# MỤC LỤC

[MỤC LỤC I](#_Toc185528869)

[DANH MỤC CÁC HÌNH VẼ III](#_Toc185528870)

[DANH MỤC CÁC BẢNG BIỂU III](#_Toc185528871)

[DANH MỤC CÁC TỪ VIẾT TẮT IV](#_Toc185528872)

[LỜI CẢM ƠN 1](#_Toc185528873)

[CHƯƠNG 1 TỔNG QUAN VỀ ĐỀ TÀI 2](#_Toc185528874)

[1.1 Lý do chọn đề tài 2](#_Toc185528875)

[1.2 Mục tiêu đề tài 2](#_Toc185528876)

[1.3 Nội dung nghiên cứu 2](#_Toc185528877)

[1.4 Mục tiêu nghiên cứu 3](#_Toc185528878)

[1.5 Đối tượng nghiên cứu 4](#_Toc185528879)

[1.6 Bố cục đề tài 4](#_Toc185528880)

[CHƯƠNG 2 CƠ SỞ LÝ THUYẾT CÔNG NGHỆ 6](#_Toc185528881)

[2.1 Giới thiệu về Hệ thống Trả lời câu hỏi 6](#_Toc185528882)

[2.2 Hệ thống Trả lời câu hỏi miền mở là gì 7](#_Toc185528883)

[2.3 Cơ chế hoạt động của Open-Domain Question Answering 7](#_Toc185528884)

[2.4 Tìm kiếm mức độ tương tự văn bản 9](#_Toc185528885)

[2.4.1 Thuật toán TF-IDF 9](#_Toc185528886)

[2.4.2 Thuật toán BM25 9](#_Toc185528887)

[2.4.3 Tìm kiếm tương đồng dựa trên vector nhúng 9](#_Toc185528888)

[2.5 Lịch sử hình thành mô hình ngôn ngữ 9](#_Toc185528889)

[2.6 Lượng tử hoá mô hình (Quantization model) 9](#_Toc185528890)

[2.7 Xây dựng Retriver 9](#_Toc185528891)

[CHƯƠNG 3 XÂY DỰNG CHATBOT 10](#_Toc185528892)

[3.1 Xây dựng Vector Database 10](#_Toc185528893)

[3.1.1 Chuẩn bị dữ liệu 10](#_Toc185528894)

[3.1.2 Làm sạch dữ liệu 10](#_Toc185528895)

[3.1.3 Khám phá dữ liệu 10](#_Toc185528896)

[3.2 Lựa chọn mô hình ngôn ngữ lớn 10](#_Toc185528897)

[3.2.1 Lượng tử hoá mô hình 10](#_Toc185528898)

[3.2.2 Huấn luyện mô hình 10](#_Toc185528899)

[3.2.3 Triển khai mô hình 10](#_Toc185528900)

[CHƯƠNG 4 KẾT QUẢ THỰC NGHIỆM 10](#_Toc185528901)

[4.1 Môi trường thực nghiệm 10](#_Toc185528902)

[4.2 Kết quả thực nghiệm 10](#_Toc185528903)

[CHƯƠNG 5 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 11](#_Toc185528904)

[5.1 Kết luận 11](#_Toc185528905)

[5.2 Hướng phát triển 11](#_Toc185528906)

[PHỤ LỤC 13](#_Toc185528907)

# DANH MỤC CÁC HÌNH VẼ

[Hình 2‑1 Minh hoạ Hệ thống trả lời câu hỏi 6](#_Toc181640119)

[Hình 2‑2 Cách thức vận hành của một hệ thống Truy xuất Thông tin 8](#_Toc181640120)

# DANH MỤC CÁC BẢNG BIỂU

[Bảng 2‑1 Tên bảng 6](#_Toc476558223)

# DANH MỤC CÁC TỪ VIẾT TẮT

|  |  |
| --- | --- |
| **Từ viết tắt** | **Diễn giải** |
| LLM | Large language models, Mô hình Ngôn ngữ lớn |
| NLP | Natural Language Processing, Xử lý ngôn ngữ tự nhiên |
|  |  |

