

# Chatbot Tra cứu Luật Giao thông Đường bộ Việt Nam

Đồ án môn học CS431.P22

## **Thực hiện bởi:**

Lê Minh Nhựt - 22521060

Trần Đình Khánh Đăng - 22520195

Thái Ngọc Quân - 22521189

**Giảng viên hướng dẫn:** TS. Nguyễn Vinh Tiệp

<sup>1</sup>Khoa Khoa học Máy tính

Trường Đại học Công nghệ Thông tin - ĐHQG TP.HCM

Báo cáo Cuối kỳ – Tháng 5, 2025

# Mục lục

- 1 Giới thiệu
- 2 Dữ liệu
- 3 Quy trình xử lý
- 4 Độ đo đánh Giá

# Mục lục

1 **Giới thiệu**

2 Dữ liệu

3 Quy trình xử lý

4 Độ đo đánh Giá

# Giới thiệu: Động lực

- Hệ thống pháp luật hiện nay **phức tạp**, gồm nhiều loại văn bản (Luật, Nghị định, Thông tư...), thường **dài dòng, khó tra cứu**.
- **Luật mới, tuyên truyền còn hạn chế**, nhiều người tham gia giao thông **chưa nắm rõ**, dẫn đến **vi phạm nghiêm trọng**.
- Một bộ phận người dân **thiếu ý thức**, nếu hiểu rõ mức phạt và tính răn đe, sẽ tuân thủ tốt hơn, góp phần giảm tai nạn.
- **Thiếu công cụ tra cứu tập trung**, người dân gặp khó khăn khi cần giải đáp nhanh các thắc mắc.

## Giải pháp

- Xây dựng **Chatbot Tra cứu Luật Giao thông Đường bộ Việt Nam**

# Giới thiệu: Mục tiêu dự án

- **Đáp ứng nhu cầu hỏi đáp:** Về quy định, mức phạt, thủ tục cơ bản, biển báo giao thông (có hình ảnh minh họa), ...
- **Đảm bảo chính xác và cập nhật:** Dựa trên văn bản pháp luật **mới nhất**, nguồn tin **chính thống, đáng tin cậy**.
- **Thân thiện người dùng:** Giao diện tương tác **tự nhiên, đơn giản**, phản hồi **mạch lạc, dễ hiểu**.
- **Truy xuất nguồn:** Chỉ rõ **điều, khoản, điểm, văn bản gốc**, kèm trích dẫn nguồn đầy đủ.

# Giới thiệu: Ví dụ



Xe máy vượt đèn đỏ phải chịu các mức phạt nào?



Người điều khiển xe máy vượt đèn đỏ sẽ phải chịu các mức phạt sau:

- Trường hợp không gây tai nạn giao thông:
  - Phạt tiền từ **4.000.000 đồng đến 6.000.000 đồng**. (Theo Điều 7, Khoản 7, Điểm c, Nghị định số 168/2024\_NĐ-CP)
  - Bị trừ **04 điểm** vào giấy phép lái xe. (Theo Điều 7, Khoản 13, Điểm b, Nghị định số 168/2024\_NĐ-CP)
- Trường hợp gây tai nạn giao thông:
  - Phạt tiền từ **10.000.000 đồng đến 14.000.000 đồng**. (Theo Điều 7, Khoản 10, Điểm b, Nghị định số 168/2024\_NĐ-CP)
  - Bị trừ **10 điểm** vào giấy phép lái xe. (Theo Điều 7, Khoản 13, Điểm d, Nghị định số 168/2024\_NĐ-CP)

Nguồn tham khảo (Văn bản pháp luật):

- [Nghị định số 168/2024/NĐ-CP](#)

Nhập câu hỏi của bạn về Luật GTĐB...



*Hình 1: Hỏi đáp về mức phạt vượt đèn đỏ của xe máy.*

# Mục lục

1 Giới thiệu

**2 Dữ liệu**

3 Quy trình xử lý

4 Độ đo đánh Giá

- **Lĩnh vực:** *Luật giao thông đường bộ Việt Nam*
- **Nguồn chính:**
  - Với các văn bản pháp luật: Thư viện Pháp Luật, Luật Việt Nam là nền tảng uy tín, cập nhật và phổ biến.
  - Với các loại biển báo: Quy chuẩn KTQG QCVN 41:2024/BGTVT, Vietnamse traffic sign wiki.
- **Số lượng:** gồm **46 văn bản** giao thông đang có hiệu lực thi hành (Ví dụ: Luật 36/2024/QH15, Nghị định 168/2024/NĐ-CP, ...) và khoảng **400 biển báo** giao thông đường bộ các loại.



# Dữ liệu: Cấu trúc văn bản Pháp luật

## Chương II

### QUY TẮC GIAO THÔNG ĐƯỜNG BỘ

#### Điều 10. Quy tắc chung

1. Người tham gia giao thông đường bộ phải đi bên phải theo chiều đi của mình, đi đúng làn đường, phần đường quy định, chấp hành báo hiệu đường bộ và các quy tắc giao thông đường bộ khác.
2. Người lái xe và người được chở trên xe ô tô phải thắt dây đai an toàn tại những chỗ có trang bị dây đai an toàn khi tham gia giao thông đường bộ.
3. Khi chở trẻ em dưới 10 tuổi và chiều cao dưới 1,35 mét trên xe ô tô không được cho trẻ em ngồi cùng hàng ghế với người lái xe, trừ loại xe ô tô chỉ có một hàng ghế; người lái xe phải sử dụng, hướng dẫn sử dụng thiết bị an toàn phù hợp cho trẻ em.

