

VieLegalRAG - Chatbot RAG văn bản luật Việt Nam

Trần Quốc Khanh

AILAB-UET

tháng 1 năm 2026

Mục lục

- ① Tổng quan và bối cảnh nghiên cứu
- ② Các thành phần và công nghệ nền tảng
- ③ Thiết kế kiến trúc và hiện thực hóa hệ thống
- ④ Thực nghiệm và đánh giá hiệu năng
- ⑤ Kết luận và hướng phát triển

1.1 Bối cảnh nghiên cứu

- Các mô hình ngôn ngữ lớn (LLMs) đạt nhiều tiến bộ vượt bậc
- Ứng dụng phổ biến trong hỏi–đáp và trợ lý ảo
- Pháp luật là lĩnh vực đặc thù:
 - Yêu cầu độ chính xác tuyệt đối
 - Không chấp nhận thông tin sai lệch

1.2 Hạn chế của LLM thuần túy

- Hallucination: sinh thông tin không tồn tại
- Tri thức bị đóng băng theo thời điểm huấn luyện
- Thiếu minh bạch, không trích dẫn được nguồn

So sánh LLM thuần túy và RAG

Bảng So Sánh: LLM Thuần và Giải pháp RAG				
Tiêu chí so sánh	Tính Xác Thực & Độ Tin Cậy	Tính Thời Gian & Cập Nhật	Phạm Vi Kiến Thức	Tính Minh Bạch & Nguồn Gốc
Vấn đề LLM thuần	<ul style="list-style-type: none">Ảo giác: Tự bịa ra thông tin không có thật.	<ul style="list-style-type: none">Kiến thức lỗi thời: Bị giới hạn bởi dữ liệu tại thời điểm huấn luyện (ví dụ: dữ liệu dừng ở năm 2020).	<ul style="list-style-type: none">Thiếu kiến thức nội bộ: Không biết về dữ liệu riêng, đặc thù của tổ chức.	<ul style="list-style-type: none">Thiếu minh bạch (Hộp đen): Không rõ câu trả lời đến từ đâu, không thể kiểm chứng.
Giải pháp RAG	<ul style="list-style-type: none">Trả lời dựa trên nguồn: Bám sát vào ngữ cảnh và tài liệu được cung cấp để đưa ra câu trả lời.	<ul style="list-style-type: none">Kiến thức cập nhật: Có khả năng truy xuất thông tin từ các tài liệu mới nhất theo thời gian thực.	<ul style="list-style-type: none">Kết nối Knowledge Base: Tích hợp trực tiếp với cơ sở tri thức riêng của doanh nghiệp/tổ chức.	<ul style="list-style-type: none">Trích dẫn nguồn: Cung cấp nguồn gốc thông tin rõ ràng, cho phép người dùng đối chiếu.

Hình 1.1. So sánh: LLM thuần túy và giải pháp RAG

Retrieval-Augmented Generation (RAG)

- Kết hợp hai thành phần:
 - Retrieval: truy xuất tài liệu liên quan
 - Generation: sinh câu trả lời bằng LLM
- LLM chỉ trả lời dựa trên ngữ cảnh truy xuất
- Giảm hallucination và tăng độ tin cậy

1.3 Mục tiêu và đóng góp

- Xây dựng hệ thống RAG cho pháp luật Việt Nam
- Đề xuất:
 - Hybrid Search
 - VietNam legal embedding
- Đánh giá trên Legal QA ALQAC benchmark

2.1 Công nghệ cốt lõi của hệ thống

Công Nghệ Cốt Lõi		
Thành phần	Công nghệ được chọn	Lý do lựa chọn
Embedding	Vietnam_legal_embeddings	Model chuyên biệt cho pháp luật VN, vector 768 chiều tối ưu tốc độ/bộ nhớ trên GPU Laptop.
Vector DB	Qdrant	Hỗ trợ mạnh mẽ Hybrid Search, hiệu năng cao và dễ dàng triển khai cục bộ (Local/Docker).
LLM	Qwen 2.5-3B (Ollama)	Mô hình mã nguồn mở nhẹ, chạy offline mượt mà trên RTX 4050, bảo mật dữ liệu tuyệt đối.
Sparse Search	BM25 + PyVi	Sử dụng tokenizer tiếng Việt (PyVi), bắt chính xác từ khóa chuyên ngành (VD: "Điều 128").
Fusion/Rerank	Reciprocal Rank Fusion (RRF)	Thuật toán hợp nhất kết quả tin cậy, cân bằng điểm số giữa Dense và Sparse mà không cần model nặng.

