HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



Hoàng Đức Việt

XÂY DỰNG HỆ THỐNG TRA CỨU CHÍNH SÁCH PHÁP LUẬT DỰA TRÊN MÔ HÌNH NGÔN NGỮ LỚN

Chuyên ngành: Hệ thống thông tin
Mã số: 8.48.01.04

TÓM TẮT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

HÀ NỘI - NĂM 2025

Đề án tốt nghiệp được hoàn thành tại:
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học:
.....

Phản biện 1: .....
Phản biện 2: .....

Đề án tốt nghiệp sẽ được bảo vệ trước Hội đồng chấm đề án tốt nghiệp thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ..... giờ ..... ngày ..... tháng ..... năm .....

Có thể tìm hiểu đề án tốt nghiệp tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

I. MỞ ĐẦU

1. Lý do chọn đề tài

An toàn thực phẩm là lĩnh vực quan trọng tại Việt Nam với hệ thống pháp luật tương đối đầy đủ, tuy nhiên việc tiếp cận và tuân thủ các quy định vẫn gặp nhiều khó khăn, đặc biệt đối với doanh nghiệp vừa và nhỏ, hộ kinh doanh, người dân nông thôn và cán bộ quản lý cơ sở, nhất là tại các khu vực hạn chế kết nối Internet. Nguyên nhân chủ yếu xuất phát từ ngôn ngữ pháp lý phức tạp, số lượng văn bản lớn và thường xuyên thay đổi, cùng với việc thiếu các công cụ tra cứu pháp luật thông minh bằng tiếng Việt. Thực trạng này làm gia tăng khoảng cách giữa quy định và thực tiễn áp dụng, dẫn đến nhiều vi phạm về an toàn thực phẩm. Trước bối cảnh đó, sự phát triển của mô hình ngôn ngữ lớn (LLM) và kỹ thuật Retrieval-Augmented Generation (RAG) mở ra hướng tiếp cận mới trong lĩnh vực pháp lý, cho phép xây dựng trợ lý ảo pháp luật ngoại tuyến, hỗ trợ tra cứu chính xác và cung cấp câu trả lời có căn cứ pháp lý rõ ràng.

2. Mục tiêu nghiên cứu

Đề án hướng tới xây dựng một hệ thống chatbot RAG ngoại tuyến chuyên biệt cho tra cứu pháp luật về an toàn thực phẩm tại Việt Nam, đáp ứng yêu cầu về độ chính xác, tốc độ phản hồi và khả năng cập nhật tri thức mà không cần kết nối Internet. Hệ thống được phát triển trên cơ sở tri thức tổng hợp từ hơn 170 văn bản quy phạm pháp luật, sử dụng pipeline RAG ngoại tuyến với mô hình nhúng và mô hình ngôn ngữ lớn cục bộ. Bên cạnh đó, đề án triển khai cơ chế cập nhật tri thức theo dạng đồng bộ gia tăng nhằm rút ngắn thời gian cập nhật, xây dựng giao diện quản trị và chatbot có khả năng trích dẫn nguồn tự động, đồng thời đánh giá hiệu năng hệ thống thông qua các bộ chỉ số phù hợp.

3. Đối tượng và phạm vi nghiên cứu

3.1 Đối tượng nghiên cứu

Đối tượng nghiên cứu của đề án là một hệ thống phần mềm chatbot pháp luật hoàn chỉnh dựa trên kiến trúc Retrieval-Augmented Generation (RAG), hoạt động hoàn toàn ngoại tuyến, nhằm hỗ trợ tra cứu, giải thích và cung cấp căn cứ pháp lý về an toàn thực phẩm tại Việt Nam. Hệ thống tích hợp các thành phần kỹ thuật từ thu thập, tiền xử lý, phân đoạn và gán metadata văn bản pháp luật, tạo embedding và lưu trữ vector, đến cơ chế truy xuất, sinh câu trả lời bằng mô hình ngôn ngữ lớn cục bộ và giao diện người dùng có khả năng trích dẫn nguồn. Hiệu năng hệ thống được đánh giá thông qua các chỉ số về

độ chính xác, tính đầy đủ, mức độ phù hợp ngữ cảnh và thời gian phản hồi trên các kịch bản truy vấn thực tế.

3.2 Phạm vi nghiên cứu

Phạm vi nghiên cứu tập trung vào các văn bản quy phạm pháp luật liên quan trực tiếp đến an toàn thực phẩm tại Việt Nam còn hiệu lực đến tháng 12/2025, với khoảng 170 văn bản chính thức thu thập từ các nguồn công khai. Hệ thống được triển khai bằng các công nghệ mã nguồn mở, vận hành trên máy cá nhân hoặc hạ tầng nội bộ, không phụ thuộc dịch vụ đám mây, nhằm đáp ứng yêu cầu bảo mật và khả năng sử dụng tại khu vực không có Internet. Đối tượng người dùng mục tiêu bao gồm cán bộ quản lý cơ sở, doanh nghiệp và hộ kinh doanh thực phẩm, cũng như người dân ở vùng nông thôn, vùng sâu vùng xa. Đề án không thay thế tư vấn pháp lý chuyên sâu và các kết quả đánh giá chỉ có giá trị trong phạm vi dữ liệu và kịch bản thử nghiệm đã xác định.

