XÂY DỰNG HỆ THỐNG HỎI ĐÁP PHÁP LUẬT VIỆT NAM DỰA TRÊN RAG

Bùi Quốc Bảo
23520091
Khoa Khoa học & Kỹ thuật thông tin
Trường Đại học Công Nghệ Thông Tin
23520091@gm.uit.edu.vn

Huỳnh Phát Đạt
24520270
Khoa Khoa học & Kỹ thuật thông tin
Trường Đại học Công Nghệ Thông Tin
24520270@gm.uit.edu.vn

Lê Minh Khôi
23520767
Khoa Khoa học & Kỹ thuật thông tin
Trường Đại học Công Nghệ Thông Tin
23520767@gm.uit.edu.vn

Tóm tắt - Đồ án này nhóm tập trung tìm hiểu và xây dựng hệ thống hỏi đáp pháp luật Việt Nam dựa trên kỹ thuật Truy xuất tăng cường (RAG). Mục tiêu chính là tinh chỉnh các mô hình ngôn ngữ nhằm trích xuất và truy xuất chính xác thông tin từ các văn bản pháp luật có cấu trúc phức tạp. Quy trình thực hiện bao gồm việc tái sử dụng mô hình Embedding đã được tinh chỉnh trên bộ dữ liệu, tinh chỉnh hai mô hình cốt lõi là mô hình BERT cho tác vụ trích xuất câu trả lời (MRC) và mô hình Rerank để sắp xếp lại các kết quả truy xuất. Bộ dữ liệu phục vụ cho việc truy xuất được nhóm lấy từ nguồn uy tín là Thư viện Pháp luật và bộ dữ liệu dùng để tinh chỉnh mô hình lấy từ cuộc thi Zalo AI Challenge 2021. Kết quả thực nghiệm cho thấy phương pháp tinh chỉnh mang lại hiệu quả rõ rệt. Cụ thể, mô hình BERT sau khi tinh chỉnh đạt chỉ số F1-score 82.39% và EM 73.09%; mô hình Rerank cũng cải thiện độ chính xác so với các mô hình chưa được tinh chỉnh đạt chỉ số F1-score 0.99, precision 0.99 và recall 1.00. Kết quả đồ án khẳng định việc tinh chỉnh mô hình chuyên biệt giúp hệ thống hỏi đáp hoạt động ổn định và chính xác hơn trong lĩnh vực pháp luật Việt Nam.

I. GIỚI THIỆU

1.1. Lý do chọn đề tài

Trong lĩnh vực xử lý ngôn ngữ tự nhiên, việc xây dựng hệ thống hỏi đáp tự động cho văn bản pháp luật luôn là một thách thức lớn. Các văn bản pháp luật Việt Nam thường có cấu trúc phức tạp, ngôn ngữ chuyên ngành chặt chẽ và các mối liên kết ngữ nghĩa tinh vi. Việc sử dụng các mô hình ngôn ngữ lớn (LLM) thuần túy đôi khi dẫn đến hiện tượng phản hồi thông tin không chính xác hoặc thiếu căn cứ pháp lý. Do đó, việc kết hợp giữa khả năng truy xuất dữ liệu và mô hình tạo sinh (RAG) là hướng đi cần thiết để đảm bảo độ tin cậy của câu trả lời.

1.2. Các bài toán trọng tâm

Đồ án này nhóm đã tập trung tìm hiểu sâu vào quy trình tối ưu hóa các thành phần trong hệ thống hỏi đáp thông qua ba bài toán chính:

- **Bài toán Trích xuất thông tin (Machine Reading Comprehension - MRC):** Đây là bài toán đòi hỏi mô hình hiểu ý nghĩa và ngữ cảnh của câu hỏi từ người dùng và đoạn văn bản pháp luật để trích xuất ra câu trả lời ngắn gọn, đầy đủ và chính xác.
- **Bài toán Biểu diễn ngữ nghĩa (Embedding) và Truy xuất tăng cường (RAG):** Hệ thống sử dụng mô hình Embedding để chuyển đổi văn bản thành các vector trong không gian đa chiều. Điều này cho phép hệ thống tìm kiếm tài liệu dựa trên sự tương đồng về mặt từ ngữ giữa câu hỏi của người dùng và các điều luật.
- **Bài toán Đánh giá và Sắp xếp (Reranking):** Sau khi truy xuất bằng sự tương đồng về mặt từ ngữ được danh sách các tài liệu tiềm năng, mô hình Rerank đóng vai trò kiểm soát chất lượng bằng cách đánh giá lại mức độ liên quan về ý nghĩa và ngữ cảnh của từng đoạn văn so với câu hỏi người dùng, giúp loại bỏ các thông tin nhiễu và cung cấp đầu vào tinh gọn và chính xác nhất cho quá trình tạo sinh câu trả lời.