Hình 2.1. Các thành phần công nghệ cốt lõi của hệ thống

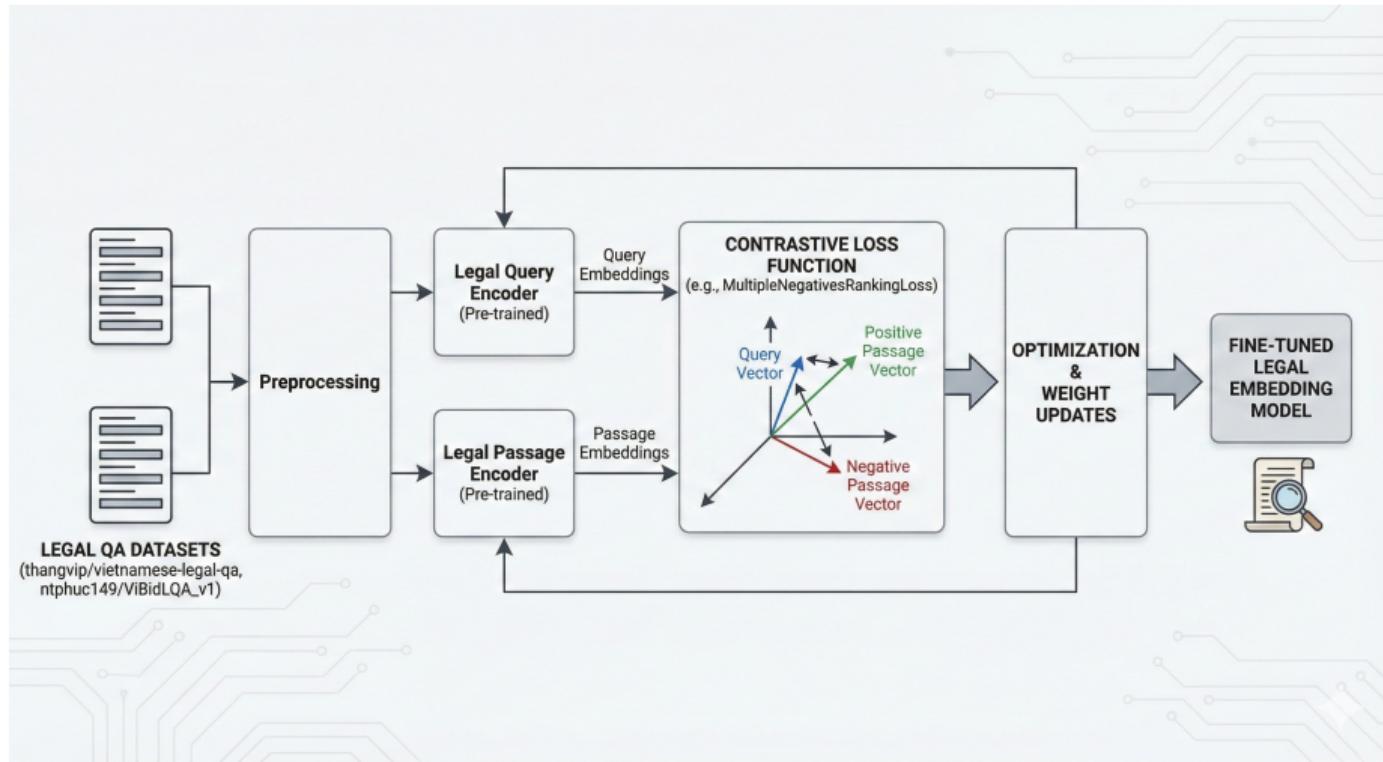
2.2 Tiền xử lý và biểu diễn dữ liệu

- Phân mảnh văn bản theo cấu trúc pháp luật:
 - Luật – Chương – Điều – Khoản
- Biểu diễn văn bản bằng vector 768 chiều

