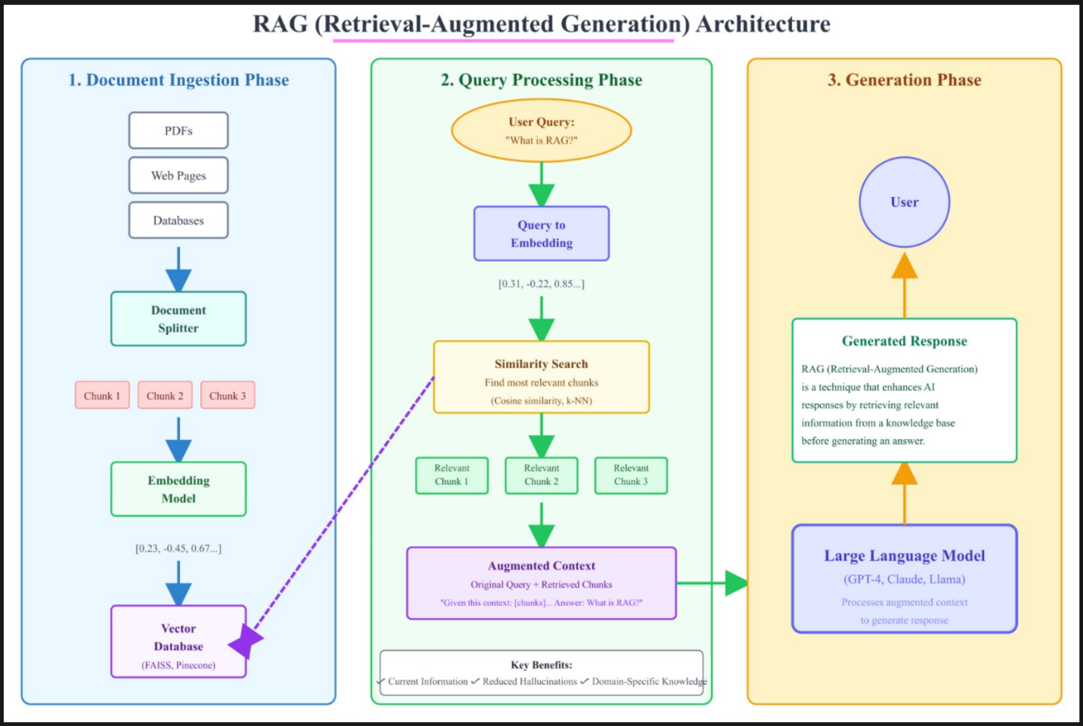
RAG

80%

RAG


RAG (Retrieval-Augmented Generation) Architecture
1. Document Ingestion Phase
PDFs
Web Pages
Databases
Document Splitter
Chunk 1
Chunk 2
Chunk 3
Embedding Model
[0.23, -0.45, 0.67...]
Vector Database
(FAISS, Pinecone)
2. Query Processing Phase
User Query:
"What is RAG?"
Query to Embedding
[0.31, -0.22, 0.85...]
Similarity Search
Find most relevant chunks
(Cosine similarity, k-NN)
Relevant Chunk 1
Relevant Chunk 2
Relevant Chunk 3
Augmented Context
Original Query + Retrieved Chunks
"Given this context: [chunks]... Answer: What is RAG?"
Key Benefits:
✓ Current Information ✓ Reduced Hallucinations ✓ Domain-Specific Knowledge
3. Generation Phase
User
Generated Response
RAG (Retrieval-Augmented Generation) is a technique that enhances AI responses by retrieving relevant information from a knowledge base before generating an answer.
Large Language Model
(GPT-4, Claude, Llama)
Processes augmented context to generate response

RAG (Retrieval-Augmented Generation) is a powerful technique that enhances AI language models by combining their generation capabilities with external knowledge retrieval.

What is RAG?

RAG is like giving an AI assistant access to a library while it's answering questions. Instead of relying solely on what it learned during training, the AI can now look up specific, current, or specialized information from external sources before generating its response.

Think of it this way: Traditional language models are like students taking a closed-book exam - they can only use what they memorized. RAG-enabled models are like students in an open-book exam - they can reference materials to provide more accurate, detailed, and up-to-date answers.

Chatbot

{web Search}

July 2023

What is the current AI news?