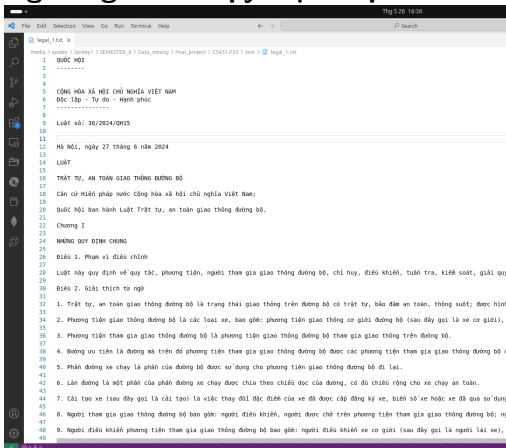
#### Điều 11. Chấp hành báo hiệu đường bộ

1. Báo hiệu đường bộ bao gồm: hiệu lệnh của người điều khiển giao thông; đèn tín hiệu giao thông; biển báo hiệu đường bộ; vạch kẻ đường và các dấu hiệu khác trên mặt đường; cọc tiêu, tường bảo vệ, rào chắn, đỉnh phản quang, tiêu phản quang, cột Km, cọc H; thiết bị âm thanh báo hiệu đường bộ.
2. Người tham gia giao thông đường bộ phải chấp hành báo hiệu đường bộ theo thứ tự ưu tiên từ trên xuống dưới như sau:
  - a) Hiệu lệnh của người điều khiển giao thông;
  - b) Tín hiệu đèn giao thông;
  - c) Biển báo hiệu đường bộ;
  - d) Vạch kẻ đường và các dấu hiệu khác trên mặt đường;
  - đ) Cọc tiêu, tường bảo vệ, rào chắn, đỉnh phản quang, tiêu phản quang, cột Km, cọc H;
  - e) Thiết bị âm thanh báo hiệu đường bộ.

### Hình 2: Cấu trúc của 1 văn bản Pháp luật

# Dữ liệu: Lưu dưới dạng text

- Nội dung của mỗi văn bản pháp luật sẽ được lưu vào file text **một cách thủ công bằng cách copy trực tiếp từ các văn bản gốc.**



```
1 QUỐC HỘI
2 -----
3
4
5 CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
6 Độc lập - Tự do - Hạnh phúc
7 -----
8
9 Luật số: 38/2024/QH15
10
11
12 Hà Nội, ngày 27 tháng 6 năm 2024
13
14 LUẬT
15
16 TRẬT TỰ, AN TOÀN GIAO THÔNG ĐƯỜNG BỘ
17
18 Căn cứ Hiến pháp nước Cộng hòa xã hội chủ nghĩa Việt Nam;
19 Quốc hội ban hành Luật Trật tự, an toàn giao thông đường bộ.
20
21 Chương I
22
23 NHỮNG QUY ĐỊNH CHUNG
24
25 Điều 1. Phạm vi điều chỉnh
26
27 Luật này quy định về quy tắc, phương tiện, người tham gia giao thông đường bộ, chỉ huy, điều khiển, tuần tra, kiểm soát, giải cứu
28
29 Điều 2. Giải thích từ ngữ
30
31 1. Trật tự, an toàn giao thông đường bộ là trạng thái giao thông trên đường bộ có trật tự, bảo đảm an toàn, thông suốt; được hình
32
33 2. Phương tiện giao thông đường bộ là các loại xe, bao gồm: phương tiện giao thông cơ giới đường bộ (sau đây gọi là xe cơ giới),
34
35 3. Phương tiện tham gia giao thông đường bộ là phương tiện giao thông đường bộ tham gia giao thông trên đường bộ.
36
37 4. Đường ưu tiên là đường mà trên đó phương tiện tham gia giao thông đường bộ được các phương tiện tham gia giao thông đường bộ
38
39 5. Phanh đường xe chạy là phần của đường bộ được sử dụng cho phương tiện giao thông đường bộ đi lại.
40
41 6. Làn đường là một phần của phần đường xe chạy được chia theo chiều dọc của đường, có đủ chiều rộng cho xe chạy an toàn.
42
43 7. Cải tạo xe (sau đây gọi là cải tạo) là việc thay đổi đặc điểm của xe đã được cấp đăng ký xe, biến số xe hoặc xe đã qua sử dụng
44
45 8. Người tham gia giao thông đường bộ bao gồm: người điều khiển, người được chở trên phương tiện tham gia giao thông đường bộ;
46
47 9. Người điều khiển phương tiện tham gia giao thông đường bộ bao gồm: người điều khiển xe cơ giới (sau đây gọi là người lái xe),
48
49
```

Hình 3: Dữ liệu được lưu ở dạng text file

# Dữ liệu: Thách thức

- Tuy nhiên **có 2 thách thức cần tốn nhiều thời gian để giải quyết**, nhằm đáp ứng một bộ dữ liệu rõ ràng cho Chatbot là:
  - ➊ Phải xử lý các Điều, Khoản, Điểm của Luật **được bổ sung, sửa đổi, thay thế giữa các văn bản liên quan nhau.**
  - ➋ Phải xử lý các Điều, Khoản, Điểm **đề cập tới Điều, Khoản, Điểm khác chứ không nêu cụ thể đối tượng, hành vi vi phạm, ...**

⇒ Cách hiệu quả nhất để giải quyết là **đọc văn bản và thực hiện thay đổi thủ công** để tạo ra các đoạn có nội dung đầy đủ hơn.

