

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ



Đỗ Minh Duy

NGHIÊN CỨU XÂY DỰNG CHATBOT
DỰA TRÊN MÔ HÌNH NGÔN NGỮ LỚN VÀ RAG

KHÓA LUẬN TỐT NGHIỆP ĐẠI HỌC CHÍNH QUY
Ngành: Công nghệ thông tin

Hà Nội - 2024

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Đỗ Minh Duy

NGHIÊN CỨU XÂY DỰNG CHATBOT
DỰA TRÊN MÔ HÌNH NGÔN NGỮ LỚN VÀ RAG

KHÓA LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY
Ngành: Công nghệ thông tin

Cán bộ hướng dẫn: PGS.TS. Nguyễn Việt Hà

Cán bộ đồng hướng dẫn: TS. Bùi Ngọc Thăng

Hà Nội - 2024

Tóm tắt

Chatbot hiện đại dựa trên các mô hình ngôn ngữ lớn (Large Language Models - LLMs) đã chứng tỏ khả năng hiểu và hội thoại gần giống con người. Tuy nhiên, chúng bị giới hạn bởi dữ liệu huấn luyện, dễ tạo thông tin không chính xác và thiếu khả năng xử lý các lĩnh vực chuyên môn. Phương pháp Tạo tăng cường truy xuất (Retrieval-Augmented Generation - RAG) được áp dụng để cải thiện vấn đề này bằng cách bổ sung thông tin từ các nguồn tri thức bên ngoài.

Khóa luận này nghiên cứu sự khác biệt giữa RAG dựa trên cơ sở dữ liệu vector và đồ thị tri thức, đồng thời đề xuất phương pháp kết hợp cả hai. Hệ thống RAG dựa trên vector cho thấy hiệu quả với các câu hỏi tổng quát, trong khi hệ thống dựa trên đồ thị lại vượt trội trong các truy vấn phức tạp cần suy luận từ nhiều tài liệu. Bằng cách tận dụng ưu điểm của cả hai loại cơ sở dữ liệu, nghiên cứu hướng tới cải thiện độ chính xác và hiệu suất phản hồi của Chatbot trong các lĩnh vực cụ thể với tài liệu chuẩn hóa.

Lời cảm ơn

Để hoàn thành khóa luận tốt nghiệp và đạt được những kết quả như hôm nay, tôi đã trải qua một hành trình học tập và nghiên cứu đầy thử thách tại trường Đại học Công nghệ. Bên cạnh sự nỗ lực không ngừng của bản thân, tôi đã nhận được sự hỗ trợ tận tình, động viên và tạo điều kiện từ các thầy cô giáo, gia đình và bạn bè.

Tôi xin gửi lời tri ân sâu sắc tới PGS.TS. Nguyễn Việt Hà và TS. Bùi Ngọc Thăng, những người đã luôn tận tâm hướng dẫn, chỉ bảo tôi hết sức chu đáo và tạo mọi điều kiện thuận lợi trong suốt quá trình thực hiện khóa luận.

Bên cạnh đó, tôi cũng xin bày tỏ lòng biết ơn chân thành tới toàn thể các thầy cô giáo tại Khoa Công nghệ thông tin và trường Đại học Công nghệ, những người đã kiên nhẫn truyền đạt cho tôi vô vàn tri thức quý báu trong suốt thời gian học tập, từ những ngày đầu nhập học cho đến nay.

Cuối cùng, tôi xin gửi lời cảm ơn sâu sắc nhất tới gia đình, bạn bè và người thân, những người luôn là chỗ dựa vững chắc, tiếp thêm sức mạnh và động lực cho tôi trên mỗi bước đường học tập.

Lời cam đoan

Tôi xin cam kết rằng toàn bộ công việc và kết quả trong khóa luận này là thành quả nghiên cứu của cá nhân tôi, thực hiện dưới sự hướng dẫn tận tình của PGS.TS. Nguyễn Việt Hà và TS. Bùi Ngọc Thăng. Mọi tài liệu, công trình nghiên cứu được sử dụng trong khóa luận đều được trích dẫn đầy đủ và ghi rõ nguồn gốc.

Tôi chịu hoàn toàn trách nhiệm về nội dung của khóa luận và cam kết sẵn sàng chấp nhận mọi hình thức xử lý từ phía nhà trường nếu có bất kỳ vi phạm nào liên quan đến lời cam đoan này.

Hà Nội, tháng 12 năm 2024

Sinh viên thực hiện

Đỗ Minh Duy

Mục lục

Bảng viết tắt thuật ngữ

Danh sách hình vẽ

Danh sách bảng

1	Đặt vấn đề	1
1.1	Bối cảnh và vấn đề	1
1.2	Mục tiêu và phạm vi nghiên cứu	4
1.2.1	Mục tiêu	4
1.2.2	Phạm vi nghiên cứu	5
1.3	Cấu trúc của khóa luận	6
2	Cơ sở lý thuyết	8
2.1	Chatbot	8
2.1.1	Định nghĩa	8
2.1.2	Góc nhìn tổng quan về Chatbot	9
2.1.3	Phân loại ý định	9
2.1.4	Quản lý tri thức	10
2.1.5	Đánh giá hiệu suất	10
2.2	Mô hình ngôn ngữ lớn	11
2.3	Truy xuất thông tin	12
2.4	Tạo tăng cường truy xuất (RAG)	13
2.4.1	RAG truyền thống dựa trên truy xuất cơ sở dữ liệu vector . .	13
2.4.2	Tạo sinh tăng cường truy xuất dựa trên đồ thị	16
2.4.3	RAG kết hợp đồ thị tri thức và văn bản	18
3	Cải thiện chất lượng phản hồi của Chatbot sử dụng RAG và đồ thị tri thức	20
3.1	Các thuật ngữ cơ bản	21

3.2	Xây dựng đồ thị tri thức, cơ sở dữ liệu từ những tài liệu chuẩn hóa .	22
3.2.1	Xây dựng cây tài liệu từ tài liệu chuẩn hóa	22
3.2.2	Xây dựng đồ thị tri thức từ tài liệu chuẩn hóa	23
3.3	Mô hình RAG tích hợp truy xuất, suy luận đồ thị tri thức với truy xuất cơ sở dữ liệu vector	25
3.3.1	Khởi tạo	25
3.3.2	Khám phá tri thức	26
3.3.3	Suy luận từ tri thức kết hợp	29
4	Thực nghiệm	30
4.1	Hệ thống thực nghiệm	30
4.1.1	Các phần của hệ thống	30
4.1.2	Cài đặt môi trường	34
4.2	Dữ liệu thực nghiệm	36
4.3	Kết quả thực nghiệm	36
4.3.1	Chỉ số đánh giá	36
4.3.2	Kết quả so sánh với điểm trung bình của học sinh từng năm	37
4.3.3	So sánh với các phương pháp RAG khác	38
4.3.4	Đánh giá sự tác động của LLMs	39
4.3.5	Đánh giá khả năng trả lời câu hỏi và đưa ra giải thích	40
5	Kết luận	44
5.1	Kết luận	44
5.2	Định hướng phát triển trong tương lai	45
	Tài liệu tham khảo	46
	A PSEUDOCODE	52
	B PROMPT	54
	C Đề thi THPT Lịch sử	58

Bảng viết tắt thuật ngữ

DRMs Dense Retrieval Models - Mô hình truy xuất dày đặc 25, 28, 32, 33

KG Knowledge Graph - Đồ thị tri thức 16, 17, 20–26, 28, 31

LLM Large Language Models - Mô hình ngôn ngữ lớn 3, 10–13, 15–17, 19, 21, 23–27, 29, 32, 33, 38, 39

NLP Natural Language Processing - Xử lý ngôn ngữ tự nhiên 11, 12, 14, 16, 18

RAG Retrieval-Augmented Generation - Tạo tăng cường truy xuất 3–6, 12, 14, 15, 18, 20, 24, 30, 32, 34, 35, 37–39

Danh sách hình vẽ

Hình 3.1: Minh họa xây dựng đồ thị tri thức, cây tài liệu từ tài liệu chuẩn hóa	22
Hình 3.2: Mô hình Retrieval-Augmented Generation (RAG) tích hợp truy xuất, suy luận đồ thị tri thức với truy xuất cơ sở dữ liệu vector	24
Hình 4.1: Minh họa cây tài liệu được xây dựng từ bộ tài liệu sách giáo khoa Lịch Sử	31
Hình 4.2: Ontology cho lịch sử	32
Hình 4.3: Minh họa một phần đồ thị tri thức được xây dựng dựa trên tài liệu sách giáo khoa Lịch Sử	33
Hình 4.4: Giao diện người dùng của hệ thống	34
Hình 4.5: Kết quả so sánh với kết quả trung bình của từng năm	38
Hình C.1: Trang 1 đề thi THPT Lịch sử 2021	59
Hình C.2: Trang 2 đề thi THPT Lịch sử 2021	60
Hình C.3: Trang 3 đề thi THPT Lịch sử 2021	61
Hình C.4: Trang 4 đề thi THPT Lịch sử 2021	62

Danh sách bảng

Bảng 4.1: Cấu hình phần cứng sử dụng trong nghiên cứu	34
Bảng 4.2: Danh sách thư viện và công cụ sử dụng trong nghiên cứu . . .	35
Bảng 4.3: So sánh hiệu suất của các phương pháp khác nhau với GPT-4o-mini.	38
Bảng 4.4: So sánh hiệu suất giữa câu trả lời trực tiếp và phương pháp đề xuất với các mô hình khác nhau. Các giá trị trong ngoặc đơn biểu thị phần trăm cải thiện.	39
Bảng 4.5: Thực nghiệm trả lời câu hỏi mở.	40
Bảng 4.6: Thực nghiệm trả lời câu hỏi mở.	42
Bảng 4.7: Thực nghiệm lý giải câu hỏi trắc nghiệm.	43
Bảng B.1: Prompt cho tác vụ nhận diện thực thể	54
Bảng B.2: Prompt cho tác vụ cắt giảm các quan hệ trong quá trình khám phá tri thức	55
Bảng B.3: Prompt cho tác vụ cắt giảm tất cả các quan hệ cùng 1 lúc trong quá trình khám phá tri thức	55
Bảng B.4: Prompt cho tác vụ cắt giảm thực thể trong quá trình khám phá	56
Bảng B.5: Prompt cho tác vụ suy luận câu trả lời	57
Bảng B.6: Prompt cho tác vụ dự đoán các manh mối	57
Bảng C.1: Câu trả lời của phương pháp đề xuất cho đề thi THPT Lịch sử 2018, mã đề 302. Với các câu in đậm là câu trả lời sai	58

Chương 1

Đặt vấn đề

1.1 Bối cảnh và vấn đề

Mặc dù việc tìm cách tạo ra một thứ có thể hiểu và giao tiếp với người tạo ra nó đã ăn sâu vào lịch sử loài người, Alan Turing được cho là người đầu tiên hình thành ý tưởng về chatbot vào năm 1950, khi ông đặt câu hỏi: “Máy móc có thể suy nghĩ không?”. Mô tả của Turing về hành vi của một cỗ máy thông minh gợi lên khái niệm chatbot mà chúng ta thường hiểu ngày nay [1].

Chatbot đã phát triển cùng với sự gia tăng dần khả năng tính toán và những tiến bộ trong các công cụ và kỹ thuật xử lý ngôn ngữ tự nhiên (NLP). Việc triển khai chatbot đầu tiên, dựa nhiều vào các quy tắc ngôn ngữ và kỹ thuật khớp mẫu, đã được thực hiện vào năm 1966 với sự ra đời của ELIZA. Chatbot này có thể giao tiếp với người dùng thông qua chương trình khớp từ khóa, tìm kiếm các quy tắc chuyển đổi thích hợp để tái cấu trúc đầu vào và đưa ra phản hồi, tức là câu trả lời cho người dùng. ELIZA là một hệ thống mang tính bước ngoặt, khuyến khích nghiên cứu sâu hơn trong lĩnh vực này. Tuy nhiên, phạm vi hiểu biết của ELIZA bị giới hạn vì nó phụ thuộc rất ít vào việc nhận diện ngữ cảnh và các quy tắc khớp mẫu thường không linh hoạt để triển khai trong các lĩnh vực mới [2–4].

Một bước tiến quan trọng trong sự phát triển của chatbot vào những năm 1980 là việc sử dụng trí tuệ nhân tạo. A.L.I.C.E. (Artificial Intelligent Internet Computer Entity) dựa trên ngôn ngữ đánh dấu trí tuệ nhân tạo (AIML), một phần mở rộng của XML. AIML được phát triển đặc biệt để cho phép thêm kiến thức về mẫu hội thoại vào phần mềm của A.L.I.C.E., giúp mở rộng cơ sở dữ liệu kiến thức. Các đối tượng dữ liệu trong AIML bao gồm các chủ đề và danh mục. Danh mục là đơn vị kiến thức cơ bản, bao gồm các quy tắc để khớp đầu vào của

người dùng với đầu ra của chatbot. Đầu vào của người dùng được biểu diễn dưới dạng mẫu quy tắc, trong khi đầu ra của chatbot được xác định bằng mẫu quy tắc trong cơ sở kiến thức của A.L.I.C.E. Việc bổ sung các đối tượng dữ liệu mới vào AIML đại diện cho một cải tiến đáng kể so với các hệ thống khớp mẫu trước đây vì cơ sở dữ liệu kiến thức dễ dàng mở rộng. Hơn nữa, ChatScript, kế thừa từ AIML, cũng là công nghệ nền tảng đằng sau các chatbot đoạt giải Loebner. Ý tưởng chính của công nghệ này là khớp các đầu vào văn bản từ người dùng với một chủ đề, và mỗi chủ đề sẽ có các quy tắc cụ thể để tạo ra phản hồi. ChatScript đã mở ra một kỷ nguyên mới trong sự phát triển công nghệ chatbot, bắt đầu chuyển trọng tâm sang phân tích ngữ nghĩa và hiểu biết [3–8].

Hạn chế chính của việc dựa vào các quy tắc và khớp mẫu trong chatbot là chúng phụ thuộc vào lĩnh vực, khiến chúng trở nên kém linh hoạt vì phải dựa vào các quy tắc được viết thủ công cho các lĩnh vực cụ thể. Với những tiến bộ gần đây trong các kỹ thuật học máy và công cụ xử lý ngôn ngữ tự nhiên, kết hợp với khả năng tính toán mạnh mẽ, các khung công việc và thuật toán mới đã được tạo ra để triển khai các chatbot “nâng cao” mà không phụ thuộc vào quy tắc và kỹ thuật khớp mẫu, đồng thời khuyến khích việc sử dụng chatbot trong thương mại. Việc áp dụng các thuật toán học máy vào chatbot đã được nghiên cứu, và những kiến trúc chatbot mới đã xuất hiện.