Accurately O/P

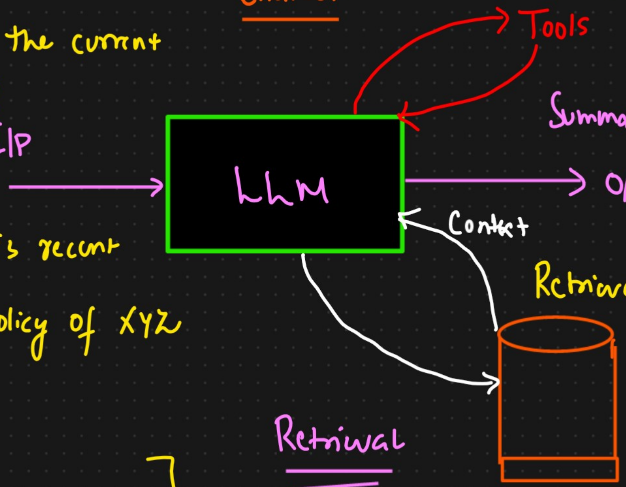
Summarized O/P

Info

XyZ

{Employee}

What is recent leave policy of XYZ



R A G

Retrieved Content ⇒ Enrich

← Vector Database → Vectors

Retrival

[Revenue increased by 10%]

⇓

Add more metadata.

Similarity Search

Augmented = {
text: "Revenue increased by 10%.
"Source:" Tesla Annual Report 2023
"date"}: —

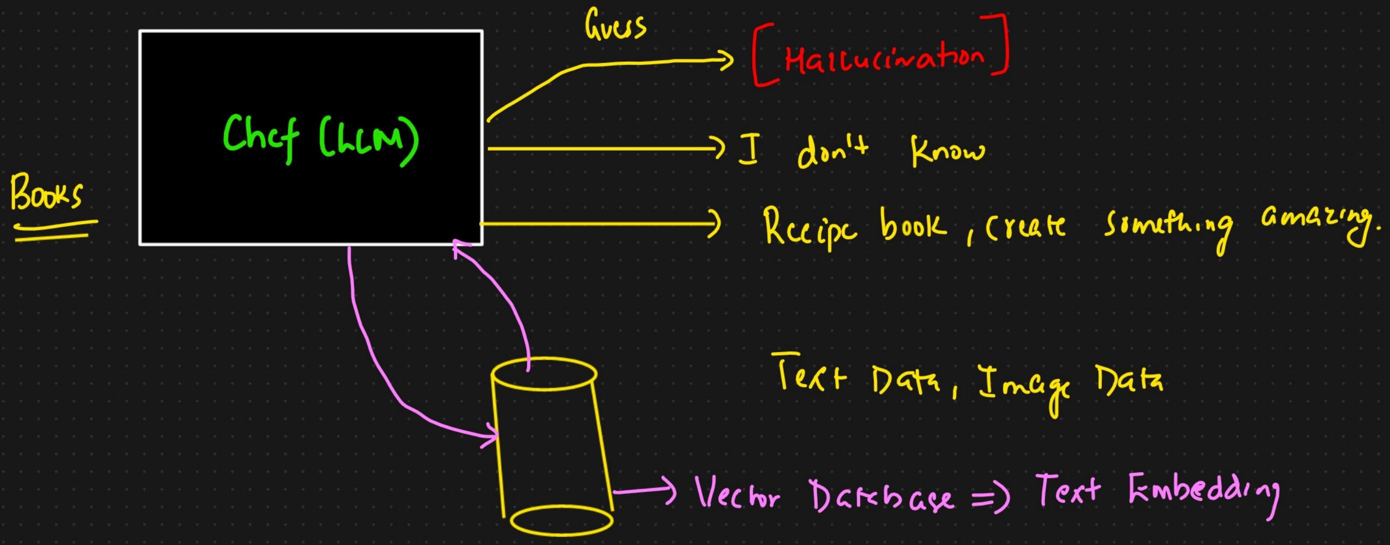
2) RAG

① Retrieval – Finding Relevant Info [Vector Databases]

② Augmentation – Enhancing Context with Metadata.

③ Generation – Producing the Answer.

Country A [India] Country B [European]


Chef (LLM)
Books
Guess
[Hallucination]
I don't know
Recipe book, create something amazing.
Text Data, Image Data
Vector Database => Text Embedding

Customer Support

[LLM] => Customer Support

Without RAG

Customer: "What's your return policy for items bought during the Black Friday sale?"
AI: "Generally, most companies offer 30-day returns, but policies may vary..."
[Generic, unhelpful response]

With RAG

AI Assistant [LLM + Vector Databases]

Customer: "What's your return policy for items bought during the Black Friday sale?"
[RAG searches company policy database]
AI: "According to our current policy (Policy Doc v3.2, updated Nov 2024), Black Friday purchases have an extended 60-day return window until January 31st. Items must be unused with original tags. Electronics have a 15-day return period due to rapid depreciation. Would you like me to start a return for a specific item?"

Cost Savings: JPMorgan saved $150M annually by implementing RAG for research analysis instead of fine-tuning models monthly
Accuracy: Microsoft reported 94% reduction in AI hallucinations after implementing RAG in their Copilot products
Flexibility: Bloomberg updates their financial AI assistant hourly with new market data - impossible with traditional LLMs
Compliance: Healthcare companies use RAG to ensure AI responses always cite approved medical sources"