR, legal, insurance, etc.

Analogy: Instead of remembering everything, the AI checks the folder you provided.

RAG Architecture

👉 4-step flow:

- User Question
- Retriever — finds matching document pieces (chunks)
- LLM — reads those pieces
- Final Answer — grounded in your data

Visualization:

- User → Search Docs → LLM → Answer

What Actually Happens Behind the Scenes

Non-technical breakdown:

- Documents are split into small readable units
- AI stores them in a searchable index (like supercharged CTRL+F)
- When a question comes, the system fetches the best matches
- LLM uses ONLY that retrieved info to answer
- Helps businesses provide consistent, accurate, compliant outputs

Key takeaway:

- RAG lets AI “look up” information instead of guessing.