1.3. Mục tiêu

Mục tiêu cốt lõi của đồ án là thực hiện tinh chỉnh (fine-tuning) các mô hình nền tảng để thích nghi với dữ liệu pháp luật Việt Nam từ nguồn dữ liệu Thư viện Pháp luật và bộ dữ liệu chuyên dụng từ cuộc thi Zalo AI Challenge 2021. Nhóm tập trung vào việc tìm ra phương pháp tối ưu nhất để tái sử dụng mô hình Embedding (Bi-Encoder) đã được tinh chỉnh và hoàn thiện, tinh chỉnh mô hình BERT cho tác vụ trích xuất câu trả lời và mô hình Reranker (Cross-Encoder) để đảm bảo khả năng xếp hạng tài liệu chính xác nhất nhằm nâng cao độ chính xác khi truy vấn trước khi thực hiện demo kết quả.

II. CƠ SỞ LÝ THUYẾT

2.1. Bài toán Trích xuất thông tin từ văn bản (Machine Reading Comprehension - MRC)

Machine Reading Comprehension là một lĩnh vực quan trọng trong xử lý ngôn ngữ tự nhiên, tập trung vào khả năng hiểu văn bản và trả lời các câu hỏi liên quan của máy tính. Trong bài toán này, mô hình được cung cấp một đoạn văn bản

ngữ cảnh (Context) và một câu hỏi (Question), nhiệm vụ của nó là xác định chính xác vị trí bắt đầu và vị trí kết thúc của câu trả lời (Answer) ngay trong đoạn văn bản đó. Đối với dữ liệu pháp luật, MRC đòi hỏi mô hình phải có khả năng lý luận logic và hiểu sâu sắc các mối quan hệ ngữ nghĩa để trích xuất thông tin chính xác nhất.

2.2. Kiến trúc mô hình BERT trong nhiệm vụ trích xuất BERT (Bidirectional Encoder Representations from Transformers)

Đây là là một kiến trúc dựa trên Transformer được thiết kế để học biểu diễn ngôn ngữ hai chiều. Đối với tác vụ trích xuất thông tin, mô hình sử dụng các lớp Attention để hiểu mối quan hệ giữa câu hỏi và ngữ cảnh. Quá trình fine-tune mô hình này hướng đến việc tối ưu hóa hàm mất mát (Loss function) dựa trên tổng Log-Likelihood của các dự đoán vị trí chính xác:

$$Loss = -\sum_{i=1}^{N} [y_{start[i]} \cdot logp_{start[i]} + y_{end[i]} \cdot logp_{end[i]}]$$

Trong đó, N là tổng số lượng mẫu dữ liệu; $y_{start[i]}, y_{end[i]}$ Là nhãn thực tế (ground truth) của vị trí bắt đầu và kết thúc của câu trả lời; $p_{start[i]}, p_{end[i]}$ là xác suất mô hình dự đoán cho vị trí bắt đầu và kết thúc.

Việc fine-tune giúp mô hình điều chỉnh các trọng số từ vựng và vị trí để thích nghi tốt hơn với cấu trúc đặc thù của dữ liệu chuyên ngành.