2.3 Finetune Embedding

- **Mô hình embedding Gốc:**
 - DEk21-hcmute-embedding
- **Dữ liệu huấn luyện:**
 - Tích hợp đa nguồn dữ liệu: *Vietnamese-Legal-QA*, *ViBidLQA*
 - Chuẩn hóa thành cặp (Query, Positive Document)
- **Chiến lược tối ưu hoá:**
 - Contrastive Learning với hàm mất mát MNRL
 - Cơ chế In-batch Negatives: (Q_i, P_i) vs. P_j ($j \neq i$)
 - Tác động: vector Q_i gần P_i , xa các văn bản khác → phân tách rõ ràng trong không gian n chiều

Quy trình finetune Embedding



2.4 Cơ Sở Dữ Liệu Vector Qdrant Và Thuật Toán HNSW

- Sau khi chuyển đổi văn bản thành vector, hệ thống cần cơ chế lưu trữ và truy xuất hiệu quả.
- Sử dụng **Qdrant Server** kết hợp với thuật toán **HNSW (Hierarchical Navigable Small World)** để đảm bảo tốc độ tìm kiếm cao trên hàng trăm nghìn vector.
- **Kiến trúc Dual-Store trong Qdrant:**
 - **Collection legal_rag:** Chứa 100,507 điều luật Đây là kho dữ liệu chính
 - **Collection user_docs:** Chứa tài liệu người dùng tải lên (PDF, DOCX, TXT), được vector hóa

2.5 Hybrid Search

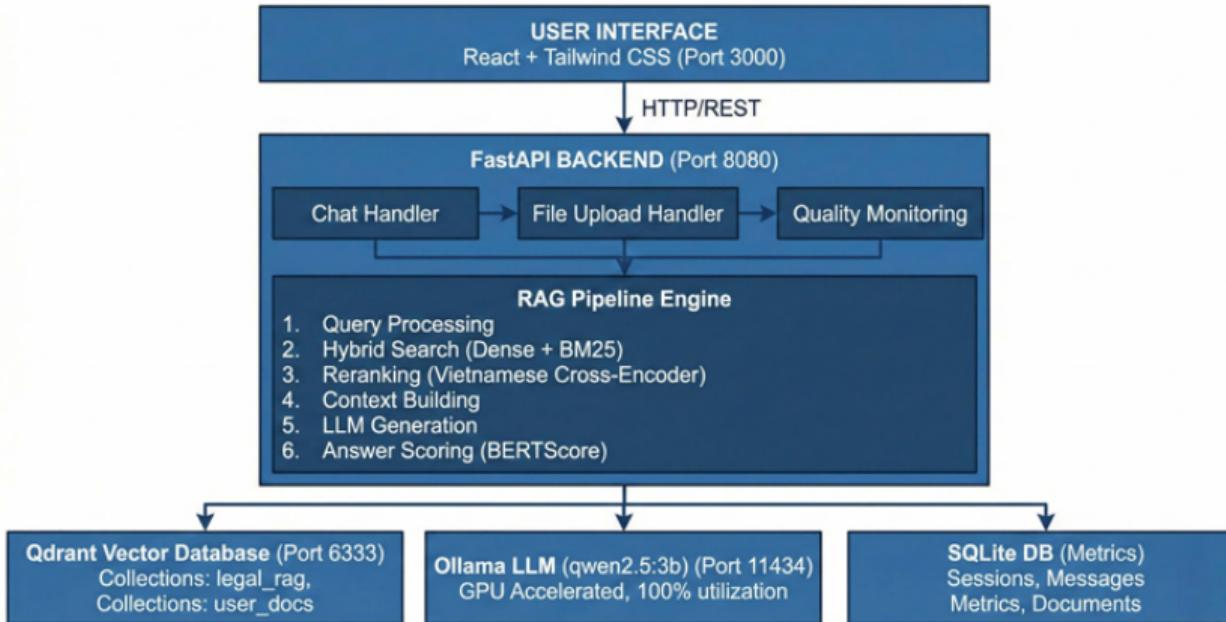
- Dense Search: tìm theo ngữ nghĩa
- Sparse Search (BM25): tìm theo từ khóa
- Hợp nhất kết quả bằng Reciprocal Rank Fusion (RRF)