# LỜI CẢM ƠN

Trước tiên, tôi xin bày tỏ lòng biết ơn sâu sắc đến Trường Đại học Mỏ - Địa chất, Khoa Công nghệ Thông tin và Bộ môn Khoa học Máy tính, nơi đã tạo điều kiện học tập, nghiên cứu và trải nghiệm cho tôi trong suốt thời gian qua. Trong suốt quá trình thực hiện đồ án tốt nghiệp với đề tài **"Nghiên cứu xây dựng Chatbot AI hỗ trợ tư vấn tuyển sinh"** tôi đã nhận được sự hỗ trợ nhiệt tình từ nhiều cá nhân và tập thể, giúp tôi hoàn thành công việc nghiên cứu của mình.

Tôi xin gửi lời cảm ơn chân thành tới các thầy cô trong khoa và bộ môn đã không ngừng nỗ lực giảng dạy, truyền đạt những kiến thức quý báu, đặt nền tảng vững chắc cho tôi trong suốt quá trình học tập. Chính nhờ nền tảng kiến thức chuyên môn này, tôi đã có thể tự tin tiếp cận các vấn đề phức tạp trong nghiên cứu và hoàn thành đề tài tốt nghiệp với nhiều kết quả đáng ghi nhận.

Đặc biệt, tôi xin bày tỏ lòng biết ơn sâu sắc tới giảng viên hướng dẫn của mình, ThS. Nguyễn Thị Phương Bắc. Cô không chỉ là người hướng dẫn, mà còn là người đồng hành tận tụy, luôn sẵn sàng chia sẻ kiến thức, kinh nghiệm, cũng như những lời khuyên quý báu giúp tôi định hướng đúng đắn trong suốt quá trình nghiên cứu. Cô đã dành rất nhiều thời gian, công sức để theo sát tiến độ, chỉnh sửa từng chi tiết và đưa ra những nhận xét chuyên môn kịp thời. Sự chỉ dẫn tận tình và trách nhiệm của cô là động lực to lớn giúp tôi hoàn thành đề tài này một cách tốt nhất có thể.

Bên cạnh đó, tôi xin gửi lời cảm ơn chân thành đến các anh chị và các bạn trong lớp Khoa học Máy tính Ứng dụng khóa 66E, những người luôn sẵn sàng lắng nghe, chia sẻ, trao đổi và hỗ trợ tôi trong những lúc khó khăn. Các bạn đã cùng tôi vượt qua nhiều thử thách trong quá trình học tập và nghiên cứu, góp phần tạo nên một môi trường học tập đoàn kết, thân thiện và giàu động lực.

Ngoài ra, tôi cũng xin cảm ơn tất cả các cá nhân và tập thể đã hỗ trợ tôi trong việc thu thập thông tin, tài liệu và kiến thức thực tiễn để xây dựng hệ thống Chatbot AI hỗ trợ tư vấn tuyển sinh. Sự hợp tác và chia sẻ thông tin từ các thầy cô, bạn bè và đồng nghiệp đã giúp tôi có thêm nguồn tài liệu phong phú, nâng cao chất lượng nghiên cứu và mở rộng hiểu biết của mình về lĩnh vực trí tuệ nhân tạo và xử lý ngôn ngữ tự nhiên.

Cuối cùng, mặc dù đã cố gắng hết sức nhưng do giới hạn về thời gian và kinh nghiệm, đề tài của tôi chắc chắn không tránh khỏi những thiếu sót. Tôi rất mong nhận được sự góp ý của các thầy cô và bạn bè để hoàn thiện hơn trong những nghiên cứu và công việc sau này. Đề tài này là kết quả của một hành trình học tập đầy ý nghĩa, và tôi hy vọng rằng những gì tôi đã nghiên cứu và xây dựng có thể đóng góp một phần nhỏ vào sự phát triển của Trường Đại học Mỏ - Địa chất cũng như lĩnh vực ứng dụng trí tuệ nhân tạo trong giáo dục.

Tôi xin chân thành cảm ơn!