Ứng dụng của chatbot đã mở rộng với sự xuất hiện của các thuật toán học sâu (Deep Learning). Một trong những ứng dụng mới và thú vị nhất là sự phát triển của các trợ lý cá nhân thông minh (như Alexa của Amazon, Siri của Apple, Google Assistant của Google, Cortana của Microsoft, và Watson của IBM). Các trợ lý cá nhân thông minh hoặc tác nhân hội thoại này thường có thể giao tiếp với người dùng thông qua giọng nói và thường được tích hợp trong điện thoại thông minh, đồng hồ thông minh, loa và màn hình gia đình chuyên dụng, thậm chí cả xe hơi. Ví dụ, khi người dùng nói một từ hoặc cụm từ đánh thức, thiết bị sẽ kích hoạt và trợ lý cá nhân thông minh bắt đầu lắng nghe. Thông qua việc hiểu ngôn ngữ tự nhiên, trợ lý có thể hiểu các lệnh và trả lời yêu cầu của người dùng, thường bằng cách cung cấp thông tin (ví dụ: “Alexa, thời tiết hôm nay ở Los Angeles thế nào?” – “Ở Los Angeles, trời nắng và nhiệt độ là 75°F”), hoặc thực hiện các nhiệm vụ (ví dụ: “Ok Google, phát danh sách nhạc buổi sáng của tôi trên Spotify”). Tuy nhiên, nhiệm vụ hiểu ngôn ngữ của con người vẫn là một thách thức lớn vì sự đa dạng về giọng điệu, vùng miền, địa phương, và thậm chí là cách nói cá nhân.

Tất cả các trợ lý cá nhân thông minh đều có các đặc điểm cốt lõi giống nhau về công nghệ sử dụng, giao diện người dùng và chức năng. Tuy nhiên, một số chatbot có tính cách phát triển hơn, và những chatbot phát triển nhất có thể cung cấp giải trí chứ không chỉ hỗ trợ các công việc hàng ngày; những chatbot này được gọi là chatbot xã hội. Một ví dụ thú vị về chatbot xã hội là XiaoIce của Microsoft. XiaoIce được thiết kế để trở thành một người bạn đồng hành lâu dài với người dùng, và để đạt được mức độ gắn kết cao, nó được xây dựng với tính cách, chỉ số thông minh (IQ) và chỉ số cảm xúc (EQ). Các khả năng IQ bao gồm mô hình hóa kiến thức và trí nhớ, hiểu hình ảnh và ngôn ngữ tự nhiên, lý luận, sáng tạo, và dự đoán. Đây là những thành phần quan trọng trong việc phát triển khả năng hội thoại, đáp ứng các nhu cầu cụ thể của người dùng và hỗ trợ họ. Khả năng quan trọng và phức tạp nhất là Core Chat, cho phép trò chuyện dài và trong các lĩnh vực mở với người dùng. Đồng cảm và kỹ năng xã hội là hai thành phần quan trọng của EQ. Công cụ hội thoại của XiaoIce sử dụng một trình quản lý hội thoại để theo dõi trạng thái của cuộc hội thoại và lựa chọn giữa Core Chat (thành phần tạo hội thoại mở) hoặc kỹ năng hội thoại để tạo phản hồi. Do đó, mô hình tích hợp cả khả năng truy xuất thông tin và khả năng tạo hội thoại [9–11].

Trong bối cảnh các mô hình ngôn ngữ lớn - Large Language Models (LLM) ngày càng phát triển, việc áp dụng chúng vào trong các Chatbot trở thành các trợ lý, công cụ hỗ trợ càng trở nên quan trọng và trở thành một trong các lĩnh vực quan trọng để nghiên cứu. Mặc dù khả năng của các Chatbot sử dụng LLMs như GPT, Gemini, Claude chứng tỏ được năng lực mạnh mẽ trong nhiệm vụ hỏi đáp nhưng vẫn còn một số vấn đề nổi cộm còn tồn đọng khi vẫn phải đối mặt với những hạn chế cố hữu, chẳng hạn như ảo giác và các kiến thức được học đã lỗi thời cũng như chưa tập trung vào trong các miền tri thức cụ thể. Với khả năng mạnh mẽ trong việc cung cấp thông tin bổ trợ mới nhất và hữu ích, các Chatbot sử dụng mô hình ngôn ngữ lớn tăng cường truy xuất (RAG) đã xuất hiện để khai thác các cơ sở kiến thức bên ngoài và có thẩm quyền, thay vì chỉ dựa vào kiến thức bên trong của mô hình, nhằm tăng cường chất lượng phản hồi của LLM.

Mặc dù vậy RAG vẫn còn có một số hạn chế như phụ thuộc vào các mô hình nhúng văn bản (Embedding models), độ chính xác của việc trích xuất thông tin, và khả năng trả lời các câu hỏi cần suy luận. Để giải quyết một phần thách thức được đề cập, trong khóa luận này tôi đề xuất một phương pháp cải tiến RAG bằng cách kết hợp hiệu quả kiến thức phi cấu trúc từ văn bản và các thông tin chi tiết có cấu trúc từ đồ thị tri thức để nâng cao chất lượng phản hồi của Chatbot cho

nhiệm vụ hỏi đáp trong một lĩnh vực cụ thể với các tài liệu đã được chuẩn hóa.

1.2 Mục tiêu và phạm vi nghiên cứu

Mục tiêu chính của khóa luận này là nâng cao khả năng trả lời câu hỏi của Chatbot trong các miền tri thức với tài liệu đã được chuẩn hóa thông qua việc phát triển và cải thiện hệ thống RAG nhằm cải thiện đáng kể độ chính xác và độ tin cậy của các phản hồi được tạo ra bởi các mô hình ngôn ngữ lớn (LLMs) - xương sống của Chatbot hiện nay - trong nhiệm vụ Q&A. Cụ thể khóa luận sẽ tập trung vào việc giải quyết các hạn chế của các mô hình ngôn ngữ lớn bao gồm: giảm sự xuất hiện của hiện tượng "hallucination" (khi mô hình đưa ra thông tin không chính xác hoặc không tồn tại), xử lý hiệu quả các truy vấn phức tạp và đòi hỏi suy luận, đảm bảo các biện pháp bảo mật mạnh mẽ để bảo vệ thông tin nhạy cảm. Phạm vi khóa luận bao gồm việc nghiên cứu và triển khai một phương pháp, hệ thống RAG trong phạm vi một miền tri thức cụ thể. Nghiên cứu đảm bảo rằng hệ thống có thể áp dụng hiệu quả trong các miền tri thức thực tế, chẳng hạn như các miền tri thức về giáo dục bao gồm Lịch Sử, Địa Lý, Sinh Học, ... với các tài liệu là sách giáo khoa đã được chuẩn hóa. (sửa lại phạm vi)

1.2.1 Mục tiêu

Khóa luận này đặt mục tiêu chính là phát triển và triển khai một hệ thống Retrieval-Augmented Generation (RAG) tích hợp cả đồ thị tri thức và văn bản, nhằm khắc phục những hạn chế của các mô hình ngôn ngữ lớn hiện tại cũng như các phương pháp RAG trước đây trong việc giải quyết các câu hỏi thuộc một miền tri thức cụ thể với các tài liệu đã được chuẩn hóa. Các hệ thống RAG truyền thống sử dụng cơ sở dữ liệu vector thường mang lại thông tin rộng rãi, phù hợp với những câu hỏi tổng quát. Ngược lại, RAG dựa trên cơ sở dữ liệu đồ thị lại cung cấp cấu trúc phong phú và logic, giúp nâng cao khả năng xử lý các câu hỏi phức tạp, yêu cầu nhiều quy tắc. Sự kết hợp giữa hai phương pháp này sẽ tạo ra một hệ thống cải thiện đáng kể độ chính xác trong câu trả lời của Chatbot. Cụ thể, các mục tiêu bao gồm:

1. **Nâng cao năng lực của mô hình ngôn ngữ:** Tích hợp các cơ chế truy xuất nâng cao để giảm thiểu tình trạng ảo giác thông tin, đảm bảo rằng mô hình ngôn ngữ tạo ra các phản hồi chính xác và đáng tin cậy về mặt dữ liệu.

2. **Cải thiện khả năng xử lý các truy vấn phức tạp:** Phát triển các khả năng xử lý dữ liệu khó (long-tail data) và các truy vấn đa bước, cho phép hệ thống suy luận dựa trên nhiều nguồn bằng chứng hỗ trợ cho các câu hỏi phức tạp trong doanh nghiệp.
3. **Giải quyết các vấn đề về quyền riêng tư:** Triển khai các quy trình xử lý và truy xuất dữ liệu mạnh mẽ để giảm thiểu rủi ro rò rỉ dữ liệu, đảm bảo tính bảo mật cho các tập dữ liệu độc quyền và nhạy cảm.
4. **Hỗ trợ các ứng dụng thực tiễn trong doanh nghiệp:** Thiết kế hệ thống linh hoạt và có khả năng ứng dụng trong nhiều nhiệm vụ doanh nghiệp, bao gồm trả lời câu hỏi dựa trên tài liệu, truy xuất thông tin và tạo nội dung.

Nghiên cứu này nhằm nâng cao độ chính xác phản hồi của Chatbot khi được triển khai với các tài liệu đã được chuẩn hóa thuộc một miền tri thức cụ thể. Bằng cách đạt được các mục tiêu này, khóa luận sẽ đóng góp vào.

1.2.2 Phạm vi nghiên cứu

Phạm vi của khóa luận này tập trung vào việc nghiên cứu sẽ tập trung vào cải thiện chất lượng phản hồi của Chatbot trong nhiệm vụ hỏi đáp trong một miền tri thức cụ thể với các tài liệu đã được chuẩn hóa. Vậy nên việc phát triển hệ thống RAG được đề xuất là chủ yếu và giao diện của người dùng là một phần không thể thiếu của Chatbot, nhưng không phải là mục tiêu chính của khóa luận này nên sẽ được xây dựng cơ bản đủ để thực hiện các thử nghiệm và đánh giá. Thêm vào đó, sẽ không tập trung vào việc cải thiện các khía cạnh khác của Chatbot như xử lý lịch sử hội thoại, tương tác giọng nói, hoặc các ứng dụng khác của Chatbot. Thông qua các nghiên cứu chi tiết này, khóa luận hướng tới việc đóng góp vào sự phát triển của các hệ thống RAG trong các ứng dụng hỏi đáp trong các miền tri thức cụ thể, đặc biệt là trong lĩnh vực giáo dục.

1. Thiết kế và kiến trúc hệ thống: Khám phá chi tiết và tài liệu hóa kiến trúc hệ thống của hệ thống RAG được đề xuất, bao gồm quy trình trích xuất dữ liệu, cập nhật dữ liệu, và các cơ chế truy vấn cho cả dữ liệu có cấu trúc và không cấu trúc.
2. Triển khai cơ chế truy xuất thông tin: Phát triển và tích hợp các kỹ thuật truy xuất thông tin tiên tiến nhằm nâng cao độ chính xác và độ tin cậy của

các phản hồi được tạo bởi mô hình ngôn ngữ. Điều này bao gồm việc thiết kế các phương pháp lập chỉ mục, thuật toán tìm kiếm hiệu quả, và quy trình bổ sung dữ liệu một cách liên mạch.

3. Thiết kế giao diện và trải nghiệm người dùng: Thiết kế một giao diện thân thiện và trực quan để người dùng tương tác với hệ thống. Điều này bao gồm việc tạo các cơ chế nhập liệu rõ ràng cho truy vấn, đảm bảo đầu ra dễ hiểu, và nâng cao trải nghiệm tổng thể của người dùng.

1.3 Cấu trúc của khóa luận

Chương mở đầu này đã cung cấp một sự hiểu biết cơ bản về vấn đề nghiên cứu và tầm quan trọng của nó. Các chương tiếp theo sẽ đi sâu vào các khía cạnh khác nhau của khóa luận:

- **Chương 2** cung cấp một cái nhìn tổng quan về các nghiên cứu trước đây liên quan đến Chatbot, hệ thống Retrieval-Augmented Generation và sự tích hợp của chúng với các mô hình ngôn ngữ lớn. Nội dung nhấn mạnh các tiến bộ quan trọng, những thách thức hiện tại, và khoảng trống trong tài liệu nghiên cứu mục tiêu của khóa luận này muốn giải quyết.
- **Chương 3** trình bày thiết kế và triển khai hệ thống RAG kết hợp với đồ thị tri thức được đề xuất. Nội dung bao gồm kiến trúc tổng thể của hệ thống, từ các quy trình truy xuất và xử lý dữ liệu, cơ chế xử lý truy vấn, đến các chiến lược tạo phản hồi. Chương cũng đề cập đến các khía cạnh kỹ thuật như lưu trữ siêu dữ liệu, chia nhỏ dữ liệu (data chunking), xây dựng cơ sở dữ liệu đồ thị (construction graph database), và lưu trữ vector nhúng (embedding vector).
- **Chương 4** cung cấp một đánh giá chi tiết về hệ thống RAG được đề xuất thông qua các thử nghiệm và kiểm tra mở rộng. Nội dung bao gồm thiết lập thí nghiệm, như môi trường phần cứng và phần mềm, bộ dữ liệu và các loại truy vấn được sử dụng. Chương phân tích hiệu suất của hệ thống về độ chính xác, độ tin cậy và khả năng xử lý các truy vấn có độ phức tạp khác nhau trên bộ tài liệu thuộc miền tri thức cụ thể. Kết quả làm nổi bật các điểm mạnh của hệ thống trong việc cung cấp thông tin chính xác và liên quan, đồng thời chỉ ra các lĩnh vực cần cải thiện. Đánh giá này chứng minh tính ứng dụng

thực tế của hệ thống trong các kịch bản kinh doanh và định hướng cho các nỗ lực phát triển trong tương lai.

- **Chương 5** tóm tắt các phát hiện chính của nghiên cứu và phân tích những điểm mạnh cũng như hạn chế của hệ thống được đề xuất. Đồng thời, nội dung cũng đưa ra các tác động tiềm năng của hệ thống trong các môi trường kinh doanh thực tế và xác định các lĩnh vực cần cải tiến và nghiên cứu trong tương lai.

Chương 2

Cơ sở lý thuyết

2.1 Chatbot

2.1.1 Định nghĩa

Trong tài liệu khoa học, chatbot thường được gọi chính thức là tác nhân hội thoại (conversational agents). Trong ngữ cảnh của khóa luận này, các thuật ngữ chatbot/tác nhân hội thoại sẽ được sử dụng thay thế cho nhau.

