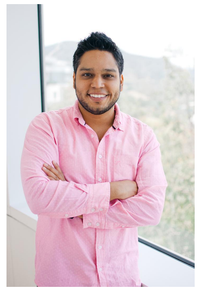
Google Developers

Experts

# Building Smart Apps with Your Data: A Firebase Genkit RAG Deep Dive



Juan Guillermo Gómez
GDE Firebase & GCP & Kotlin & AI/ML
@jggomezt

# Juan Guillermo Gómez

- Co-leader and co-founder of GDG Cali.
- Tech Lead at **WordBox** & Founder DevHack.
- Google Developer Expert (GDE) in Firebase & GCP & AI/ML
- BS in System Engineering and a MS in Software Engineering.
- @jggomezt
- devhack.co

SCAN ME

Youtube

Experts

# Networking


Linkedin


X


Instagram


Onlyfans

# Workshop Genkit

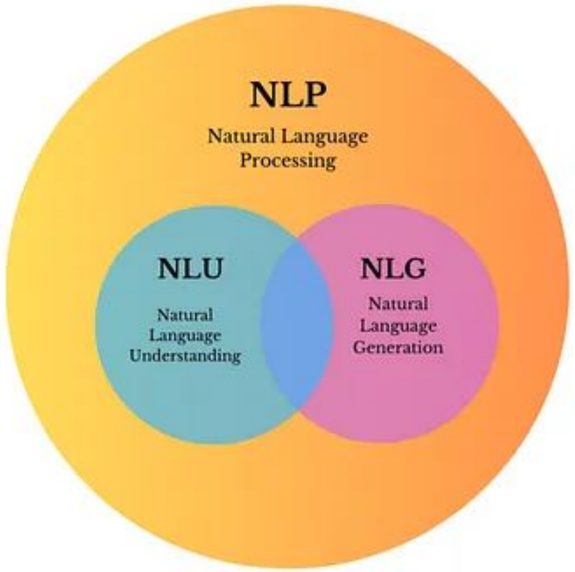
# Agenda

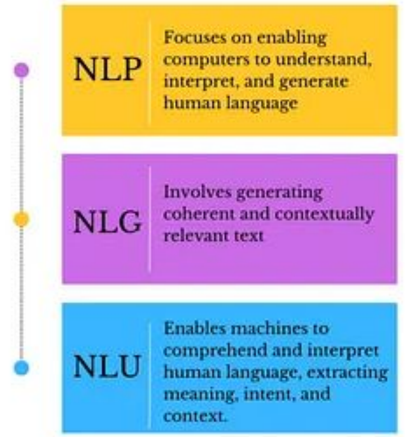
- LLMs
- Limitations of LLMs
- Firebase Genkit
- What is RAG?
- Demo