4. Phương pháp nghiên cứu

4.1 Phương pháp nghiên cứu lý thuyết

Đề án sử dụng kết hợp phương pháp nghiên cứu lý thuyết và thực nghiệm nhằm xây dựng cơ sở khoa học và kiểm chứng tính khả thi của hệ thống chatbot pháp luật dựa trên kiến trúc RAG ngoại tuyến. Nghiên cứu lý thuyết tập trung phân tích kiến trúc RAG, mô hình ngôn ngữ lớn, kỹ thuật embedding và truy xuất tri thức, đồng thời xem xét đặc thù ngôn ngữ pháp luật tiếng Việt và các yêu cầu triển khai trong môi trường không có Internet. Trên cơ sở đó, đề án xác định các tiêu chí thiết kế và hướng tối ưu hóa hệ thống phù hợp với điều kiện hạn chế tài nguyên.

4.2 Phương pháp nghiên cứu thực nghiệm

Nghiên cứu thực nghiệm được thực hiện thông qua việc xây dựng tập dữ liệu pháp luật an toàn thực phẩm, thử nghiệm và lựa chọn mô hình ngôn ngữ lớn cục bộ phù hợp, triển khai nguyên mẫu hệ thống RAG ngoại tuyến và tích hợp cơ chế cập nhật tri thức theo dạng đồng bộ gia tăng. Chất lượng hệ thống được đánh giá bằng các kịch bản nghiệp vụ thực tế, dựa trên các tiêu chí về độ chính xác pháp lý, tính đầy đủ và mức độ trung thành với văn bản gốc, qua đó phân tích và điều chỉnh mô hình.

4.3 Ý nghĩa và đóng góp của đề tài

Về ý nghĩa và đóng góp, đề án có giá trị khoa học trong việc đề xuất mô hình RAG

của câu trả lời. Kết quả đánh giá cho thấy hệ thống đạt chất lượng tốt, có thể ứng dụng trong tư vấn pháp lý cơ bản, nhưng để triển khai trong môi trường pháp lý chuyên nghiệp cần bổ sung các cơ chế kiểm chứng nghiêm ngặt hơn.

Đề án đề xuất mở rộng và cập nhật kho dữ liệu pháp luật, tích hợp thuật toán reranking để nâng cao độ chính xác truy xuất, phát triển cơ chế kiểm chứng pháp lý bắt buộc dẫn chiếu điều khoản cụ thể, cải thiện khả năng theo dõi lỗi sinh câu trả lời và tăng tính minh bạch thông qua giao diện hiển thị nguồn văn bản. Đồng thời, khuyến nghị các hệ thống AI pháp lý phải đảm bảo minh bạch, truy vết nguồn và cập nhật dữ liệu thường xuyên để duy trì độ tin cậy.

4. Hướng nghiên cứu tiếp theo

Đề tài có thể được mở rộng theo nhiều hướng nhằm nâng cao chất lượng và phạm vi ứng dụng của hệ thống chatbot pháp luật. Các định hướng chính gồm phát triển kiến trúc RAG đa tầng để truy xuất theo thứ bậc văn bản pháp luật; huấn luyện hoặc fine-tune mô hình ngôn ngữ chuyên biệt cho pháp lý Việt Nam nhằm tăng khả năng hiểu thuật ngữ và độ chính xác; tích hợp cơ chế suy luận có cấu trúc và chuỗi lập luận kèm căn cứ pháp lý rõ ràng. Ngoài ra, hệ thống có thể được nâng cấp lên mức chuyên sâu với khả năng so sánh, đối chiếu và phân tích mâu thuẫn giữa các văn bản, hỗ trợ dự đoán cách áp dụng luật trong thực tiễn. Cuối cùng, việc xây dựng bộ benchmark pháp lý Việt Nam chuẩn hóa và phát triển API/chatbot tích hợp sẽ tạo nền tảng cho nghiên cứu, triển khai và ứng dụng rộng rãi trong các hệ thống pháp lý và hành chính công.

ngoại tuyến cho tiếng Việt và cơ chế cập nhật tri thức hiệu quả, đồng thời mang ý nghĩa thực tiễn khi cung cấp công cụ tra cứu pháp luật dễ triển khai tại các khu vực hạn chế Internet và có khả năng mở rộng sang các lĩnh vực pháp luật khác.

4.4 Bố cục đề án

Đề án được chia thành 4 chương:

Chương 1: Cơ sở lý luận và tổng quan các hệ thống hiện có

Chương 2: Xử lý và tổ chức nguồn dữ liệu

Chương 3: Thiết kế kiến trúc và các thành phần của hệ thống

Chương 4: Thực nghiệm và đánh giá kết quả

II. NỘI DUNG

CHƯƠNG 1: CƠ SỞ LÝ LUẬN VÀ TỔNG QUAN VỀ HỆ THỐNG HIỆN CÓ

1.1 Tổng quan về xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP)

1.1.1 Khái niệm

Xử lý ngôn ngữ tự nhiên (NLP) là lĩnh vực thuộc trí tuệ nhân tạo và khoa học máy tính, tập trung phát triển các hệ thống có khả năng hiểu, phân tích và sinh ngôn ngữ con người, hướng tới các ứng dụng thực tiễn trong đời sống và sản xuất.

1.1.2 Các nhánh chính của NLP

NLP gồm hai nhánh chính là hiểu ngôn ngữ tự nhiên (NLU) và sinh ngôn ngữ tự nhiên (NLG), đồng thời thường được kết hợp với nhận dạng giọng nói để xử lý và chuyển đổi giữa ngôn ngữ nói và văn bản.

1.1.3 Vai trò và tầm quan trọng

NLP đóng vai trò quan trọng trong nhiều lĩnh vực như thương mại – dịch vụ, y tế, giáo dục, truyền thông xã hội và khai thác dữ liệu, thông qua các ứng dụng như chatbot, phân tích văn bản, hỗ trợ ra quyết định và hệ thống hỏi – đáp tự động.