## ĐIỀU KHOẢN THI HÀNH

Điều 52. Sửa đổi, bổ sung một số điều của Nghị định số **100/2019/NĐ-CP** ngày 30 tháng 12 năm 2019 của Chính phủ quy định xử phạt vi phạm hành chính trong lĩnh vực giao thông đường bộ và đường sắt đã được sửa đổi, bổ sung một số điều theo Nghị định số **123/2021/NĐ-CP** ngày 28 tháng 12 năm 2021 của Chính phủ sửa đổi, bổ sung một số điều của các Nghị định quy định xử phạt vi phạm hành chính trong lĩnh vực hàng hải; giao thông đường bộ, đường sắt; hàng không dân dụng

1. Bổ sung khoản 2a vào sau **khoản 2 Điều 1** như sau:

"2a. Hình thức, mức xử phạt, biện pháp khắc phục hậu quả đối với từng hành vi vi phạm hành chính; thẩm quyền lập biên bản, thẩm quyền xử phạt, mức phạt tiền cụ thể đối với từng chức danh về trật tự, an toàn giao thông trong lĩnh vực giao thông đường bộ thì áp dụng quy định tại Nghị định quy định xử phạt vi phạm hành chính về trật tự, an toàn giao thông trong lĩnh vực giao thông đường bộ; trừ điểm, phục hồi điểm giấy phép lái xe".

*Hình 4: Minh họa Luật cần sửa đổi bổ sung*

- Phải dựa vào thông tin của **Điều 52, Khoản 1** (thuộc văn bản mới hơn) trong *Hình 4* để cập nhật lại nội dung của **Nghị định số 100/2019/NĐ-CP** (vẫn còn hiệu lực) nếu chưa có văn bản hợp nhất.

# Dữ liệu: Thách thức 2

- Ta phải cập nhật lại **Khoản 12, Điểm a (Hình 5a)** dựa trên thông tin của **Khoản 2, Điểm d (Hình 5b)** như sau: Khoản 12, Điểm a) Thực hiện hành vi **xe không được quyền ưu tiên lắp đặt sử dụng thiết bị phát tín hiệu của xe được quyền ưu tiên** còn bị tịch thu thiết bị phát tín hiệu ưu tiên lắp đặt, sử dụng trái quy định.

12. Ngoài việc bị áp dụng hình thức xử phạt chính, người điều khiển xe thực hiện hành vi vi phạm còn bị áp dụng các hình thức xử phạt bổ sung sau đây:

a) Thực hiện hành vi quy định tại điểm đ khoản 2 Điều này còn bị tịch thu thiết bị phát tín hiệu ưu tiên lắp đặt, sử dụng trái quy định;

b) Thực hiện hành vi quy định tại điểm a, điểm b, điểm h, điểm i, điểm k khoản 9 Điều này bị tước quyền sử dụng giấy phép lái xe từ 10 tháng đến 12 tháng;

## *Hình 5a: Minh họa khi Luật chi trích dẫn Điều, Khoản, Điểm khác*

2. Phạt tiền từ 400.000 đồng đến 600.000 đồng đối với người điều khiển xe thực hiện một trong các hành vi vi phạm sau đây:

a) Dừng xe, đỗ xe trên phần đường xe chạy ở đoạn đường ngoài đô thị nơi có lề đường;

b) Điều khiển xe chạy quá tốc độ quy định từ 05 km/h đến dưới 10 km/h;

c) Điều khiển xe chạy tốc độ thấp mà không đi bên phải phần đường xe chạy gây cản trở giao thông;

d) Dừng xe, đỗ xe ở lòng đường gây cản trở giao thông; tụ tập từ 03 xe trở lên ở lòng đường, trong hầm đường bộ; đỗ, để xe ở lòng đường, vỉa hè trái phép;

đ) Xe không được quyền ưu tiên lắp đặt, sử dụng thiết bị phát tín hiệu của xe được quyền ưu tiên;

## *Hình 5b: Thông tin đầy đủ nằm ở Khoản 2, Điểm đ trước đó*

# Dữ liệu: JSON

- **Chunking:** chia văn bản thành các đoạn nhỏ theo cấp độ **Chương** → **Điều** → **Khoản** → **Điểm** để không làm mất ngữ cảnh đầy đủ của luật, lưu vào trường **'text'**.
- Ngoài ra lưu trữ thêm **'metadata'** gồm:
  - **'source'**: số hiệu văn bản
  - **'effective\_date'**: ngày hiệu lực
  - **'url'**: đường dẫn gốc
  - **'context'**: {chuong, dieu, khoan, diem}
  - **'traffic\_sign'**: tên ảnh của biển báo (nếu có)
- **Định dạng JSON:** mỗi đoạn (chunk) lưu theo một **'id'** duy nhất, phục vụ cho việc truy vấn hiệu quả.  
→ Cách lưu này không chỉ giúp giữ được ngữ cảnh trong văn bản pháp luật, mà còn giúp cho các đoạn không vượt quá kích thước tối đa mà các mô hình dùng để xử lý có thể nhận (512 token).