Nguyên tắc cơ bản của mọi chatbot là tương tác với người dùng (trong hầu hết các trường hợp) thông qua tin nhắn văn bản và hành xử như thể nó có khả năng hiểu cuộc trò chuyện và trả lời người dùng một cách phù hợp. Nguồn gốc của việc máy tính giao tiếp với con người lâu đời như chính lĩnh vực khoa học máy tính. Thật vậy, Alan Turing đã định nghĩa một bài kiểm tra đơn giản, hiện được gọi là bài kiểm tra Turing, vào năm 1950, trong đó một giám khảo là con người sẽ dự đoán xem thực thể mà họ đang giao tiếp qua tin nhắn có phải là một chương trình máy tính hay không [12]. Tuy nhiên, tham vọng của bài kiểm tra này lớn hơn nhiều so với trường hợp sử dụng thông thường của chatbot; điểm khác biệt chính là kiến thức chuyên môn của chatbot thường hẹp, trong khi bài kiểm tra Turing giả định rằng một người có thể trò chuyện về bất kỳ chủ đề nào với tác nhân. Điều này giúp ích trong việc thiết kế các tác nhân hội thoại vì chúng không cần phải có một kiến thức chuyên môn (có thể là) vô hạn mà có thể tập trung vào các chủ đề rất cụ thể, chẳng hạn như giúp người dùng đặt bàn tại một nhà hàng.

Hơn nữa, một giả định chung khác mà các nhà thiết kế chatbot thường lưu ý là người dùng thường có một mục tiêu mà họ muốn đạt được vào cuối cuộc trò chuyện khi họ bắt đầu tương tác với chatbot. Điều này sau đó ảnh hưởng đến luồng

và chủ đề của cuộc trò chuyện để đạt được mục tiêu đã chọn. Các nhà phát triển có thể khai thác điều này vì các mô hình hành vi nhất định có xu hướng xuất hiện như một kết quả.

Do đó, định nghĩa về chatbot được sử dụng trong tài liệu này là một chương trình máy tính giao tiếp bằng văn bản theo cách giống con người và cung cấp dịch vụ cho người dùng nhằm hoàn thành một mục tiêu được xác định rõ ràng.

2.1.2 Góc nhìn tổng quan về Chatbot

Về mặt khái niệm, một chatbot được cấu thành từ nhiều thành phần hoạt động đồng bộ nhằm đạt được một mục tiêu chung.

Khi nhận được một tin nhắn mới, bước đầu tiên là xử lý nó thông qua mô-đun nhận diện ngôn ngữ. Quá trình này có thể đơn giản như việc truy xuất một thẻ (tag) hoặc phức tạp hơn với các phương pháp thống kê. Tin nhắn mới, cùng với thông tin ngôn ngữ và các tin nhắn trước đó được lấy từ hệ thống backend, sẽ được đưa vào mô-đun phân loại ý định. Vai trò của mô-đun này là suy luận ý định mà người dùng muốn truyền đạt.

Tiếp theo, metadata của tin nhắn, ý định suy luận được, và các thông tin khác từ backend sẽ được sử dụng để xác định một hành động hoặc chuỗi hành động phù hợp. Ví dụ, chatbot có thể quyết định trả lời bằng một câu hỏi nếu ý định của người dùng chưa rõ ràng, hoặc kích hoạt lại tài khoản người dùng nếu ý định của họ là yêu cầu khôi phục tài khoản.

Cuối cùng, mô-đun xử lý hành động nhận đầu vào là hành động được xác định và thực hiện hành động đó một cách phù hợp. Việc thiết kế theo cách này là hữu ích vì một hành động có thể được thực thi theo nhiều cách khác nhau tùy thuộc vào môi trường hoạt động của chatbot. Cách thực hiện hành động có thể hoàn toàn khác biệt nếu chatbot hoạt động trên nền tảng Messenger so với trên website của một công ty.

2.1.3 Phân loại ý định

Khi nhận được một tin nhắn mới, tác nhân hội thoại cần có khả năng xác định mục tiêu mà người dùng đang cố gắng đạt được. Điều này thường được mô hình hóa như một bài toán phân loại đa lớp, trong đó các nhãn đại diện cho tên của các ý định khả thi từ phía người dùng. Các kỹ thuật để giải quyết vấn đề này dao

động từ phương pháp trích xuất từ khóa đơn giản đến suy luận Bayes nhằm xác định yêu cầu của người dùng dựa trên nhiều tin nhắn.

Các mạng mô hình ngôn ngữ lớn (LLM) đã được chứng minh là hoạt động hiệu quả trong lĩnh vực này trước đây [13]. Chúng là nền tảng, xương sống của khóa luận này

2.1.4 Quản lý tri thức

Một tác nhân thông minh chỉ có thể hoạt động hiệu quả trong giới hạn nếu thiếu kiến thức. Lĩnh vực cho phép máy tính xử lý kiến thức đã có những tiến bộ đáng kể trong thập niên 1980, với tên gọi kỹ thuật tri thức (knowledge engineering). Các kỹ thuật ban đầu thường sử dụng một bộ suy luận (inference engine) để xử lý các dữ kiện và suy ra kiến thức mới bằng cách sử dụng logic bậc nhất và bậc hai. Đây là một cách để suy diễn câu trả lời cho những câu hỏi không đầy đủ và thường dễ dàng được chuyển thành các lệnh gọi API.

Đối với các tác nhân đối thoại (conversational agents), kỹ thuật tri thức rất hữu ích, chẳng hạn để trả lời các câu hỏi cơ bản về các sự kiện tổng quát. Siri và Amazon Alexa sử dụng các phương pháp suy luận tri thức nội bộ để truy xuất thông tin từ web và các nguồn khác (ví dụ, khi hỏi Alexa về các chuyến tàu khởi hành từ Brussels hôm nay, hệ thống có thể kích hoạt một phép suy luận nội bộ dưới dạng `train(brussels, D, today)`, trong đó D là một biến ẩn danh đại diện cho điểm đến).

Ngày nay, việc quản lý tri thức chủ yếu được thực hiện thông qua các lệnh gọi API và các truy vấn cơ sở dữ liệu tối ưu. Mặc dù vậy, các phương pháp đặc biệt hơn lấy cảm hứng từ các ontology có cấu trúc đồ thị đôi khi vẫn được sử dụng trong các cơ sở tri thức [14].

2.1.5 Đánh giá hiệu suất

Một lĩnh vực cần cải thiện trong lĩnh vực tác nhân hội thoại là việc đánh giá hiệu suất và các chỉ số được sử dụng để định lượng chất lượng hành vi của chatbot. Trong nghiên cứu “How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation”, Liu và cộng sự chỉ ra rằng các chỉ số chuẩn trong lĩnh vực này, chẳng hạn như điểm số BLEU [15] và ROUGE [16], thường không tương quan với đánh giá trực quan của con người [17].

Vấn đề cốt lõi nằm ở chỗ ngôn ngữ và sự trôi chảy vốn mang tính chủ quan, do đó rất khó để định lượng chính xác, tương tự như bất kỳ thuộc tính mang tính chủ quan nào khác của một hệ thống.

2.2 Mô hình ngôn ngữ lớn

Các mô hình ngôn ngữ tìm cách giải quyết nhiệm vụ dự đoán từ tiếp theo dựa trên một chuỗi các từ đứng trước đó [18]. Nhiệm vụ này tạo thành cốt lõi của việc hiểu ngôn ngữ và đặt nền tảng cho các nhiệm vụ Natural Language Processing (NLP) cụ thể hơn. Nghiên cứu về các mô hình ngôn ngữ bắt đầu từ những năm 1980 với sự phát triển của các mô hình ngôn ngữ thống kê (Statistical Language Models - SLM) [19]. Những mô hình này dựa trên các phương pháp xác suất n-gram, được thiết kế để tính toán khả năng xuất hiện của từ tiếp theo trong một câu chưa hoàn chỉnh.

Bước tiến đáng kể tiếp theo trong các mô hình ngôn ngữ đến từ việc giới thiệu mạng Nơ-ron nhân tạo (Neural Networks) và sự phát triển của các Mô hình ngôn ngữ dựa trên mạng Nơ-ron (Neural-based Language Models - NLM). NLM mở ra cơ hội cho các nghiên cứu NLP sâu hơn bằng cách chuyển từ các biểu diễn n-gram thưa thớt sang các biểu diễn vectơ dày đặc có kích thước thấp hơn của văn bản [20]. Những đổi mới như Word2Vec của Mikolov, K. Chen, Corrado, và Dean [21] và GloVe của Pennington, Socher, và Manning [22] đã giúp thúc đẩy lĩnh vực này bằng cách giới thiệu các kỹ thuật hiệu quả để nắm bắt các biểu diễn ngữ nghĩa của từ, cải thiện khả năng hiểu sự tương đồng và ngữ cảnh của từ. Khi sức mạnh tính toán trở nên mạnh mẽ và dễ tiếp cận hơn, nghiên cứu về Mạng Nơ-ron tiếp tục phát triển.

Các mô hình ngôn ngữ lớn (Large Language Models - LLM), thường ám chỉ các mô hình dựa trên kiến trúc Transformer được giới thiệu bởi Vaswani, Shazeer, Parmar, và cộng sự [23], đại diện cho một cột mốc khác trong NLP. Các mô hình dựa trên bộ mã hóa (Encoder) như BERT [24] sử dụng Transformers để tạo ra các biểu diễn từ nhúng ngữ cảnh bằng cách xem xét toàn bộ ngữ cảnh của câu. Mặt khác, các mô hình dựa trên bộ giải mã (Decoder) như GPT [12], Llama [25], và PaLM [26] xử lý cốt lõi của các mô hình ngôn ngữ bằng cách cải thiện nhiệm vụ dự đoán từ tiếp theo dựa trên một tập hợp các từ trước đó. Những mô hình này được tinh chỉnh thêm để tạo ra các chuỗi văn bản mạch lạc khi nhận được một gợi ý.

Các mô hình Seq2Seq như BART [27] và T5 [28] bao gồm cả thành phần mã hóa và giải mã. Khác với các mô hình chỉ sử dụng giải mã, vốn xử lý đầu vào theo hướng đơn chiều, các mô hình encode-decode mã hóa đầu vào bằng các cơ chế chú ý hai chiều để hiểu rõ ngữ cảnh. Sau đó, biểu diễn vectơ được đưa vào bộ giải mã để tạo văn bản, xử lý các nhiệm vụ như tóm tắt, dịch thuật, và nhiều nhiệm vụ tạo ngôn ngữ phức tạp khác.

Mặc dù đã đạt được nhiều tiến bộ, các LLM tiêu tốn nhiều tài nguyên và đòi hỏi sức mạnh tính toán lớn để huấn luyện, điều này làm cho việc tích hợp và cập nhật dữ liệu thời gian thực trở nên không khả thi. Hơn nữa, vì thông tin được chứa trong các tham số tĩnh, các LLM không đáng tin cậy trong việc duy trì độ chính xác về mặt thông tin. Những hạn chế này đã dẫn đến các nghiên cứu sâu hơn, bao gồm RAG, vốn kết hợp thông tin liên quan một cách hiệu quả trong quá trình tạo nội dung [13]. Mục tiêu chính của RAG là cải thiện phản hồi của LLM, đảm bảo chúng vừa chính xác về mặt thông tin vừa có căn cứ bằng cách tận dụng dữ liệu được chọn lọc.

2.3 Truy xuất thông tin

Trước khi được sử dụng để tăng cường LLMs, Truy xuất thông tin (Information Retrieval - IR) đã được giới thiệu nhằm đơn giản hóa việc điều hướng các cơ sở dữ liệu lớn bằng cách cho phép người dùng truy cập thông tin thông qua các truy vấn văn bản. Nhiệm vụ NLP này đã được sử dụng rộng rãi trong các chương trình như công cụ tìm kiếm và hệ thống gợi ý [14]. Các tiến bộ trong IR song hành với sự phát triển của các mô hình ngôn ngữ, với các kỹ thuật ban đầu dựa vào đếm từ thừa thớt và n-grams. Các phương pháp nổi bật bao gồm Boolean Retrieval [29], Term Frequency-Inverse Document Frequency (TF-IDF) [30], và BM25 [17], những phương pháp này cố gắng trích xuất các tài liệu liên quan bằng cách khớp các từ trong truy vấn.

Với sự trỗi dậy của mạng nơ-ron, IR đã phát triển song song với các mô hình ngôn ngữ và chuyển hướng sang các biểu diễn vectơ dày đặc. Hiện nay, kỹ thuật truy xuất phổ biến nhất cho các nhiệm vụ NLP khác nhau là Dense Passage Retrieval (DPR) được giới thiệu bởi Karpukhin, O'guir, S. Min, và cộng sự [15], sử dụng các biểu diễn vectơ dày đặc để truy xuất dữ liệu. DPR là một mô hình sử dụng hai bộ mã hóa: một bộ mã hóa các tài liệu để lưu trữ, và một bộ mã hóa các câu hỏi để liên kết với các vectơ tài liệu. Cách tiếp cận này đã chứng minh sự cải

thiện đáng kể so với các biểu diễn vectơ thưa truyền thống.

Nhiều nghiên cứu đã tập trung vào việc tinh chỉnh kết quả truy xuất bằng cách xếp hạng dữ liệu thu thập được để mang lại kết quả chính xác hơn. RankNet của Burges, Shaked, Renshaw, và cộng sự [16] và ListNet của Cao, Qin, T.-Y. Liu, và cộng sự [31] là hai kỹ thuật nền tảng được thiết kế để sắp xếp lại danh sách thông tin đã truy xuất nhằm cải thiện mức độ liên quan. RankNet sử dụng phương pháp pairwise, trong đó một truy vấn và hai tài liệu được đưa vào để xác định tài liệu nào liên quan hơn trong cặp. Ngược lại, ListNet sử dụng phương pháp listwise và xử lý toàn bộ danh sách các tài liệu đã truy xuất cùng với truy vấn để đánh giá mức độ liên quan của từng tài liệu và sắp xếp dữ liệu tương ứng.

Những tiến bộ trong mạng Nơ-ron sâu (Deep Neural Network) và LLMs đã cho phép triển khai các kỹ thuật cross-encoding để tái xếp hạng tài liệu, như được chứng minh bởi Nogueira và Cho [32]. Trong phương pháp này, một mô hình tái xếp hạng với cơ chế cross-encoding xử lý đồng thời truy vấn và các tài liệu để đánh giá chính xác hơn, so với việc chỉ sử dụng điểm tương đồng của các vectơ dày đặc độc lập. Phương pháp này đã được chứng minh là cải thiện đáng kể độ chính xác của việc truy xuất nhờ khả năng phân tích kỹ lưỡng mối quan hệ giữa truy vấn và tài liệu. Tuy nhiên, các mô hình tái xếp hạng đòi hỏi nhiều tài nguyên tính toán, đặc biệt khi phải đánh giá tất cả các tài liệu trong cơ sở dữ liệu lớn. Do đó, các mô hình tái xếp hạng thường được sử dụng để tinh chỉnh kết quả truy xuất sau khi sử dụng DPR hoặc các phương pháp khác để lọc ban đầu.

