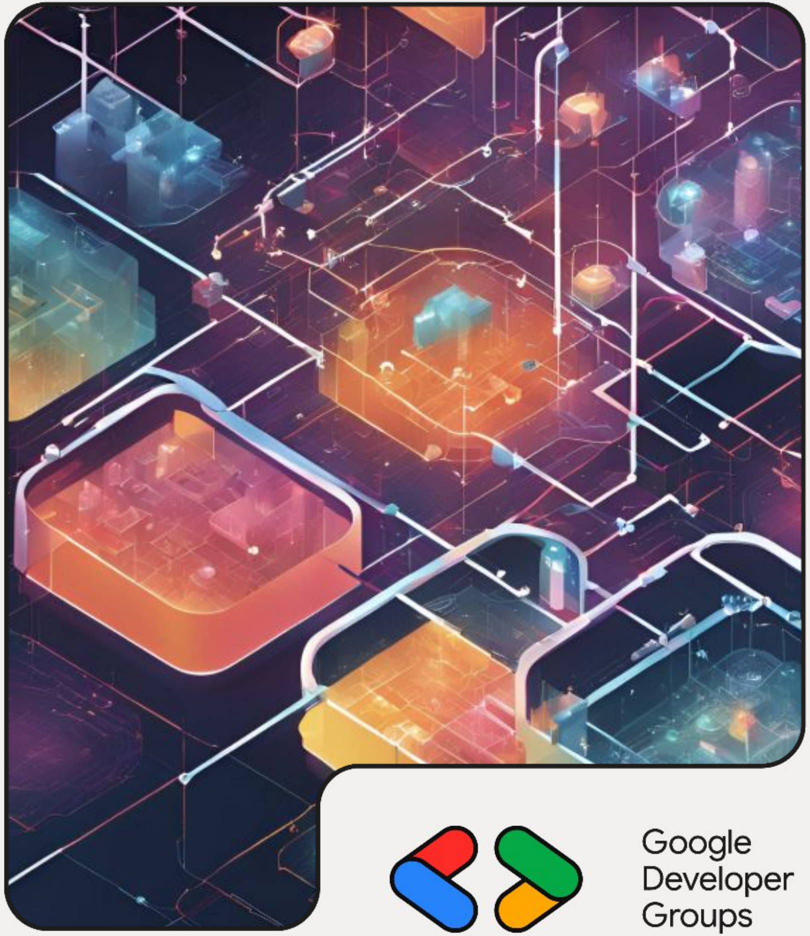
{ DevFest }

Kathmandu

Unlocking the power of RAG with Gemini : Building AI that thinks smarter

Rashika Karki



Google Developer Groups

{ DevFest }
Kathmandu

- Software Engineer | MLH
- Organizer | CNCF Kathmandu
- Msc Student | Pulchowk Campus

- in/rashikakarki
- rashikakarki9841@gmail.com

Google Developer Groups

AI
@DevFest

https://bit.ly/slido-devfest

Google
Developer
Groups

Large Language Models (LLMs)

