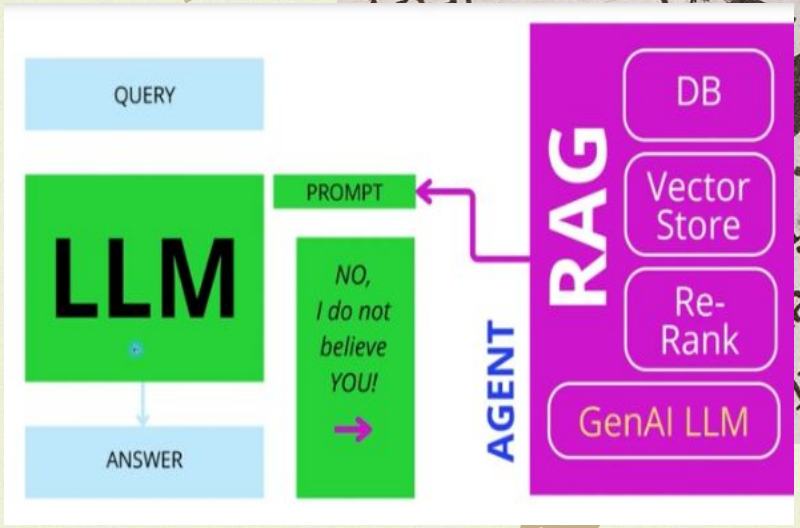
Building an AI-Powered App: Learning AI through Practical Applications

LLM + RAG



Project goal

Create an AI-powered app that extracts structured data (e.g., titles, summaries, authors, publication years) from research papers.

Source - [1]

Real-world Applications

Business reports, credit card statements, or even invoices.

Why this project matters

01

Unstructured data is everywhere: PDFs, invoices, reports, and research papers.

02

Manual extraction of key information is **time-consuming, tedious,** and **error-prone**.

03

AI provides a **smarter**, **faster**, and **more reliable** alternative.

TABLE OF CONTENTS

01

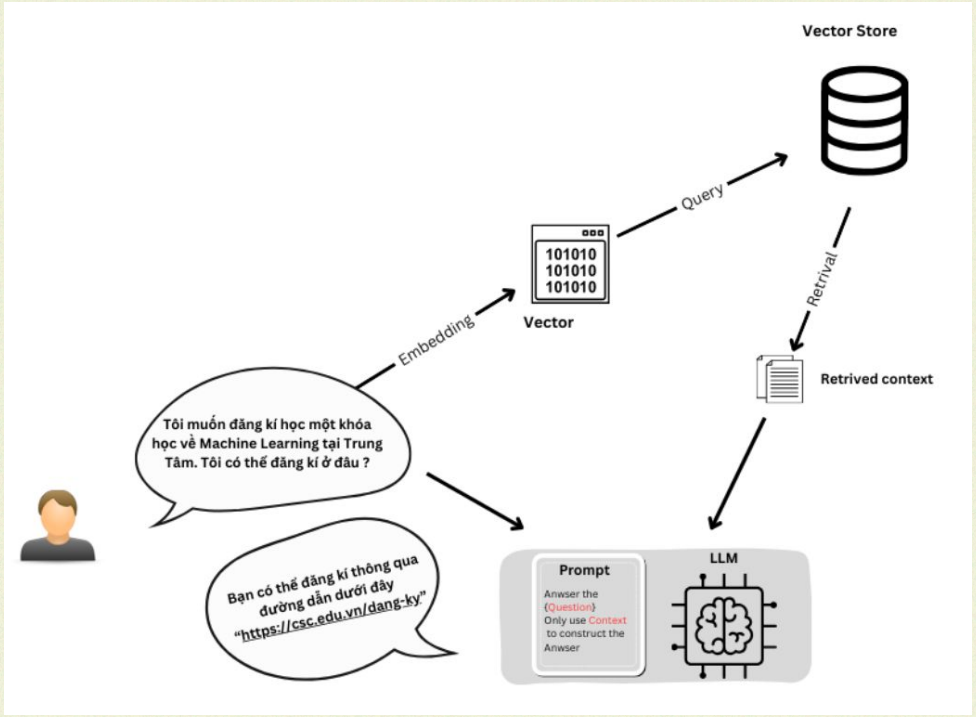
Key Concepts

LLM (Large Language Models), LangChain, và RAG (Retrieval-Augmented Generation)

