

RAG stands for *Retrieval-Augmented Generation*. It might sound complicated at first, but it's actually a pretty simple and smart idea.

When you ask a normal AI a question, it tries to answer based only on what it learned during its training. Imagine it like a person who read a lot of books a while ago but isn't allowed to look anything up anymore. Sometimes, it gets things right. But other times, it might guess or give outdated information.

That's where RAG comes in. With RAG, the AI doesn't just guess — it actually goes and searches through your own collection of information, like PDFs, documents, or notes you've given it. It finds the most relevant pieces of information first and then uses that to answer your question.

You can think of it like giving the AI its own private library. When you ask something, it quickly flips through the pages to find helpful facts and then explains them to you in simple language. This way, the answers are not only more accurate, but also based on real sources you've chosen.

The great thing is, you're already using RAG in your chatbot! Your code takes the PDFs you upload, breaks them into smaller pieces, and turns them into searchable data. When you ask your assistant a question, it searches through that data to find the best answers, instead of just guessing.

This makes your assistant smarter, more reliable, and able to handle custom knowledge that regular AI might not know. Plus, you can keep adding new documents anytime to keep it learning!