Retrieval-Augmented Generation System for Vietnamese Legal and Tax Documents

Đặng Trần Long
BLUE TECHNOLOGY
22520805@gm.uit.edu.vn

Ngày 20 tháng 9 năm 2025

Tóm tắt nội dung

Báo cáo này trình bày việc áp dụng LightRAG–một framework RAG open-source để xây dựng hệ thống truy xuất và tổng hợp thông tin liên quan đến thuế tại Việt Nam. Em đã tận dụng kiến trúc sẵn có của LightRAG (insert, embedding, indexing, entity–relation extraction, query, reranking, generation) và tuỳ biến các thành phần: cơ sở dữ liệu **Postgres** cho metadata & lịch sử hội thoại, vector store `Qdrant`, embedding `AITeamVN/Vietnamese-Embedding`, LLM `Qwen2.5-72B-Instruct`, cùng reranker `cohere-reranker-v3-multilingual`. Kết quả thực nghiệm cho thấy cấu hình trên đem lại chất lượng truy xuất cao, trích xuất Entity–Relationship tốt và câu trả lời có trích dẫn nguồn rõ ràng.

1 Giới thiệu

Retrieval-Augmented Generation (RAG) giúp giảm "ảo tưởng" của LLM bằng cách kết nối tex response với bằng chứng truy xuất được. Bài toán đặt ra là xây dựng hệ thống hỏi–đáp cho kho văn bản pháp luật Việt Nam (luật, nghị định, thông tư, quy định thuế), yêu cầu trả lời tự nhiên, chính xác và có trích dẫn.

Em lựa chọn LightRAG[1] nhờ kiến trúc mở, hỗ trợ đa backend lưu trữ, mô-đun hoá LLM/embedding/reranker và sẵn sàng tích hợp mô hình từ Hugging Face, openAI, Gemini hoặc vLLM, localhost qua Docker, Ollama.

2 Tổng quan

Các hướng tiếp cận hiện có gồm tìm kiếm ngữ nghĩa dựa trên vector, trích xuất cấu trúc (entity–relation) và trợ lý pháp lý dùng RAG. Với tiếng Việt, thách thức chính là embedding/LLM tối ưu hoá cho ngôn ngữ và ngữ cảnh pháp luật. LightRAG nổi bật ở việc kết hợp *vector retrieval* và *knowledge graph* trong một pipeline thống nhất, kèm khả năng hoán đổi backend (Postgres/pgvector, Faiss, `Qdrant`, Neo4j) và reranker.