# Dữ liệu: Cấu trúc của một đoạn (chunk)

```
{
  "id": "41_2024_BGTVT_chunk_510",
  "text": "Phụ lục B\nB.2 Biển số P.102 “Cấm đi ngược chiều”\nna) Để báo đường cấm các loại xe (cơ giới và thô sơ)...",
  "metadata": {
    "source": "Quy chuẩn kỹ thuật số 41_2024_BGTVT",
    "effective_date": "1-1-2025",
    "url": "https://luatvietnam.vn/giao-thong/q...",
    "context": {
      "khoan": "a",
      "dieu": "B.2",
      "phu_luc": "B"
    }
  },
  "traffic_sign": ["P_102.png"]
}
```

# Mục lục

1 Giới thiệu

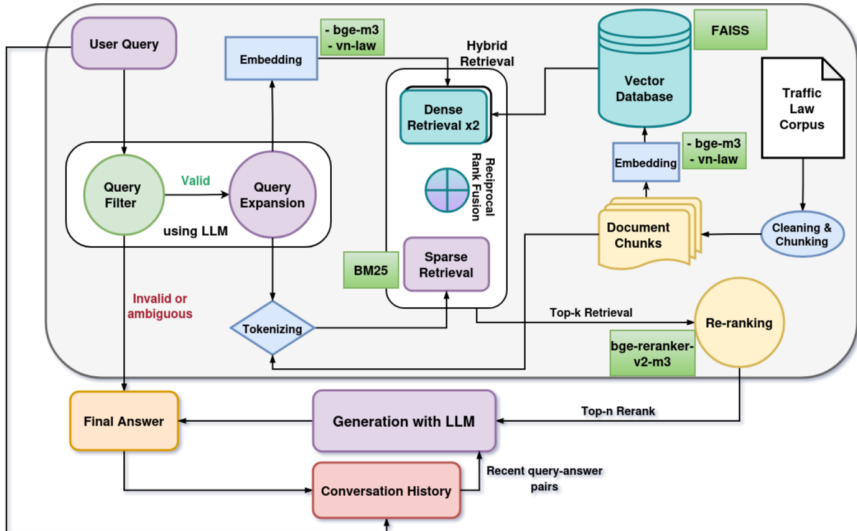
2 Dữ liệu

3 Quy trình xử lý

4 Độ đo đánh Giá



# Quy trình: Pipeline



Hình 6: Pipeline chính của chatbot giao thông đường bộ

# Quy trình: Tổng quan về RAG

**Retrieval-Augmented Generation (RAG)** là mô hình kết hợp giữa:

- **Truy xuất thông tin (Retrieval):** Tìm văn bản liên quan từ tập dữ liệu lớn.
- **Sinh văn bản (Generation):** Sinh câu trả lời tự nhiên dựa trên văn bản truy xuất và câu hỏi.

**Tại sao chọn RAG cho Chatbot Luật Giao thông:**

- Luật rất dài, nhiều điều khoản — không thể nhét hết vào prompt một cách hiệu quả.
- RAG giúp chatbot **truy xuất đúng điều khoản liên quan** thay vì dựa vào trí nhớ mô hình.
- Tránh "hallucination"— chatbot chỉ sinh câu trả lời từ dữ liệu đã truy xuất.
- Dễ cập nhật khi luật thay đổi: chỉ cần cập nhật cơ sở dữ liệu.

# Quy trình: 1. Chuẩn bị Dữ liệu & Embedding

**Mục tiêu:** Chuyển đổi các câu hỏi và đoạn văn bản pháp luật thành các **vector số (embedding)** nhằm hỗ trợ **Truy vấn theo ngữ nghĩa (Dense Retrieval)**. Chỉ chuyển đổi trường **'text'** chứa nội dung chính.

- **Ví dụ:** “Xe không có biển số sẽ bị xử phạt ...”  
→ [0.13, -0.02, 0.27, 0.22 ...]

# Quy trình: 1. Chuẩn bị Dữ liệu & Embedding

- **Công cụ:** Sử dụng SentenceTransformer với ba mô hình thử nghiệm:
  - vn-law-embedding – một mô hình được huấn luyện chuyên biệt cho tiếng Việt và văn bản pháp lý (768 chiều)
  - bge-m3 – mô hình mạnh mẽ đa ngôn ngữ, hỗ trợ cả văn bản ngắn và dài (1024 chiều)
  - multilingual-e5-large – một mô hình embedding đa ngôn ngữ tiên tiến, hỗ trợ hơn 100 ngôn ngữ, bao gồm tiếng Việt (1024 chiều)

Bảng 1: So sánh hiệu suất các mô hình embedding

Độ đo	vn-law-embedding				bge-m3				multilingual-e5-large			
	@3	@5	@10	@15	@3	@5	@10	@15	@3	@5	@10	@15
Recall	0.4677	0.5381	0.5938	0.6124	0.5477	0.6308	0.7058	0.7346	0.5173	0.6124	0.6864	0.7126
MRR	0.6717	0.6812	0.6849	0.6862	0.7117	0.7203	0.7244	0.7260	0.6967	0.7083	0.7132	0.7148
nDCG	0.5450	0.5484	0.5633	0.5701	0.6150	0.6255	0.6490	0.6597	0.5919	0.6087	0.6306	0.6407

- bge-m3 cho kết quả tốt nhất.