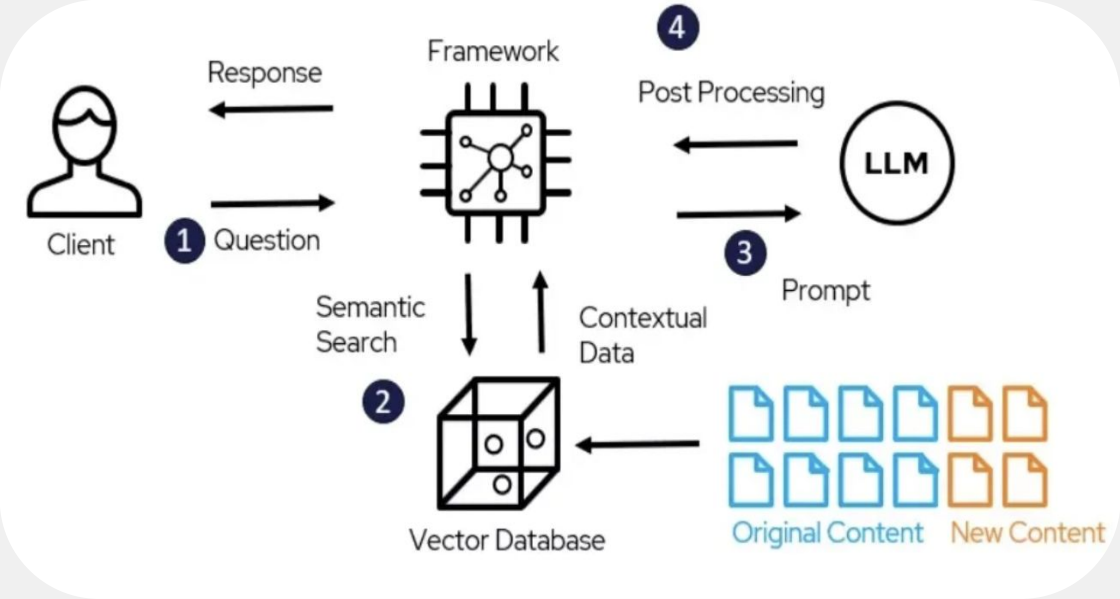
What are LLMs?

- Advanced AI models like GPT or Gemini trained on massive datasets.
- Designed to generate human-like text.
- Capable of tasks like summarization, Q&A, creative writing, and more.
- Pretrained on diverse data: They "**know a little about a lot.**"
- Ideal for general-purpose tasks.

Limitations of LLMs

- Limited Knowledge Cutoff
- Hallucination Problem
- High Cost for Fine-Tuning
- Scalability and Size

The Rise of RAG: Why It Matters



Source: [Advanced RAG for LLMs/SLMs]

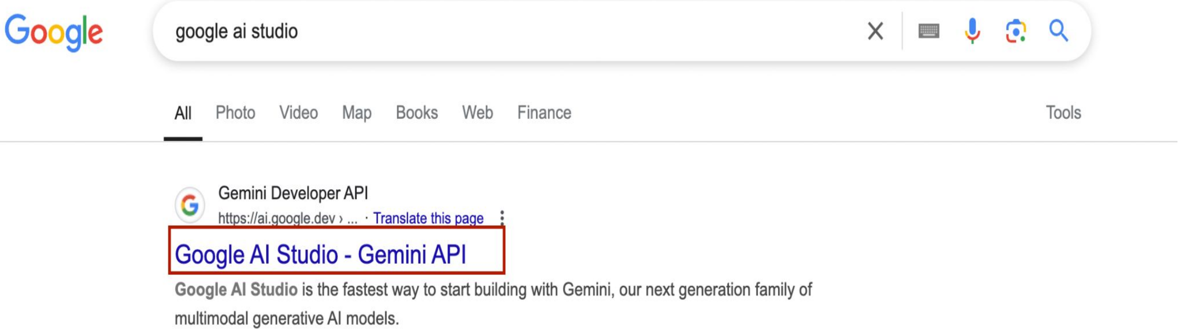
- Retrieval-Augmented Generation (RAG) bridges the gap by combining LLMs with real-time retrieval systems.
- Bridges the gap between **static model knowledge** and **dynamic external information.**
- **Key Components of RAG**
  - Retriever
  - Generator
  - Knowledge Base (Database)
  - Orchestrator