1.1.4 Các tác vụ điển hình trong NLP

NLP bao gồm nhiều tác vụ phổ biến như phân tích và phân loại văn bản, dịch máy, nhận diện thực thể, phát hiện thư rác, sinh và tóm tắt văn bản, chatbot và hệ thống hỏi – đáp tự động.

1.1.5 Cách thức hoạt động

Quy trình xử lý ngôn ngữ tự nhiên thường gồm ba giai đoạn chính: tiền xử lý dữ liệu để chuẩn hóa văn bản, trích xuất đặc trưng nhằm biểu diễn ngữ nghĩa dưới dạng vector, và mô hình hóa bằng các thuật toán học máy truyền thống hoặc mô hình học sâu để thực hiện các tác vụ NLP.

PHẦN KẾT LUẬN

1. Kết quả nghiên cứu của đề tài

Đề án đã xây dựng thành công hệ thống hỏi đáp pháp lý về an toàn thực phẩm dựa trên kiến trúc RAG, cho phép khai thác và sử dụng hiệu quả tri thức từ các văn bản pháp luật. Nghiên cứu đã hình thành bộ dữ liệu pháp lý có cấu trúc, với quy trình thu thập, tiền xử lý, chia đoạn chồng lấn và gán metadata nhằm nâng cao độ chính xác khi truy xuất.

Một pipeline RAG hoàn chỉnh được triển khai, kết hợp mô hình embedding cho tiếng Việt, cơ chế tìm kiếm ngữ nghĩa và mô hình ngôn ngữ lớn để sinh câu trả lời có căn cứ pháp lý rõ ràng. Đồng thời, đề án đề xuất bộ tiêu chí đánh giá gồm độ chính xác, tính đầy đủ và mức độ phù hợp ngữ cảnh, kèm theo công thức định lượng cụ thể.

Kết quả thực nghiệm trên 15 câu hỏi pháp lý cho thấy hệ thống đạt hiệu năng tốt với độ chính xác khoảng 91%, tính đầy đủ 88%, độ phù hợp ngữ cảnh 91% và thời gian phản hồi từ 1,8 đến 4 giây. Những kết quả này khẳng định tính ổn định, khả năng ứng dụng thực tế của hệ thống, đồng thời đóng góp một quy trình và khung đánh giá có thể làm cơ sở cho các nghiên cứu và ứng dụng RAG pháp lý tiếng Việt trong tương lai.

2. Hạn chế của đề tài

- Mặc dù đạt được kết quả khả quan, đề tài vẫn tồn tại một số hạn chế. Về phạm vi tri thức, hệ thống mới chỉ khai thác văn bản pháp luật trong lĩnh vực an toàn thực phẩm và chưa xử lý được các mối quan hệ liên văn bản như sửa đổi, bãi bỏ hay thay thế. Về suy luận pháp lý, hệ thống phù hợp với các câu hỏi định nghĩa và trích dẫn trực tiếp nhưng hiệu quả giảm đối với các yêu cầu suy luận nhiều bước, đồng thời vẫn có nguy cơ sinh thông tin ngoài văn bản gốc.
- Ở tầng truy xuất, độ chính xác chưa tối ưu với các câu hỏi dài hoặc phức tạp do thiếu cơ chế reranking chuyên sâu. Ngoài ra, quá trình đánh giá còn hạn chế về quy mô và độ đa dạng kịch bản, đồng thời chưa có phương pháp đánh giá riêng cho hiện tượng hallucination theo tiêu chuẩn pháp lý.

3. Nhận xét, đề xuất và khuyến nghị

Kiến trúc RAG được chứng minh là phù hợp cho truy vấn văn bản pháp luật tiếng Việt, vừa đảm bảo độ chính xác dựa trên nguồn gốc pháp lý, vừa giữ được tính tự nhiên

Thực hiện chạy 15 câu lệnh bên trên mà không sử dụng kỹ thuật RAG (gọi trực tiếp qua model)

Nhận xét:

- LLM-only trả lời kém chính xác do không nắm rõ nội dung của từng điều khoản pháp luật Việt Nam.
- RAG cải thiện Accuracy gấp 3.25 lần, Coverage gấp 2.5 lần, và Relevance gấp 2.2 lần.
- Tốc độ phản hồi tăng nhẹ do chi phí truy xuất vector.

Bảng 4.2 Kết quả so sánh giữa hai mô hình LLM-only và LLM RAG

Mô hình	Accuracy	Coverage	Relevance	Response Time
LLM-only (không RAG)	~28%	~34%	~41%	1.2–1.6 s
RAG-enabled	~91%	~88%	~91%	1.8–4.0 s

1.1.6 Các mô hình tiêu biểu

Trong lịch sử phát triển, nhiều mô hình NLP đã tạo dấu ấn:

- Eliza (1960s): chatbot đầu tiên.
- Tay (2016): chatbot của Microsoft, minh chứng cho rủi ro dữ liệu thiên lệch.
- BERT, ELMo, RoBERTa: mô hình embedding ngữ cảnh.
- GPT-2, GPT-3: mô hình sinh văn bản tự nhiên dựa trên Transformer.
- LaMDA: chatbot của Google, tập trung vào hội thoại.
- Mixture of Experts (MoE): tối ưu hiệu suất xử lý bằng cách phân bổ tham số theo ngữ cảnh.

1.2 Tổng quan về mô hình ngôn ngữ lớn (Large Language Model - LLM)

1.2.1 Giới thiệu

Mô hình ngôn ngữ lớn (LLM) là bước đột phá của trí tuệ nhân tạo, được huấn luyện trên dữ liệu văn bản quy mô lớn, cho phép hiểu và sinh ngôn ngữ tự nhiên hiệu quả, đồng thời được ứng dụng rộng rãi trong nhiều lĩnh vực nhờ khả năng xử lý các tác vụ phức tạp.

