

RAG (Retrieval-Augmented Generation)

Hãy cùng tìm hiểu chi tiết về **RAG (Retrieval-Augmented Generation)** — một kỹ thuật rất quan trọng trong việc xây dựng hệ thống AI có khả năng trả lời chính xác dựa trên kiến thức bên ngoài, ví dụ như tài liệu nội bộ, cơ sở tri thức công ty, v.v.

1. RAG là gì?

RAG (Retrieval-Augmented Generation) là một **kiến trúc kết hợp giữa hai thành phần**:

- Retrieval (Truy xuất)**
→ Truy vấn và tìm kiếm thông tin phù hợp từ cơ sở dữ liệu, tài liệu, văn bản, v.v.
- Generation (Sinh văn bản)**
→ Dựa vào thông tin đã truy xuất, mô hình ngôn ngữ (LLM như GPT) sẽ **sinh ra câu trả lời tự nhiên và có ngữ cảnh**.

“ Mục tiêu: Giúp mô hình trả lời **chính xác và theo ngữ cảnh cụ thể**, thay vì chỉ dựa vào kiến thức đã được huấn luyện từ trước.

2. Tại sao cần RAG?

Mô hình LLM như GPT có **hạn chế**:

- Không biết thông tin mới hoặc riêng biệt (ví dụ: chính sách nội bộ, hướng dẫn sử dụng của công ty).
- Dễ “bịa ra” câu trả lời khi không chắc chắn.

RAG khắc phục điều đó bằng cách thêm một bước tìm kiếm thông tin thật, rồi mới trả lời.

3. Quy trình hoạt động của RAG

[Câu hỏi người dùng]

↓

[1] Truy vấn → Tìm văn bản liên quan trong cơ sở dữ liệu

↓

[2] Sinh → Đưa thông tin tìm được vào prompt → Mô hình tạo câu trả lời



[Trả lời chính xác và rõ ràng]

4. Ví dụ cụ thể về RAG

Tình huống:

Bạn có một tài liệu nội bộ hướng dẫn sử dụng phần mềm quản lý kho. Người dùng hỏi:

“Làm sao để kiểm kê tồn kho định kỳ?”

Bước 1: Truy xuất tài liệu liên quan

- Hệ thống tìm được đoạn văn bản trong tài liệu có nội dung:

“Để kiểm kê tồn kho định kỳ, người quản lý cần truy cập vào module 'Báo cáo kho', chọn chức năng 'Kiểm kê', và thực hiện quy trình đếm thực tế, sau đó đối chiếu với hệ thống.”

Bước 2: Sinh câu trả lời tự nhiên

Mô hình được "bơm" đoạn tài liệu vào prompt, và trả lời:

“Để kiểm kê tồn kho định kỳ, bạn hãy vào module 'Báo cáo kho', sau đó chọn 'Kiểm kê'. Thực hiện đếm hàng hóa thực tế, nhập số liệu và hệ thống sẽ đối chiếu với tồn kho ghi nhận trong phần mềm.”

Đây là câu trả lời sinh ra từ mô hình nhưng có **nội dung chính xác, không bịa đặt**, vì dựa trên văn bản có thật.

5. Các công cụ/công nghệ để triển khai RAG

Thành phần	Công cụ phổ biến
Truy xuất dữ liệu	FAISS, Weaviate, Elasticsearch
Tách đoạn văn	LangChain, Haystack
LLM trả lời	GPT-4, Claude, Mistral, LLaMA, v.v.

Thành phần	Công cụ phổ biến
Framework RAG	LangChain, LlamaIndex, Haystack

6. Một số ứng dụng thực tế của RAG

Ứng dụng	Mô tả
Hỗ trợ khách hàng (chatbot)	Trả lời câu hỏi từ khách dựa trên tài liệu công ty
Trợ lý tài liệu kỹ thuật	Giải thích hướng dẫn sử dụng, quy trình phức tạp
Tìm kiếm có ngữ cảnh	Cải thiện kết quả tìm kiếm với câu trả lời sinh ngôn ngữ tự nhiên
Tóm tắt báo cáo dài	Tìm đoạn liên quan rồi tóm tắt hoặc giải thích cho người dùng

Tóm tắt

Mục	Nội dung
❑ RAG là gì?	Kết hợp tìm kiếm thông tin và sinh văn bản
❑ Lợi ích	Trả lời chính xác, cập nhật, không bịa
❑ Hoạt động	Truy xuất → Tạo câu trả lời
❑ Công cụ	LangChain, FAISS, GPT, Haystack
❑ Ứng dụng	Trợ lý nội bộ, chatbot thông minh, tìm kiếm nâng cao

Tác giả: **Đỗ Ngọc Tú**
Công Ty Phần Mềm **VHTSoft**

Phiên bản #2
Được tạo 4 tháng 5 2025 05:18:54 bởi Đỗ Ngọc Tú
Được cập nhật 6 tháng 5 2025 09:52:34 bởi Đỗ Ngọc Tú