Tools We’ll Be Using

- **Orchestration:** LangChain
- **LLM**: Google Gemini
- **Vector DB**: Chroma

AI
@DevFest

Gemini API Key



Gemini Developer API

https://ai.google.dev › ... · Translate this page

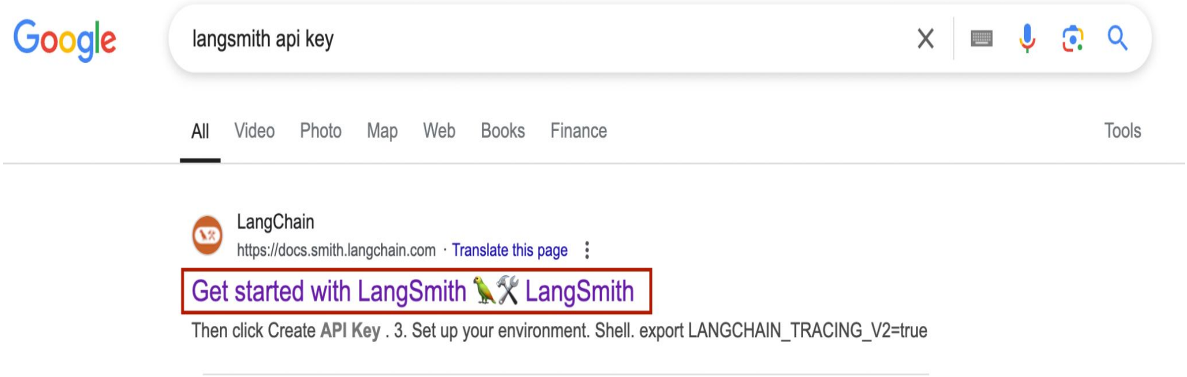
Google AI Studio - Gemini API

Google AI Studio is the fastest way to start building with Gemini, our next generation family of multimodal generative AI models.

Google
Developer
Groups

AI
@DevFest

Langsmith API Key



langsmith api key

LangChain
https://docs.smith.langchain.com · Translate this page

Get started with LangSmith 🦜🛠️ LangSmith

Then click Create **API Key** . 3. Set up your environment. Shell. export LANGCHAIN_TRACING_V2=true