# Quy trình: 2. Xây dựng Chỉ mục Tìm kiếm

- **Chỉ mục Vector (Dense Index):**

- Sử dụng **vector database FAISS** với cấu hình IndexFlatL2 để lưu trữ embedding của văn bản.
- Truy vấn bằng cách tìm các đoạn văn bản (chunk) có **khoảng cách Euclid gần nhất** với embedding của câu hỏi.
- Phù hợp với **tìm kiếm theo ngữ nghĩa**, khi câu hỏi và văn bản có **nội dung tương đồng** nhưng không dùng từ giống nhau.

- **Chỉ mục Từ khóa (Sparse Index):**

- Dựa trên thuật toán **BM25** (cải tiến từ TF-IDF) để xếp hạng mức độ liên quan dựa trên tần suất từ.
- Tiền xử lý bao gồm: **tách từ (tokenize)**, **loại bỏ từ dừng (stop words)**, và **chuẩn hoá** nội dung trong 'text'.
- Phù hợp với **truy vấn ngắn**, có từ khóa rõ ràng và **trùng khớp** với văn bản.

## Quy trình: 3. Nhận truy vấn & Kiểm tra liên quan

**Nhận truy vấn** Giao diện người dùng nhập câu hỏi (ví dụ: “Chạy xe máy trên lề đường có bị phạt không?”).

Dùng **LLM (Gemini)** để xử lý 2 bước: **Kiểm tra liên quan (Query Filtering)** và **Mở rộng truy vấn (Query Expansion)**

### Kiểm tra liên quan (Query Filtering)

- Nếu truy vấn nằm ngoài luật giao thông hoặc mơ hồ → trả lời từ chối hoặc hỏi lại.
- Nếu hợp lệ → tiếp tục sang bước **Mở rộng truy vấn**. Mục tiêu là tăng tính đa dạng cho các trường ngữ nghĩa.

# Quy trình: 3. Mở rộng truy vấn (Query Expansion)

**Câu hỏi gốc:** *"Chạy xe máy trên lề đường có bị phạt không?"*

**1. Đơn giản:** Truy vấn bằng đúng câu hỏi gốc, không thay đổi:

- *"Chạy xe máy trên lề đường có bị phạt không?"*

**2. Mở rộng:** Tạo một câu hỏi dài, tổng hợp nhiều từ liên quan để tăng khả năng khớp:

- *"Đi xe máy, xe mô tô hai bánh và các loại xe tương tự trên lề đường, vỉa hè có bị xử phạt theo Luật Giao thông đường bộ không?"*

**3. Đa dạng:** Tạo danh sách nhiều biến thể khác nhau từ câu hỏi gốc + giữ lại câu hỏi gốc:

- *"Chạy xe máy trên lề đường có bị phạt không?"* (gốc)
- *"Đi xe máy trên vỉa hè có vi phạm luật không?"*
- *"Có bị xử lý nếu lái xe máy trên lề đường?"*
- *"Luật nói gì về chạy xe máy lên vỉa hè?"*

# Quy trình: 3. So sánh mở rộng truy vấn (Query Expansion)

## Cấu hình chung:

- Chiến lược truy vấn: *Hybrid (bge-m3 + vn-law + sparse(BM25))*
- Reranker: *bge-reranker-v2-m3*

Bảng 2: So sánh hiệu suất các theo loại câu hỏi truy vấn.

Độ đo	Đơn giản (Gốc)				Đa dạng				Mở rộng			
	@3	@5	@10	@15	@3	@5	@10	@15	@3	@5	@10	@15
Recall	0.5771	0.6733	0.7630	0.7922	0.5808	0.6698	0.7576	0.7978	0.5675	0.6707	0.7582	0.7958
MRR	0.7211	0.7296	0.7349	0.7359	0.7483	0.7570	0.7630	0.7647	0.7306	0.7414	0.7465	0.7479
nDCG	0.6347	0.6559	0.6882	0.6998	0.6496	0.6604	0.6892	0.7036	0.6400	0.6571	0.6861	0.7004

- Truy vấn **Đa dạng** cho kết quả tốt nhất. Tuy nhiên, tốn thời gian và tài nguyên tìm kiếm hơn, nên hầu hết các thực nghiệm chúng em chọn truy vấn **Mở rộng** để cân bằng hiệu suất hơn.