# TỔNG QUAN VỀ ĐỀ TÀI

## Lý do chọn đề tài

Trong thời đại cách mạng công nghiệp 4.0, công nghệ trí tuệ nhân tạo (AI) đã và đang trở thành một phần không thể thiếu trong hầu hết các lĩnh vực, từ kinh doanh, y tế, đến giáo dục. Đặc biệt, trong ngành giáo dục, sự gia tăng mạnh mẽ của nhu cầu cá nhân hóa trải nghiệm học tập và tư vấn đã thúc đẩy các trường đại học tìm kiếm các giải pháp công nghệ nhằm cải thiện chất lượng dịch vụ của mình. Một trong những giải pháp hiệu quả nhất là ứng dụng Chatbot AI để tự động hóa quá trình tư vấn, hỗ trợ và tương tác với người dùng.

Trường Đại học Mỏ - Địa chất là một trong những cơ sở giáo dục hàng đầu về khoa học địa chất và kỹ thuật tại Việt Nam. Hàng năm, trường tiếp nhận hàng ngàn câu hỏi từ thí sinh và phụ huynh, xoay quanh các chủ đề như ngành học, học phí, ký túc xá, cơ hội nghề nghiệp, và thông tin tuyển sinh. Tuy nhiên, hệ thống tư vấn hiện tại của trường đang bộc lộ nhiều hạn chế. Đầu tiên, các kênh tư vấn truyền thống chủ yếu dựa vào nhân sự, hoạt động theo khung giờ hành chính, gây bất tiện cho những thí sinh ở xa hoặc không thể tiếp cận dịch vụ trong khung giờ đó. Thứ hai, với khối lượng lớn câu hỏi lặp lại, đội ngũ tư vấn thường gặp phải áp lực công việc, dẫn đến giảm chất lượng phản hồi. Thêm vào đó, hệ thống hiện tại khó có thể đáp ứng nhu cầu cá nhân hóa thông tin, một yếu tố quan trọng trong việc cung cấp trải nghiệm người dùng tốt hơn.

Những hạn chế này đặt ra nhu cầu cấp thiết về một hệ thống tự động hóa hiện đại, có khả năng giải đáp thông tin nhanh chóng, chính xác và không giới hạn thời gian. Việc xây dựng một Chatbot AI hỗ trợ tư vấn tuyển sinh không chỉ giúp giảm tải khối lượng công việc cho đội ngũ nhân sự mà còn nâng cao trải nghiệm người dùng. Hệ thống Chatbot sẽ đảm bảo khả năng phản hồi 24/7, cung cấp thông tin một cách tức thời mà không cần phụ thuộc vào sự hiện diện của con người. Đồng thời, nhờ vào việc tích hợp mô hình ngôn ngữ lớn (LLM) và cơ sở dữ liệu vector tiên tiến, hệ thống có thể đạt được độ chính xác cao trong việc xử lý và trả lời câu hỏi của người dùng. Đây cũng là một bước tiến lớn trong việc áp dụng công nghệ hiện đại vào lĩnh vực giáo dục, thúc đẩy sự đổi mới sáng tạo và tối ưu hóa quy trình.

## Mục tiêu đề tài

Đề tài sử dụng những kiến thức đã học trong lĩnh vực học máy và học sâu để xây dựng, đào tạo và áp dụng một số phương pháp thống kê nhằm kiểm thử, đánh giá mức độ hiệu quả của mô hình Đồng thời cũng áp dụng những công nghệ bao gồm: Đề tài sử dụng các công nghệ như sau:

* **Thư viện học máy, học sâu:** Pytorch, Hugging Face, Langchain
* **Triển khai Back-end:** vLLM, FastAPI
* **Triển khai Front-end:** ReactJS, TailwindCSS, HTML.

## Nội dung nghiên cứu

Trong bối cảnh cuộc Cách mạng Công nghiệp 4.0 đang lan rộng, công nghệ trí tuệ nhân tạo (AI) và xử lý ngôn ngữ tự nhiên (NLP) ngày càng được ứng dụng rộng rãi vào nhiều lĩnh vực khác nhau, trong đó có lĩnh vực giáo dục và tư vấn tuyển sinh. Việc cung cấp thông tin và giải đáp thắc mắc cho các thí sinh, phụ huynh trong quá trình tuyển sinh là một phần quan trọng trong hoạt động của các trường đại học, giúp họ hiểu rõ hơn về ngành học, chương trình đào tạo và các thông tin liên quan khác.