Google
Developer
Groups

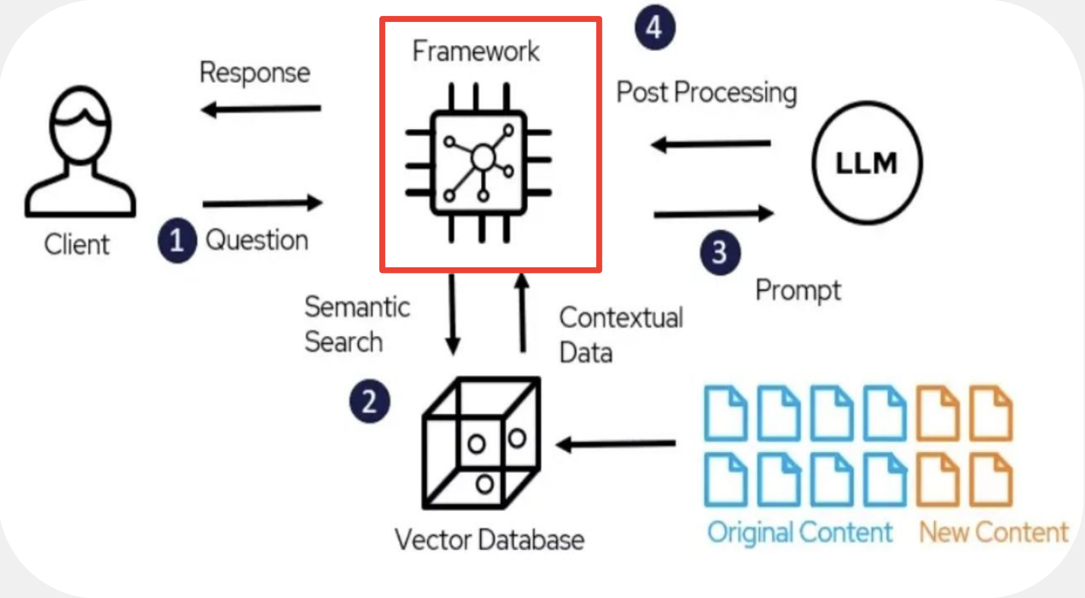
AI
@DevFest

Skeleton Code

https://github.com/rashikakarki/rag-devfest-2024

Google
Developer
Groups

Orchestrator : LangChain

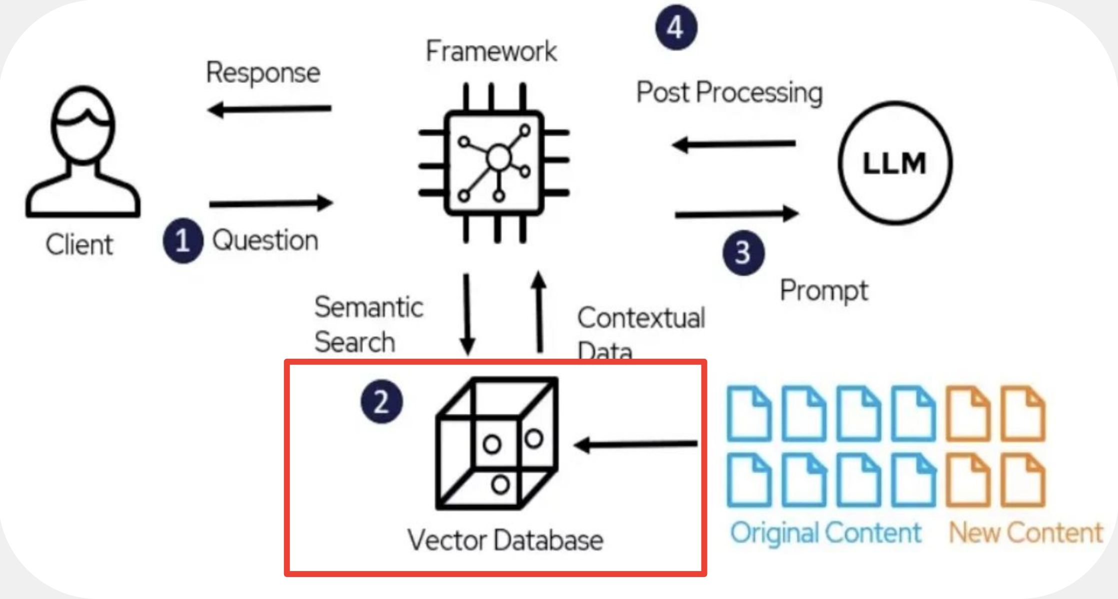


- Orchestrator (LangChain) manages the workflow/pipeline by connecting the retriever, vector database, and generator to deliver accurate and contextual responses.
- **LangChain** is a powerful **orchestration framework** to build applications that integrate LLMs, tools, and external data sources.
- Acts as the glue connecting various components like retrievers, generators, and databases.

Why LangChain?

- Simplifies workflows by providing prebuilt modules for retrieval, generation, and chaining tasks.
- Enables seamless orchestration of multiple tools for enhanced functionality.

Vector Database

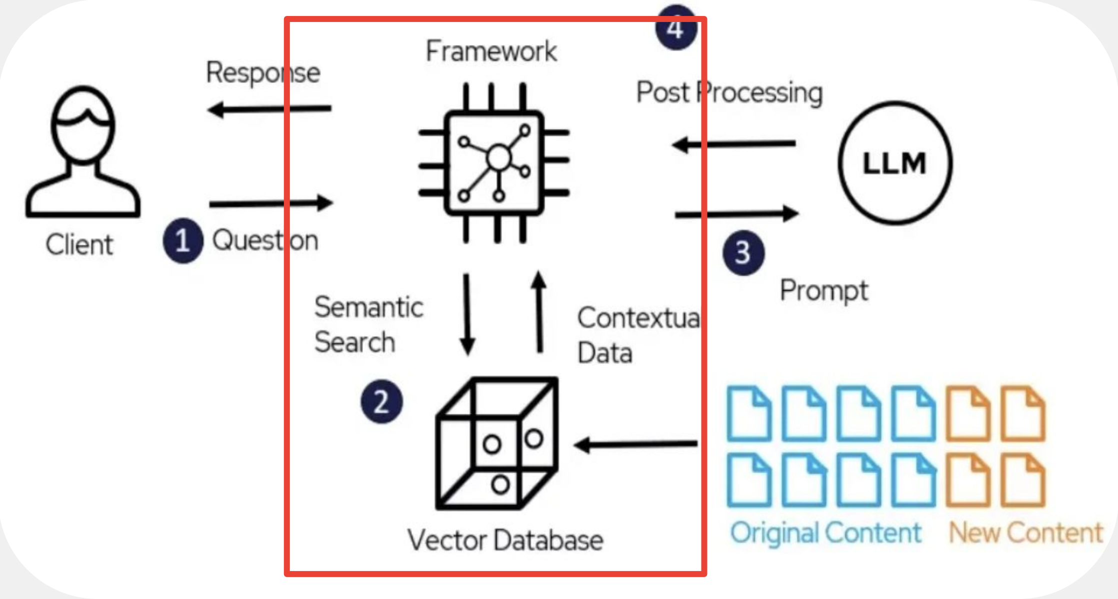


- A **database** that stores and searches data as numerical vectors, optimized for similarity-based retrieval.