- **LLM (Large Language Models):**
  - Là các mô hình AI mạnh mẽ như GPT-4, được huấn luyện trên lượng dữ liệu khổng lồ để hiểu và tạo ra ngôn ngữ tự nhiên.
  - Ứng dung: Trả lời câu hỏi tạo nội dung, tóm tắt văn bản, và nhiều hơn nữa.
  - Han chế: LLM có thể "bịa" thông tin (**hallucination**) nếu không được cung cấp ngữ cảnh chính xác.
- **LangChain**:
  - Là một **framework** giúp tích hợp LLM vào các ứng dụng thực tế.
  - Vai trò: Kết nối LLM với các công cụ khác như cơ sở dữ liệu, API, hoặc các tài liệu cụ thể.
  - Lơi ích: Giúp xây dựng các ứng dụng AI phức tạp dễ dàng hơn, như tự động hóa quy trình làm việc hoặc xử lý dữ liệu.
- **RAG (Retrieval-Augmented Generation)**:
  - Là một phương pháp kết hợp **truy xuất thông tin** (retrieval) từ tài liệu với khả năng **tạo nội dung** (generation) của LLM.
  - Cách hoat đông**:**
    - Truy xuất các phần tài liệu liên quan từ cơ sở dữ liệu (retrieval).
    - Sử dụng LLM để tạo câu trả lời dựa trên thông tin đã truy xuất (generation).
  - Lơi ích: Đảm bảo câu trả lời chính xác, đáng tin cậy, và có thể trích dẫn nguồn.

Mối liên hệ:

- LLM là "bộ não" xử lý ngôn ngữ.
- LangChain là "công cụ" giúp kết nối LLM với các tài liệu và cơ sở dữ liệu.
- RAG là "phương pháp" sử dụng cả hai để tạo ra câu trả lời chính xác và đáng tin cậy từ dữ liệu cụ thể.



Source - [6]

Vai trò của Facebook AI (Meta AI):

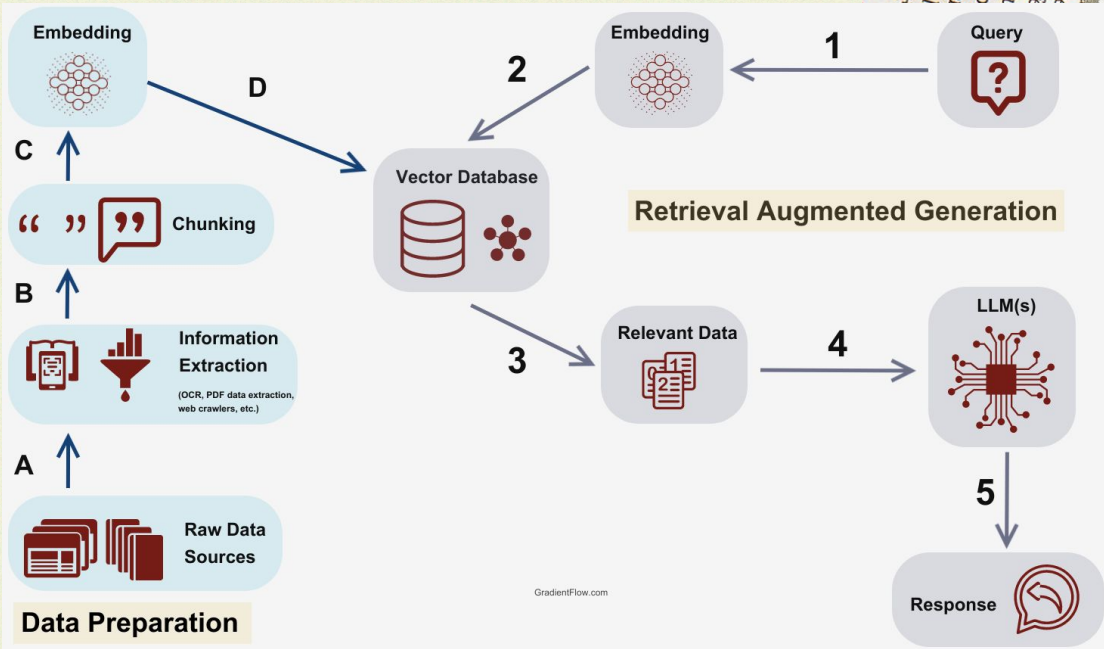
- Facebook AI (nay là Meta AI) đã công bố nghiên cứu về RAG vào năm 2020 trong đó họ trình bày cách kết hợp **retrieval** (truy xuất thông tin) với **generation** (tạo văn bản) để cải thiện độ chính xác và tính đáng tin cậy của các mô hình ngôn ngữ.
- Nghiên cứu của họ tập trung vào **việc sử dụng cơ sở dữ liệu vector để lưu trữ và truy xuất thông tin** sau đó sử dụng LLM để tạo câu trả lời dựa trên thông tin đã truy xuất.