2.3. Mô hình Embedding và Kiến trúc Bi-Encoder

Mô hình Embedding đóng vai trò then chốt trong việc chuyển đổi các đoạn văn bản pháp luật thành các vectơ đặc trưng trong không gian đa chiều, cho phép hệ thống nắm bắt được ý nghĩa ngữ nghĩa thực tế thay vì chỉ so khớp từ khóa đơn thuần. Đồ án sử dụng các mô hình pre-trained (đã được huấn luyện trước) và tối ưu hóa cho tiếng Việt dựa trên kiến trúc Bi-Encoder. Trong kiến trúc này, câu hỏi của người dùng và các tài liệu pháp luật được mã hóa độc lập qua các nhánh BERT song song để tạo ra các vectơ biểu diễn cố định. Sau khi các văn bản được chuyển đổi sang dạng số (vectơ), mức độ liên quan giữa chúng sẽ được xác định thông qua hàm Cosine Similarity:

$$sim(a, b) = \frac{a \cdot b}{\|a\| \cdot \|b\|}$$

Trong đó a và b lần lượt là vectơ của câu hỏi và tài liệu pháp luật. Việc sử dụng mô hình pre-trained đã qua tinh chỉnh giúp hệ thống tận dụng được khả năng hiểu ngôn ngữ mạnh mẽ, từ đó truy xuất nhanh chóng và chính xác các điều luật có nội dung sát nhất với câu hỏi mà người dùng đặt ra.

2.4. Mô hình Rerank và kiến trúc Cross-Encoder

Mô hình Reranker đóng vai trò quan trọng trong việc đánh giá lại và sắp xếp các đoạn văn bản theo mức độ phù hợp nhất với truy vấn. Kiến trúc Cross-Encoder nhận đầu vào là cặp câu hỏi và tài liệu nối với nhau, sau đó trích xuất biểu diễn ngữ nghĩa toàn cục để dự đoán điểm số liên quan. Quá trình fine-tune mô hình này sử dụng hàm mất mát Cross-Entropy để tối ưu hóa khả năng phân biệt giữa các cặp liên quan và không liên quan:

$$L = -\frac{1}{N} \sum_{i=1}^{N} y_i \cdot log(a_i)$$

Trong đó, y_i là nhãn thực tế thể hiện mức độ liên quan và a_i là xác suất dự đoán của mô hình.

2.5. Kiến trúc Truy xuất tăng cường (Retrieval-Augmented Generation - RAG)

Truy xuất tăng cường (RAG) là một khung kiến trúc kết hợp giữa khả năng truy xuất thông tin chính xác từ kho dữ liệu và khả năng tạo sinh ngôn ngữ tự nhiên của các mô hình ngôn ngữ lớn (LLM). Trong lĩnh vực pháp luật, nơi yêu cầu sự chính xác tuyệt đối về căn cứ pháp lý, RAG giúp mô hình tránh được hiện tượng ảo tưởng thông tin bằng cách cung cấp các đoạn văn bản luật thực tế làm ngữ cảnh bổ sung. Kiến trúc RAG trong đồ án này không chỉ dừng lại ở việc tìm kiếm đơn thuần mà là một quy trình tối ưu hóa đa bước, bao gồm: truy xuất ngữ nghĩa dựa trên các mô hình embedding đã tinh chỉnh, lọc tài liệu qua các bộ lọc metadata, xếp hạng lại (Reranking) để chọn lọc thông tin phù hợp nhất và trích xuất nội dung trọng tâm (Extraction) nhằm tối ưu hóa ngữ cảnh đầu vào cho mô hình tạo sinh.