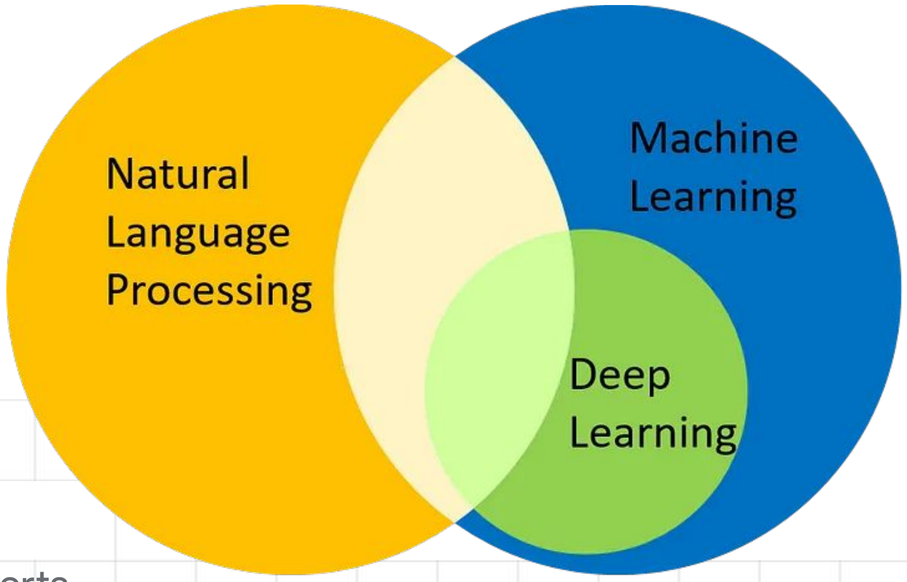

Experts

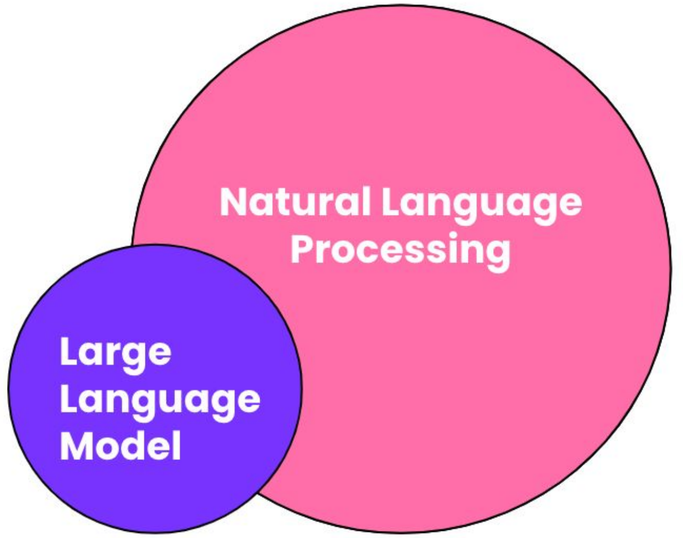
# Now, let's go deeper...

# Large Language Models (LLM)



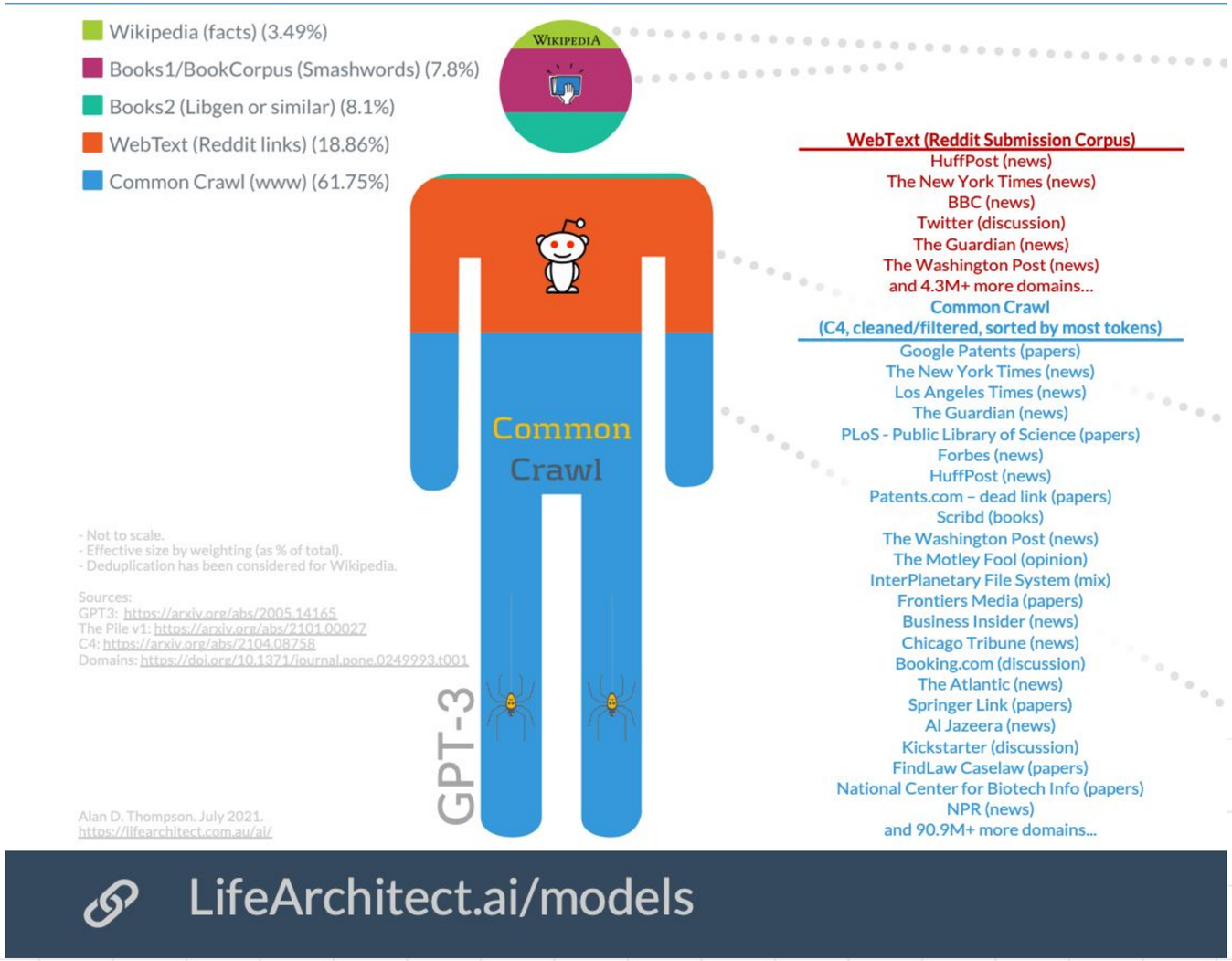

Experts

https://medium.com/@vipra_singh/llm-architectures-explained-nlp-fundamentals-part-1-de5bf75e553a

