Mô hình **RAG (Retrieval-Augmented Generation)** là một kiến trúc AI đột phá, kết hợp những gì tốt nhất của hai thế giới: **khả năng tìm kiếm thông tin mạnh mẽ** và **khả năng tạo sinh văn bản sáng tạo**.

Hãy tưởng tượng bạn đang làm một bài kiểm tra "mở sách".

* Một **Mô hình Ngôn ngữ Lớn (LLM)** thông thường giống như một học sinh cực kỳ thông minh, cố gắng trả lời câu hỏi chỉ dựa vào những gì đã học và ghi nhớ. Đôi khi, học sinh này có thể "chế" ra câu trả lời nếu không chắc chắn.
* Một **mô hình RAG** cũng là một học sinh thông minh, nhưng trước khi trả lời, học sinh này được phép **mở sách (cơ sở kiến thức) ra, tìm đúng trang có thông tin liên quan, đọc nó, và sau đó dùng kiến thức vừa đọc để viết ra câu trả lời**.

Cách tiếp cận này giúp câu trả lời của AI không chỉ sáng tạo mà còn **chính xác, đáng tin cậy và có thể kiểm chứng được**.

**Cách hoạt động của RAG**

Mô hình RAG hoạt động theo hai giai đoạn chính:

**1. Retrieval (Truy xuất) 🔎**

Khi nhận được một câu hỏi hoặc yêu cầu từ người dùng, hệ thống sẽ không trả lời ngay. Thay vào đó, nó thực hiện các bước sau:

* **Vector hóa câu hỏi:** Chuyển câu hỏi của bạn thành một vector số bằng mô hình embedding.
* **Tìm kiếm trong Vector DB:** Dùng vector câu hỏi này để tìm kiếm trong cơ sở dữ liệu vector (đã được xây dựng từ tài liệu của bạn).
* **Lấy ngữ cảnh:** Hệ thống sẽ lấy ra những đoạn văn bản (chunks) có vector gần nhất, tức là liên quan nhất về mặt ngữ nghĩa với câu hỏi của bạn.

**2. Augmented Generation (Tạo sinh Tăng cường) ✍️**

* **Tăng cường Prompt:** Những đoạn văn bản liên quan vừa được truy xuất sẽ được "đính kèm" vào câu hỏi ban đầu của người dùng để tạo thành một prompt mới, đầy đủ ngữ cảnh hơn.
* **Tạo câu trả lời:** Prompt tăng cường này sau đó được gửi đến một Mô hình Ngôn ngữ Lớn (như GPT, Gemini). Dựa vào ngữ cảnh được cung cấp, LLM sẽ tạo ra một câu trả lời chính xác và phù hợp.

**Tại sao RAG lại quan trọng?**

* **Giảm thiểu "ảo giác" (Hallucination):** Vì câu trả lời dựa trên một nguồn kiến thức cụ thể, mô hình ít có khả năng bịa đặt thông tin.
* **Cập nhật kiến thức dễ dàng:** Thay vì phải huấn luyện lại toàn bộ mô hình (tốn kém), bạn chỉ cần cập nhật cơ sở kiến thức (Vector DB). LLM có thể truy cập thông tin mới ngay lập tức.
* **Tăng tính minh bạch và tin cậy:** Hệ thống có thể trích dẫn chính xác nguồn thông tin đã sử dụng để tạo ra câu trả lời, cho phép người dùng kiểm chứng.
* **Bảo mật và cá nhân hóa:** Cho phép các doanh nghiệp xây dựng các ứng dụng AI trên dữ liệu nội bộ, độc quyền của họ mà không cần chia sẻ dữ liệu đó ra bên ngoài.

**Ứng dụng thực tế**

* **Chatbot hỗ trợ khách hàng:** Trả lời các câu hỏi về sản phẩm dựa trên tài liệu hướng dẫn sử dụng.
* **Hệ thống hỏi-đáp nội bộ:** Giúp nhân viên nhanh chóng tìm kiếm thông tin trong các quy định, báo cáo của công ty.
* **Công cụ nghiên cứu:** Tóm tắt và trả lời các câu hỏi từ hàng ngàn tài liệu nghiên cứu khoa học.