Hiện nay, Trường Đại học Mỏ - Địa chất là một trong những cơ sở giáo dục hàng đầu trong đào tạo các ngành khoa học địa chất và kỹ thuật. Tuy nhiên, mỗi mùa tuyển sinh, nhà trường phải đối mặt với số lượng lớn các yêu cầu tư vấn từ thí sinh và phụ huynh. Những câu hỏi thường xoay quanh nhiều lĩnh vực như điều kiện xét tuyển, chương trình đào tạo, cơ hội nghề nghiệp, và các thủ tục nhập học. Việc xử lý một lượng lớn yêu cầu tư vấn này theo phương pháp thủ công không chỉ gây áp lực cho bộ phận tuyển sinh mà còn có thể làm giảm chất lượng dịch vụ hỗ trợ do thiếu tính tức thời và chính xác.

Sự phát triển của công nghệ Chatbot AI mang lại giải pháp tối ưu để đáp ứng nhu cầu tư vấn nhanh chóng, chính xác và hiệu quả. Một chatbot có khả năng xử lý ngôn ngữ tự nhiên có thể hỗ trợ nhà trường trong việc cung cấp thông tin tự động 24/7, giúp giảm tải cho đội ngũ tư vấn, đồng thời cải thiện trải nghiệm của thí sinh và phụ huynh. Ngoài ra, Chatbot AI còn có khả năng cá nhân hóa câu trả lời, đáp ứng tốt hơn nhu cầu và mong muốn cụ thể của từng người dùng, từ đó giúp nhà trường nâng cao hiệu quả truyền thông và quảng bá hình ảnh của mình.

Với những lý do trên, việc nghiên cứu và xây dựng Chatbot AI hỗ trợ tư vấn tuyển sinh Trường Đại học Mỏ - Địa chất là một hướng đi cần thiết và có ý nghĩa thực tiễn cao. Đề tài này không chỉ giúp cải thiện quy trình tư vấn của nhà trường mà còn đóng góp vào việc ứng dụng công nghệ AI trong lĩnh vực giáo dục, đồng thời đáp ứng xu hướng chuyển đổi số trong hoạt động của các cơ sở đào tạo đại học hiện nay.

## Mục tiêu nghiên cứu

Đề tài **"Nghiên cứu xây dựng Chatbot AI hỗ trợ tư vấn tuyển sinh**" hướng đến việc phát triển một hệ thống Chatbot thông minh, có khả năng hoạt động tự động để hỗ trợ và nâng cao chất lượng quá trình tư vấn tuyển sinh của nhà trường. Mục tiêu đầu tiên là xây dựng một Chatbot có thể giao tiếp bằng tiếng Việt, sử dụng công nghệ xử lý ngôn ngữ tự nhiên (NLP) để trả lời các câu hỏi phổ biến từ thí sinh và phụ huynh. Chatbot này sẽ cung cấp các thông tin cần thiết về chương trình đào tạo, điều kiện tuyển sinh, cơ hội nghề nghiệp, cũng như thủ tục nhập học và học phí một cách chính xác và kịp thời.

Mục tiêu tiếp theo là nâng cao trải nghiệm người dùng. Chatbot được thiết kế để hoạt động 24/7, cho phép thí sinh và phụ huynh truy cập thông tin tuyển sinh bất cứ lúc nào và từ bất kỳ đâu. Hệ thống sẽ được tối ưu để dễ dàng sử dụng, thân thiện với người dùng và có khả năng cá nhân hóa nội dung trả lời dựa trên các nhu cầu cụ thể của từng người, nhằm mang lại trải nghiệm tư vấn toàn diện và phù hợp.

Ngoài ra, đề tài hướng đến việc giảm tải cho bộ phận tuyển sinh bằng cách tự động hóa một phần quy trình tư vấn. Chatbot sẽ đảm nhận vai trò là kênh hỗ trợ đầu tiên, xử lý các câu hỏi phổ biến và đơn giản, giúp đội ngũ tư vấn tập trung vào các trường hợp phức tạp, yêu cầu sự can thiệp và tư vấn từ con người.