2.4 Tạo tăng cường truy xuất (RAG)

Tạo tăng cường truy xuất (Retrieval-Augmented Generation - RAG) đã nổi lên như một phương pháp quan trọng để cải thiện khả năng của các mô hình ngôn ngữ bằng cách kết hợp thông tin bên ngoài vào quá trình mô hình hóa sinh văn bản. Phương pháp này gồm 3 giai đoạn chính: Truy xuất (Retrieval), kết hợp truy xuất (Fusion Retrieval), Tạo sinh (Generation). Nhưng trong phạm vi bài nghiên cứu này sẽ chỉ tập trung vào việc cải thiện khả năng truy xuất, suy luận.

2.4.1 RAG truyền thống dựa trên truy xuất cơ sở dữ liệu vector

Nhiều kỹ thuật đã được giới thiệu nhằm tích hợp dữ liệu văn bản được truy xuất và các mô hình ngôn ngữ, một số kỹ thuật trong đó xuất hiện trước khi thuật ngữ RAG được giới thiệu. Ví dụ, bài báo Retrieval Augmented Language Model

Pre-Training (REALM) của Guu, Lee, Tung, và cộng sự [33] đánh dấu một trong những kỹ thuật sớm nhất tích hợp việc truy xuất tri thức trong quá trình huấn luyện một mô hình ngôn ngữ. Cách tiếp cận này chia nhiệm vụ thành hai phần: một bộ truy xuất tri thức nơ-ron để truy xuất tài liệu và một bộ mã hóa tăng cường tri thức để diễn giải các tài liệu được truy xuất. Dựa trên kiến trúc BERT, bộ truy xuất tri thức nơ-ron mã hóa các truy vấn và tài liệu để xác định và truy xuất tài liệu phù hợp nhất cho từng truy vấn. Sau khi truy xuất, bộ mã hóa tăng cường tri thức sẽ lấy đầu vào là sự kết hợp giữa truy vấn gốc và tài liệu đã truy xuất. Đầu vào kết hợp này được xử lý thông qua kiến trúc Transformer, cho phép sự tương tác chéo phong phú giữa truy vấn và tài liệu. Không giống như nhiều mô hình sinh, REALM tập trung vào các nhiệm vụ trích xuất, huấn luyện hệ thống một cách từ đầu tới cuối để tận dụng tri thức truy xuất và khả năng hiểu ngôn ngữ một cách liền mạch.

Thuật ngữ RAG lần đầu được giới thiệu bởi P. Lewis, Perez, Piktus, và cộng sự [34]. RAG đánh dấu một bước tiến đáng kể trong NLP bằng cách tích hợp bộ truy xuất thần kinh và bộ sinh Seq2Seq để cải thiện các nhiệm vụ đòi hỏi nhiều tri thức. Không giống như REALM, tập trung vào hỏi đáp trích xuất (extractive QA) sử dụng một mô hình duy nhất cho việc truy xuất và mã hóa, RAG giới thiệu một kiến trúc khác biệt, kết hợp một DPR để truy xuất tài liệu và một BART để tạo sinh chuỗi. Ngoài ra, các mô hình truy xuất và sinh được huấn luyện đầu-cuối trong khi giữ nguyên các vector embedding của mỗi tài liệu, bởi việc cập nhật embedding trong quá trình huấn luyện vừa tốn kém về mặt tính toán vừa không cần thiết.

Sau đó, nhiều thuật toán khác đã xuất hiện và tích hợp các phương pháp truy xuất hoặc mô hình sinh khác nhau để cải thiện kỹ thuật RAG. Ví dụ, trong một nghiên cứu của Izacard và Grave [35], cách xử lý dữ liệu truy xuất khác biệt. Mô hình tạo sinh câu trả lời của thuật toán được thiết kế bằng mô hình Fusion-in-Decoder (FiD), có khả năng mở rộng để xử lý số lượng dữ liệu văn bản lớn hơn. Mô hình này mã hóa sự kết hợp giữa câu hỏi và từng tài liệu riêng biệt, sau đó tổng hợp các mã hóa vector này trong bộ giải mã. Hệ thống tạo ra các câu trả lời phong phú về ngữ cảnh thông qua các cơ chế attention được áp dụng trên các mã hóa kết hợp.

Retrieval-Enhanced Transformer (RETRO) của Borgeaud, Mensch, và cộng sự [36] là một mô hình ngôn ngữ khác được đề xuất nhằm cải thiện các kỹ thuật

RAG và DPR. RETRO cho phép việc truy xuất lặp lại trong suốt quá trình sinh, không chỉ dựa trên gợi ý ban đầu như RAG hay FiD mà liên tục khi dữ liệu truy xuất được mở rộng. Khả năng truy xuất liên tục này cho phép RETRO thích ứng động với ngữ cảnh được thu thập từ văn bản. RETRO cũng được huấn luyện từ đầu, sử dụng kiến trúc Transformer và một cơ chế cross-attention theo từng đoạn; tuy nhiên, embedding được tính toán bằng cách sử dụng mô hình BERT cố định. Các tài liệu được chia thành nhiều phân đoạn, mã hóa và lưu trữ trong cơ sở dữ liệu vector để truy xuất nhanh hơn. RETRO hoạt động tương đương với GPT-3 và Jurassic-1 nhưng sử dụng ít hơn 25 lần số lượng tham số.

Nhiều bài báo khác đã được xuất bản để nghiên cứu sự phù hợp của việc tích hợp nhiều nguồn và loại dữ liệu vào hệ thống RAG. Bài báo "Unified representations of structured and unstructured knowledge for open-domain question answering" (UniK-QA) của Oguz, X. Chen, Karpukhin, và cộng sự [37] phân tích sự liên quan của việc kết hợp các loại dữ liệu khác nhau để tăng cường các mô hình ngôn ngữ. Mô hình FiD, được giới thiệu bởi Izacard và Grave [35], đã được sử dụng cho RAG với một cơ sở dữ liệu kết hợp dữ liệu có cấu trúc, không cấu trúc và bán cấu trúc, kết hợp dữ liệu văn bản và cơ sở tri thức.

Tất cả các loại dữ liệu đều được biến đổi theo cách heuristic thành văn bản, và quá trình truy xuất được thực hiện bằng phương pháp DPR. Y. Li, Peng, Shen, và cộng sự [38] cũng tích hợp các loại dữ liệu và nguồn khác nhau để truy xuất bằng cách giới thiệu mô hình ngôn ngữ PLUG. Mô hình này được huấn luyện sử dụng nhiều nguồn tri thức khác nhau, từ văn bản đến dữ liệu từ đồ thị tri thức (KG). Quá trình truy xuất khá đơn giản, dựa trên các truy vấn tìm kiếm và khớp từ khóa để tìm dữ liệu liên quan. Sau đó, dữ liệu được biến đổi thành văn bản và lọc qua hai giai đoạn: xếp hạng thống kê và xếp hạng ngữ nghĩa. Xếp hạng thống kê sử dụng TF-IDF để tạo một tập dữ liệu ban đầu. Sau đó, tập này được lọc thêm bằng xếp hạng ngữ nghĩa dựa trên điểm tương đồng của sentence-BERT và một ngưỡng được đặt trước.

Không giống như các nghiên cứu khác tinh chỉnh các mô hình để tạo câu trả lời từ các tài liệu được truy xuất và sử dụng cơ chế cross-attention giữa các tài liệu và đầu vào, Re-plug của Weijia, Sewon, Michihiro, và cộng sự [39] coi mô hình ngôn ngữ lớn sinh (LLM) như một hộp đen. Mỗi tài liệu được truy xuất từ hệ thống RAG được sử dụng riêng để tăng cường LLM và tạo ra xác suất cho các token đầu ra, sau đó các xác suất này được tổng hợp trên các tài liệu để tính toán

xác suất token cuối cùng. Kỹ thuật này ngăn chặn việc cắt bớt các tài liệu được thêm vào nếu chúng vượt quá kích thước ngữ cảnh của LLM.

2.4.2 Tạo sinh tăng cường truy xuất dựa trên đồ thị

GraphRAG (Tạo sinh tăng cường truy xuất dựa trên đồ thị) [31–33, 35–37] là một phương pháp mới tận dụng đồ thị tri thức để cải thiện hiệu suất của các tác vụ NLP như hệ thống hỏi đáp. Bằng cách tích hợp Knowledge Graph (KG)s với các kỹ thuật RAG, GraphRAG cho phép tạo ra các phản hồi chính xác và có ngữ cảnh hơn dựa trên thông tin có cấu trúc được trích xuất từ các tài liệu tài chính. Tuy nhiên, GraphRAG thường hoạt động kém hiệu quả trong các tác vụ hỏi đáp mang tính trừu tượng hoặc khi câu hỏi không đề cập rõ ràng đến thực thể cụ thể nào.

Nhiều nghiên cứu đã khám phá các cách để trích xuất thông tin từ đồ thị tri thức và cơ sở dữ liệu đồ thị, tận dụng định dạng có cấu trúc của chúng. Một trong những hệ thống hỏi đáp dựa trên đồ thị tri thức sớm nhất thực hiện một quá trình nhiều giai đoạn để liên kết các câu hỏi của người dùng với các thực thể trong đồ thị và trích xuất câu trả lời. Đầu tiên, các thực thể trong câu hỏi của người dùng được trích xuất và liên kết với các nút tương ứng trong đồ thị tri thức. Sau đó, việc phát hiện mối quan hệ được thực hiện để tìm ra cạnh thích hợp trong đồ thị dẫn đến câu trả lời chính xác. Ví dụ, Berant, Chou, Frostig và Liang [40] đã đề xuất một phương pháp phân tích ngữ nghĩa để chuyển đổi các câu hỏi văn bản thành các dạng logic hoặc truy vấn có thể thực thi, từ đó trích xuất câu trả lời từ cơ sở dữ liệu đồ thị. Các kỹ thuật khác, chẳng hạn như những kỹ thuật được giới thiệu bởi Z. Wang, Ng, Nallapati và Xiang [41], sử dụng phương pháp truy xuất, tái xếp hạng và học đa nhiệm để cải thiện độ chính xác trong việc phát hiện thực thể, liên kết và xếp hạng.

Các tiến bộ gần đây nhằm loại bỏ các quy trình nhiều bước của việc truy xuất dữ liệu dựa trên đồ thị truyền thống để tránh sự tích lũy lỗi. Một trong những cách tiếp cận như vậy là phương pháp Direct Fact Retrieval - DiFaR do Baek, Aji, Lehmann và Hwang [42] đề xuất. Phương pháp này trích xuất tất cả các bộ ba từ cơ sở dữ liệu và mã hóa chúng thành một embedding vector dày đặc tương tự như kỹ thuật DPR. Phương pháp này vì vậy lưu trữ các mối quan hệ logic trong cơ sở dữ liệu vector thay vì chỉ lưu trữ dữ liệu văn bản. Nghiên cứu cũng sử dụng một mô hình tái xếp hạng (re-ranker) nhận đầu vào là câu hỏi và tài liệu, sau đó đưa ra một điểm xếp hạng chỉ ra tính hữu ích của tài liệu trong việc trả lời câu hỏi.

Các bộ ba hàng đầu K được truy xuất từ cơ sở dữ liệu vector, và chỉ có các bộ ba này được đưa qua mô hình tái xếp hạng để sắp xếp lại mức độ liên quan của chúng với câu hỏi.

Y. Wang, Lipka, Rossi và cộng sự [43] cũng tận dụng tính có cấu trúc của cơ sở dữ liệu đồ thị trong quá trình truy xuất của một hệ thống RAG. Bài báo này xây dựng một đồ thị tri thức sử dụng một tập hợp các tài liệu và cấu trúc dữ liệu khác. Sau đó, một phương pháp duyệt đồ thị dựa trên mô hình ngôn ngữ lớn (LLM) được thiết kế để truy xuất thông tin liên quan từ cơ sở dữ liệu đồ thị. Ban đầu, TF-IDF được sử dụng để so sánh nội dung của các nút với truy vấn đã cho. Nút có điểm tương đồng cao nhất được trích xuất và coi là điểm khởi đầu cho việc duyệt đồ thị. Sau đó, chức năng này xếp hạng và điều hướng qua các nút lân cận để thu thập thông tin quý giá cho việc gợi ý LLM. Một mô hình LLM đã được tinh chỉnh được sử dụng để xếp hạng các nút lân cận, trích xuất nút có dữ liệu liên quan nhất và xác định các bước duyệt tiếp theo. LLM được huấn luyện để xem xét cả câu hỏi và các nút đã duyệt trước đó để chọn nút hứa hẹn nhất tiếp theo, đảm bảo rằng việc hiểu biết dữ liệu đã được truy xuất tích lũy sẽ thông báo cho từng bước duyệt. Kỹ thuật này làm tăng khả năng chọn các nút với ít sự lặp lại thông tin và chất lượng cao hơn như một bổ sung vào thông tin đã thu thập. So với các kỹ thuật embedding vector, thường dựa vào dữ liệu không liên kết với các biểu diễn vector cô đặc, duyệt đồ thị nắm bắt các kết nối logic tiềm ẩn trong KGs, tận dụng các thuộc tính cấu trúc thường bị mất trong không gian vector. Điều này làm cho các phương pháp dựa trên đồ thị đặc biệt mạnh mẽ trong việc trả lời câu hỏi, đặc biệt là những câu hỏi cần đến lý luận quan hệ và nhận thức về ngữ cảnh.

Tiếp đó Darren Edge, Ha Trinh và cộng sự [44] cũng đã tận dụng tính có cấu trúc của cơ sở dữ liệu đồ thị dựa trên việc tóm tắt toàn cục từ đồ thị tri thức được tạo bởi LLM. Bài báo này xây dựng một phương pháp sử dụng một LLM để xây dựng chỉ mục văn bản dựa trên đồ thị qua hai giai đoạn: giai đoạn đầu tiên là tạo ra một đồ thị tri thức thực thể từ các tài liệu nguồn, giai đoạn sau đó là tiền tạo các bản tóm tắt cộng đồng cho tất cả các nhóm thực thể có mối liên hệ chặt chẽ. Với một truy vấn sẽ được sử dụng để tìm kiếm trong đồ thị tri thức nhằm truy xuất các nút (thực thể) và cạnh (quan hệ) liên quan đến truy vấn. Một đồ thị con, bao gồm các nút và cạnh liên quan này, được trích xuất từ toàn bộ KG để cung cấp ngữ cảnh. Đồ thị con này sau đó được tích hợp với tri thức nội tại của mô hình ngôn ngữ bằng cách mã hóa cấu trúc đồ thị thành các biểu diễn nhúng (embeddings) mà mô hình có thể hiểu được. Các bản tóm tắt do LLM tạo ra từ

các mô tả đồ thị con này cung cấp phạm vi bao phủ đầy đủ của chỉ mục đồ thị cơ bản và các tài liệu đầu vào mà nó đại diện. Việc tóm tắt tập trung vào truy vấn của toàn bộ tập dữ liệu sau đó được thực hiện bằng cách sử dụng cách tiếp cận map-reduce: đầu tiên sử dụng từng bản tóm tắt đồ thị con để trả lời truy vấn một cách độc lập và song song, sau đó tóm tắt tất cả các câu trả lời từng phần liên quan thành một câu trả lời toàn cục cuối cùng.