How Does It Work?

- Step 1: Document Embedding
- Step 2: Store Vectors in Database
- Step 3: Query Embedding
- Step 4: Similarity Search
- Step 5: Retrieve Relevant Documents

Retriever

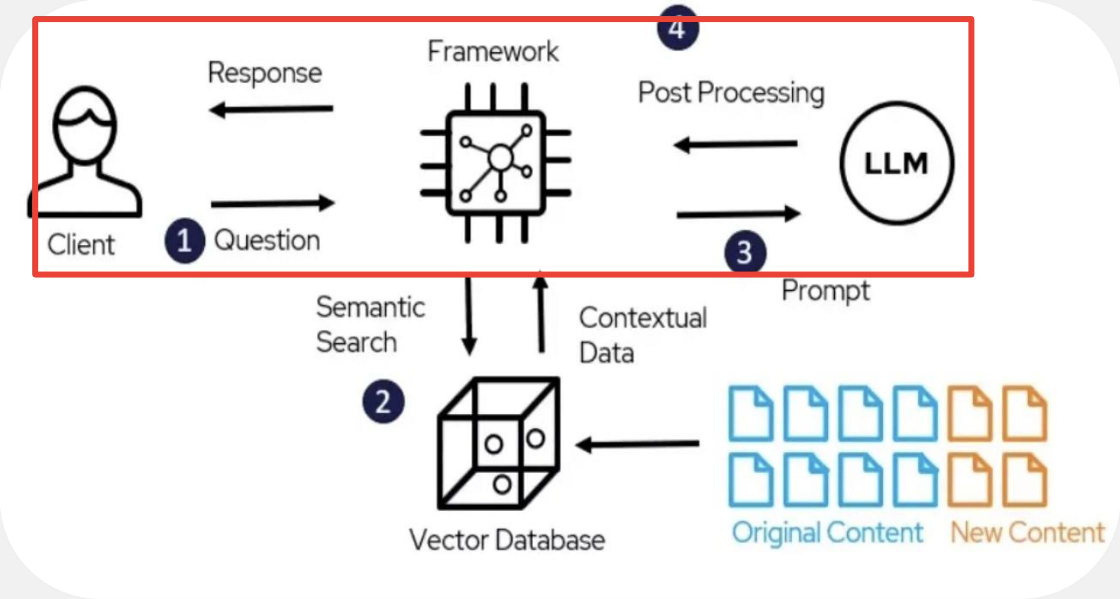


- A **retriever** fetches the most relevant information from a knowledge base or external data source based on the query.
- Bridges the gap between the user query and the knowledge base by focusing on **relevance**.

How Does It Work?

- Step 1: Embed the Query
- Step 2: Search the Vector Database
- Step 3: Retrieve Top Results
- Step 4: Provide Context to Generator

Generator



- Powered by LLMs like Gemini to generate human-like, contextual responses.
- Takes the query and the retrieved context as inputs.
- Outputs coherent and relevant answers based on the provided information.

How Does It Work?

- Step 1: Input Query
- Step 2: Receive Retrieved Context
- Step 3: Combine Query and Context
- Step 4: Generate Response

Additional Resources

- Beyond Word Embedding: Document Embedding (link)
- Explanation of RAG by DeepLearning.AI (link)
- Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (link)
- RAFT: Adapting Language Model to Domain Specific RAG (link)
- Build a RAG by langchain (link)

AI
@DevFest

Feedback Survey

🌐 https://forms.gle/yCLRC5WKRQrFmF116

Google
Developer
Groups