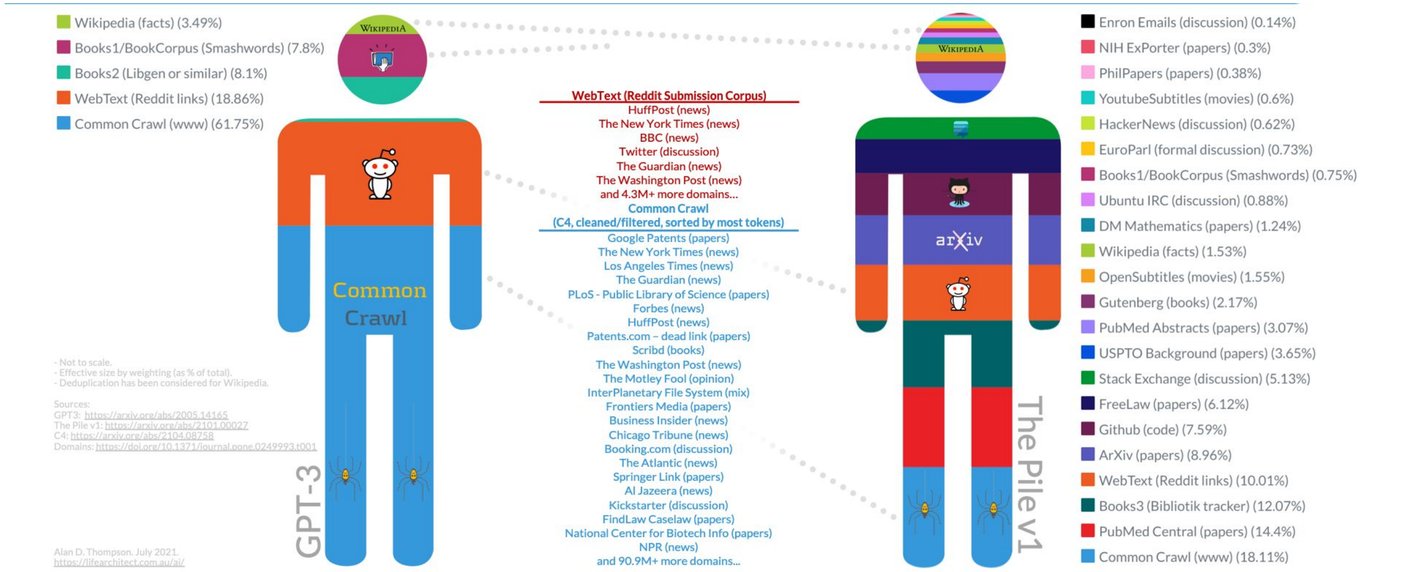
# What are LLMs?

- A large language model (LLM) is a deep learning algorithm (≅ 100 million of parameters) capable to process and understand text, equipped to summarize, translate, predict, and generate text to convey ideas and concepts.

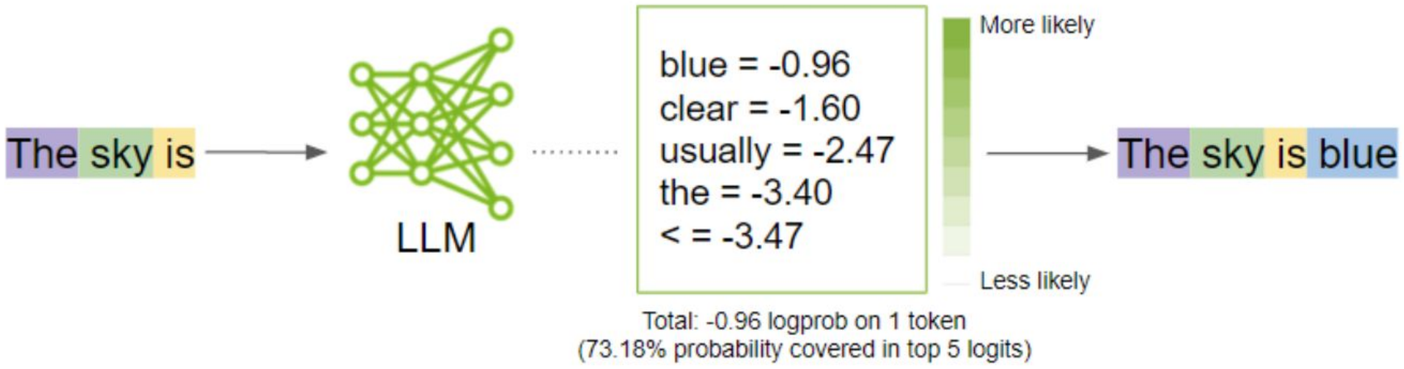
- Large Language Models (LLMs) leverage extensive datasets to identify linguistic patterns and gain a nuanced understanding of written language.