2.4.3 RAG kết hợp đồ thị tri thức và văn bản

Đây là một đề tài nghiên cứu được chú trọng trong thời gian gần đây, tập trung vào việc nâng cao khả năng của các nhiệm vụ NLP bằng cách tối ưu khả năng của truy xuất thông tin dựa trên cơ sở dữ liệu vector. Tuy nhiên khả năng khi áp dụng của phương pháp này đối với các văn bản dài vẫn còn nhiều hạn chế khi độ chính xác trong việc truy xuất các phản hồi liên quan vẫn là một thách thức. Trong khi đó, các phương pháp GraphRAG bằng cách tích hợp đồ thị tri thức với các kỹ thuật RAG đã cho phép tạo ra các phản hồi chính xác hơn và có nhận thức về ngữ cảnh dựa trên thông tin có cấu trúc được trích xuất từ các tài liệu chính. Nhưng thường hoạt động kém trong các nhiệm vụ hỏi đáp trừu tượng hoặc khi không có thực thể rõ ràng nào được đề cập trong câu hỏi.

Từ những ưu điểm và nhược điểm của cả hai phương pháp truy xuất phía trên, nhiều nghiên cứu nổi về RAG gần đây đã đề xuất nhiều phương án kết hợp cả hai phương pháp nhằm tận dụng những khả năng và giảm bớt hạn chế của cả hai phương pháp. Các nghiên cứu tiêu biểu gần đây có thể kể đến HybridRAG được đề xuất bởi Bhaskarjit Sarmah, Benika Hall, và đồng nghiệp [45]. Cụ thể, phương pháp này giải quyết các hạn chế của VectorRAG (dựa trên cơ sở dữ liệu vector) và GraphRAG (sử dụng đồ thị tri thức). Với VectorRAG có thể mạnh trong việc xử lý văn bản không cấu trúc với quy mô lớn thông qua biểu diễn vector. Tuy nhiên, nó thường gặp khó khăn trong việc hiểu mối quan hệ giữa các thực thể. Còn trong khi đó, GraphRAG mang lại hiệu quả trong việc lập luận dựa trên dữ liệu có cấu trúc nhưng hạn chế về khả năng mở rộng và yêu cầu mối quan hệ được xác định trước giữa các thực thể. HybridRAG việc tận dụng cả hai phương pháp bằng cách kết hợp các bộ mã hóa: bộ mã hóa đồ thị tri thức để lý giải mối quan hệ phức tạp giữa các thực thể và bộ mã hóa vector để trích xuất thông tin từ văn bản tự do. Kết quả sau đó được xử lý bởi một bộ giải mã lai, tạo ra các phản hồi chính xác và phù hợp với ngữ cảnh. Cách tiếp cận này đã cho thấy hiệu quả vượt trội trong các bài toán như trích xuất thực thể, phát hiện mối quan hệ, và hỏi đáp, đặc biệt

trong các lĩnh vực như tài chính hoặc các tài liệu phức tạp.

Jiashuo Sun²¹, Chengjin Xu và cộng sự [46] đã đề xuất Think-on-Graph 2.0 mang đến một cải tiến đáng kể đối với các mô hình ngôn ngữ lớn (LLMs) bằng cách kết hợp khả năng suy luận với truy xuất có sự hướng dẫn từ tri thức. Mô hình này xây dựng trên phương pháp Think-on-Graph [47] ban đầu, kết hợp việc tích hợp sâu hơn với các nguồn tri thức bên ngoài, đặc biệt là đồ thị tri thức (KGs), nhằm nâng cao độ chính xác và độ tin cậy của các phản hồi. Điểm đổi mới chính trong khuôn khổ này là việc tăng cường khả năng suy luận của LLM với một hệ thống truy xuất hiệu quả. Bằng cách khai thác dữ liệu liên quan từ một đồ thị tri thức, mô hình không chỉ tận dụng các mẫu ngôn ngữ mà còn làm phong phú thêm khả năng suy luận bằng thông tin có cấu trúc và chính xác. Quá trình này hỗ trợ các nhiệm vụ suy luận nhiều bước và giải quyết các lỗ hổng tri thức trong thời gian thực bằng cách truy xuất thông tin giúp hình thành các phản hồi mạch lạc và chính xác về ngữ cảnh. Điều này đặc biệt hữu ích đối với các câu hỏi phức tạp, yêu cầu nhiều tri thức, nơi mà các mô hình ngôn ngữ truyền thống có thể gặp khó khăn nếu không có một nền tảng tri thức bên ngoài đáng tin cậy. Ngoài ra, Think-on-Graph 2.0 cũng chú trọng vào khả năng giải thích, cung cấp cái nhìn về cách thức ra quyết định của mô hình dựa trên tri thức đã được truy xuất. Sự minh bạch này rất quan trọng đối với các nhiệm vụ cần giải thích hoặc biện minh cho các câu trả lời, giúp phương pháp này trở nên mạnh mẽ hơn trong các ứng dụng như chăm sóc sức khỏe, phân tích pháp lý và nghiên cứu khoa học, nơi độ chính xác và khả năng giải thích là yếu tố quan trọng.

Chương 3

Cải thiện chất lượng phản hồi của Chatbot sử dụng RAG và đồ thị tri thức

Hiện nay, với khả năng mạnh mẽ của LLM, Chatbot đã trở thành một công cụ hữu ích trong việc hỗ trợ người dùng truy cập thông tin, giải đáp thắc mắc. Tuy nhiên, mặc dù có khả năng học và tự cải thiện, Chatbot vẫn còn một số hạn chế trong việc cập nhật thông tin kịp thời cũng như đôi lúc vẫn còn hiện tượng ảo giác với tác vụ Q&A trong miền tri thức với tài liệu đã được chuẩn hóa. Trong chương này, tôi đề xuất một phương pháp kết hợp giữa RAG và đồ thị tri thức để cải thiện khả năng hỗ trợ hỏi đáp của Chatbot từ tài liệu. Phương pháp này sẽ giúp cải thiện khả năng truy xuất thông tin, suy luận thông qua việc kết hợp giữa việc mở rộng chuỗi logic dựa trên các liên kết trong KG với thông tin ngữ cảnh liên kết với các thực thể liên quan bằng cách thực hiện lặp đi lặp lại việc truy xuất ngữ cảnh dựa tri thức và sử dụng ngữ cảnh tăng cường truy xuất đồ thị. Từ đó tích hợp và sử dụng hiệu quả hơn kiến thức bên ngoài từ các dạng cấu trúc khác nhau.

Nội dung chính của phương pháp đề xuất gồm có 2 phần, đầu tiên là phần 3.2 thảo luận về phương pháp xây dựng đồ thị tri thức, cơ sở dữ liệu từ những tài liệu được cung cấp. Tiếp đó phần 3.3 mô tả cốt lõi khả năng của một Chatbot là phương pháp RAG tích hợp truy xuất, suy luận đồ thị tri thức với truy xuất cơ sở dữ liệu vector để nâng cao khả năng của mô hình ngôn ngữ lớn (LLMs). Phương pháp đề xuất kết hợp việc mở rộng chuỗi logic dựa trên các liên kết trong KG với thông tin ngữ cảnh liên kết với các thực thể liên quan bằng cách thực hiện lặp đi

lập lại việc truy xuất ngữ cảnh dựa tri thức và sử dụng ngữ cảnh tăng cường truy xuất đồ thị. từ đó tích hợp và sử dụng hiệu quả hơn kiến thức bên ngoài từ các dạng cấu trúc khác nhau.

3.1 Các thuật ngữ cơ bản

Trước khi đi vào chi tiết phương pháp đề xuất, tôi xin trình bày một số thuật ngữ cơ bản mà sẽ được sử dụng trong phần này:

- **Knowledge Graph - Đồ thị tri thức** chứa đựng, biểu diễn các kiến thức trong thế giới thực dưới dạng tập các bộ ba:

$$G = \{(e, r, e') | e, e' \in \mathcal{E}, r \in \mathcal{R}\},$$

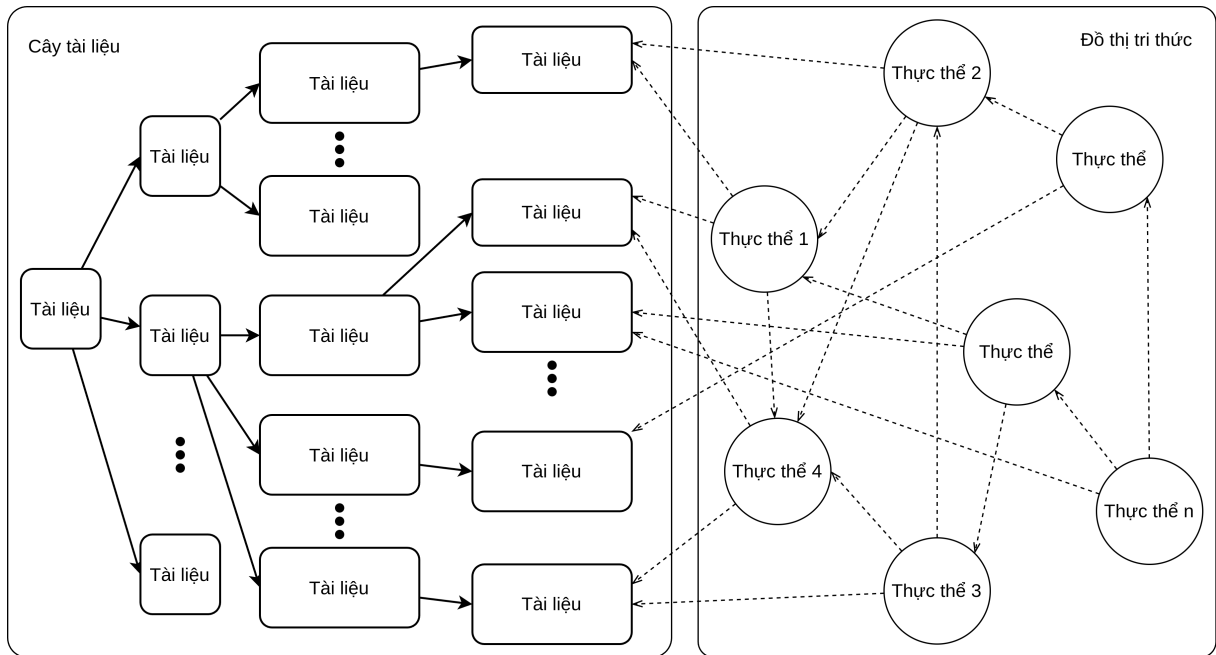
Trong đó \mathcal{E} và \mathcal{R} tương ứng biểu thị cho tập thực thể và quan hệ.

- **Relation Path - Đường dẫn quan hệ** là một chuỗi các quan hệ $z = \{r_1, r_2, \dots, r_l\}$ trong đó $r_i \in \mathcal{R}$ biểu thị cho mỗi quan hệ thứ i trong đường dẫn có độ dài là l .
- **Reasoning Path - Đường dẫn suy luận** là các thể hiện của đường dẫn quan hệ z trong KGs:

$$p_{\{z\}} = e_0 \xrightarrow{r_1} e_1 \xrightarrow{r_2} \dots \xrightarrow{r_l} e_l,$$

trong đó $e_i \in \mathcal{E}$ biểu thị cho thực thể thứ i và r_i biểu thị cho mỗi quan hệ thứ i trong đường dẫn suy luận p .

- **Prompt - lời nhắc** được sử dụng trong khóa luận về cơ bản là đầu vào được cung cấp cho LLMs để tạo ra các đầu ra có định hướng. Mục đích của một prompt là chỉ định rõ LLMs cần tạo ra gì để giúp chúng có thể hiểu ngữ cảnh và thực hiện các tác vụ cụ thể rồi đưa ra đầu ra mong muốn.
- **Tài liệu chuẩn hóa** là các tài liệu được xây dựng một cách có cấu trúc, với các chỉ mục được phân chia một cách rõ ràng và hợp lý. Từ đó chia nội dung của tài liệu thành các đoạn nhỏ có nội dung cụ thể về một ngữ cảnh và các mục nhỏ trong cùng một mục lớn hơn sẽ có sự liên quan về ngữ cảnh.
- **Ontology** là một mô hình học thức học được sử dụng để biểu diễn kiến thức trong một lĩnh vực cụ thể. Ontology thường bao gồm các thực thể, mối quan hệ, hành động, sự kiện, và các khái niệm khác trong một lĩnh vực cụ thể.



Hình 3.1: Minh họa xây dựng đồ thị tri thức, cây tài liệu từ tài liệu chuẩn hóa

3.2 Xây dựng đồ thị tri thức, cơ sở dữ liệu từ những tài liệu chuẩn hóa

Việc lưu trữ các tài liệu cũng như tri thức từ các tài liệu dưới dạng đồ thị tri thức (Knowledge Graph - KG) và các đoạn văn là một phần quan trọng trong phương pháp được đề xuất. Trong phần này, tôi sẽ trình bày chi tiết phương pháp xây dựng, lưu trữ đồ thị tri thức và cây tài liệu từ những tài liệu đó. Kết quả của quá trình này được minh họa trong hình 3.1.