1.2.2 Kiến trúc và nguyên lý hoạt động

Một hệ thống mô hình ngôn ngữ lớn (LLM) gồm các thành phần chính như tokenization, embedding, kiến trúc Transformer, decoder, context window và lớp sinh đầu ra. Trong đó, Transformer với cơ chế self-attention và multi-head attention đóng vai trò trung tâm, cho phép mô hình học các quan hệ ngữ nghĩa và cú pháp trên toàn bộ ngữ cảnh văn bản, xử lý hiệu quả hơn các mô hình tuần tự truyền thống.

Quy trình xử lý của LLM bắt đầu từ việc chia văn bản thành token, ánh xạ token thành vector ngữ nghĩa và bổ sung thông tin vị trí, sau đó đưa qua các lớp Transformer để tạo biểu diễn ngữ cảnh hóa. Đối với các mô hình decoder-only như GPT, decoder sử dụng toàn bộ ngữ cảnh trong cửa sổ ngữ cảnh để suy luận và sinh từng token đầu ra thông qua lớp xác suất, đảm bảo câu trả lời mạch lạc và tự nhiên.

1.2.3 Quá trình huấn luyện và tinh chỉnh

Quá trình phát triển mô hình ngôn ngữ lớn gồm ba giai đoạn chính: tiền huấn luyện trên dữ liệu văn bản quy mô lớn để học quy luật ngôn ngữ, tinh chỉnh theo lĩnh vực chuyên ngành nhằm nâng cao hiệu quả tác vụ cụ thể, và áp dụng học tăng cường từ phản hồi của con người (RLHF) để cải thiện chất lượng, độ an toàn và mức độ phù hợp của câu trả lời.

1.2.4 Đánh giá hiệu năng

Hiệu năng của mô hình ngôn ngữ lớn thường được đánh giá thông qua các chỉ số về chất lượng ngôn ngữ, khả năng suy luận, tốc độ phản hồi và mức độ công bằng, nhằm phản ánh toàn diện độ chính xác và tính tin cậy của mô hình.

1.2.5 Thách thức và hạn chế

Mô hình ngôn ngữ lớn còn tồn tại các hạn chế như khả năng sinh thông tin sai lệch, giới hạn bộ nhớ ngữ cảnh, nguy cơ thiên lệch dữ liệu và các vấn đề về bảo mật.

1.2.6 Ứng dụng trong thực tiễn

Mô hình ngôn ngữ lớn có thể được ứng dụng rộng rãi trong tự động hóa quy trình, dịch vụ khách hàng, phân tích dữ liệu, sáng tạo nội dung và hỗ trợ nghiên cứu – giáo dục.

1.3 Tổng quan về kỹ thuật RAG (Retrieval-Augmented Generation)

1.3.1 Giới thiệu

Kỹ thuật Retrieval-Augmented Generation (RAG) cho phép mô hình ngôn ngữ lớn khai thác dữ liệu từ cơ sở tri thức bên ngoài để tạo câu trả lời chính xác và cập nhật, giúp tùy chỉnh theo dữ liệu riêng của tổ chức mà không cần đào tạo lại hoặc tinh chỉnh mô hình với chi phí cao.

1.3.2 Khái niệm RAG

RAG là một khung AI kết hợp giữa khả năng sinh ngôn ngữ của LLM và cơ chế truy xuất dữ liệu.

Thay vì chỉ sử dụng các mô hình huấn luyện ban đầu, RAG cho phép truy cập vào cơ sở dữ liệu vector chứa thông tin mới nhất, từ đó cải thiện rõ rệt độ chính xác và tính liên quan của phản hồi.

1.3.3 Lợi ích của RAG

chuyên ngành và sinh phản hồi chính xác của hệ thống chatbot pháp luật.

4.4. Đánh giá hiệu năng của hệ thống

Hệ thống được đánh giá dựa trên các tiêu chí cốt lõi gồm độ chính xác, tính đầy đủ, mức độ phù hợp ngữ cảnh, tốc độ phản hồi và mức độ hài lòng của người dùng, phản ánh yêu cầu đặc thù của bài toán hỏi–đáp pháp luật. Kết quả cho thấy độ chính xác trung bình trên 80%, đặc biệt cao với các câu hỏi trích dẫn trực tiếp; tính đầy đủ đạt trên 85% nhờ cơ chế chunking chồng lấn và truy xuất ngữ nghĩa; mức độ phù hợp ngữ cảnh vượt 90% đối với đa số câu hỏi rõ ràng. Thời gian phản hồi dao động từ 1,8–4 giây, đáp ứng tốt yêu cầu hệ thống RAG chạy cục bộ. Khảo sát người dùng cho thấy mức độ hài lòng cao, hệ thống được đánh giá hữu ích và dễ sử dụng, dù vẫn cần cải thiện khả năng xử lý các tình huống pháp lý phức tạp và bổ sung cơ chế dẫn nguồn chi tiết hơn.

4.5. Thực nghiệm

4.5.1 Thực nghiệm với hệ thống hoàn chỉnh

Thực nghiệm được tiến hành trên 15 câu hỏi ngẫu nhiên về an toàn thực phẩm, chia thành 3 nhóm (truy xuất định nghĩa – yêu cầu pháp lý; xác định trách nhiệm và cơ quan quản lý; câu hỏi tổng hợp và khuyến nghị). Kết quả được chấm điểm độc lập bởi bên thứ ba (Prompt ChatGPT) và xác nhận bởi các chuyên gia pháp luật an toàn thực phẩm.