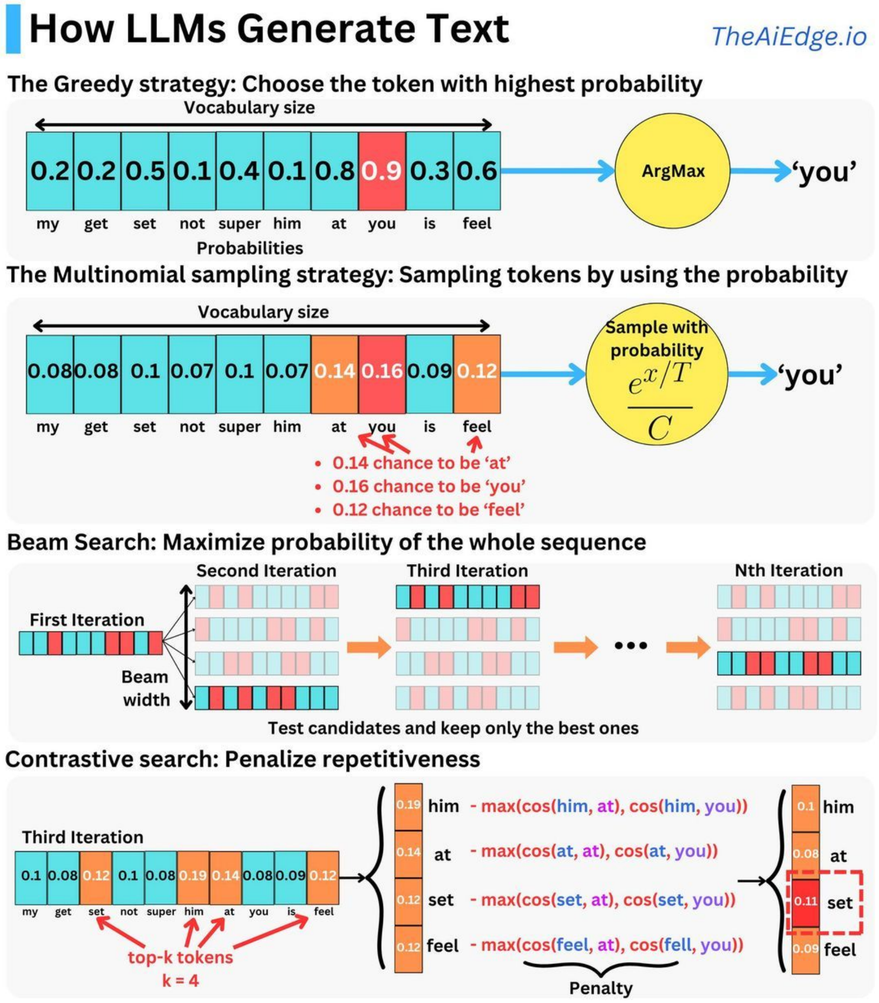


# How Do Large Language Models (LLMs) Work?




LifeArchitect.ai/models




Credits to: https://newsletter.theaiedge.io/about

# LLMs


Top 15 LLM Terms You Need to Know
1 Transformers
Self-attention to analyze relationships between words, enabling a deeper understanding of sentences
2 Token
Basic units of text an LLM processes, like words or sub-words.
3 Chunking
Breaking down text into smaller, manageable segments for LLM to analyze.
4 Indexing
catalog for the massive datasets for efficient retrieval
5 Embedding
Represent words in numerical code and such that lets the LLM understand their relationships to each other.
6 Vector Search
Helps LLMs find similar information within their vast datasets using embeddings
7 LLM Agent
In Agent LLM is the central processing unit, orchestrating the sequence of actions required to fulfill a task
8 Vector Database
Stores embeddings allowing for efficient vector search
9 Prompt Engineering
Art of crafting clear and concise instructions for the LLM to achieve the desired outcome
10 Shot Learning
How much instruction an LLM needs to learn a new task.
Zero-Shot, One-Shot, N-Shot
11 Fine Tuning
Training a smaller model on top of a larger one, focusing on a specific task while keeping resource usage in check.
12 Indexing
Machines that can think and learn like humans
13 RAG
RAG teams up large language models with external knowledge bases for more accurate and up-to-date responses.
14 MoE
Allows an LLM to leverage multiple smaller expert models for improved performance on specific tasks
15 LoRA
technique for compressing large LLM models, making them smaller and faster to run on devices
@pvergadia
Follow for more!

# LLM


Applications of
Natural Language Processing
Language translation
Email filtering
Smart assistant
Sentiment Analysis
Document analysis
Natural Language Processing
Chatbots
Online searches
Social media monitoring
Predictive text
Automatic summarization

# From LLMs to AI Agents




https://www.mongodb.com/pt-br/library/resources/ai-agents?x=inokiPMs

# Firebase Genkit

- Genkit is a framework designed to help you build AI-powered applications and features
- It provides open source libraries for Node.js and Go, plus developer tools for testing and debugging