Mục tiêu của đề tài còn bao gồm việc cải thiện hiệu quả truyền thông và quảng bá hình ảnh của nhà trường. Nhờ khả năng cung cấp thông tin nhanh chóng và chính xác, Chatbot sẽ giúp thí sinh hiểu rõ hơn về chương trình đào tạo và môi trường học tập tại Trường Đại học Mỏ - Địa chất. Đồng thời, việc áp dụng công nghệ hiện đại này sẽ góp phần xây dựng hình ảnh chuyên nghiệp, tiên tiến và thân thiện của nhà trường trong mắt thí sinh và phụ huynh.

Cuối cùng, đề tài còn hướng đến việc phân tích dữ liệu từ các tương tác của người dùng với Chatbot để thu thập và nắm bắt các xu hướng thông tin mà thí sinh và phụ huynh quan tâm. Việc phân tích này không chỉ giúp cải thiện chất lượng dịch vụ tư vấn, mà còn cung cấp dữ liệu quan trọng để tối ưu hóa và nâng cấp các tính năng của Chatbot trong tương lai, đảm bảo tính hiệu quả và độ chính xác của hệ thống.

## Đối tượng nghiên cứu

Đối tượng nghiên cứu của đề tài bao gồm các phương pháp và kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) hiện đại, đóng vai trò quan trọng trong việc xây dựng một hệ thống Chatbot AI hỗ trợ tư vấn tuyển sinh hiệu quả. Đặc biệt, đề tài tập trung vào các mạng học sâu phục vụ cho các tác vụ NLP như Transformer, BERT và GPT, những mô hình tiên tiến giúp Chatbot có khả năng hiểu và xử lý ngôn ngữ tự nhiên một cách tự động và chính xác.

Ngoài ra, nghiên cứu cũng đi sâu vào các công nghệ tối ưu mô hình, bao gồm kỹ thuật lượng tử hóa mô hình và phương pháp triển khai mô hình lên môi trường server cloud. Những công nghệ này giúp Chatbot hoạt động hiệu quả hơn, đảm bảo khả năng xử lý nhanh chóng trong môi trường thực tế mà vẫn tiết kiệm tài nguyên tính toán.

Cuối cùng, các phương pháp đánh giá mô hình ngôn ngữ lớn cũng được xem là đối tượng nghiên cứu quan trọng. Đề tài sẽ áp dụng các kỹ thuật và tiêu chuẩn đánh giá hiệu quả của mô hình ngôn ngữ nhằm đảm bảo Chatbot hoạt động với độ chính xác cao và mang lại trải nghiệm tốt nhất cho người dùng.

## Bố cục đề tài

Bố cục của đồ án được trình bày với các nội dung chính như sau:

**Chương 1:** Giới thiệu chung về đề tài

Trình bày tổng quan đề tài, lý do chọn nghiên cứu, mục tiêu và phạm vi của đề tài, tạo cơ sở cho các chương tiếp theo.

**Chương 2:** Cơ sở lý luận về vấn đề nghiên cứu

Cung cấp nền tảng lý thuyết về xử lý ngôn ngữ tự nhiên (NLP), các mô hình học sâu như Transformer, BERT, GPT, và các phương pháp triển khai, tối ưu hóa mô hình trên cloud.

**Chương 3:** Xây dựng hệ thống

Mô tả quá trình thiết kế và xây dựng hệ thống Chatbot AI, từ tiền xử lý dữ liệu, chọn mô hình ngôn ngữ, đến tích hợp và triển khai trên môi trường server cloud.

**Chương 4:** Thực nghiệm và đánh giá

Trình bày quá trình kiểm thử và đánh giá Chatbot qua các tiêu chí như độ chính xác, hiệu suất và khả năng đáp ứng yêu cầu người dùng.

**Chương 5:** Kết luận và hướng phát triển

Tổng kết kết quả nghiên cứu, nêu rõ đóng góp của đề tài và đưa ra hướng phát triển trong tương lai nhằm nâng cao hiệu quả của Chatbot.