Kết quả cho thấy hệ thống hoạt động tốt với các câu hỏi định nghĩa và trích dẫn trực tiếp, nhiều trường hợp đạt 100% ở cả ba tiêu chí. Một số hạn chế xuất hiện ở các câu hỏi yêu cầu phân tích trách nhiệm pháp lý chi tiết hoặc tình huống phức tạp, nơi độ chính xác và tính đầy đủ giảm xuống (60–80%), chủ yếu do thiếu chi tiết về khung xử phạt hoặc nghĩa vụ cụ thể.

Bảng 4.1 Tổng hợp kết quả thử nghiệm của ứng dụng

Tiêu chí	Kết quả trung bình
Độ chính xác	91%
Tính đầy đủ	88%
Độ phù hợp ngữ cảnh	91%
Tốc độ phản hồi	1.8–4 giây

4.5.2 Thực nghiệm với mô hình không sử dụng RAG

CHƯƠNG 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

4.1. Môi trường mô phỏng thực nghiệm

4.1.1 Cấu hình phần cứng

Hệ thống được triển khai và thử nghiệm trên hai môi trường riêng biệt: giao diện web chạy trên máy Windows 10 Pro với CPU Intel Core i7-12700H, RAM 16GB và SSD NVMe, đảm bảo hiệu năng truy xuất và hiển thị; trong khi hệ thống LLM cục bộ được triển khai trên máy Ubuntu 22.04 với CPU Xeon, RAM ~13GB, GPU NVIDIA 15GB VRAM và SSD NVMe, phục vụ suy luận mô hình qua Ollama với hiệu suất ổn định.

4.1.2 Công nghệ và phần mềm sử dụng

Hệ thống sử dụng các thành phần phần mềm mã nguồn mở gồm Python 3.12.7, Ollama để chạy LLM cục bộ, LangChain cho kiến trúc RAG, ChromaDB làm cơ sở dữ liệu vector, Flask triển khai API chatbot, cùng BeautifulSoup/Requests cho thu thập dữ liệu và SentenceTransformers cho tạo embedding với mô hình BAAI/bge-small-en-v1.5. Kiến trúc được tổ chức theo mô hình client–server, trong đó giao diện chạy trên Flask và suy luận LLM được xử lý riêng qua Ollama.

4.2. Dữ liệu thực nghiệm

Bộ dữ liệu thực nghiệm gồm 170 văn bản pháp luật Việt Nam về an toàn thực phẩm, được thu thập tự động từ vbpl.vn, làm sạch và loại bỏ trùng lặp. Dữ liệu bao phủ đầy đủ các loại văn bản pháp luật, trong đó thông tư và quyết định chiếm tỷ trọng lớn, phản ánh đúng thực tiễn quản lý nhà nước. Mỗi văn bản được chuẩn hóa, chuyển sang dạng văn bản thuần và chia thành các đoạn nhỏ có chồng lấn, gán định danh duy nhất bằng SHA-256 để hỗ trợ đồng bộ và truy vết. Tập dữ liệu có quy mô phù hợp, đáp ứng yêu cầu đánh giá hiệu năng hệ thống RAG và phản ánh đặc thù ngôn ngữ, cấu trúc pháp lý trong lĩnh vực an toàn thực phẩm.

4.3. Các kịch bản thực nghiệm

Các kịch bản thử nghiệm được xây dựng dựa trên nhu cầu tra cứu pháp luật thực tế, bao gồm cả câu hỏi cơ bản và chuyên sâu, nhằm đánh giá toàn diện hệ thống. Các kịch bản tập trung vào ba nhóm chính: (i) truy xuất thông tin trực tiếp từ văn bản pháp luật để đánh giá độ chính xác của thành phần truy xuất; (ii) xác định trách nhiệm pháp lý và cơ quan quản lý nhằm kiểm tra khả năng hiểu ngữ cảnh và lập luận pháp lý; và (iii) tổng hợp thông tin, đưa ra hướng dẫn hoặc khuyến nghị để đánh giá mức độ đầy đủ và phù hợp của câu trả lời. Thiết kế kịch bản đa dạng giúp kiểm chứng năng lực tìm kiếm, hiểu ngôn ngữ

RAG mang lại nhiều lợi ích như giảm chi phí và thời gian triển khai, nâng cao độ chính xác câu trả lời, hỗ trợ cập nhật thông tin linh hoạt và đảm bảo bảo mật dữ liệu thông qua việc quản lý cơ sở tri thức riêng.

1.3.4 Quy trình hoạt động của RAG

Quy trình RAG gồm việc xây dựng và tiền xử lý cơ sở kiến thức, chuyển đổi dữ liệu thành vector và lưu trữ trong cơ sở dữ liệu vector, sau đó truy xuất thông tin liên quan để kết hợp với mô hình ngôn ngữ lớn nhằm tạo ra câu trả lời cuối cùng.

1.3.5 Ứng dụng thực tiễn của RAG

Kỹ thuật RAG đã chứng minh hiệu quả trong nhiều lĩnh vực như sản xuất, tài chính và doanh nghiệp, khi hỗ trợ khai thác dữ liệu chuyên sâu, cập nhật thông tin kịp thời và nâng cao năng suất. RAG đồng thời giúp đảm bảo bảo mật, giảm chi phí triển khai và mở ra hướng tiếp cận mới cho cả nghiên cứu học thuật lẫn ứng dụng thực tiễn trong bối cảnh dữ liệu trở thành tài sản chiến lược.