3.2.1 Xây dựng cây tài liệu từ tài liệu chuẩn hóa

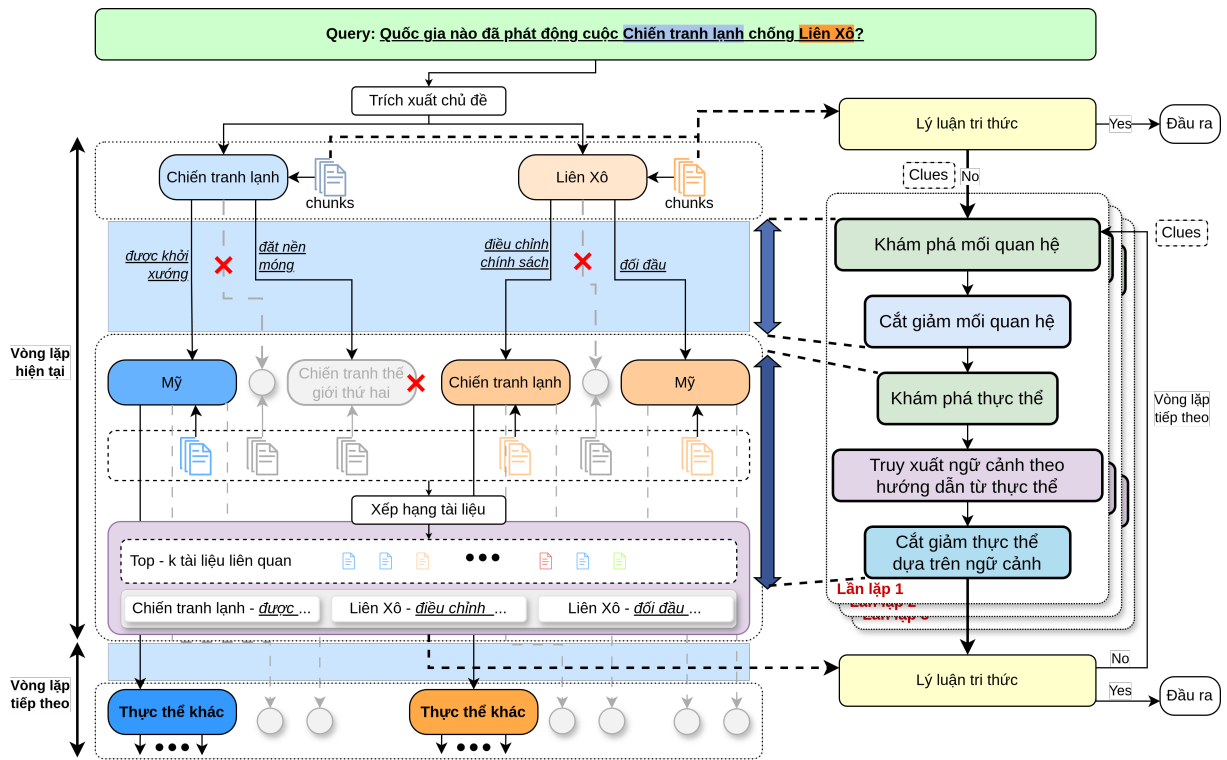
Trong phạm vi là các tài liệu chuẩn hóa, thì các tài liệu này đều sẽ là dạng văn bản bán cấu trúc với các tiêu đề, đoạn văn, câu, từ, ... Để tận dụng được cấu trúc của các tài liệu này, một cây tài liệu sẽ được xây dựng từ các tài liệu. Cây tài liệu này sẽ được xây dựng dựa trên mục lục của tài liệu. Mỗi nút trong cây tài liệu sẽ biểu diễn một phần của tài liệu, từ đó giúp cho việc truy xuất thông tin từ tài liệu trở nên dễ dàng hơn. Với việc xây dựng cây tài liệu, việc truy xuất thông tin từ tài liệu sẽ trở nên dễ dàng hơn cũng như có thể phản ánh được mối liên quan giữa các tài liệu với nhau thông qua khoảng cách giữa các nút trong cây tài liệu. Càng gần nhau thì mối liên quan về ngữ cảnh giữa các tài liệu càng cao.

3.2.2 Xây dựng đồ thị tri thức từ tài liệu chuẩn hóa

Một đồ thị tri thức (KG) là một biểu diễn có cấu trúc của các thực thể trong thế giới thực, các thuộc tính của chúng và mối quan hệ giữa chúng. Một KG được biểu diễn dưới dạng một tập hợp các bộ ba có hướng (thực thể, mối quan hệ, thực thể). Một bộ ba đồ thị là một đơn vị thông tin cơ bản trong một KG, bao gồm một chủ ngữ, một vị ngữ và một tân ngữ. Với mỗi miền tri thức cụ thể ứng với tài liệu chuẩn hóa, một Ontology tương ứng sẽ được xây dựng nhằm tối ưu hóa việc trích xuất tri thức từ tài liệu. Dựa trên cây tài liệu và Ontology đã xây dựng, một KG sẽ được xây dựng và liên kết với tài liệu tương ứng. Quá trình này sẽ bao gồm hai phần chính là trích xuất tri thức và cải thiện tri thức.

Trích xuất tri thức: Mục tiêu của bước này đến việc chuyển đổi thông tin, tri thức từ dữ liệu không có cấu trúc như văn bản thành một KG có cấu trúc có thể được truy vấn sau này. Quá trình này chủ yếu phụ thuộc vào việc trích xuất các bộ ba có định dạng là (thực thể) – [mối quan hệ] → (thực thể). Để giúp cho các thực thể được trích xuất là những thực thể tổng quan và tránh những thông tin dư thừa, một Ontology sẽ được sử dụng để hướng dẫn quá trình trích xuất tri thức. Ontology sẽ được xây dựng dựa trên kiến thức về miền tri thức cụ thể, thông qua thực nghiệm kèm theo hướng dẫn từ các chuyên gia trong lĩnh vực. Hỗ trợ điều này, phương pháp few-shot prompt được áp dụng vào một LLM, yêu cầu nó trích xuất càng nhiều bộ ba càng tốt từ các đoạn văn bản, đồng thời đưa các ví dụ về chuyển đổi văn bản thành bộ ba vào trong câu lệnh. Các tác vụ chính trong bước này sẽ bao gồm nhận dạng thực thể (ER), trích xuất mối quan hệ thông qua sử dụng LLM với prompt tương ứng (được trình bày tại B.1).

Cải thiện tri thức: Bước này nhằm nâng cao chất lượng và độ hoàn thiện của KG bằng cách loại bỏ các thông tin dư thừa và khắc phục những lỗ hổng trong dữ liệu đã trích xuất. Các nhiệm vụ chính trong bước này bao gồm hoàn thiện KG và hợp nhất tri thức. Kỹ thuật hoàn thiện KG tìm ra các thực thể, mối quan hệ còn thiếu trong đồ thị bằng phương pháp dự đoán liên kết và xác định thực thể. Dự đoán liên kết dự đoán sự tồn tại và loại mối quan hệ giữa hai thực thể dựa trên cấu trúc và đặc điểm của đồ thị. Trong khi đó xác định thực thể khớp và hợp nhất các thực thể cùng biểu diễn một thực thể, khái niệm trong thế giới thực. Quá trình hợp nhất tri thức kết hợp thông tin từ nhiều nguồn khác nhau để tạo ra một KG



Hình 3.2: Mô hình RAG tích hợp truy xuất, suy luận đồ thị tri thức với truy xuất cơ sở dữ liệu vector

hoàn chỉnh và thống nhất. Các nguồn thông tin này có thể bao gồm các KG khác, dữ liệu từ các nguồn bên ngoài, ... Quá trình này bao gồm giải quyết xung đột và trùng lặp giữa các nguồn và tổng hợp hoặc điều hòa thông tin dựa trên các quy tắc, xác suất, hoặc sự tương đồng ngữ nghĩa.

Một chuỗi LLM được triển khai gồm 2 tầng để tinh chỉnh nội dung và trích xuất tri thức. Tầng thứ nhất sử dụng một LLM để tạo ra biểu diễn tóm tắt cho từng đoạn tài liệu. Quá trình tinh chỉnh này rất quan trọng vì nó chất lọc thông tin cốt lõi đồng thời giữ nguyên ý nghĩa ban đầu và các mối quan hệ chính giữa các khái niệm. Điều này cung cấp một đầu vào tập trung hơn cho các bước xử lý tiếp theo, nâng cao hiệu quả và độ chính xác của quy trình trích xuất bộ ba. Tầng thứ hai là một LLM dành riêng cho việc trích xuất thực thể và xác định mối quan hệ. Tất cả các bước đều được thực hiện thông qua kỹ thuật gợi ý (prompt engineering) một cách cẩn thận.

3.3 Mô hình RAG tích hợp truy xuất, suy luận đồ thị tri thức với truy xuất cơ sở dữ liệu vector

Phần này sẽ trình bày chi tiết từng bước của mô hình được đề xuất trong khóa luận này. Mô hình này sẽ bắt đầu với phần khởi tạo với nhiệm vụ trích xuất các thực thể từ câu hỏi đã cho làm các thực thể chủ đề ban đầu. Sau đó, nó thực hiện một quy trình lặp đi lặp lại việc khám phá tri thức và lý luận các tri thức được khám phá thông qua việc sử dụng LLM. Tại mỗi vòng lặp, bước “khám phá” sẽ truy xuất, tìm kiếm tri thức có chọn lọc bao gồm các quan hệ, thực thể liên kết với thực thể chủ đề hiện tại dựa trên KG được xây dựng trong phần 3.2, sử dụng các thực thể mới gặp phải để tinh chỉnh phạm vi truy xuất, từ đó nâng cao cả hiệu quả và độ chính xác. Rồi sau đó, sẽ xếp hạng và chọn lọc các thực thể dựa trên truy vấn và các ngữ cảnh thu thập được từ các tài liệu có liên quan để giảm thiểu sự mơ hồ từ đó tìm ra $top - N$ đường dẫn suy luận chứa thông tin liên quan nhất đến câu hỏi. Tiếp đến tại bước “lý luận”, sử dụng LLM đánh giá các thông tin, kiến thức có được từ các đường dẫn lập luận và ngữ cảnh dựa trên nó có đủ thông tin để trả lời câu hỏi. Quá trình này tiếp tục cho đến khi thu thập đủ thông tin thông qua $top - N$ đường lý luận để trả lời câu hỏi (được đánh giá bởi LLM trong bước "lý luận") hoặc đạt đến độ sâu tìm kiếm tối đa được định trước. Mô hình được mô tả chi tiết các bước thực hiện kèm theo ví dụ minh họa trong hình 3.2.

3.3.1 Khởi tạo

Với đầu vào là câu hỏi q , bước đầu tiên là xác định thực thể xuất hiện trong q và liên kết chúng với các thực thể tương ứng trong đồ thị tri thức. Bước này có thể được hoàn thành dựa vào nhiều phương pháp liên kết thực thể (Entity Linking - EL) khác nhau, tiêu biểu có thể sử dụng LLMs hoặc các công cụ, mô hình chuyên về EL. Tiếp theo là bước đánh giá thực thể (Topic Evaluate - TE) để chọn ra những thực thể phù hợp nhất làm điểm bắt đầu cho việc khám phá trong một KG. Bước này sẽ sử dụng LLM để đánh giá câu hỏi q và các thực thể xuất hiện, từ đó chọn ra N thực thể chủ đề $\mathcal{E}_{\text{topic}}(e_1, e_2, \dots, e_N)$ để làm điểm khởi đầu cho n đường dẫn suy luận $\mathcal{P}(p_1, p_2, \dots, p_N)$. (Với N là siêu tham số chiều rộng khám phá của mô hình hay số lượng đường dẫn suy luận tối đa được giữ lại tại mỗi vòng lặp)

Trước khi bước vào lần đầu tiên truy xuất đồ thị, một mô hình Dense Retrieval

Models (DRMs) được sử dụng để chọn ra $top - k$ đoạn văn Ctx^0 từ các tài liệu liên kết với các chủ đề ban đầu \mathcal{E}_{topic} . Sau đó LLM sẽ đánh giá thông tin này có đủ để trả lời câu hỏi hay không dựa vào kiến thức đã được huấn luyện của nó. Nếu LLM kết luận rằng thông tin hiện có đủ để trả lời câu hỏi, các bước tiếp theo là không cần thiết và có thể trực tiếp trả về câu hỏi cho người dùng. Ngược lại, nếu thông tin hiện có không đủ, mô hình sẽ sử dụng LLM với prompt B.6 để dự đoán các manh mối, bằng chứng cần bổ sung để trả lời câu hỏi và tạo ra truy vấn phù hợp - *Clues* - để lấy những thông tin hữu ích này.

$$Clues = Prompt_{request}(q, Ctx^0) \quad (1)$$

3.3.2 Khám phá tri thức

Phần này sẽ trình bày chi tiết cách mô hình được đề xuất lặp lại quy trình khám phá tri thức để thống nhất và gắn kết chặt chẽ các tri thức từ KG và các đoạn văn được xây dựng trong phần 3.2.

Tại thời điểm bắt đầu của vòng lặp thứ I , mỗi đường dẫn p_n (với $n \in \{1, N\}$) bao gồm $I - 1$ bộ 3, một bộ ba ở đây tức là: $p_n = \{(e_{s,n}^i, r_j^i, e_{o,n}^i)\}_{i=1}^{I-1}$, trong đó $e_{s,n}^I$ và $e_{o,n}^I$ lần lượt biểu thị các thực thể chủ ngữ và thực thể tân ngữ, r_j^I là một quan hệ cụ thể giữa chúng, các bộ ba $(e_{s,n}^I, r_j^I, e_{o,n}^I)$ và $(e_{s,n}^{I+1}, r_j^{I+1}, e_{o,n}^{I+1})$ được kết nối với nhau. Tập hợp các thực thể đuôi và các quan hệ trong \mathcal{P} lần lượt được ký hiệu là $\mathcal{E}^{I-1} = e_1^{I-1}, e_2^{I-1}, \dots, e_n^{I-1}$ và $\mathcal{R}^{I-1} = r_1^{I-1}, r_2^{I-1}, \dots, r_n^{I-1}$. Quá trình lặp lại này gồm có 2 phần chính: nâng cao truy xuất đồ thị dựa trên ngữ cảnh và truy xuất ngữ cảnh thông qua hướng dẫn từ tri thức:

3.3.2.1 Nâng cao truy xuất đồ thị thông qua ngữ cảnh:

Bằng cách tận dụng sự kết nối phong phú có cấu trúc của kiến thức trên đồ thị tri thức (KG), việc tìm kiếm trên đồ thị nhằm khám phá và thiết lập các khái niệm cấp cao cũng như mối quan hệ giữa câu hỏi và thông tin mục tiêu, vốn dường như cách xa nhau trong không gian ngữ nghĩa. Mô hình được đề xuất trong khóa luận bao gồm các bước sau đây.

Khám phá mối quan hệ (Relation Exploration): tại thời điểm bắt đầu của vòng lặp thứ I , mô hình sẽ tìm kiếm toàn bộ quan hệ được liên kết tới thực thể

cuối cùng của mỗi reasoning path thông qua sử dụng hàm:

$$\text{Edge}(e_j^{I-1}) = \{(r_{j,m}^I, h_m) | h_m \in \{True, False\}\} \quad (2)$$

Hàm $\text{Edge}()$ là hàm tìm kiếm các quan hệ (relationship) của thực thể. Trong đó h biểu thị chiều của quan hệ $r_{j,m}^I$ đối với thực thể e_j^{I-1} .