2.6 Reranking và sinh câu trả lời

- Cross-Encoder Reranker cho tiếng Việt thanhtantran/VietnameseReranker
- Chọn Top-k ngữ cảnh liên quan nhất
- LLM sinh câu trả lời có ràng buộc theo ngữ cảnh

3.1 Kiến trúc triển khai hệ thống

2.1 System Architecture



Hình 3.1. Sơ đồ kiến trúc triển khai của hệ thống RAG Pháp luật

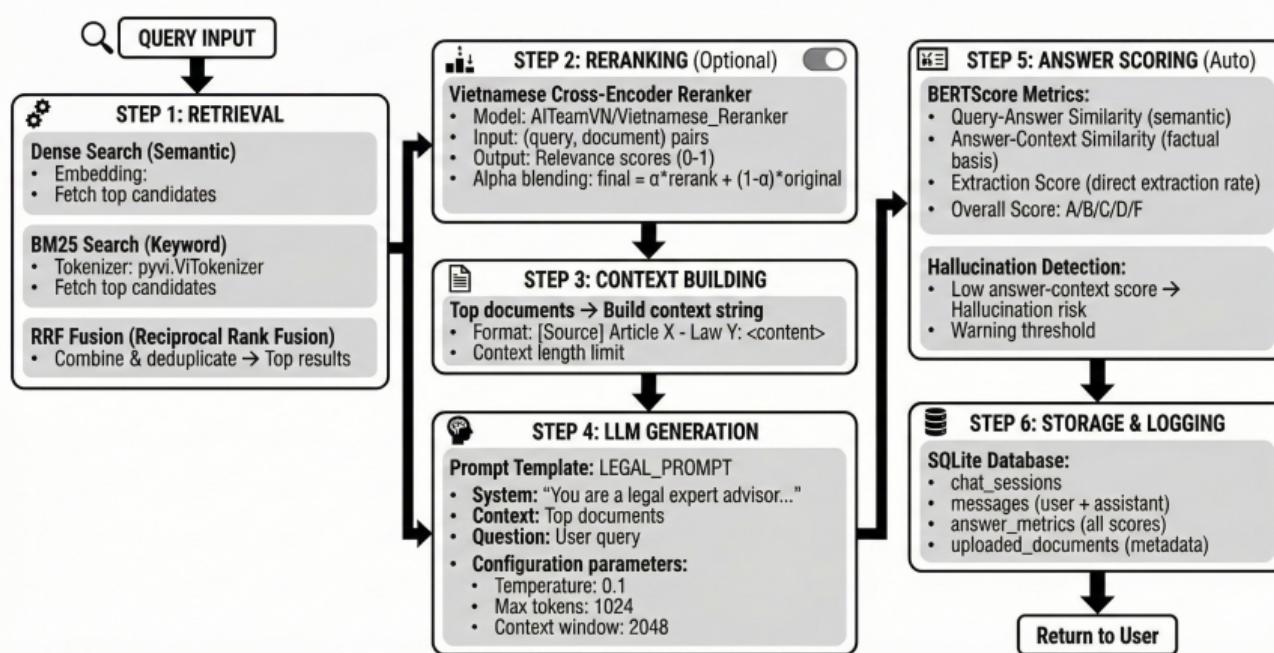
- **Router phân loại ý định:** Phân tích câu hỏi để chọn chiến lược phù hợp:
 - Tìm kiếm song song (luật + tài liệu người dùng)
 - Tìm kiếm chuyên sâu trên kho luật
 - Tìm kiếm tập trung vào tài liệu cá nhân
- **Cơ chế song song:** Các tiến trình tìm kiếm được khởi tạo đồng thời, kết quả hợp nhất tại điểm cuối. → Giảm độ trễ, tăng trải nghiệm mượt mà.