**Kết luận**

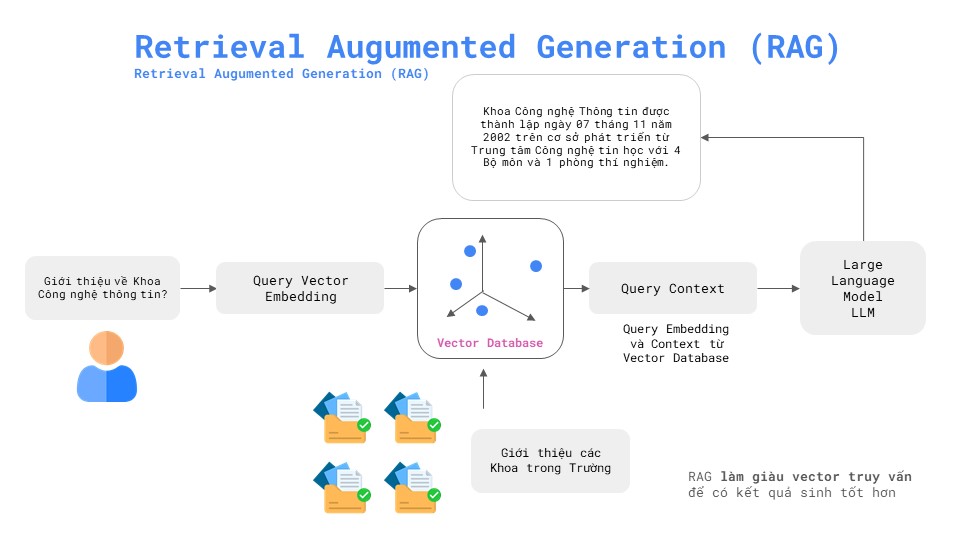
Phần này tôi tổng kết lại các kết quả và những đóng góp mà việc thực hiện đề tài này đem lại. Ngoài ra, tổng kết những việc chưa làm được cần khắc phục, đề xuất các phương hướng nghiên cứu tiếp theo, làm cho đề tài trở lên hoàn thiện và hữu ích hơn.

# CƠ SỞ LÝ THUYẾT CÔNG NGHỆ

## Giới thiệu về Hệ thống Trả lời câu hỏi

Từ thuở ban đầu, con người luôn tìm kiếm câu trả lời cho những thắc mắc của mình thông qua các phương pháp thực nghiệm và suy luận. Qua thời gian, những tri thức quý giá đã được tích lũy, hệ thống hóa và truyền lại dưới nhiều hình thức, từ những cuốn sách cổ xưa đến các tài liệu nghiên cứu hiện đại. Ngày nay, kho tàng tri thức khổng lồ ấy đã được số hóa và lưu trữ trên các thiết bị phần cứng hiện đại, cho phép chúng ta truy cập và tìm kiếm thông tin một cách thuận tiện hơn bao giờ hết.

Tuy nhiên, với sự phát triển của công nghệ và mạng Internet, khối lượng thông tin ngày càng đồ sộ, tăng trưởng theo cấp số nhân qua từng năm, dẫn đến nhu cầu cấp thiết về các hệ thống có khả năng tổng hợp và xử lý thông tin một cách nhanh chóng và chính xác. Những hệ thống sơ khai nhất trong việc đáp ứng nhu cầu này có thể kể đến như Google, hay gần đây hơn là ChatGPT – một mô hình ngôn ngữ do OpenAI phát triển, đã tạo ra một cuộc cách mạng trong việc truy vấn và tìm kiếm thông tin. ChatGPT và các mô hình tương tự giúp việc tìm kiếm thông tin trở nên đơn giản và dễ dàng hơn, mang lại cho người dùng trải nghiệm tìm kiếm gần gũi và tự nhiên như trò chuyện với con người.