# Quy trình 4: Truy vấn Thông tin

- Thử nghiệm hiệu quả giữa từng chiến lược truy vấn:
  - Truy vấn ngữ nghĩa (Dense Retrieval)
  - Truy vấn từ khóa (Sparse Retrieval)
  - Truy vấn kết hợp (Hybrid Retrieval) dùng RRF (Reciprocal Rank Fusion) để:
    - Kết hợp Dense + Sparse
    - Kết hợp 2 Dense + 1 Sparse
- Thu thập ứng viên:
  - Lấy **top-k** chunk để chuyển sang bước Reranking ( $k = 30$ )
  - Nếu dùng **Truy vấn kết hợp**, mỗi thành phần truy vấn sẽ lấy ra **top-k** sau đó được kết hợp lại bằng **RRF**. Điều này làm cho nhiều đoạn hơn được tạo ra, do đó, sau khi kết hợp xong sẽ lấy ra **top-k** một lần nữa.
  - Mục đích của việc fusion là để giảm thiểu chi phí rerank thường rất tốn kém. Vì nếu không Fusion, từng khối phải được rerank riêng biệt, thay vào đó ta có fusion trước để chọn lọc ra những ứng viên tốt để chỉ phải rerank 1 lần.

## Quy trình 4: Weighted RRF

- **Gán trọng số:** Mỗi retriever có weight riêng: (sparse,  $w_s = 0.5$ ), (dense,  $w_d = 0.4$ )
- **Fusion bằng RRF:**

$$\text{score}(d) = \sum_i \frac{w_i}{r + \text{rank}_i(d)}$$

với  $r = 10$  giúp làm trơn điểm.

- **Ví dụ:**

- $d_1$ : rank 3 (sparse), 1 (dense)

$$\text{score}(d_1) = \frac{0.5}{10 + 3} + \frac{0.4}{10 + 1} \approx 0.0749$$

- $d_2$ : rank 1 (sparse), 10 (dense)

$$\text{score}(d_2) = \frac{0.5}{10 + 1} + \frac{0.4}{10 + 10} \approx 0.0655$$

→  $d_1$  được ưu tiên hơn vì có điểm số (vị trí) cao ở cả hai retriever.

# Quy trình: 4. So sánh các chiến lược truy vấn

**Cấu hình chung:** Loại câu hỏi: *Mở rộng*, Reranker: *bge-reranker-v2-m3*

**Cấu hình chi tiết:**

- Dense dùng kết quả tốt nhất (*bge-m3*).
- Hybrid(1 Dense, 1 Sparse) dùng kết quả tốt nhất (*bge-m3 + BM25*).
- Hybrid(2 Dense, 1 Sparse) dùng kết quả tốt nhất (*bge-m3 + vn-law + BM25*).

**Bảng 3:** So sánh hiệu suất các mô hình theo chiến lược truy vấn.

Độ đo	Sparse				Dense				Hybrid (1 Dense, 1 Sparse)				Hybrid (2 Dense, 1 Sparse)			
	@3	@5	@10	@15	@3	@5	@10	@15	@3	@5	@10	@15	@3	@5	@10	@15
Recall	0.4926	0.5682	0.6312	0.6570	0.5477	0.6308	0.7058	0.7346	0.5657	0.6606	0.7455	0.7813	0.5675	0.6707	0.7582	0.7958
MRR	0.6428	0.6486	0.6516	0.6521	0.7117	0.7203	0.7244	0.7260	0.7189	0.7286	0.7332	0.7342	0.7306	0.7414	0.7465	0.7479
nDCG	0.5619	0.5705	0.5925	0.6021	0.6150	0.6255	0.6490	0.6597	0.6348	0.6490	0.6775	0.6904	0.6400	0.6571	0.6861	0.7004

- Ta thấy khi kết hợp Hybrid với 2 Dense và 1 Sparse hiệu suất đạt tốt nhất nhờ vào sự kết hợp thông tin đa dạng giữa các chiến lược truy vấn.

# Quy trình: 5. Xếp hạng lại (Reranking)

*Tinh chỉnh thứ tự các đoạn văn bản ứng viên để nâng độ chính xác.*

- **Chọn ứng viên ban đầu:**

- Lấy Top **K** chunk từ kết quả truy xuất ban đầu.

- **Áp dụng CrossEncoder để đánh giá lại:**

- Kết hợp từng cặp (**query, chunk**) làm đầu vào cho mô hình.
- Mô hình (CrossEncoder như `namdp-ptit/ViRanker`) trả về **score** phản ánh độ liên quan.
- Ưu điểm: chính xác hơn so với truy xuất thô.

- **Chọn kết quả cuối cùng:**

- Sắp xếp các chunk theo điểm số.
- Lấy Top **N** chunk để đưa vào LLM sinh câu trả lời. Dựa vào quá trình làm sạch dữ liệu và thực nghiệm, chúng em thấy  $N = 15$  là con số khá hợp lí để cân bằng giữa sự chính xác và tài nguyên tính toán.

