

LongRAG và LightRAG

1. LongRAG là gì?

Định nghĩa:

LongRAG là phiên bản mở rộng của RAG để **xử lý các tài liệu dài** hơn thông qua:

- Truy xuất **nhiều đoạn dài**
- Kết hợp với mô hình **Long-context LLM** (như Claude, Gemini, GPT-4-128k...)

Ưu điểm:

- Phù hợp cho các tài liệu lớn như:
 - Luận văn
 - Báo cáo kỹ thuật
 - Tài liệu y tế, pháp lý

Cách hoạt động:

1. **Chia nhỏ tài liệu dài thành các đoạn lớn hơn thông thường** (ví dụ 1000-3000 tokens).
2. Truy xuất các đoạn liên quan nhất từ cơ sở dữ liệu vector.
3. Nạp vào LLM có thể xử lý ngữ cảnh dài để tạo ra câu trả lời chính xác.

Ví dụ thực tế:

“📄 Truy vấn: "Nêu các biện pháp kiểm soát rủi ro tài chính được mô tả trong phần 4 của báo cáo?"”

- LongRAG có thể truy xuất **phần 4 (có thể dài 2000 tokens)** và đưa thẳng vào LLM để trả lời.

2. LightRAG là gì?

Định nghĩa:

LightRAG là phiên bản RAG **nhẹ hơn, nhanh hơn và tiết kiệm chi phí hơn**, phù hợp với các ứng dụng thực tế cần độ phản hồi nhanh.

Ưu điểm:

- Sử dụng LLM nhỏ (ví dụ Mistral, LLaMA2, Phi-2...)
- Tối ưu hóa quy trình truy xuất bằng cách:
 - Giảm số lần truy vấn
 - Nén thông tin đầu vào
 - Trích lọc phần cốt lõi (summarization trước khi đưa vào LLM)

Ứng dụng:

- Bot tư vấn nhẹ trên website
- Ứng dụng di động AI trả lời câu hỏi từ cơ sở dữ liệu nhỏ

Cách hoạt động:

- Truy xuất đoạn ngắn hoặc tóm tắt nội dung từ tài liệu.
- Ghép vào prompt tối giản.
- Gửi đến LLM nhỏ → trả lời nhanh với chi phí thấp.

Ví dụ thực tế:

“ Truy vấn: “Địa chỉ công ty được nhắc đến trong hợp đồng ở đâu?”

- LightRAG chỉ cần tìm và trích dẫn đoạn có địa chỉ mà không phải xử lý toàn bộ văn bản dài.

3. So sánh nhanh LongRAG và LightRAG

Tiêu chí	LongRAG	LightRAG
Dữ liệu đầu vào	Văn bản dài	Văn bản ngắn hoặc đã được tóm tắt
Loại mô hình	LLM có context lớn (GPT-4, Claude)	LLM nhỏ (Mistral, LLaMA2, Phi-2)
Tốc độ xử lý	Chậm hơn	Nhanh hơn
Chi phí	Cao hơn	Thấp hơn
Độ chính xác	Rất cao với tài liệu dài phức tạp	Tốt cho thông tin đơn giản
Ứng dụng phù hợp	Legal, medical, research	Chatbot, mobile app, automation tools

4. Thực hành cơ bản (ý tưởng)

Giả sử bạn có tài liệu 100 trang PDF và muốn xây chatbot AI:

- LongRAG:** Dùng `LangChain` + `GPT-4` để xử lý toàn bộ nội dung chi tiết.

- **LightRAG:** Dùng `sentence-transformers` để trích xuất những đoạn có địa chỉ email, số điện thoại, rồi trả lời bằng `Mistral-7B`.

Tác giả: Đỗ Ngọc Tú
Công Ty Phần Mềm VHTSoft

Phiên bản #1

Được tạo 5 tháng 5 2025 03:19:45 bởi Đỗ Ngọc Tú

Được cập nhật 6 tháng 5 2025 09:52:34 bởi Đỗ Ngọc Tú