Firebase Genkit

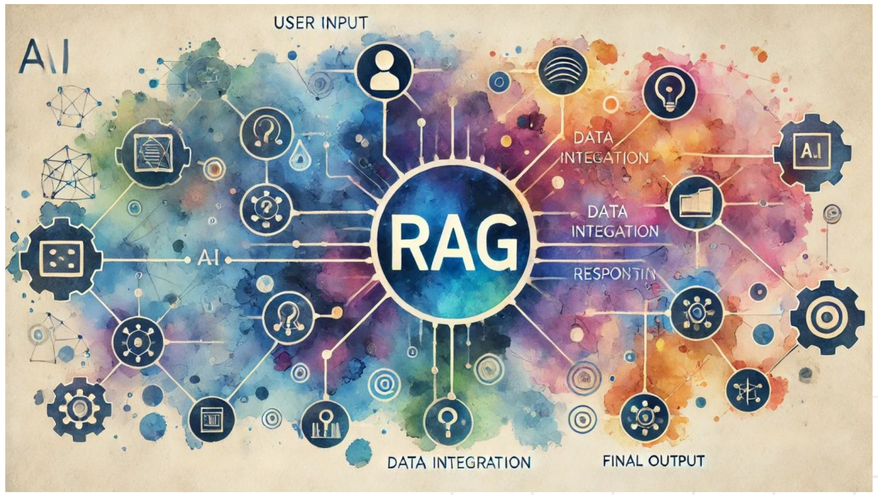
| | |
|---|---|
| Unified API for AI generation | Use one API to generate or stream content from various AI models. Works with multimodal input/output and custom model settings. |
| Structured output | Generate or stream structured objects (like JSON) with built-in validation. Simplify integration with your app and convert unstructured data into a usable format. |
| Tool calling | Let AI models call your functions and APIs as tools to complete tasks. The model decides when and which tools to use. |
| Chat | Genkit offers a chat-specific API that facilitates multi-turn conversations with AI models, which can be stateful and persistent. |
| Agents | Create intelligent agents that use tools (including other agents) to help automate complex tasks and workflows. |
| Data retrieval | Improve the accuracy and relevance of generated output by integrating your data. Simple APIs help you embed, index, and retrieve information from various sources. |

# Limitations of LLM

- LLM responds with non-current data
- LLM responds with data that they were trained
- “Hallucinations” are unreliable responses
- Responses senseless
- LLMs can be Limited by their internal knowledge

# RAG

- Retrieval Augmented Generation
- RAG is a technique for augmenting LLM knowledge with additional data
- RAG incorporates data from external sources allowing LLMs to access relevant information in real-time
- No need to fine-tune or retrain a model
- LLMs can answer questions about specific source information



# Benefits of RAG

## Improved Accuracy

More precise and reliable answers based on objective information.