Hình ‑ Minh hoạ Hệ thống trả lời câu hỏi

Từ đó, chúng ta có thể hiểu rằng hệ thống trả lời câu hỏi (Question-Answering) là một tác vụ quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên và trí tuệ nhân tạo. Mục tiêu của tác vụ này là xây dựng một hệ thống có khả năng tự động trả lời câu hỏi của người dùng dựa trên kỹ thuật đọc hiểu của máy tính (machine reading comprehension). Cụ thể, hệ thống sẽ tiếp nhận câu hỏi từ người dùng, trích xuất thông tin từ nguồn dữ liệu lớn và đa dạng, sau đó đưa ra câu trả lời phù hợp và chính xác. Đây là một lĩnh vực nghiên cứu thu hút rất nhiều sự quan tâm trong cộng đồng trí tuệ nhân tạo, đặc biệt là nhờ sự phát triển mạnh mẽ của các mô hình ngôn ngữ lớn (large language models) [1] có khả năng đọc hiểu ngôn ngữ tự nhiên. Các hệ thống này hiện nay được ứng dụng rộng rãi trong nhiều lĩnh vực như y tế, giáo dục, tài chính, và các ngành dịch vụ khách hàng.

Hệ thống trả lời câu hỏi hiện nay chủ yếu được phân thành hai loại chính: hệ thống trả lời câu hỏi miền đóng (closed-domain) và hệ thống trả lời câu hỏi miền mở (open-domain). Với sự tiến bộ của công nghệ AI, các hệ thống trả lời câu hỏi đang ngày càng hoàn thiện, trở thành công cụ quan trọng để hỗ trợ con người trong việc tìm kiếm tri thức và ra quyết định.

## Hệ thống Trả lời câu hỏi miền mở là gì

Hệ thống Trả lời câu hỏi miền mở (Open-Domain Question Answering) là loại hệ thống có khả năng trả lời các câu hỏi thuộc nhiều lĩnh vực khác nhau, miễn là dữ liệu liên quan đã có sẵn trong cơ sở dữ liệu [2]. Khác với hệ thống miền đóng chỉ tập trung vào một lĩnh vực nhất định, hệ thống miền mở có thể xử lý câu hỏi từ các lĩnh vực đa dạng nhờ vào cơ chế Truy xuất thông tin (Information Retrieval - IR) và các kỹ thuật Xử lý ngôn ngữ tự nhiên (NLP).

Quy trình hoạt động của hệ thống này bao gồm hai bước chính:

* Truy xuất thông tin (IR): Tìm kiếm và sắp xếp tài liệu có liên quan nhất từ kho dữ liệu lớn dựa trên mức độ phù hợp với câu hỏi.
* Trích xuất câu trả lời (NLP): Sử dụng các kỹ thuật NLP để rút trích nội dung cụ thể, đưa ra câu trả lời ngắn gọn và sát với yêu cầu của người dùng.

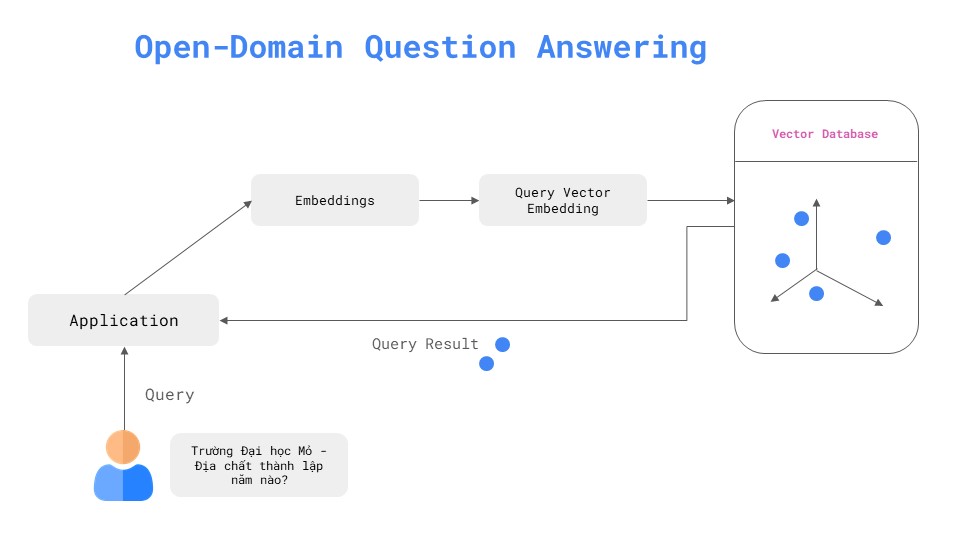
Hệ thống trả lời câu hỏi miền mở có nhiều ứng dụng quan trọng trong giáo dục, y tế, dịch vụ khách hàng và nghiên cứu khoa học, giúp người dùng truy cập thông tin một cách nhanh chóng và tiện lợi. Tuy nhiên, việc duy trì độ chính xác, khả năng hiểu ngữ cảnh và cập nhật dữ liệu liên tục là những thách thức lớn đối với hệ thống này.