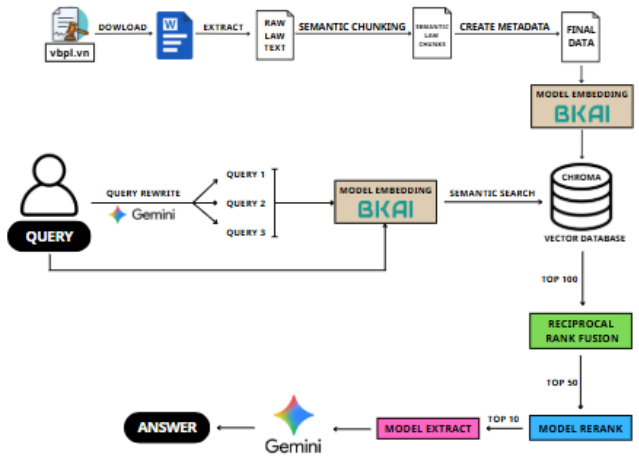
III. DỮ LIỆU

Bộ dữ liệu ViNLI_Zalo, được cung cấp từ cuộc thi Zalo AI Challenge 2021, là nguồn dữ liệu quan trọng dùng để tinh chỉnh các mô hình thành phần trong hệ thống. Cấu trúc bộ dữ liệu bao gồm ba trường thông tin chính: *query* (câu hỏi về lĩnh vực pháp luật), *positive* (đoạn văn bản luật chứa câu trả lời chính xác) và *hard_neg* (các đoạn văn bản luật không liên quan nhưng có độ tương đồng cao về mặt từ ngữ). Tập dữ liệu này đóng vai trò chủ đạo trong việc huấn luyện mô hình Rerank để sắp xếp kết quả và mô hình Trích xuất thông tin (Extract/MRC) để tìm câu trả lời ngắn.

Bên cạnh đó, đồ án còn sử dụng bộ dữ liệu thangvip/legal-documents-splitted được cung cấp bởi nền tảng Hugging Face. Đây là tập hợp các văn bản pháp lý tiếng Việt quy mô lớn, bao gồm các luật, nghị định, thông tư và quyết định chính thống đã được tiền xử lý và chia nhỏ thành các đoạn văn (chunks) giàu ngữ cảnh. Nguồn dữ liệu này được nhóm sử dụng làm ngữ liệu để tinh chỉnh mô hình Embedding, giúp tăng khả năng biểu diễn ngữ nghĩa cho các văn bản pháp luật đặc thù.

Cuối cùng là các văn bản luật được thu thập từ trang 'Cơ sở dữ liệu Quốc gia về văn bản quy phạm pháp luật'. Các văn bản, nghị định tại đây được thu thập, xử lý chuẩn hóa và lưu trữ dưới dạng cơ sở dữ liệu tri thức (Knowledge Base). Đây chính là nền tảng dữ liệu gốc để hệ thống thực hiện quy trình truy xuất và phản hồi thông tin trong kiến trúc RAG.

IV. KIẾN TRÚC MÔ HÌNH



Kiến trúc mô hình

4.1. Tổng quan về kiến trúc hệ thống

Kiến trúc mô hình được thiết kế theo quy trình xử lý đa tầng nhằm tối ưu hóa việc truy xuất và phản hồi thông tin pháp luật chính xác. Hệ thống được chia làm hai luồng hoạt động chính: luồng chuẩn bị dữ liệu để xây dựng kho tri thức tri thức và luồng xử lý truy vấn từ người dùng. Quy trình này kết hợp linh hoạt giữa các mô hình ngôn ngữ lớn (LLM), các mô hình nhúng (Embedding), xếp hạng (Rerank) và trích xuất (Extract) đã được tinh chỉnh để đảm bảo thông tin đầu ra bám sát văn bản pháp luật hiện hành.

4.2. Giai đoạn chuẩn bị dữ liệu và xây dựng kho tri thức

Quá trình xây dựng kho tri thức bắt đầu từ việc thu thập các văn bản quy phạm pháp luật từ nguồn chính thống như Cơ sở dữ liệu Quốc gia về Văn bản pháp luật (vbpl.vn). Các văn bản sau khi tải về được chuyển đổi sang định dạng .docx và tiến hành trích xuất nội dung thô. Để đảm bảo tính liền mạch về mặt ngữ nghĩa, hệ thống áp dụng kỹ thuật chia đoạn ngữ nghĩa (Semantic Chunking) thay vì cắt đoạn theo độ dài cố định và theo từng điều luật nhất định. Mỗi đoạn văn bản sau đó được gán kèm các thông tin metadata bổ trợ để phục vụ cho việc lọc dữ liệu. Cuối cùng, dữ liệu được mã hóa thành các vector thông qua mô hình Embedding (BKAI) và lưu trữ trong cơ sở dữ liệu vector (Chroma - Vector Database) để sẵn sàng cho các bước truy vấn tiếp theo.