02 RAG Components

Cách hoạt động của RAG

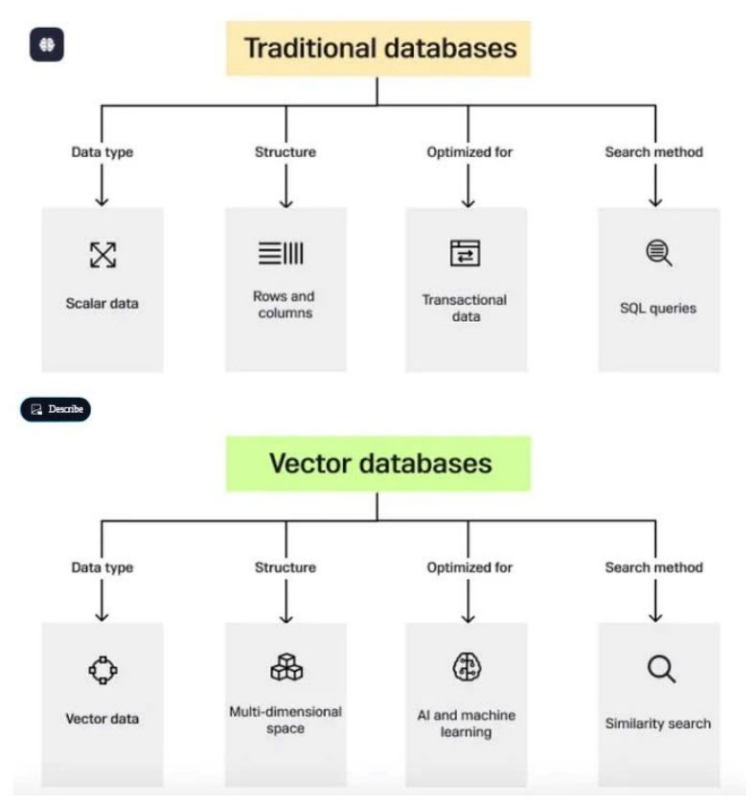
RAG hoạt động qua 3 bước chính:

1. Truy xuất thông tin (Retrieval):
   - Tìm kiếm các đoạn văn bản liên quan từ cơ sở dữ liệu vector dựa trên câu hỏi của người dùng.
   - Sử dụng các công cụ như **ChromaDB** hoặc **FAISS** để lưu trữ và truy xuất embeddings (vector hóa văn bản).
2. Tăng cường ngữ cảnh (Augmentation):
   - Kết hợp các đoạn văn bản đã truy xuất với câu hỏi của người dùng để tạo ngữ cảnh đầy đủ.
3. Tạo câu trả lời (Generation):
   - Sử dụng LLM (như GPT-4) để tạo câu trả lời dựa trên ngữ cảnh đã được tăng cường.