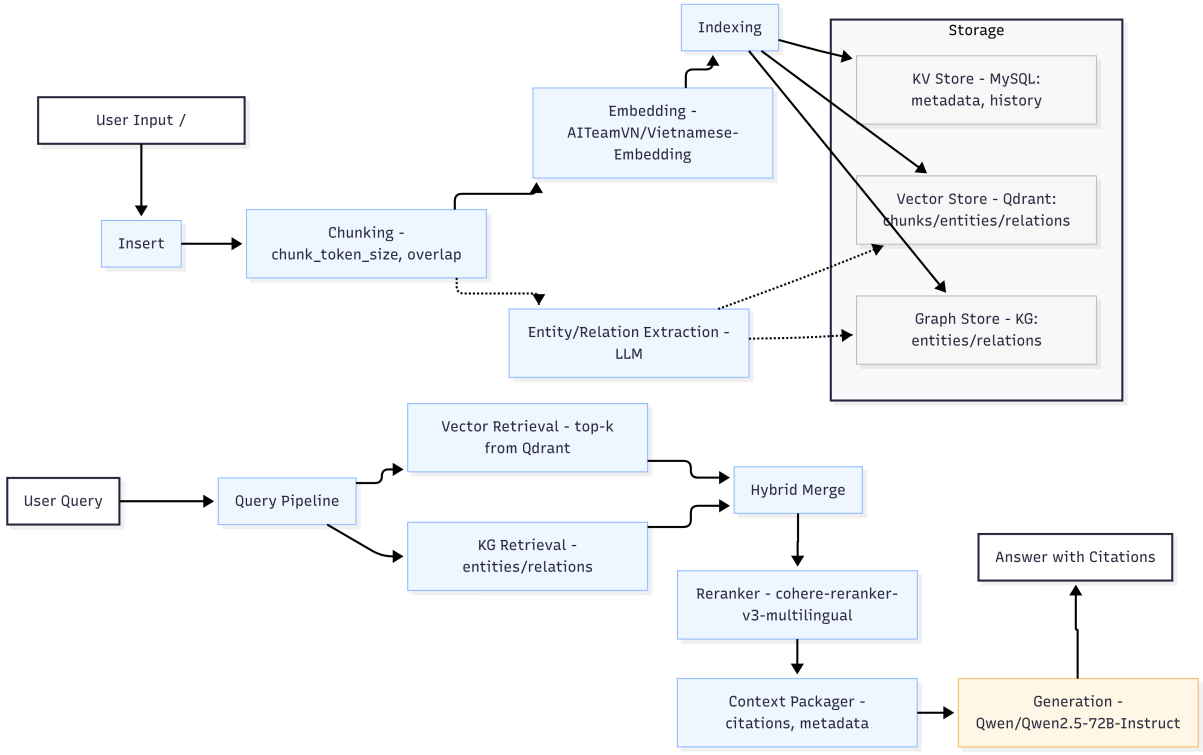
3 Phương pháp và kiến trúc

3.1 Pipeline áp dụng (bám sát LightRAG)

1. **Insert**: nạp văn bản vào hệ thống.
2. **Chunking**: chia nhỏ theo cấu hình (`chunk_token_size`, `overlap`).

[1] LightRAG: `https://github.com/HKUDS/LightRAG`

3. **Embedding**: dùng `AITeamVN/Vietnamese-Embedding`.
4. **Indexing**: lưu vector vào `Qdrant`; lưu *metadata & chat history* trong `Postgres`.
5. **Entity/Relation Extraction**: LLM trích xuất thực thể và quan hệ để xây dựng KG.
6. **Query**: nhận câu hỏi tự nhiên.
7. **Retriever + Reranker**: hybrid retrieval (vector + KG) và rerank bằng `cohere-reranker-v3-multilingual`.
8. **LLM Generation**: tổng hợp câu trả lời bởi `Qwen2.5-72B-Instruct`, kèm *citation* theo metadata.



Hình 1: Architecture LightRAG cho LAW–TAX

3.2 Tuỳ biến hệ thống

- **Vector Store**: `Qdrant` để bảo đảm hiệu năng truy vấn và khả năng mở rộng, self-host.
- **Database**: `Postgres` lưu metadata & lịch sử hội thoại (thống nhất toàn báo cáo).
- **Embedding**: thử nhiều lựa chọn, kết quả tốt nhất với `AITeamVN/Vietnamese-Embedding`.
- **LLM**: thử `gpt-oss:20b`, `VINAI/PhoGPT`, `kimi-k2-instruct`; chọn `Qwen2.5-72B-Instruct`.
- **Reranker**: `cohere-reranker-v3-multilingual` cải thiện đáng kể độ chính xác top-k.

4 Kết quả

4.1 So sánh LLM (định tính)

Bảng 1: Đánh giá rút gọn các LLM trên LAW–TAX (định tính)

Mô hình	Entity	Liên quan	Tổng thể
gpt-oss:20b	Trung bình	Trung bình	Khá
VINAI/PhoGPT	Trung bình	Thấp	Hạn chế
kimi-k2-instruct	Trung bình	Trung bình	Khá
Qwen2.5-72B-Instruct	Cao	Cao	Tốt nhất

4.2 Embedding

`AITeamVN/Vietnamese-Embedding` cho kết quả ổn định nhất trên văn bản pháp luật tiếng Việt; embedding đa ngữ phổ dụng cho độ liên quan thấp hơn ở các truy vấn tinh tế.

4.3 Hiệu năng hệ thống

- Độ trễ truy vấn ở mức chấp nhận được cho môi trường doanh nghiệp (sau khi bật reranker).
- Reranker nâng chất lượng top-k rõ rệt; câu trả lời luôn đính kèm *citation*.

5 Ứng dụng trong doanh nghiệp

- **Giảm thời gian tra cứu**: từ hàng giờ xuống vài phút với kết quả có nguồn dẫn.
- **Tăng minh bạch**: mọi câu trả lời gắn *metadata*/citation từ văn bản gốc.
- **Tiết kiệm chi phí**: self-host **Postgres** + `Qdrant` + mô hình open-source.
- **Mở rộng dễ dàng**: phục vụ đồng thời nhiều nhóm pháp chế–thuế.

6 Kết luận và hướng phát triển

Em đã áp dụng LightRAG cho LAW–TAX tiếng Việt với cấu hình *Qdrant + Postgres + Vietnamese-Embedding + Qwen2.5-72B* và thu được kết quả tốt về chất lượng truy xuất lẫn tính giải thích (citation). Hướng tiếp theo: tích hợp KG sâu hơn và tối ưu độ trễ.

Tài liệu tham khảo

- LightRAG Project: `https://github.com/HKUDS/LightRAG`
- AITeamVN Vietnamese Embedding: `https://huggingface.co/AITeamVN/Vietnamese_Embedding`
- Qwen2.5-72B-Instruct: `https://huggingface.co/Qwen`
- Qdrant: `https://qdrant.tech` Postgres: `https://www.postgresql.org`
- Cohere Reranker v3: `https://docs.cohere.com/docs/rerank`