## Reduced Hallucinations

Fewer absurd or irrelevant results

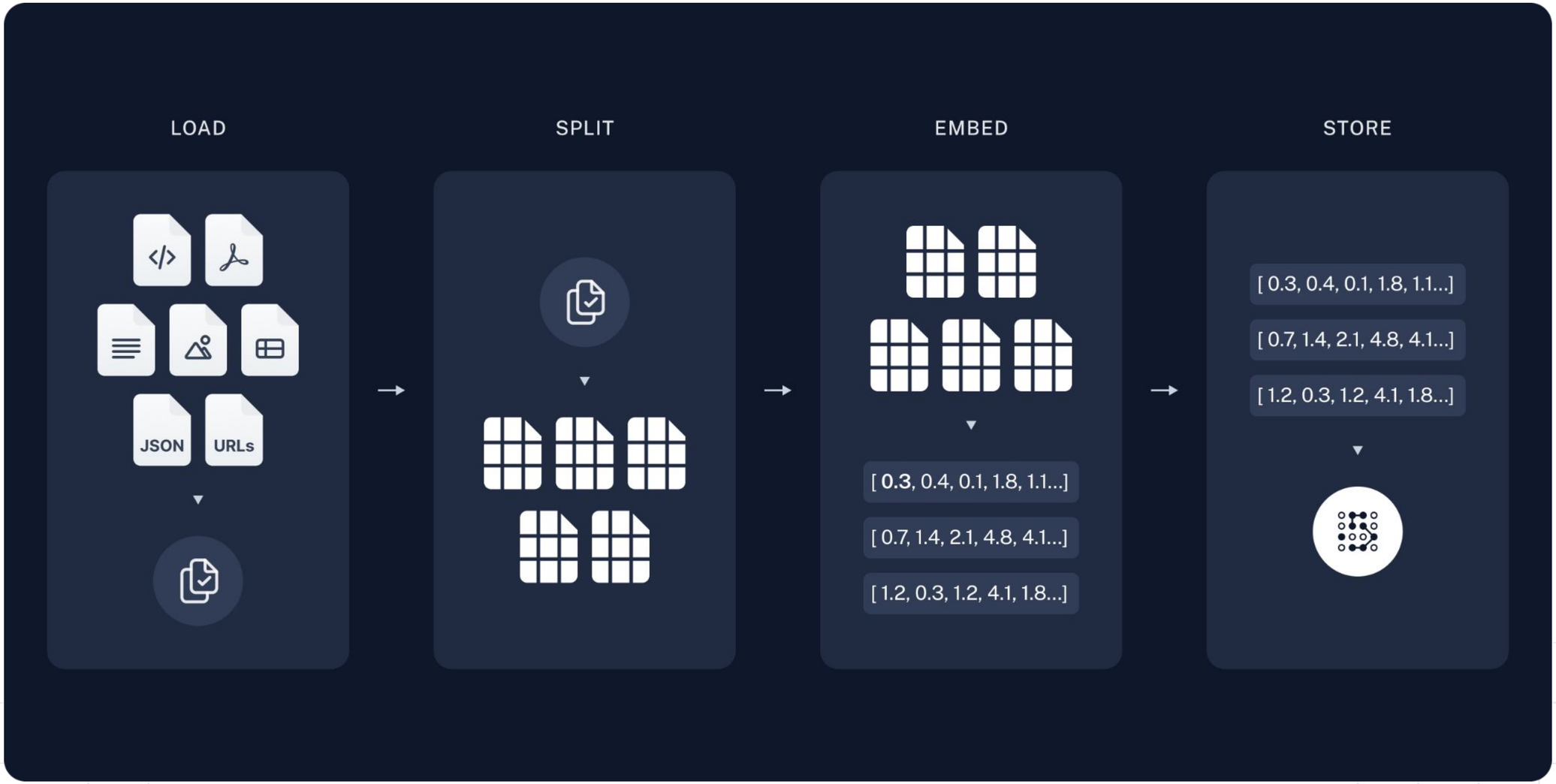
## Updated Information

Answers are as current as the stored data, overcoming the model's knowledge limitations.

## Explainability

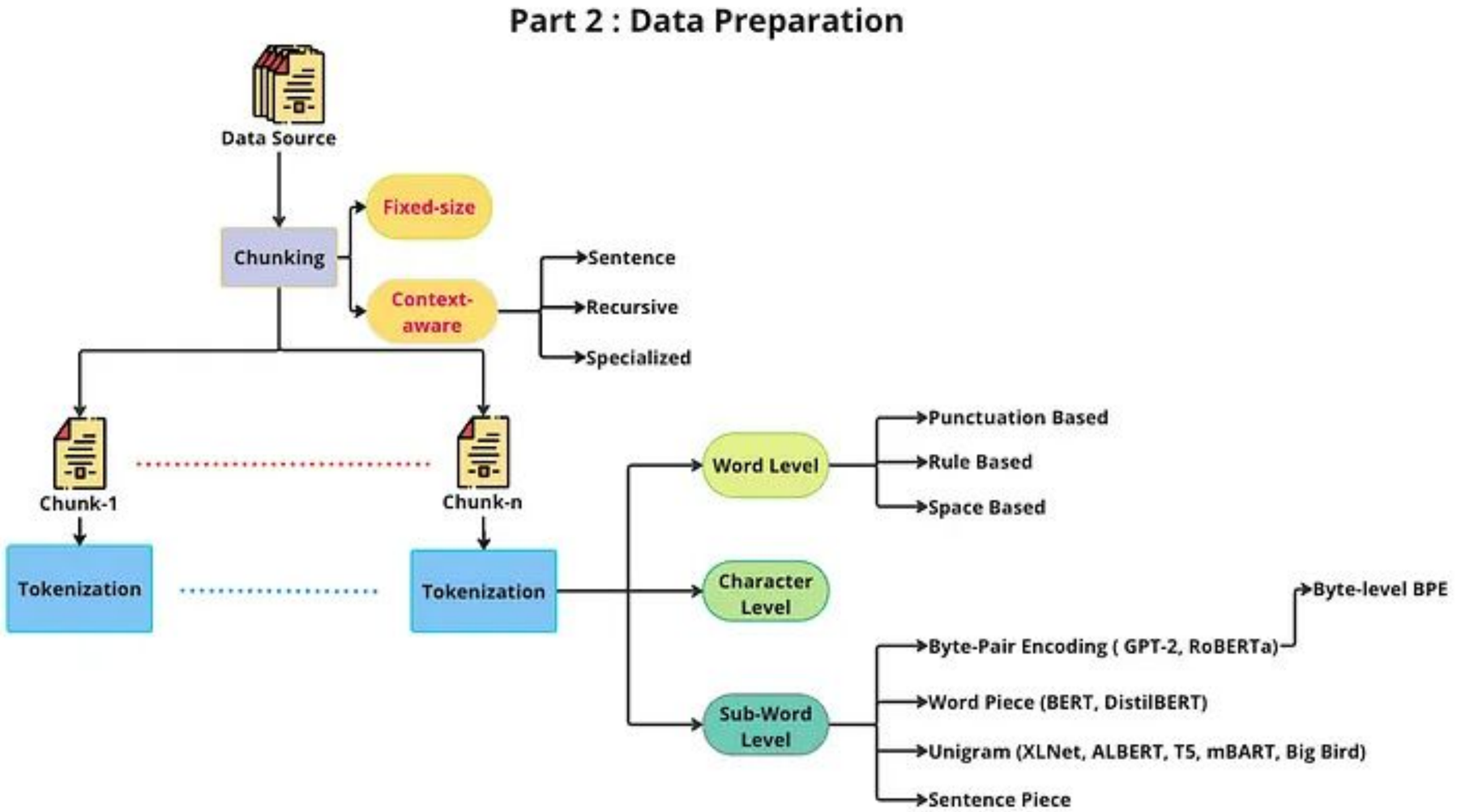
Retrieved sources help explain how the LLM generated its answers