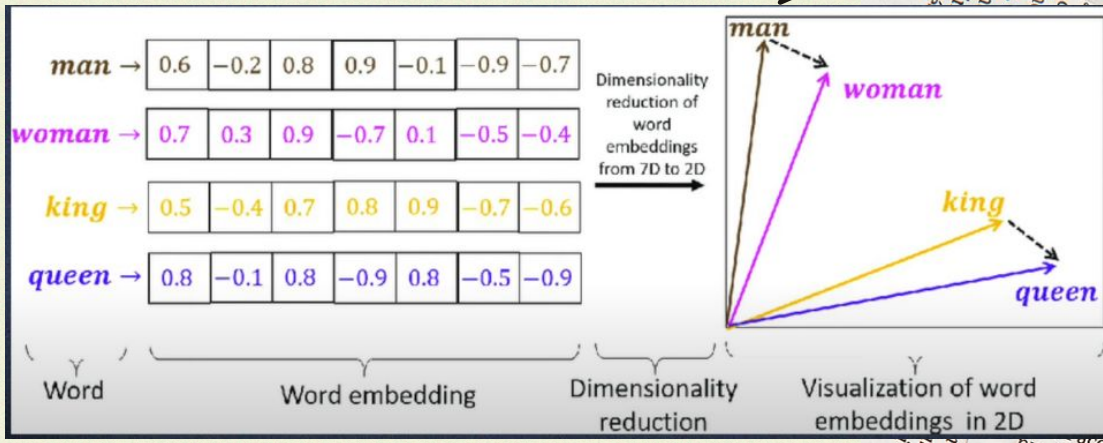


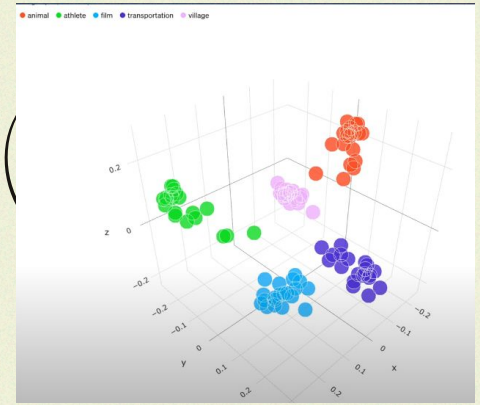
Source - [3]

02 RAG Components



Source - [6]





Source - [1]

Real-World Applications

Automated Document Processing

Extract key details from contracts, invoices, PDFs, and reports.

Customer Support Automation

Provide accurate answers to customer queries using company-specific documents.

Research Summarization

Summarize lengthy reports or papers into actionable insights.

Data Organization

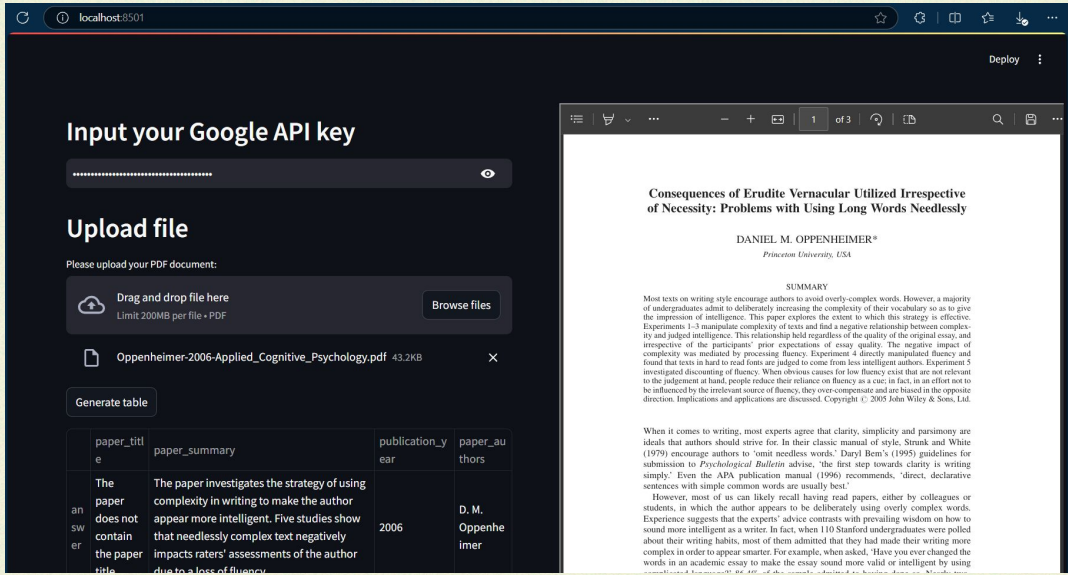
Transform unstructured data into structured Excel or database formats.