4.3. Quy trình xử lý và mở rộng truy vấn

Khi người dùng nhập câu hỏi (Query), hệ thống không thực hiện tìm kiếm ngay lập tức mà đi qua một bước tối ưu hóa truy vấn. Thông qua mô hình Gemini, hệ thống thực hiện kỹ thuật tái viết câu hỏi (Query Rewrite) để sinh ra các phiên bản truy vấn tương đương, giúp mở rộng không gian tìm kiếm ngữ nghĩa. Các truy vấn mở rộng này sau đó được mã hóa bởi cùng một mô hình Embedding đã sử dụng trong giai đoạn chuẩn bị dữ liệu. Việc này đảm bảo tính đồng nhất giữa câu hỏi và kho tri thức, cho phép thực hiện tìm kiếm ngữ nghĩa (Semantic Search) để thu thập danh sách 100 tài liệu tiềm năng nhất từ cơ sở dữ liệu vector.

4.4. Cơ chế truy xuất, xếp hạng và trích xuất thông tin

Sau khi có danh sách các tài liệu ban đầu, hệ thống áp dụng thuật toán hợp nhất thứ hạng (Reciprocal Rank Fusion - RRF) để sàng lọc và chọn ra 50 tài liệu có mức độ phù hợp cao nhất. Các tài liệu này tiếp tục được đưa qua mô hình xếp hạng lại (Model Rerank) đã được tinh chỉnh để đánh giá sâu hơn mối tương quan ngữ nghĩa, từ đó rút gọn xuống 10 tài liệu chất lượng nhất. Để tối ưu hóa ngữ cảnh đầu vào cho bước cuối cùng, mô hình trích xuất (Model Extract) dựa trên kiến trúc BERT sẽ quét qua các tài liệu này để lấy ra những phân đoạn thông tin cốt lõi nhất, loại bỏ các thành phần dư thừa không liên quan đến câu hỏi.

4.5. Tổng hợp và tạo sinh câu trả lời

Ở bước cuối cùng, phần nội dung đã được trích xuất tinh gọn cùng với câu hỏi gốc của người dùng được đưa vào mô hình ngôn ngữ lớn Gemini. Tại đây, mô hình thực hiện vai trò tổng hợp và diễn đạt thông tin thành một câu trả lời hoàn chỉnh, tự nhiên và chính xác về mặt pháp lý. Quy trình đa lớp này giúp hệ thống không chỉ cung cấp câu trả lời có căn cứ mà còn tối ưu hóa được chi phí vận hành thông qua việc giảm thiểu lượng ngữ cảnh không cần thiết khi gọi mô hình sinh.

V. THỰC NGHIỆM

5.1. Mô hình Embedding

Thực hiện nhiệm vụ mã hóa văn bản (Embedding) và truy vấn (Retrieval) trong hệ thống là mô hình bkai-foundation-models/vietnamese-bi-encoder. Đây là một mô hình dạng Sentence-Transformers được phát triển bởi Trung tâm Quốc tế Nghiên cứu về Trí tuệ Nhân tạo (BKAI) của Đại học Bách Khoa Hà Nội.

Nhóm quyết định lựa chọn mô hình này dựa trên những ưu điểm vượt trội về kiến trúc và tập dữ liệu huấn luyện:

- Kiến trúc nền tảng mạnh mẽ: Mô hình được tinh chỉnh (fine-tuned) từ PhoBERT v2, đây là kiến trúc ngôn ngữ hàng đầu cho tiếng Việt hiện nay giúp tối ưu hóa khả năng hiểu ngữ pháp và ngữ nghĩa đặc thù của tiếng Việt.
- Dữ liệu huấn luyện phù hợp với các văn bản pháp luật: Điểm đặc biệt của mô hình này là đã được huấn luyện trên sự kết hợp của nhiều tập dữ liệu quy mô lớn, bao gồm MS MARCO và SQuAD v2 (được dịch sang tiếng Việt) và đặc biệt là bộ dữ liệu ViNLI_Zalo giúp model đạt hiệu quả cao trong việc tìm kiếm các đoạn luật có sự tương đồng về mặt ngữ nghĩa thực tế thay vì chỉ dựa trên các từ khóa thuần túy.
- Hiệu năng thực nghiệm: Theo các báo cáo của tác giả, mô hình đạt được kết quả ấn tượng trên tập dữ liệu luật với chỉ số Accuracy@1 đạt 73.28% và Accuracy@10 lên tới 93.59%, chứng minh sự phù hợp của mô hình cho hệ thống hỏi đáp và truy vấn pháp luật.