# RAG - Data Preparation

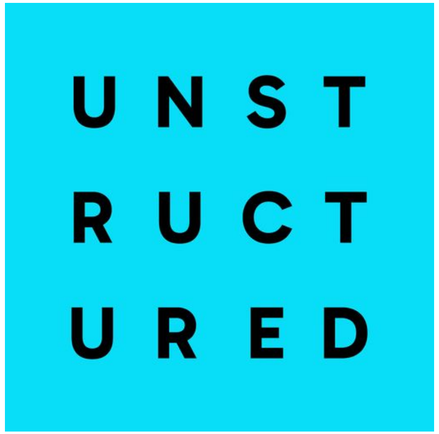



https://python.langchain.com/docs/tutorials/rag/

# Data - Preparation



# Data - Preparation

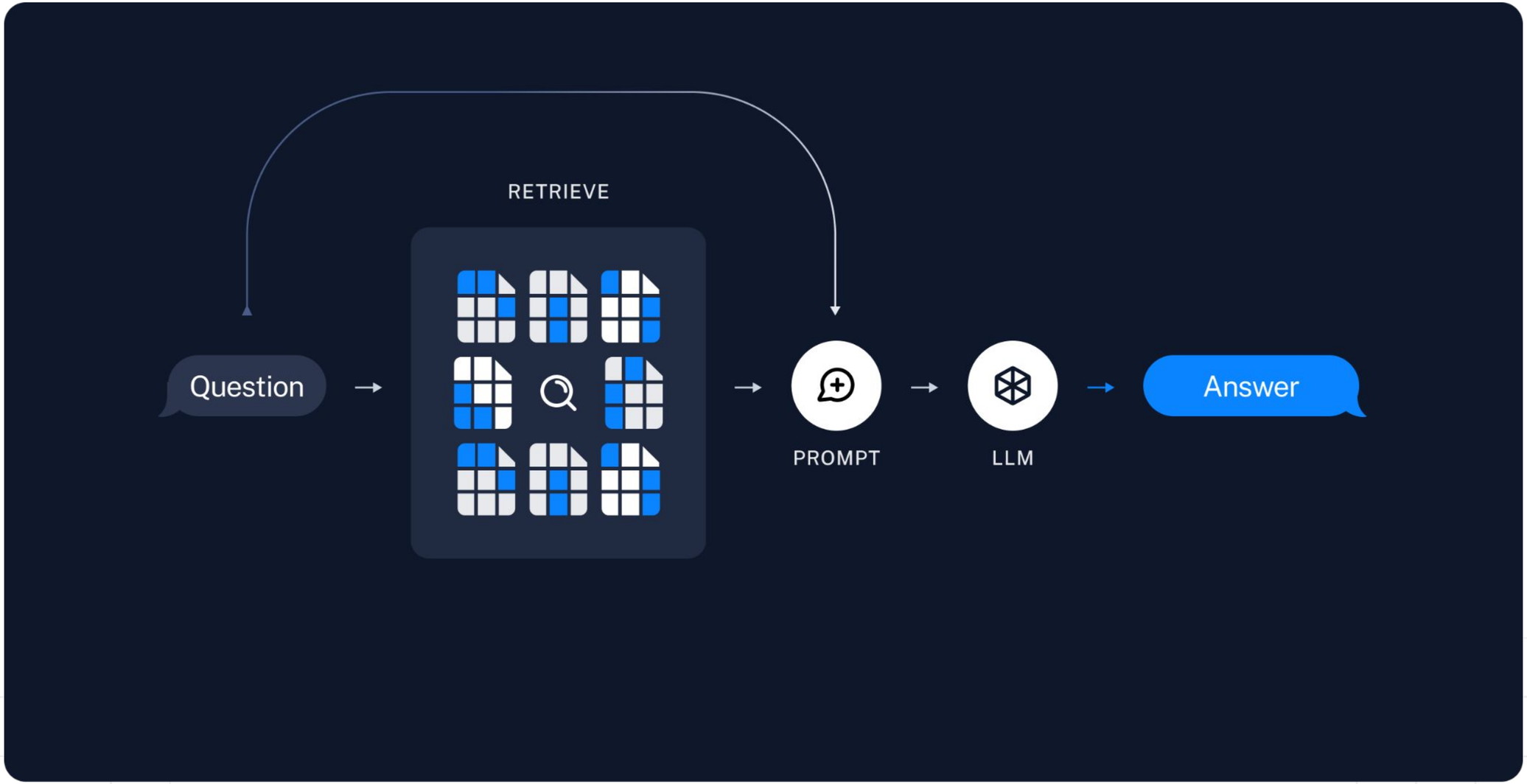


# Vector Database



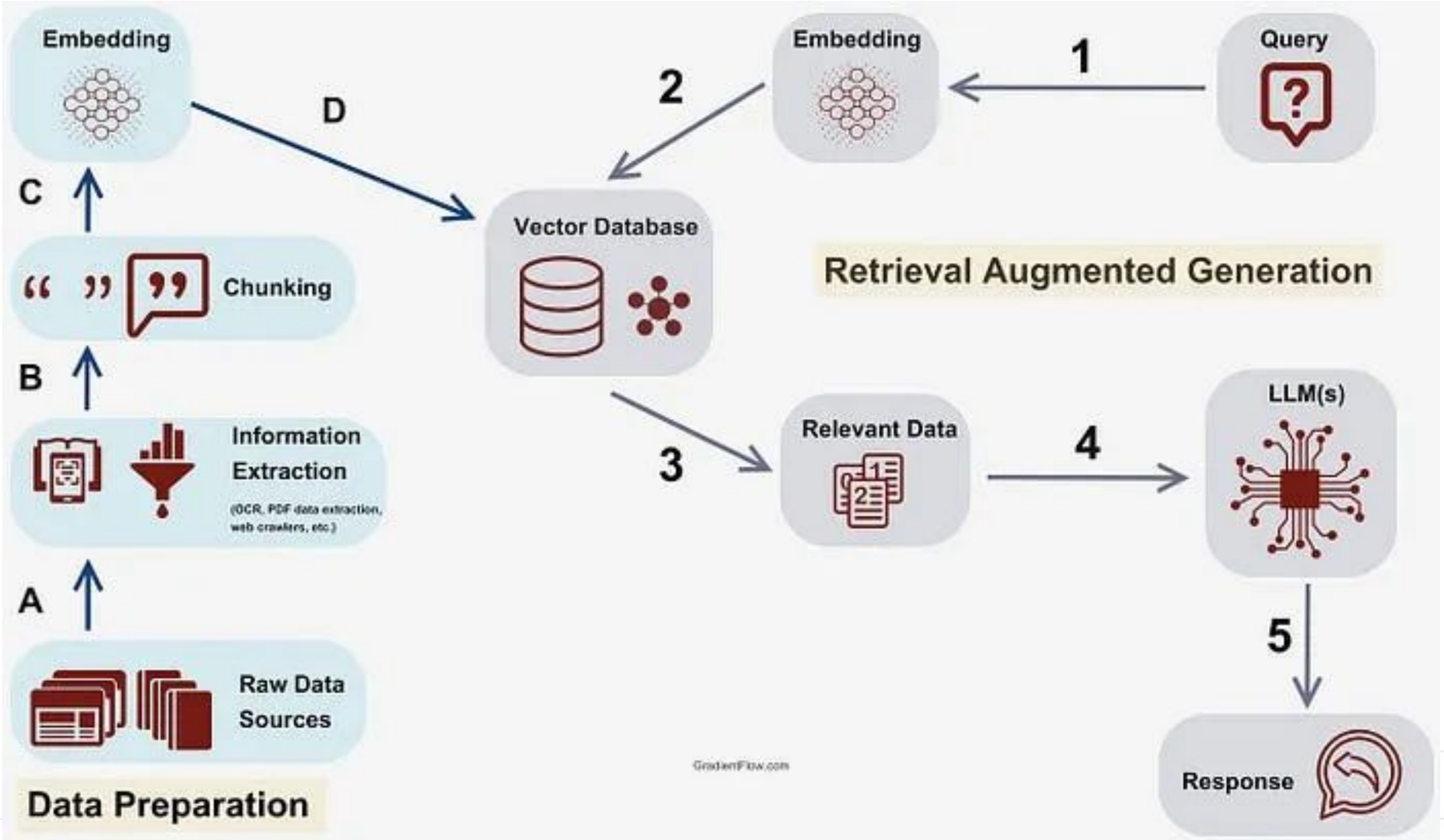


# RAG Application




https://python.langchain.com/docs/tutorials/rag/

# RAG Application




https://medium.com/@vipra_singh/building-llm-applications-introduction-part-1-1c90294b155b

# LLMs

OpenAI

ANTHROP\C

Gemini


Hugging Face

# DEMO

# Networking


Linkedin


X

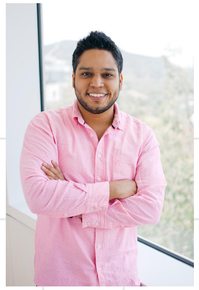

Instagram


Onlyfans

Google Developers

Experts

# Thank You!
# Gracias !



Juan Guillermo Gómez
@jggomezt

SCAN ME