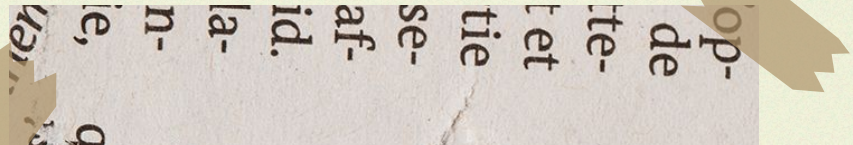
03

What Did We Build?

Summary of the AI-Powered App

- Extracts structured information (e.g., title, authors, summary) from research papers.
- Outputs data in a user-defined JSON or table format.
- Provides source citations for trust and transparency.
- User-friendly interface built with Streamlit.
- Deployed as a fully portable Docker container.

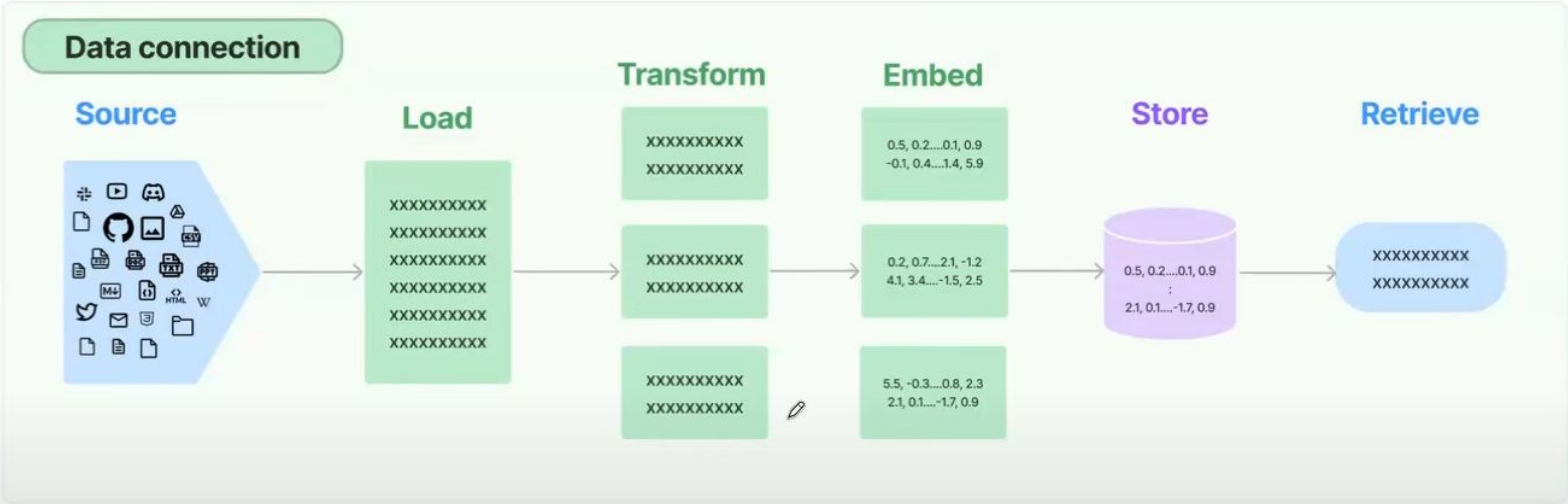




04

How Does the System Work?

❖ LangChain: Document Loaders



In some LLMs application (e.g: RAG...), we might have to utilize a source of documents.

Source - [2]

RESOURCES

[1] - Thu Vu - Extracting Structured Data From PDFs | Full Python AI project for beginners (ft Docker)

[2] - LLMs: Xây dựng ứng dụng RAG với LangChain (AIO2023)

[3] - Best Practices in Retrieval Augmented Generation

[4] - BankStatement-Data-Extractor

[5] - Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

[6] - Khám phá RAG - Hướng dẫn xây dựng chatbot với RAG

Thank you for your attention!