*Nếu không dùng Reranker: Lấy trực tiếp Top N từ kết quả truy xuất. (Cắt ngắn lại theo K)*

# Quy trình: 5. So sánh các reranker

- **Công cụ:** Sử dụng các mô hình reranker để tinh chỉnh kết quả truy vấn ban đầu:
  - ViRanker
  - bge-reranker-v2-m3
  - Vietnamese\_Reranker

**Bảng 4:** So sánh hiệu suất các mô hình reranker.

Độ đo	Không sử dụng				ViRanker				bge-reranker-v2-m3				Vietnamese_Reranker			
	@3	@5	@10	@15	@3	@5	@10	@15	@3	@5	@10	@15	@3	@5	@10	@15
Recall	0.4837	0.5715	0.6881	0.7569	0.5448	0.6290	0.7439	0.7895	0.5675	0.6707	0.7582	0.7958	0.5661	0.6589	0.7495	0.7952
MRR	0.6017	0.6138	0.6242	0.6275	0.6956	0.7037	0.7128	0.7142	0.7306	0.7414	0.7465	0.7479	0.7128	0.7219	0.7279	0.7298
nDCG	0.5171	0.5333	0.5780	0.6030	0.5990	0.6126	0.6543	0.6717	0.6400	0.6571	0.6861	0.7004	0.6225	0.6369	0.6683	0.6851

- bge-reranker-v2-m3 cho kết quả tốt nhất, trong khi đó không sử dụng rerank cho kết quả rất thấp.

# Quy trình: 6. Tạo & Hoàn thiện Câu trả lời

**1. Mục tiêu:** Trả lời câu hỏi về luật giao thông dựa **duy nhất** vào "Ngữ cảnh được cung cấp".

## 2. Cấu trúc Prompt:

- **Lịch sử trò chuyện (tùy chọn):** Tóm tắt các tương tác trước đó (nếu có).
- **Ngữ cảnh cung cấp:** Chứa các đoạn văn bản pháp luật liên quan.
- **Câu hỏi hiện tại:** Do người dùng đặt ra.
- **Yêu cầu trả lời:**
  - Chỉ dùng ngữ cảnh, không suy đoán.
  - Trích dẫn nguồn rõ ràng, gộp nhóm nếu cùng văn bản.
  - Trình bày súc tích, mạch lạc.
  - Phân biệt từ ngữ pháp lý tương đương.
  - Nếu không có thông tin phù hợp → trả lời "Không tìm thấy".
  - Ưu tiên thứ tự điểm, khoản, điều khi so sánh.
  - Hiển thị biển báo (nếu có) bằng placeholder:  
[DISPLAY\_TRAFFIC\_SIGN\_INDEX\_x]


**2 phiên bản prompt:** chi tiết và ngắn gọn tùy mục tiêu ứng dụng.

# Quy trình: 7. Ghi nhớ ngữ cảnh

- Lưu lịch sử truy vấn, câu trả lời ( lưu khoảng **4 query và answer gần nhất** để giữ được độ ổn định của câu trả lời).
- Hỗ trợ đối thoại liên tục.

Hình ảnh biển báo cấm đi ngược chiều

Biển báo cấm đi ngược chiều là biển số P.102. (Theo Phụ lục B, B.2, Quy chuẩn kỹ thuật số 41\_2024\_BGTVT)



P.102

Nguồn tham khảo (Văn bản pháp luật):

- [Quy chuẩn kỹ thuật số 41/2024/BGTVT](#)

Xe máy nếu vi phạm biển này bị phạt bao nhiêu?

Đối với xe máy (xe mô tô, xe gắn máy) vi phạm biển cấm đi ngược chiều:

- Phạt tiền từ **4.000.000 đồng đến 6.000.000 đồng** và bị trừ **02 điểm** giấy phép lái xe (Theo Điều 7, Khoản 7, Điểm a và Khoản 13, Điểm a, Nghị định số 168\_2024\_NĐ-CP).
- Nếu hành vi này gây tai nạn giao thông, phạt tiền từ **10.000.000 đồng đến 14.000.000 đồng** và bị trừ **10 điểm** giấy phép lái xe (Theo Điều 7, Khoản 10, Điểm d và Khoản 13, Điểm d, Nghị định số 168\_2024\_NĐ-CP).

Nguồn tham khảo (Văn bản pháp luật):

- [Nghị định số 168/2024/NĐ-CP](#)

Hình 7. Ghi nhớ ngữ cảnh

# Mục lục

1 Giới thiệu

2 Dữ liệu

3 Quy trình xử lý

**4 Độ đo đánh Giá**



# Đánh Giá Truy Vấn

## Mục tiêu Đánh giá

- Đo hiệu quả truy vấn trong RAG.

## Bộ dữ liệu Đánh giá

- 300 câu hỏi tổng hợp thủ công về Luật GTDB.
- Ground truth: các ID của các chunk liên quan.

## Cách thức đánh giá

- Dựa trên các độ đo Recall@k, MRR@k, nDCG@k khi tính trên kết quả truy vấn được với ground truth.

# Đánh giá truy vấn: Dữ liệu

## Ví dụ dữ liệu đánh giá truy vấn

```
[
  {
    "query_id": "eval_001",
    "query": "Mức phạt không đội mũ bảo hiểm xe máy ...?",
    "relevant_chunk_ids": ["168_2024_NĐ-CP_161", ...]
  },
  {
    "query_id": "eval_002",
    "query": "Chạy quá tốc độ 10 km/h ô tô bị phạt ...",
    "relevant_chunk_ids": [
      "168_2024_NĐ-CP_chunk_103",
      "168_2024_NĐ-CP_chunk_127"
    ]
  },
  ...
]
```

# Đánh giá truy vấn: Độ đo Recall@k

$$\text{Recall@}k = \frac{|S_q \cap R_q|}{|R_q|}$$

- Đo lường **tỉ lệ** tài liệu liên quan ( $R_q$ ) thực sự được trả về trong top- $k$  kết quả ( $S_q$ ).
- Quan trọng để đánh giá **độ bao phủ** (không bỏ sót) của hệ thống tìm kiếm.
- Thường tính ở các ngưỡng  $k = 3, 5, 10$  để so sánh hiệu quả. Thêm  $k=15$  vì có những điều luật có rất nhiều Điểm.