## Cơ chế hoạt động của Open-Domain Question Answering

Một hệ thống trả lời câu hỏi miền mở thường bao gồm hai thành phần chính: Retriever và Reader.

**Retriever** là thành phần đảm nhiệm việc truy xuất các tài liệu liên quan từ kho lưu trữ, sắp xếp chúng theo mức độ liên quan cao nhất. Các tài liệu này được chọn lọc dựa trên phương pháp đo lường độ tương đồng văn bản, như TF-IDF, BM25, hoặc các mô hình vector dày đặc (dense vector), để nhanh chóng xác định và chọn ra top-K tài liệu có nội dung phù hợp nhất với câu hỏi.

**Reader** tiếp nhận những tài liệu này và trích xuất đoạn văn bản ngắn hoặc câu có khả năng cao chứa câu trả lời. Reader thực hiện việc này bằng cách xác định chính xác vị trí (index) của đoạn bắt đầu và kết thúc của đáp án trong ngữ cảnh. Các mô hình thường được dùng cho thành phần Reader bao gồm BERT, RoBERTa, DeBERTa, và XLM, có khả năng phân tích ngữ cảnh để đảm bảo độ chính xác cao nhất cho câu trả lời.



Hình ‑ Cách thức vận hành của một hệ thống Truy xuất Thông tin

Trong một số hệ thống hiện đại, Reader có thể được thay thế bằng Generator như GPT-2, BLOOM hoặc LLaMA [3]. Các mô hình Generator này không chỉ trích xuất mà còn tạo câu trả lời mang tính tự nhiên và sát với ngữ cảnh hơn, nhờ vào khả năng tạo sinh văn bản dựa trên thông tin đã được huấn luyện trước đó. Đề tài này lựa chọn Generator thay cho Reader, nhằm mang lại câu trả lời không chỉ chính xác mà còn tự nhiên và gần gũi như lời nói của con người.

## Tìm kiếm mức độ tương tự văn bản

### Thuật toán TF-IDF

### Thuật toán BM25

### Tìm kiếm tương đồng dựa trên vector nhúng

## Lịch sử hình thành mô hình ngôn ngữ

## Lượng tử hoá mô hình (Quantization model)

## Xây dựng Retriver

# XÂY DỰNG CHATBOT

## Xây dựng Vector Database

### Chuẩn bị dữ liệu

### Làm sạch dữ liệu

### Khám phá dữ liệu

## Lựa chọn mô hình ngôn ngữ lớn

### Lượng tử hoá mô hình

### Huấn luyện mô hình

### Triển khai mô hình

# KẾT QUẢ THỰC NGHIỆM

## Môi trường thực nghiệm

## Kết quả thực nghiệm

# CHƯƠNG 5 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

## 5.1 Kết luận

## 5.2 Hướng phát triển

**TÀI LIỆU THAM KHẢO**

|  |  |
| --- | --- |
| [1] | AWS, "What is RAG (Retrieval-Augmented Generation)," 2024. [Online]. Available: https://aws.amazon.com/vi/what-is/retrieval-augmented-generation/. |
| [2] | Pinecone, "An Introduction to Open Domain Question-Answering," 2022. [Online]. Available: https://www.pinecone.io/learn/series/nlp/question-answering. |
| [3] | Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, Guillaume Lample, "LLaMA: Open and Efficient Foundation Language Models," 27 2 2023. [Online]. Available: https://arxiv.org/abs/2302.13971. |

# PHỤ LỤC