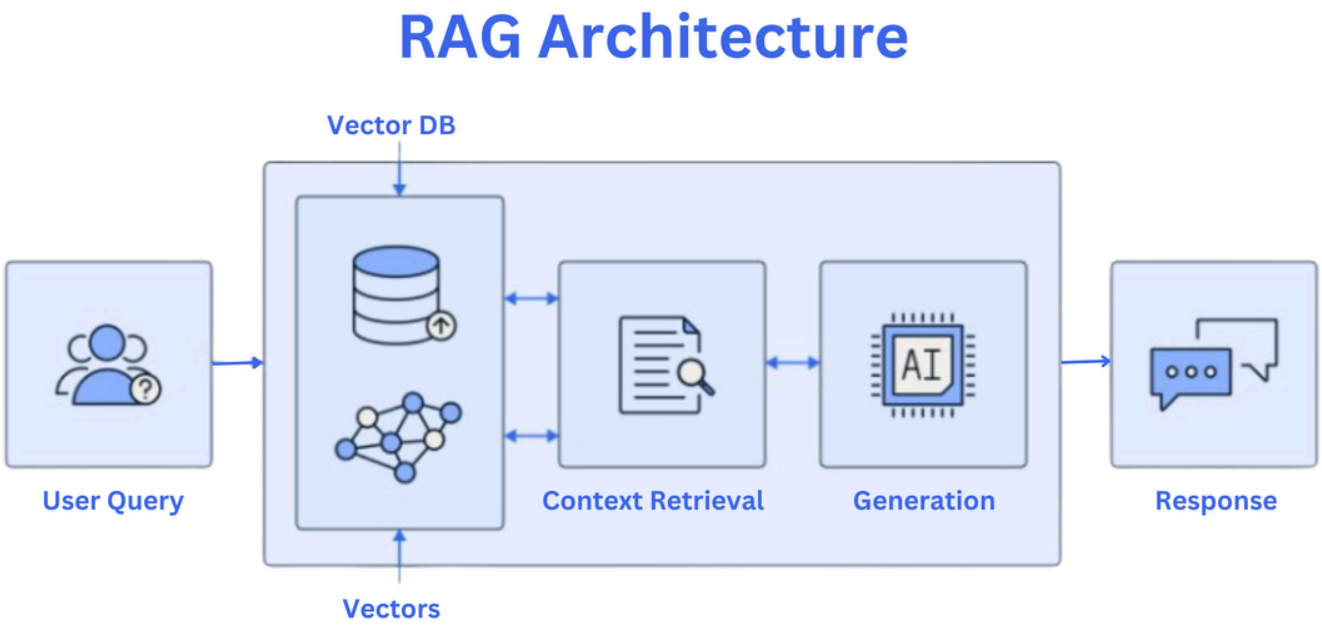
🔍 What is RAG (Retrieval-Augmented Generation)?

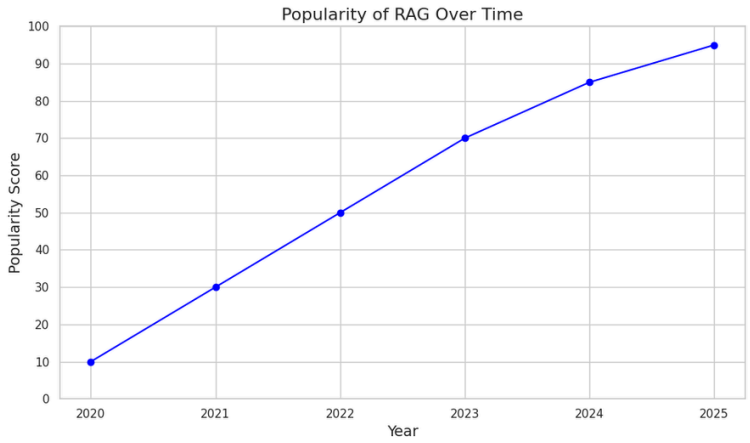
Retrieval-Augmented Generation (RAG) is a powerful architecture in Generative AI that enhances the capabilities of large language models (LLMs) by integrating real-time information retrieval. Traditional LLMs generate responses based solely on their training data, which can be outdated or limited. RAG solves this by connecting the model to external knowledge sources—like databases, APIs, or document repositories—allowing it to fetch relevant information before generating a response.

This hybrid approach works in three key steps: first, the system retrieves relevant data based on the user's query; second, it augments the prompt with this data; and finally, the LLM generates a response that's more accurate, current, and context-aware. RAG is especially useful in domains like customer support, academic research, and enterprise applications where precision and up-to-date knowledge are critical



🌐 Why RAG Is Commonly Used in Generative AI

Retrieval-Augmented Generation (RAG) has become a go-to architecture in modern AI systems because it solves some of the biggest limitations of traditional language models. Here's why it's so widely adopted:



✅ Improves Accuracy and Reduces Hallucinations

🔄 Keeps Responses Up-to-Date

💰 Cost-Effective Domain Adaptation

🧠 Context-Aware and Specific Answers

🔐 Greater Control Over Output

Company	Use Case Description
DoorDash	RAG-powered chatbot for delivery support; retrieves internal docs to assist Dashers
Hugging Face	Hosts RAG models and datasets; used by developers to build custom retrieval pipelines
Evidently AI	Provides RAG evaluation tools for benchmarking and improving LLM reliability
Signity Solutions	Implements RAG in virtual assistants, healthcare, and customer support systems
Google DeepMind	Uses RAG-like architectures in research for grounded and factual generation
Meta AI	Developed the original RAG architecture for combining retrieval with generation