Cắt giảm các quan hệ (Relation Prune): Thông qua phương trình 2 ta thu được các tập các mối quan hệ ứng viên $\{\text{Edge}(e_j^I)\}_{j=1}^W$. Tại đây sẽ sử dụng phương pháp few-shot prompt để yêu cầu LLM đánh giá, lựa chọn và chấm điểm các mối quan hệ với các đường dẫn suy luận tương ứng có khả năng tìm thấy thực thể chứa thông tin ngữ cảnh phù hợp hữu ích cho việc giải quyết câu hỏi q . Rồi sau đó chọn ra $top - N$ mối quan hệ với các đường dẫn suy luận tương ứng có số điểm cao nhất để sử dụng cho bước tiếp theo. Tại đây, có 2 prompt được xây dựng để sử dụng:

$$PROMPT_{RP}(p_n, q, \text{Edge}(e_n^{I-1})) \quad (3)$$

và

$$PROMPT_{RP_all}(q, \{p_n, \text{Edge}(e_n^{I-1})\}_{n=1}^N) \quad (4)$$

Chi tiết của các prompt được trình bày tại phần phụ lục B.2 và B.3. phương trình 3 được xây dựng với mục tiêu là gọi LLM nhiều lần cho việc cắt giảm mối quan hệ trên từng trên từng đường dẫn suy luận, trong khi đó phương trình 4 được xây dựng để xử lý việc chọn quan hệ phù hợp cho tất cả các reasoning paths trong 1 lần sử dụng LLM. Phương trình 3 đơn giản hóa nhiệm vụ của LLM và luôn giữ được các đường dẫn suy luận trước đó được tiếp tục, tuy nhiên nó lại thiếu hiệu quả khi phải sử dụng LLM nhiều lần và nhiều trường hợp giữ lại các đường dẫn không chứa thông tin hữu ích. Còn tại phương trình 4, việc xử lý tất cả trong 1 lần giúp cho việc sử dụng LLM giảm đi từ đó nâng cao tốc độ suy luận và cho phép xem xét một cách tổng quan các mối liên kết giữa nhiều đường dẫn suy luận cùng lúc cùng lúc, tạo điều kiện cho việc đánh giá trở nên khách quan, chính xác hơn. Tuy nhiên, việc đưa toàn bộ các quan hệ với số lượng lớn đối với từng đầy các đường dẫn suy luận là một thách thức trong khả năng hiểu toàn bộ ngữ cảnh và đưa ra lựa chọn chính xác của LLM.

Khám phá thực thể (Entity Discovery): Tại đây chúng ta có tập các đường dẫn suy luận với được chọn từ bước cắt giảm mối quan hệ với các mối quan hệ tương ứng được thêm vào. Cho một đường dẫn suy luận p_n trong \mathcal{P} với e_n^{I-1} là

thực thể đuôi của p_n và các mối quan hệ tương ứng r_n^{I-1} . Tại đây mô hình sử dụng hàm sau:

$$Tail(e_n^{I-1}, (r_n^{I-1}, h)) = \mathcal{C}_n^{I-1} \quad (5)$$

từ đó xác định được tập các thực thể ứng viên \mathcal{C}_n^{I-1} kết nối với e_n^{I-1} thông qua quan hệ (r_n^{I-1}, h) .

3.3.2.2 Truy xuất ngữ cảnh thông qua hướng dẫn từ tri thức:

Trong bước này, mô hình sẽ tập trung khai thác các thông tin hữu ích dựa theo các tri thức trích xuất được từ KG thông qua phần 3.3.2.1.

Sau khi xác định được tất cả các tập thực thể ứng viên \mathcal{C}_n^{I-1} thông qua thực thi hàm 5. Mô hình sẽ thu thập các tài liệu được liên kết với các thực thể ứng viên, rồi kết hợp với $top-k$ tài liệu Ctx^{I-1} để đánh giá, xếp hạng những tài liệu và chọn ra $top-k$ tài liệu có điểm cao nhất. Say đây là các bước chi tiết thực hiện:

Truy xuất ngữ cảnh, đoạn văn dựa trên các thực thể: để tìm kiếm những thông tin hữu dụng từ trong các tài liệu ngữ cảnh liên kết với các thực thể ứng viên, mô hình sẽ sử dụng DRMs để tính điểm liên quan của các tài liệu. Thay vì tính toán trực tiếp điểm liên quan giữa tài liệu ngữ cảnh và câu hỏi (từ đó bỏ qua ngữ cảnh giữa tài liệu và thực thể tương ứng của nó) điểm sẽ được tính thông qua việc chuyển đổi các đường dẫn suy luận p_n tương ứng với từng \mathcal{C}_n^{I-1} thành một câu ngắn gọn và thêm nó vào ngữ cảnh rồi từ đó sẽ sử dụng mô hình DRMs để đánh giá điểm liên quan giữa ngữ cảnh mới này và câu hỏi. Đây là phương trình tính điểm liên quan giữa tài liệu thứ z của thực thể ứng viên $chunk_{n,z}^{I-1}$ như sau:

$$srl(chunk_{n,z}^{I-1}) = \text{DRM}(q, [brief_sentence(p_n), : chunk_{n,z}^{I-1}]) \quad (6)$$

Vì các tài liệu được lưu trữ dưới dạng cây tài liệu, nên giữa các tài liệu luôn có mối quan hệ về ngữ cảnh, vậy nên để điểm của mỗi tài liệu sẽ chịu tác động từ các tài liệu khác. Hiện tại chưa có nghiên cứu nào về cách tính sự tương tác giữa các tài liệu. Vậy nên sau nhiều lần thực nghiệm, tôi đã xây dựng phương trình tính toán sự liên quan giữa tài liệu u và tài liệu v trong việc trả lời câu hỏi q dựa trên kết quả của phương trình 6 như sau (với α là một siêu tham số để điều chỉnh mức độ tương tác giữa các tài liệu):

$$interaction(u, v) = \frac{1}{Distance(u, v)^2 + \alpha} \times srl(u) \times srl(v) \quad (7)$$

Điểm cuối cùng của mỗi tài liệu sẽ là trung bình cộng của điểm liên quan giữa tài liệu đó với tất cả các tài liệu có liên kết với chúng cộng với điểm phương trình 6 của chúng:

$$c_score_{j,z}^{I-1} = \frac{1}{|linked_chunks|} \times \sum_{v \in linked_chunks} interaction(chunk_{n,z}^{I-1}, v) + srl(chunk_{n,z}^{I-1}) \quad (8)$$

trong đó $linked_chunks$ là tập hợp các tài liệu liên kết với $chunk_{n,z}^{I-1}$. Sau đó dựa trên điểm, mô hình sẽ chọn ra $top - k$ tài liệu

$$Ctx^I = \{chunk_i \mid i \in \arg \max_{i_1, i_2, \dots, i_k} (Chunk_{related}^I)\}, \quad (9)$$

với $Chunk_{related}^I$ là toàn bộ điểm của tất cả các tài liệu liên kết với các thực thể ứng viên C_n^{I-1} để sử dụng cho bước tiếp theo.

Cắt giảm thực thể (Entity Prune): Dựa vào Ctx^I được chọn ra từ bước truy xuất ngữ cảnh, mô hình sẽ sử dụng mô hình LLM để đánh giá chấm điểm các đường dẫn suy luận với đuôi là các thực thể ứng viên tương ứng để chọn ra những đường dẫn suy luận có khả năng trả lời câu hỏi tốt nhất và để sử dụng cho các vòng lặp tiếp theo. Dưới đây là prompt được sử dụng cho việc cắt giảm thực thể:

$$PROMPT_{EP}(Ctx^I, q, p_n, C_n^{I-1}) \quad (10)$$

Sau đó sẽ chọn ra $top - N$ đường dẫn suy luận tốt nhất để sử dụng cho vòng lặp kế tiếp. Chi tiết về prompt được trình bày tại phần phụ lục B.4.

3.3.3 Suy luận từ tri thức kết hợp

Cuối vòng lặp thứ i , mô hình sẽ sử dụng LLM với toàn bộ tri thức tìm được từ các bước phía trên, bao gồm các đường dẫn suy luận, manh mối $Clue^{I-1}$ và Ctx^I để đánh giá liệu rằng các thông tin đó có đủ để trả lời câu hỏi, với $Clue^{I-1}$ là phản hồi truy xuất từ lần lặp phía trước với mục tiêu là để duy trì tri thức hữu ích từ bối cảnh lịch sử. Nếu LLM đánh giá rằng kiến thức được cung cấp là đủ để đưa ra câu trả lời, sẽ trực tiếp trả về câu trả lời. Ngược lại, thì sẽ tạo ra truy vấn $Clue^I$ và bắt đầu vòng lặp tiếp theo. Để hướng dẫn LLM đánh giá, mô hình sẽ sử dụng prompt sau:

$$PROMPT_{RS}(q, \mathcal{P}, Clue^{I-1}, Ctx^I) = \begin{cases} \text{Ans.}, & (\text{Có}) \\ Clues^i, & (\text{Không}) \end{cases} \quad (11)$$

Chi tiết về prompt sẽ được trình bày tại phần phụ lục B.5.

Chương 4

Thực nghiệm

Trong chương này, tôi trình bày quá trình cài đặt hệ thống, dữ liệu thực nghiệm và kết quả thực nghiệm của hệ thống Chatbot được đề xuất. Mục tiêu chính của chương này là đánh giá cách tiếp cận được đề xuất nhằm cải thiện phản hồi của chatbot thông qua mô hình RAG tích hợp, kết hợp truy xuất thông tin dựa trên văn bản với lập luận đồ thị tri thức. Chương này nhằm chứng minh hiệu quả của hệ thống được cải tiến trong việc nâng cao mức độ phù hợp, độ chính xác và ngữ cảnh của các phản hồi so với các phương pháp cơ bản.

4.1 Hệ thống thực nghiệm

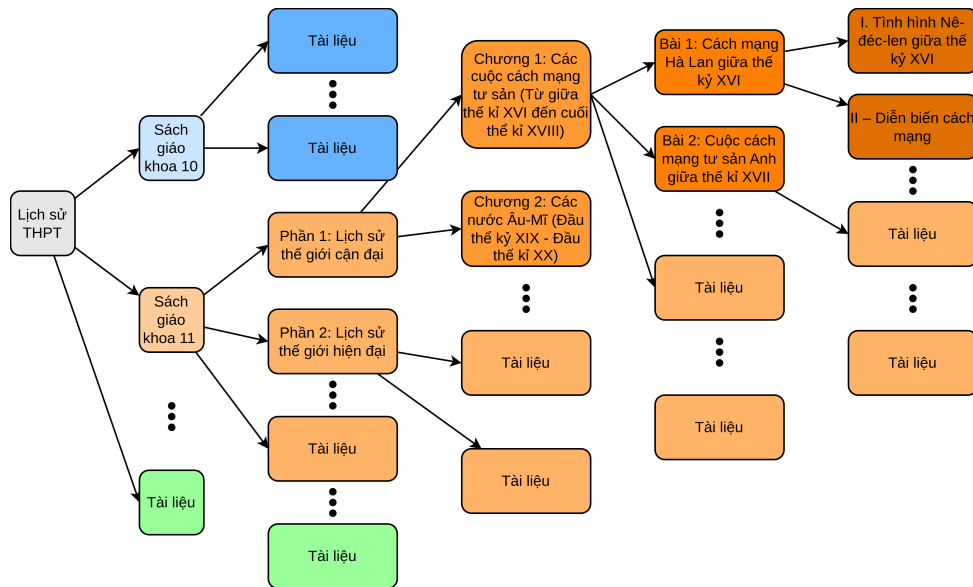
Trong phần này, chúng tôi trình bày chi tiết cách thiết lập thử nghiệm được sử dụng để đánh giá hệ thống Lookinglass Retrieval-Augmented Generation (RAG). Nội dung bao gồm môi trường phần cứng và phần mềm, các bộ dữ liệu, các loại truy vấn, phương pháp đo lường và ghi nhận kết quả, cũng như các bước được thực hiện để đảm bảo tính tái lập của thử nghiệm.

4.1.1 Các phần của hệ thống

Tôi đã phát triển hệ thống thực nghiệm cho phương pháp được đề xuất tại chương 3. Hệ thống bao gồm 3 phần chính sau:

4.1.1.1 Xây dựng đồ thị tri thức, cơ sở dữ liệu tri thức từ tài liệu văn bản

Phần này sẽ trình bày chi tiết các bước xây dựng đồ thị tri thức và cây tài liệu văn bản từ dữ liệu thực nghiệm. Cụ thể, tôi sẽ trình bày cách xây dựng đồ thị tri



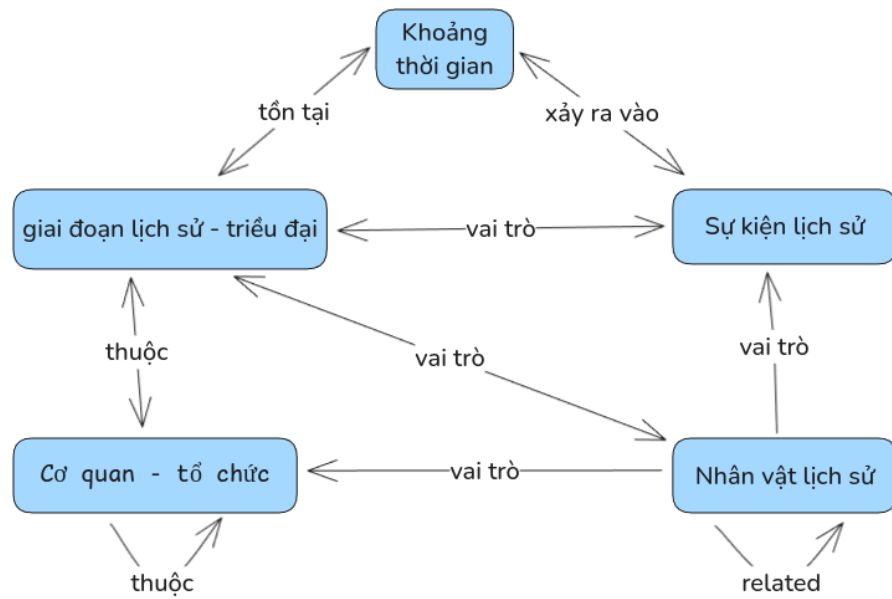
Hình 4.1: Minh họa cây tài liệu được xây dựng từ bộ tài liệu sách giáo khoa Lịch Sử

thức từ tập dữ liệu sách giáo khoa Lịch sử, cách xây dựng cây tài liệu văn bản từ tập dữ liệu sách giáo khoa Lịch sử theo phương pháp được đề xuất tại phần 3.2.

Xây dựng cây tài liệu: Vì sách giáo khoa Lịch sử là tài liệu văn bản được chia đầu mục một cách cụ thể và chi tiết. Vậy nên tôi sẽ sử dụng phương pháp được nêu tại phần 3.2.1 để xây dựng cây tài liệu văn bản từ tập dữ liệu sách giáo khoa Lịch sử. Các bước xây dựng cây tài liệu văn bản cũng sử dụng mô hình ngôn ngữ lớn GPT-4o với prompt được hiển thị tại phụ lục. Cây tài liệu văn bản được lưu vào cơ sở dữ liệu Neo4j.