5.2. Tinh chỉnh mô hình Rerank

Mục tiêu của thực nghiệm là tạo ra một mô hình có phân biệt rõ các văn bản phù hợp và không phù hợp với với câu truy vấn đầu vào. Từ đó cải thiện độ chính xác của hệ thống truy xuất.

5.2.1. Quá trình chuẩn bị dữ liệu

Dữ liệu sử dụng để tinh chỉnh mô hình Rerank là tập dữ liệu anti-ai/ViNLI-Zalo-supervised. Từ định dạng gốc, dữ liệu được chuyển đổi thành các cặp (query, document) đi kèm với nhãn nhị phân (label). Trong đó, các văn bản liên quan trực tiếp được gán nhãn 1 (Positive) và các văn bản không

liên quan nhưng có độ gây nhiễu cao (Hard Negative) được gán nhãn 0.

5.2.2. Thiết lập bài toán

Mô hình Rerank được thiết lập theo kiến trúc Cross-Encoder nhằm tối ưu hóa sự tương tác giữa câu hỏi và văn bản pháp luật. Đầu vào của mô hình là một chuỗi nối theo định dạng [CLS] query [SEP] document [SEP], được mã hóa qua các lớp Attention của Transformer để rút trích biểu diễn ngữ nghĩa toàn cục. Tại đây, vector biểu diễn tại vị trí token [CLS] (nơi hội tụ thông tin tương tác của cả cặp câu) sẽ được đưa qua một lớp đầu ra (output layer) để dự báo điểm số liên quan. Quá trình tối ưu hóa sử dụng hàm mất mát Binary Cross-Entropy Loss.

5.2.3. Huấn luyện

Nhóm thực hiện fine-tune trên mô hình nền gte-multilingual-reranker-base với độ dài chuỗi tối đa là 512 tokens. Quá trình huấn luyện sử dụng learning rate là 2e-5, kết hợp kỹ thuật tích lũy gradient để đạt được batch size hiệu dụng là 64. Để tăng tốc độ tính toán và tiết kiệm bộ nhớ GPU, kỹ thuật FP16 được áp dụng cùng với trọng số giảm dần (weight_decay) là 0.25. Nhóm cũng thiết lập cơ chế Early Stopping với patience là 3, cho phép dừng huấn luyện và tự động tải lại phiên bản tốt nhất khi hàm mất mát trên tập thẩm định không còn cải thiện, giúp tránh hiện tượng quá khớp (overfitting).

5.2.4. Kết quả thực nghiệm

Sau khi fine-tune, mô hình đạt hiệu suất vượt trội trên tập kiểm thử với độ chính xác (Accuracy) và chỉ số F1-Score đạt xấp xỉ 99,91%. Đặc biệt, chỉ số Recall đạt mức tuyệt đối (100%), cho thấy hệ thống không bỏ sót bất kỳ tài liệu liên quan nào. So với các mô hình pre-trained mạnh như *bge-reranker-v2-m3* (92,95%) hay *jina-reranker-v2-base* (87,49%), mô hình của nhóm đề xuất cho thấy sự cải thiện rõ rệt, khẳng định hiệu quả vượt trội khi được tinh chỉnh chuyên biệt cho dữ liệu pháp luật Việt Nam.

5.3. Tinh chỉnh mô hình Extract

Mục tiêu của thực nghiệm này là xây dựng mô hình có khả năng xác định chính xác vị trí bắt đầu và kết thúc của câu trả lời trong đoạn văn bản pháp luật, phục vụ cho bước trích xuất thông tin trọng tâm trong hệ thống RAG.

5.3.1. Quá trình chuẩn bị dữ liệu

Để huấn luyện mô hình MRC, dữ liệu cần có ba thành phần chính: đoạn văn chứa thông tin (context), câu hỏi về thông tin trong đoạn văn (question) và câu trả lời cụ thể kèm vị trí bắt đầu trong đoạn văn (answer). Bộ dữ liệu ViNLI Zalo ban đầu chỉ có 3 cột là query (câu hỏi), positive (đoạn văn bản chứa câu trả lời) và hard_neg (đoạn văn bản không chứa câu trả lời). Nhóm tiến hành bỏ đi 2 cột là query và hard_neg, sau đó đổi tên cột positive thành context. Sau đó dựa vào cột context, sử dụng các mô hình Generative AI như gemini để sinh ra câu hỏi và câu trả lời dựa vào đoạn context. Kết quả cuối cùng nhóm sẽ có được bộ dữ liệu có đủ 3 thành phần chính như trên.