1.4 Khái quát các thành phần chính trong hệ thống RAG pháp lý

Hệ thống Retrieval-Augmented Generation (RAG) được thiết kế gồm ba thành phần cốt lõi là Retriever, Vector Store và Generator, phối hợp chặt chẽ để phục vụ bài toán tra cứu pháp luật an toàn thực phẩm theo hướng ngoại tuyến, chính xác và có trích dẫn nguồn rõ ràng. Retriever chuyển câu hỏi của người dùng thành vector và truy xuất các điều khoản pháp luật liên quan nhất; Vector Store lưu trữ các vector và metadata với cơ chế định danh ổn định, cho phép cập nhật dữ liệu theo dạng gia tăng; Generator sử dụng mô hình ngôn ngữ lớn cục bộ để sinh câu trả lời tiếng Việt mạch lạc, kèm dẫn chiếu pháp lý chính xác. Toàn bộ quy trình được điều phối bởi framework LangChain, hình thành một pipeline RAG ngoại tuyến, bảo mật, cập nhật nhanh và đáp ứng hiệu quả yêu cầu tra cứu pháp luật trong thực tiễn.

1.5 Các công trình nghiên cứu liên quan

Trên thế giới, Legal AI đã phát triển mạnh với các hệ thống chatbot dựa trên LLM và RAG nhằm hỗ trợ tra cứu, suy luận pháp lý chính xác và có trích dẫn nguồn. Các công trình tiêu biểu như LawLLM chứng minh khả năng xây dựng hệ thống pháp lý đa nhiệm, áp dụng lập luận pháp lý có cấu trúc (legal syllogism), truy xuất và xác minh tri thức, qua đó đạt hiệu năng cao và độ tin cậy vượt trội so với nhiều mô hình pháp lý khác. Tuy

nhiên, các hệ thống này chủ yếu hoạt động trên dữ liệu tiếng Anh và môi trường đám mây, còn hạn chế về bảo mật và triển khai ngoại tuyến.

Tiếp nối, SaulLM-7B là mô hình LLM chuyên biệt cho pháp lý, nổi bật với cơ chế legal chain of thought và truy xuất có kiểm chứng, giúp tăng tính minh bạch, giảm hiện tượng ảo giác và đạt kết quả tốt trên nhiều bộ đánh giá pháp lý. Song song đó, các nghiên cứu thực tiễn triển khai RAG (Milvus, Ollama, LangChain, LlamaIndex, Pinecone...) cho thấy tính linh hoạt và hiệu quả của RAG trong pháp lý, nhưng phần lớn vẫn phụ thuộc môi trường trực tuyến và ngôn ngữ tiếng Anh.

Tại Việt Nam, chatbot pháp lý còn sơ khai, chủ yếu dựa trên tìm kiếm từ khóa, phụ thuộc Internet, thiếu khả năng suy luận thông minh và trích dẫn nguồn đáng tin cậy. Mặc dù đã có nền tảng NLP tiếng Việt như PhoBERT, vẫn chưa tồn tại hệ thống RAG pháp lý ngoại tuyến chuyên biệt. Khoảng trống này tạo động lực cho đề tài nghiên cứu, hướng tới địa phương hóa công nghệ RAG kết hợp LLM cục bộ để hỗ trợ tra cứu pháp luật an toàn thực phẩm tại Việt Nam theo hướng ngoại tuyến, bảo mật và hiệu quả.

tải lên các file .txt/.pdf/.docx, tự động chuyển đổi sang văn bản, lưu trữ cục bộ và kích hoạt đồng bộ vector store chỉ với một thao tác, đồng thời hiển thị trạng thái xử lý theo thời gian thực.

3.3.2 Retriever – Cơ chế truy xuất tài liệu pháp luật chính thức

Retriever là thành phần trung tâm của tầng RAG Logic, đảm nhiệm việc chuyển câu hỏi người dùng thành vector ngữ nghĩa bằng mô hình nhúng BAAI/bge-small-en-v1.5 chạy ngoại tuyến, sau đó truy xuất các đoạn văn bản pháp luật liên quan nhất từ ChromaDB bằng tìm kiếm tương đồng cosine. Cơ chế truy xuất được tối ưu với tìm kiếm gần đúng, lọc theo ngưỡng tương đồng và sắp xếp theo mức độ liên quan, đảm bảo tốc độ và độ chính xác cao. Việc sử dụng stable chunk ID dựa trên SHA-256 cho phép cập nhật delta hiệu quả khi văn bản thay đổi, giúp retriever luôn ưu tiên nguồn pháp lý chính thức và góp phần giảm sai sót trong suy luận pháp lý của hệ thống.

3.3.3 Generator – Thành phần sinh phản hồi (LLM)

Generator là thành phần cuối của chuỗi RAG, sử dụng mô hình Qwen2.5-7B-Instruct chạy cục bộ qua Ollama để sinh câu trả lời tiếng Việt pháp lý dựa hoàn toàn trên ngữ cảnh đã truy xuất. Thông qua prompt template nghiêm ngặt, generator chỉ sử dụng dữ liệu được cung cấp, luôn kèm trích dẫn rõ ràng và áp dụng tham số suy luận bảo thủ nhằm đảm bảo độ chính xác, giảm hallucination. Cơ chế hậu xử lý giúp kiểm tra tính nhất quán và liên kết ngược tới văn bản gốc, từ đó tạo ra các câu trả lời đáng tin cậy, minh bạch và có giá trị thực tiễn cao cho tra cứu pháp luật an toàn thực phẩm.

lại văn bản gốc từ kho tài liệu cục bộ. Các đoạn truy xuất được tổng hợp thành ngữ cảnh và đưa vào prompt tăng cường, sau đó gửi đến LLM chạy cục bộ qua Ollama để sinh câu trả lời. Kết quả cuối cùng được trả về cho người dùng qua giao diện, đảm bảo phản hồi chính xác, có căn cứ và hoàn toàn ngoại tuyến.