Cơ chế sàng lọc đa tầng (The Funnel)

- **Tầng 1 – Thu thập:** Lấy Top-20 kết quả từ mỗi luồng tìm kiếm song song.
- **Tầng 2 – Hợp nhất:** Dùng Reciprocal Rank Fusion (RRF) để trộn kết quả, khử trùng lặp.
- **Tầng 3 – Tinh chỉnh:** Vietnamese-Reranker đánh giá lại, chọn Top-k văn bản liên quan nhất.

Kiến trúc luồng hoạt động

2.2 RAG Pipeline Flow



4.1 Đánh giá mô hình Embedding

Bảng 2.1. Hiệu năng các mô hình embedding trên Legal QA benchmark

Mô hình	Loại	NDCG@10	MRR@10	NDCG@3	MRR@3
AITeamVN/Vietnamese_EMBEDDING	Dense	0.8650	0.8334	0.8427	0.8221
BAAI/bge-m3	Hybrid	0.8120	0.7713	0.7752	0.7477
BAAI/bge-m3	Dense	0.8170	0.7803	0.7841	0.7633
Quockhanh05/Vietnam_legal	Dense	0.8020	0.7557	0.7653	0.7370
huuydangg/DEk21 (Mô hình gốc)	Dense	0.7851	0.7411	0.7522	0.7247
hieu/halong_embedding	Hybrid	0.7792	0.7320	0.7363	0.7104
dangvantuan/vietnamese-embedding	Dense	0.7634	0.7189	0.7190	0.6964
BM25 (Từ khóa truyền thống)	Lexical	0.7616	0.7157	0.7281	0.6995

4.2 Thiết lập thực nghiệm

- Dataset: ALQAC (530 câu hỏi)
- So sánh các chiến lược:
 - BM25
 - Dense
 - Hybrid
 - VieLegalRAG

4.3 Kết quả truy xuất

Bảng 5.1. Bảng so sánh hiệu năng truy xuất giữa các phương pháp (N=530)

Metric	Sparse (BM25)	Dense (Vector)	Std. Hybrid	Proposed (VN-Rerank)
MRR	0.752	0.801	0.873	0.9499
Recall@1	0.680	0.750	0.830	0.9377
Recall@5	0.810	0.860	0.920	0.9641
Recall@10	0.860	0.900	0.930	0.9736

- **MRR:** 0.9499
- **Recall@10:** 97.36%
- VieLegalRAG vượt trội so với các baseline

4.4 Kết quả sinh câu trả lời

Bảng 4.2. Chất lượng sinh câu trả lời của hệ thống đề xuất

Metric	Kết quả	Ý nghĩa thực tiễn
BERTScore F1	0.8468	Câu trả lời "hiểu" và bám sát ngữ nghĩa đáp án chuẩn.
ROUGE-L	0.6359	Cấu trúc câu mạch lạc, văn phong pháp lý chuẩn xác.
Extractive Rate	98.49%	98.5% thông tin được trích xuất trực tiếp từ luật.
Avg. Time	3.75s	Tốc độ phản hồi hợp lý cho trải nghiệm người dùng.

- **BERTScore F1:** 0.8468
- **Extractive Rate:** 98.49%
- Giảm đáng kể hiện tượng hallucination

5.1 Kết luận

- Đề xuất thành công hệ thống Advanced RAG
- Phù hợp cho bài toán hỏi–đáp pháp luật Việt Nam
- Đảm bảo độ chính xác và minh bạch

5.2 Hướng phát triển

- Fine-tuning LLM cho suy luận pháp lý
- GraphRAG
- Agentic RAG

Cảm ơn mọi người đã lắng nghe!