5.3.2. Thiết lập bài toán

Bài toán được xác định là Trích xuất câu trả lời (Extractive Question Answering). Mục tiêu là xây dựng một mô hình có khả năng đọc hiểu một đoạn văn bản pháp luật (Context) và một câu hỏi (Question), sau đó xác định chính xác vị trí bắt đầu (Start position) và kết thúc (End position) của đoạn văn bản chứa câu trả lời.

5.3.3. Huấn luyện

Nhóm thực hiện fine-tune trên mô hình nền google-bert/bert-base-multilingual-cased. Quá trình huấn luyện sử dụng learning rate là 3e-5, batch size hiệu dụng là 8, weigh_decay là 0.01 để giảm thiểu overfitting. Nhóm cũng thiết lập cơ chế Early Stopping với patience là 3, nếu chỉ số kiểm thử (validation loss) không cải thiện sau 3 epoch liên tiếp, quá trình huấn luyện sẽ dừng lại để tiết kiệm tài nguyên và chọn mô hình tốt nhất.

5.3.4. Kết quả thực nghiệm

Mô hình sau khi huấn luyện được đánh giá trên tập kiểm thử (Test set) sử dụng bộ đo chuẩn cho bài. Kết quả đạt được rất khả quan, Exact Match đạt 73.09% (độ đo cho biết tỷ lệ câu trả lời do mô hình dự đoán trùng khớp hoàn toàn với câu trả lời mẫu), F1 Score đạt 82.39% (độ đo trung bình điều hòa giữa độ chính xác Precision và độ phủ Recall), cho thấy khả năng mô hình tìm được vùng thông tin chồng lấp tốt với đáp án chuẩn.

5.4. Đánh giá tổng thể hệ thống

Sau khi thực nghiệm các thành phần riêng lẻ, nhóm tiến hành đánh giá tổng thể hệ thống trên nhiều cấu hình khác nhau để xác định sự tác động của từng module (Query Rewrite, Reranker và Extract) đối với chất lượng câu trả lời và hiệu suất tính toán. Kết quả được đánh giá trên các chỉ số: ROUGE-L (độ tương đồng văn bản), Semantic (độ tương đồng ngữ nghĩa) và Lượng token đầu vào trung bình (chi phí).

	ROUGE - L	Semantic	Average Input Token on 10 Queries
Baseline	0.4332	0.6807	158111
Baseline + Rewrite	**0.4499**	0.6632	78347
Baseline + Rewrite + Rerank	0.4416	**0.7232**	19362
Baseline + Rewrite + Rerank + Extract	0.3914	0.6748	919

Kết quả đánh giá tổng thể

Cấu hình Baseline + Rewrite + Rerank đạt điểm số Semantic cao nhất (0.7232). Điều này chứng minh rằng việc tích hợp mô hình Rerank đã được fine-tune giúp lọc bỏ các văn bản nhiễu từ bộ truy xuất, chỉ giữ lại những căn cứ pháp lý có độ liên quan cao nhất để đưa vào mô hình ngôn ngữ lớn (LLM). Việc tập trung vào ngữ nghĩa thay vì chỉ khớp từ vựng giúp câu trả lời cuối cùng chính xác và sát với thực tế pháp luật hơn.

Module Rewrite giúp cải thiện chỉ số ROUGE-L lên mức cao nhất (0.4499) do khả năng làm rõ câu hỏi người dùng, giúp bộ truy xuất tìm được các văn bản có cấu trúc từ vựng gần nhất. Tuy nhiên, khi kết hợp thêm Rerank, mặc dù ROUGE-L giảm nhẹ nhưng độ tương đồng ngữ nghĩa lại tăng mạnh. Điều này cho thấy hệ thống đã chuyển dịch từ việc khớp từ ngữ sang hiểu ý nghĩa, một yếu tố then chốt trong tư vấn pháp luật.