3.2.2 Kiến trúc tổng thể của hệ thống

Hệ thống chatbot pháp luật được xây dựng theo kiến trúc ba tầng nhằm đảm bảo khả năng bảo trì, mở rộng và hiệu năng ổn định trong môi trường hoạt động hoàn toàn ngoại tuyến.

- Tầng giao diện người dùng (UI) sử dụng Flask, cung cấp môi trường hỏi–đáp thân thiện bằng tiếng Việt, hiển thị câu trả lời kèm trích dẫn nguồn rõ ràng, hỗ trợ đăng nhập bảo mật và hoạt động hiệu quả trên thiết bị cấu hình thấp với thời gian phản hồi trung bình dưới 5 giây.
- Tầng xử lý logic RAG đóng vai trò trung tâm, điều phối toàn bộ quy trình từ tiền xử lý câu hỏi, tạo embedding, truy xuất các đoạn luật liên quan trong ChromaDB, đến sinh câu trả lời bằng mô hình Qwen2.5-7B-Instruct chạy cục bộ qua Ollama. Tầng này áp dụng chiến lược chunking tối ưu, stable ID để cập nhật delta, prompt nghiêm ngặt để giảm hallucination, đồng thời tự động đồng bộ dữ liệu và xử lý lỗi nhằm duy trì độ tin cậy cao.
- Tầng lưu trữ và truy xuất tri thức sử dụng ChromaDB persistent để lưu trữ vector embedding và metadata của các văn bản pháp luật, hỗ trợ tìm kiếm ngữ nghĩa nhanh, cập nhật từng phần và mở rộng linh hoạt cho các lĩnh vực pháp luật khác, đồng thời đảm bảo bảo mật dữ liệu nhờ lưu trữ hoàn toàn cục bộ.

Sự kết hợp này tạo nên một hệ thống chatbot pháp luật ngoại tuyến, chính xác, dễ triển khai và phù hợp với điều kiện thực tế tại Việt Nam.

3.3. Thiết kế các thành phần của hệ thống

3.3.1 UI Layer – Tầng giao diện người dùng

Tầng giao diện người dùng được triển khai bằng Flask 2.3.3, chạy hoàn toàn trên localhost và chỉ đảm nhiệm vai trò tương tác, không chứa logic xử lý pháp lý. Tầng này thu thập câu hỏi, hiển thị câu trả lời và hỗ trợ quản trị dữ liệu. Giao diện chatbot cung cấp ô nhập tiếng Việt có dấu, gửi truy vấn qua endpoint /ask kèm session ID để duy trì lịch sử hội thoại theo từng phiên. Kết quả được hiển thị dưới dạng bong bóng hội thoại, kèm danh sách trích dẫn nguồn rõ ràng. Ngoài ra, giao diện còn hỗ trợ quản trị tài liệu bằng cách cho phép

CHƯƠNG 2: XỬ LÝ VÀ TỔ CHỨC NGUỒN DỮ LIỆU

2.1 Nguồn dữ liệu chính thức và tiêu chí lựa chọn

Nguồn dữ liệu được thu thập chủ yếu từ vbpl.vn, cơ sở dữ liệu pháp luật chính thức do Bộ Tư pháp quản lý, bảo đảm độ tin cậy, tính pháp lý và khả năng cập nhật. Dữ liệu được lựa chọn theo từ khóa "an toàn thực phẩm", bao gồm nhiều loại văn bản từ luật đến thông tư, ưu tiên các văn bản còn hiệu lực hoặc hết hiệu lực một phần, loại bỏ văn bản đã hết hiệu lực hoàn toàn để phù hợp cho xây dựng cơ sở tri thức ngoại tuyến và cập nhật định kỳ.

Tập dữ liệu thực nghiệm gồm 170 văn bản pháp luật tiếng Việt, phản ánh đầy đủ và toàn diện khung pháp lý về an toàn thực phẩm tại Việt Nam. Trong đó, Thông tư chiếm tỷ lệ lớn nhất (55,29%), tiếp theo là Quyết định (21,18%) và Nghị định (10,59%), cho thấy tính đa dạng và mức độ chi tiết cao của hệ thống văn bản pháp luật được sử dụng.

Bảng 2.1 Cơ cấu thành phần dữ liệu

Loại văn bản	Số lượng	Tỷ lệ (%)
Chỉ thị	5	2,94%
Luật	3	1,76%
Nghị định	18	10,59%
Nghị quyết	1	0,59%
Quyết định	36	21,18%
Thông tư	94	55,29%
Thông tư liên tịch	13	7,65%

2.2. Quy trình thu thập tự động (web crawling) từ vbpl.vn

Quy trình thu thập dữ liệu được tự động hóa bằng script Python, sử dụng requests để gửi yêu cầu HTTP, BeautifulSoup để phân tích HTML và logging để giám sát tiến trình. Script truy vấn endpoint tìm kiếm của vbpl.vn với các tham số về loại văn bản, trạng thái hiệu lực và từ khóa "an toàn thực phẩm", kết hợp phân trang nhiều mức nhằm đảm bảo độ bao phủ dữ liệu rộng.

2.3. Tiền xử lý văn bản pháp luật

Sau khi thu thập, văn bản pháp luật được tiền xử lý nhằm nâng cao chất lượng dữ liệu cho hệ thống RAG. Quy trình bao gồm làm sạch nội dung, loại bỏ ký tự nhiễu và chuẩn hóa khoảng trắng, đồng thời chuẩn hóa Unicode để đảm bảo tính nhất quán tiếng Việt khi tạo vector nhúng. Văn bản được tách theo cấu trúc pháp lý (chương – điều – khoản) bằng regex, giúp bảo toàn ngữ cảnh và cải thiện độ chính xác truy xuất. Các bước tiền xử lý này góp phần giảm nhiễu dữ liệu và tăng hiệu quả truy xuất, tạo nền tảng cho các bước xử lý tiếp theo.