# Đánh giá truy vấn: Độ đo MRR@k & nDCG@k

$$\text{MRR} = \frac{1}{|Q|} \sum_{q=1}^{|Q|} \frac{1}{\text{rank}_q} \quad , \quad \text{nDCG@k} = \frac{1}{\text{IDCG@k}} \sum_{i=1}^k \frac{\text{rel}_i}{\log_2(i+1)}$$

- **MRR:** Tập trung vào vị trí xuất hiện của kết quả liên quan đầu tiên. Giá trị càng cao  $\rightarrow$  kết quả đúng xuất hiện càng sớm.
- **nDCG@k:** Ưu tiên xếp những đoạn có *độ liên quan* cao ở đầu. Tính điểm giảm giá (discount) theo vị trí để phản ánh chất lượng ranking.
- Cả hai chỉ số giúp đánh giá *chất lượng thứ tự* của kết quả, không chỉ số lượng.

# Đánh Giá Câu Trả lời

## Mục tiêu Đánh giá

- Đo hiệu quả trả lời của chatbot.

## Bộ dữ liệu Đánh giá

- 300 câu hỏi tổng hợp thủ công về Luật GTDB.
  - Ground truth: các thông tin đầy đủ của các chunk liên quan.
- 282 câu hỏi (không chứa hình ảnh) trong bộ đề 600 câu hỏi GPLX 2025.

## Cách thức đánh giá

- Với 300 câu hỏi thủ công, đánh giá bán tự động bằng GPT4o.
- Với 282 câu hỏi lí thuyết, đánh giá bằng số đáp án chatbot trả lời đúng.

# Đánh giá câu trả lời: Dữ liệu đánh giá thủ công

## Ví dụ dữ liệu đánh giá câu trả lời

```
[{
  "query_id": "eval_001",
  "query": "Mức phạt không đội mũ BH xe máy 2025?",
  "relevant_chunk_ids": {
    "text": "Chương I NHỮNG QUY ĐỊNH CHUNG\nĐiều 9. ...",
    "metadata": {
      "source": "Luật số 36_2024_QH15",
      "effective_date": "1-1-2025",
      "url": "https://thuvienphapluat.vn/van-ban/...",
      "context": { "khoan": "2", "dieu": "9", "chuong": "1"
    }
  }
},
...
]
```

# Đánh giá câu trả lời: Dữ liệu lí thuyết GPLX

**Câu 25. Hành vi của người điều khiển xe ô tô và các loại xe tương tự khi tham gia giao thông đường bộ mà trong máu hoặc hơi thở có nồng độ cồn thì bị áp dụng hình thức xử phạt vi phạm hành chính nào dưới đây?**

1. Bị phạt tiền.
2. Có thể bị tước giấy phép lái xe.
3. Cả hai ý trên.

**Câu 26. Theo Luật Phòng chống tác hại của rượu, bia, đối tượng nào dưới đây bị cấm sử dụng rượu, bia khi tham gia giao thông?**

1. Người điều khiển xe ô tô, xe mô tô, xe đạp, xe gắn máy.
2. Người được chở trên xe cơ giới.
3. Cả hai ý trên.

*Hình 8. Câu hỏi trắc nghiệm lí thuyết GPLX*

# Đánh giá câu trả lời: Đánh giá bán tự động

Câu trả lời từ chatbot được kết hợp với các dữ liệu đánh giá tương ứng cho câu hỏi và **đưa vào một prompt chuẩn. 4 tiêu chí chấm điểm chính (thang điểm: 1 – rất kém, 5 – xuất sắc):**

- ❶ **Tính chính xác và đầy đủ:** – Trả lời đúng theo luật, không thêm/sai thông tin.
- ❷ **Sự liên quan:** – Trả lời đúng trọng tâm, không lan man.
- ❸ **Cấu trúc và ngôn ngữ:** – Trình bày rõ ràng, dễ đọc, hợp lý.
- ❹ **Khả năng trích dẫn nguồn:** – Dẫn nguồn đúng định dạng, đúng nội dung pháp lý.



# Đánh giá câu trả lời: Kết quả

- Với 300 câu hỏi tổng hợp thủ công:

**Bảng 5:** Điểm đánh giá 300 câu trả lời bởi GPT4o

Tiêu chí	Tính Chính Xác và Đầy Đủ	Sự Liên Quan	Cấu Trúc và Ngôn Ngữ	Khả năng trích dẫn nguồn
Điểm số	3.041	4.048	4.696	3.027

- Với 282 câu hỏi trắc nghiệm: chatbot trả lời đúng 267 câu (đạt độ chính xác 94,68%)

**Cảm ơn thầy và các bạn  
đã lắng nghe!**

Q&A