Một trong những kết quả ấn tượng nhất là khả năng tiết kiệm token đầu vào:

- Sử dụng Reranker giúp giảm lượng token từ 78,347 xuống còn 19,362 (giảm khoảng 75%) mà vẫn tăng chất lượng câu trả lời. Điều này là do Reranker đã loại bỏ các đoạn văn bản dư thừa trước khi gửi đến LLM.
- Khi tích hợp thêm module Extract, lượng token giảm xuống mức cực thấp (919 tokens). Tuy nhiên, chỉ số Semantic cũng giảm theo (từ 0.7232 xuống 0.6748). Điều này cho thấy việc trích xuất quá sâu có thể làm mất đi một số ngữ cảnh quan trọng cần thiết để LLM tổng hợp câu trả lời đầy đủ.

VI. KẾT LUẬN

6.1. Kết luận

Nhóm đã xây dựng thành công hệ thống chatbot hỏi đáp về văn bản quy phạm pháp luật Việt Nam dựa trên kiến trúc Truy xuất tăng cường (RAG) nâng cao. Bằng cách kết hợp linh hoạt giữa kiến trúc RAG cơ bản với các mô hình Rerank (Tái sắp xếp) và Extract (Trích xuất thông tin), hệ thống đã đạt được những kết quả thực nghiệm khả quan:

- Tối ưu hóa hiệu năng và độ chính xác: Việc tích hợp module Rerank giúp tinh lọc các kết quả truy xuất, đảm bảo những đoạn văn bản luật có giá trị nhất được đưa vào mô hình trích xuất. Kết quả thực nghiệm cho thấy sự kết hợp này nâng cao đáng kể độ chính xác của câu trả lời cuối cùng.
- Tiết kiệm tài nguyên và chi phí: Thông qua việc sử dụng mô hình Extract chuyên biệt để xử lý các tác vụ trung gian, hệ thống đã giảm thiểu tối đa tần suất và khối lượng dữ liệu cần gửi tới các mô hình ngôn ngữ lớn (LLM) như Gemini. Điều này không chỉ giúp tối ưu hóa chi phí vận hành (API calls) mà còn giảm thời gian phản hồi (latency) của hệ thống.
- Tính thực tiễn cao: Hệ thống đã khai thác hiệu quả nguồn dữ liệu chính thống từ Cơ sở dữ liệu Quốc gia về văn bản quy phạm pháp luật, kết hợp cùng kỹ thuật Semantic Chunking để duy trì tính toàn vẹn của ngữ nghĩa trong các điều luật phức tạp.

6.2. Hướng phát triển

Dù đã đạt được những mục tiêu đề ra, hệ thống vẫn có thể được cải tiến thông qua các hướng nghiên cứu sau:

- Tối ưu hóa Vector Database: Nghiên cứu áp dụng kỹ thuật Matryoshka Representation Learning (MRL) để linh hoạt hóa kích thước vector embedding, giúp tăng tốc độ truy xuất trên quy mô dữ liệu cực lớn mà vẫn đảm bảo hiệu năng.
- Mở rộng kho tri thức: Tích hợp thêm các dạng văn bản bổ trợ như án lệ, quyết định hành chính và các văn bản hướng dẫn thi hành luật để chatbot có thể đưa ra những tư vấn pháp lý chuyên sâu và đa chiều hơn.
- Cải thiện giao diện người dùng: Phát triển thêm các tính năng gợi ý câu hỏi liên quan và hiển thị trực tiếp dẫn dẫn chứng (trích lục điều khoản) trong giao diện để tăng tính minh bạch và độ tin cậy của thông tin pháp luật.

TÀI LIỆU THAM KHẢO

[1] AWS, What is RAG (Retrieval-Augmented Generation)? Link

[2] kirouane Ayoub, “Fine-tune Re-ranking Models : A Beginner’s Guide”. Link

[3] Nils Reimers, Iryna Gurevych ,“ Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks ” ,arXiv:1908.10084, 27 Aug 2019. [Online]. Link

[4] AI VietNam, Vector Storage in RAG. Link