2.4. Phân đoạn văn bản (chunking) và tạo chunk ID

Văn bản được phân đoạn bằng kỹ thuật sliding window với kích thước chunk 1000 ký tự và overlap 200 ký tự nhằm giữ liên tục ngữ cảnh, phù hợp với đặc thù văn bản pháp luật. Mỗi chunk được đánh dấu vị trí bằng chỉ số chunk để tránh cắt rời các điều khoản liên quan. Đồng thời, hệ thống sử dụng định danh chunk ổn định dựa trên hàm băm SHA-256 của nguồn file, vị trí và nội dung chuẩn hóa, hỗ trợ cập nhật delta hiệu quả. Nhờ đó, chỉ các đoạn thay đổi mới được cập nhật, giúp rút ngắn đáng kể thời gian đồng bộ dữ liệu.

2.5. Sinh embedding và tổ chức dữ liệu trong vector store

Sau phân đoạn, mỗi chunk được chuyển đổi thành vector ngữ nghĩa bằng mô hình BAAI/bge-small-en-v1.5 tải cục bộ, áp dụng masked mean pooling để tính trung bình có trọng số theo attention mask, tạo ra vector 384 chiều đại diện cho ngữ nghĩa pháp lý tiếng Việt. Các vector này được tổ chức trong ChromaDB persistent, với mỗi bản ghi bao gồm vector, nội dung chunk và metadata đầy đủ (loại văn bản, trạng thái hiệu lực, ngày ban hành, file nguồn).

2.6. Đánh giá dữ liệu và đảm bảo chất lượng

Để đảm bảo chất lượng, dữ liệu được đánh giá qua các chỉ số độ chính xác (kiểm tra thủ công 10% mẫu so với nguồn gốc), độ bao phủ (tỷ lệ văn bản còn hiệu lực đạt 92%) và tính logic liên hệ (sử dụng cosine similarity giữa các chunk liên quan để xác nhận mối quan hệ chương-điều). Các bước này phát hiện và loại bỏ 5% dữ liệu nhiễu, đồng thời xây dựng chỉ mục kiểm tra để theo dõi cập nhật định kỳ. Việc xử lý toàn diện này không chỉ nâng cao hiệu quả truy xuất mà còn đảm bảo hệ thống cung cấp câu trả lời pháp lý có căn cứ, liên kết chặt chẽ, phù hợp với yêu cầu của một trợ lý pháp lý đáng tin cậy.

CHƯƠNG 3: THIẾT KẾ KIẾN TRÚC VÀ CÁC THÀNH PHẦN CỦA HỆ THỐNG

3.1. Lựa chọn công nghệ và mô hình thiết kế

Việc lựa chọn công nghệ và kiến trúc cho hệ thống chatbot tra cứu pháp luật ngoại tuyến được thực hiện dựa trên các yêu cầu khắt khe về độ chính xác pháp lý tiếng Việt, khả năng hoạt động hoàn toàn không cần Internet, chi phí thấp, tài nguyên phù hợp và dễ bảo trì lâu dài. Qua thử nghiệm thực tế trên 200 câu hỏi pháp lý an toàn thực phẩm, Qwen2.5-7B-Instruct-Q4_0 được lựa chọn nhờ khả năng hiểu sâu văn bản luật tiếng Việt, tuân thủ chặt chẽ ngữ cảnh, giảm hallucination và đặc biệt là context window rất lớn (131.072 token), cho phép xử lý toàn văn các điều luật mà không làm mất ngữ cảnh.

Đối với mô hình nhúng, BAAI/bge-small-en-v1.5 được chọn vì đạt độ chính xác truy xuất cao nhất, tốc độ nhanh và tiêu thụ tài nguyên thấp, phù hợp cho hệ thống RAG chạy cục bộ. ChromaDB được sử dụng làm vector store nhờ khả năng lưu trữ bền vững, hỗ trợ cập nhật delta và tích hợp dễ dàng với LangChain mà không cần hạ tầng phức tạp. Ollama được chọn để chạy LLM cục bộ do cài đặt đơn giản, quản lý mô hình thuận tiện và API ổn định. LangChain 0.3.x đóng vai trò điều phối pipeline RAG, trong khi Flask được sử dụng để xây dựng giao diện và API nội bộ gọn nhẹ, dễ kiểm soát.

Về kiến trúc triển khai, hệ thống được thiết kế theo mô hình ngoại tuyến cục bộ hoàn toàn, cài đặt trực tiếp trên máy tính hoặc server nội bộ, nhằm đảm bảo bảo mật dữ liệu pháp lý, phù hợp với điều kiện hạ tầng và quy định tại Việt Nam. Giải pháp này cho phép triển khai nhanh, chi phí vận hành gần như bằng không, dễ bảo trì và nhân rộng, đồng thời vẫn đáp ứng yêu cầu tốc độ phản hồi và độ tin cậy cao trong tra cứu pháp luật.

3.2. Mô hình tổng thể của hệ thống chatbot pháp luật

3.2.1 Luồng hoạt động của hệ thống

Luồng xử lý của hệ thống bắt đầu khi người dùng nhập câu hỏi tiếng Việt qua giao diện web hoặc dòng lệnh, sau đó được Application Core tiếp nhận và điều phối. Câu hỏi được chuyển thành vector ngữ nghĩa, rồi gửi đến ChromaDB để tìm kiếm các đoạn văn bản pháp luật liên quan nhất kèm metadata. Khi cần hiển thị nguồn chi tiết, hệ thống truy xuất