Kết quả thu được một cây tài liệu với 1006 nút với mỗi nút là một đầu mục trong sách giáo khoa Lịch sử. Hình 4.1 mô tả một phần của cây tài liệu văn bản sau khi xây dựng.

Xây dựng đồ thị tri thức: Vì tài liệu thực nghiệm là lịch sử, vậy nên tôi sẽ trích xuất các thực thể từ các tài liệu thực nghiệm dựa theo ontology được mô tả trong hình 4.2. Sử dụng phương pháp được nêu tại phần 3.2.2 Các bước phát hiện thực thể và truy xuất mối quan hệ đều sử dụng mô hình ngôn ngữ lớn GPT-4o với prompt được hiển thị tại phụ lục. Rồi chúng được lưu vào cơ sở dữ liệu đồ thị tri thức Neo4j. Sau khi tri thức là các thực thể và mối quan hệ giữa chúng được lưu vào cơ sở dữ liệu, tôi thực hiện việc cải thiện tri thức để giúp hoàn thiện KG và giảm thiểu sai sót với việc cũng sử dụng model GPT-4o để kiểm tra mối quan hệ



Hình 4.2: Ontology cho lịch sử

giữa các thực thể và mối quan hệ.

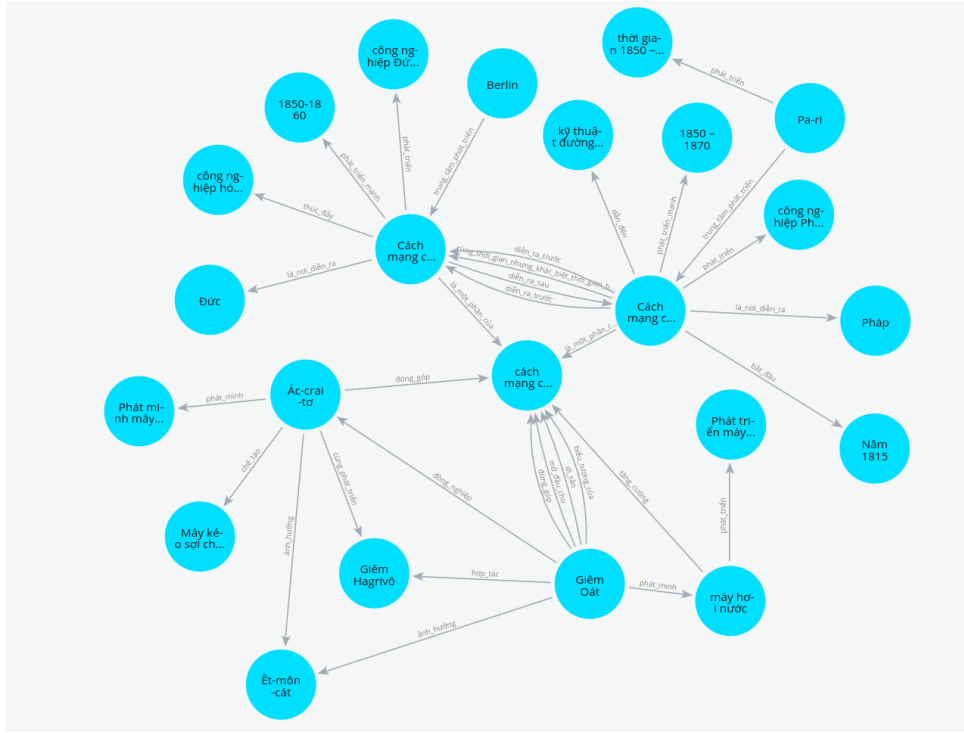
Kết quả thu được 6600 nút thực thể và 10500 mối quan hệ giữa chúng. Hình 4.3 mô tả một phần của đồ thị tri thức sau khi xây dựng.

4.1.1.2 Xây dựng hệ thống RAG

Để đánh giá hiệu quả của hệ thống RAG được đề xuất, tôi đã thiết lập thử nghiệm xây dựng mô hình được đề xuất tại phần 3.3 với các thành phần chính gồm LLM, DRMs và các thông số cài đặt.

Mô hình ngôn ngữ lớn là thành phần chính của mô hình. Trong khóa luận này, tôi chủ yếu sử dụng GPT-4o-mini là LLM xương sống của mô hình. Do các mô hình này có kích thước lớn và cần tài nguyên phần cứng mạnh mẽ, vậy nên tôi sẽ sử dụng chúng thông qua API từ các dịch vụ cung cấp mô hình như OpenAI API, ... để thực hiện thử nghiệm.

Mô hình truy xuất dày đặc đóng vai trò quan trọng trong tác vụ đánh giá tài liệu. Trong quá trình thực nghiệm, tôi đã thử nghiệm sử dụng các DRMs khác nhau bao gồm BGE-Embedding, BGE-reranker [48], Minilm [49] . Kết quả thực nghiệm cho thấy mô hình BGE-reranker cho kết quả tốt hơn so với các mô hình còn lại, tuy nhiên sẽ gặp vấn đề trong tốc độ tính toán, vậy nên tôi sẽ sử dụng



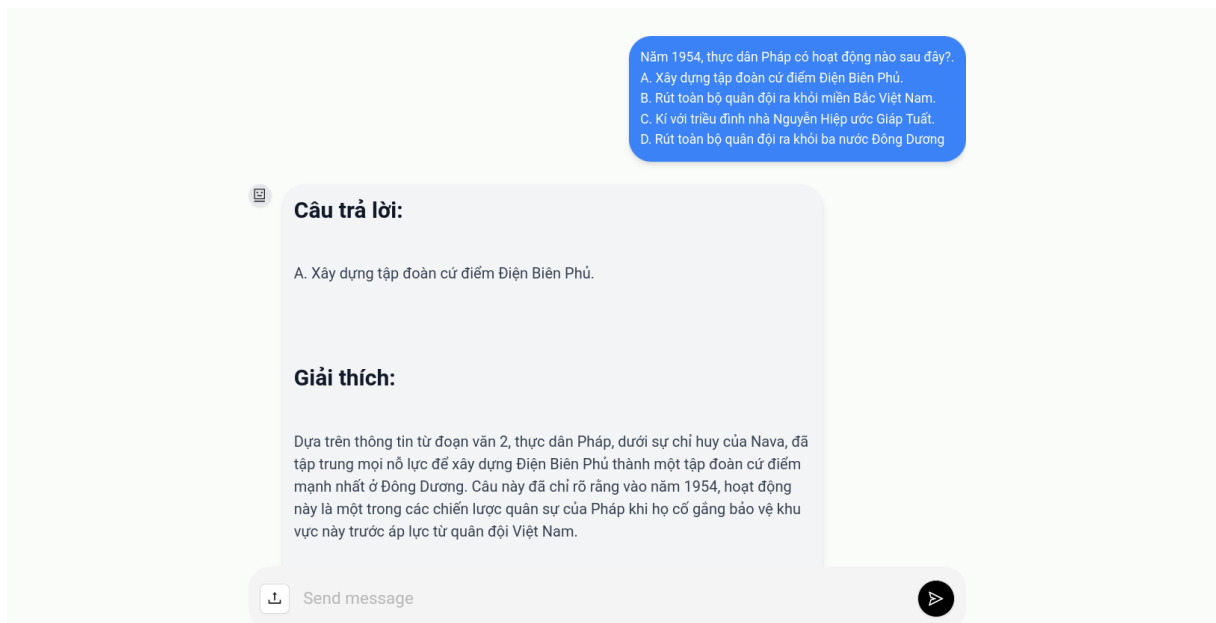
Hình 4.3: Minh họa một phần đồ thị tri thức được xây dựng dựa trên tài liệu sách giáo khoa Lịch Sử

kết hợp cùng mô hình BGE-Embedding để thực hiện thử nghiệm. Do đó, tôi sẽ sử dụng mô hình bge-m3 để chọn ra số lượng lớn các tài liệu phù hợp nhất với câu hỏi và sau đó sử dụng mô hình bge-reranker để xếp hạng lại các tài liệu đó để chọn ra tài liệu phù hợp nhất với câu hỏi (các mô hình DRMs được sử dụng đều là mô hình được đào tạo trước và không có bất kỳ sự tinh chỉnh nào). Vì cả hai mô hình này đều có kích thước nhỏ vậy nên tôi sẽ triển khai chúng trên Colab Pro.

Các thông số cài đặt của mô hình được đề xuất như sau: tôi đặt tham số temperature của LLM là 0.4 xuyên suốt quá trình khám phá tri thức và 0 cho quá trình lý luận tri thức. Độ rộng $N = 3$ là số lượng đường dẫn truy xuất và số lần tối đa của vòng lặp là 5 để đảm bảo thời gian trả lời của mô hình. Tại bước đánh giá tài liệu, tôi duy trì $k = 5$ tài liệu phù hợp nhất với câu hỏi. Tất cả các thông số cài đặt đều được chọn dựa trên kinh nghiệm và thực nghiệm trước đó.

4.1.1.3 Giao diện hệ thống

Để tạo ra một giao diện tương tác với người dùng, tôi đã xây dựng một ứng dụng web demo cho hệ thống được đề xuất. Ứng dụng web demo này sẽ cho phép người dùng nhập câu hỏi và nhận câu trả lời từ hệ thống. Ứng dụng web demo được xây



Hình 4.4: Giao diện người dùng của hệ thống

dựng dựa trên mô hình Client-Server, trong đó Server sẽ là hệ thống RAG được xây dựng và Client sẽ là giao diện người dùng, tuy nhiên trong phạm vi khóa luận này tôi chỉ tập trung vào cải thiện phản hồi của Chatbot. Hệ thống sẽ sử dụng FastAPI để xây dựng API cho hệ thống và NextJS để xây dựng giao diện người dùng. Hình 4.4 mô tả giao diện người dùng của hệ thống.

4.1.2 Cài đặt môi trường

4.1.2.1 Phần cứng

Nghiên cứu sử dụng 2 môi trường phần cứng khác nhau cho các giai đoạn khác nhau:

Bảng 4.1: Cấu hình phần cứng sử dụng trong nghiên cứu

Môi trường	Cấu hình	Mục đích sử dụng
Kaggle	10 instance với NVIDIA Tesla T4 GPU (16GB VRAM/instance)	Xây dựng đồ thị tri thức
Google Colab Pro	NVIDIA A100-SXM4-40GB	Triển khai hệ thống RAG chính

4.1.2.2 Phần mềm

Dưới đây là danh sách công cụ , thư viện phục vụ cho quá trình huấn luyện và xây dựng ứng dụng dịch máy:

Bảng 4.2: Danh sách thư viện và công cụ sử dụng trong nghiên cứu

Thư viện	Phiên bản	Chức năng
pandas	1.3.3	Thư viện xử lý dữ liệu dạng bảng, được sử dụng để xử lý dữ liệu thô từ các tài liệu sách giáo khoa Lịch sử
NumPy	1.21.2	Thư viện xử lý mảng nhiều chiều, được sử dụng để xử lý dữ liệu thô từ các tài liệu sách giáo khoa Lịch sử
python-docx	1.1.2	Thư viện xử lý tài liệu Word, được sử dụng để trích xuất dữ liệu từ các tài liệu sách giáo khoa Lịch sử
pdfplumber	0.11.4	Thư viện xử lý tài liệu PDF, được sử dụng để trích xuất dữ liệu từ các tài liệu sách giáo khoa Lịch sử
Neo4j	5.26.0	Hệ quản trị cơ sở dữ liệu đồ thị, được sử dụng để lưu trữ và truy vấn dữ liệu đồ thị tri thức
Huggingface	0.23.1	Thư viện cung cấp các mô hình nhúng, được sử dụng để nhúng câu hỏi và tài liệu văn bản
LangChain	0.3.7	Thư viện cung cấp các mô hình ngôn ngữ lớn, được sử dụng để xây dựng hệ thống RAG
FastAPI	0.115.3	Framework xây dựng hệ thống API, sử dụng để xây dựng ứng dụng web demo
NextJS	14.0.3	Framework UI Web, sử dụng để xây dựng giao diện tương tác với người dùng

4.2 Dữ liệu thực nghiệm

Dữ liệu cho việc xây dựng đồ thị tri thức: Chúng tôi đã sử dụng để bộ dữ liệu thực nghiệm là bộ sách giáo khoa Lịch Sử. Lịch sử là một đề tài chứa nhiều câu hỏi cần suy luận cũng như kiến thức chuyên ngành. Đây là một lĩnh vực phù hợp để kiểm tra khả năng hiểu ngôn ngữ và truy xuất thông tin của hệ thống chatbot được đề xuất. Các tài liệu sách giáo khoa đều là PDF, vậy nên chúng sẽ được tiền xử lý trước và chỉ trích xuất toàn bộ văn bản có trong các tài liệu đó.

Dữ liệu cho việc đánh giá hệ thống RAG: Để đánh giá chất lượng trả lời câu hỏi của hệ thống Chatbot, tôi sẽ sử dụng tập dữ liệu là tập đề thi tốt nghiệp THPT quốc gia môn Lịch sử từ năm 2018 đến 2022 với 4 mã đề khác nhau của từng năm. Đề thi này sẽ có 40 câu hỏi trắc nghiệm với 4 đáp A B C D, trong đó có 1 đáp án đúng duy nhất. Kiến thức của đề và trải dài kiến thức từ cơ bản đến nâng cao, từ đó giúp đánh giá khả năng trả lời câu hỏi của hệ thống Chatbot trong nhiều trường hợp khác nhau. Minh họa đề thi tốt nghiệp THPT quốc gia môn Lịch sử năm 2021 được mô tả trong phần phụ lục C.

4.3 Kết quả thực nghiệm

Với mỗi đề thi trong tập dữ liệu đề thi tốt nghiệp THPT quốc gia môn Lịch sử, tôi sẽ thử nghiệm trả lời trả lời 10 lần mỗi đề thi và đánh giá trung bình kết quả trả lời của hệ thống. Kết quả trả lời của hệ thống sẽ được đánh giá dựa trên các tiêu chí được nêu tại phần 4.3.1.

4.3.1 Chỉ số đánh giá

Trong phạm vi khóa luận này, tôi sử dụng hai chỉ số chuẩn để đánh giá hệ thống QA: Exact Match (EM) và F1 Score [50]. Ngoài ra tôi sử dụng thêm chỉ số Cohen's Kappa [51], đây là một chỉ số thống kê được sử dụng để đo lường mức độ thỏa thuận giữa hai bộ phân loại trong các tình huống có thể xảy ra ngẫu nhiên. Không giống như các chỉ số như EM hay F1, Cohen's Kappa không chỉ tính toán sự tương đồng tuyệt đối mà còn tính đến khả năng thỏa thuận ngẫu nhiên, do đó cung cấp một đánh giá khách quan hơn về hiệu suất của mô hình. Cohen's Kappa cũng hữu ích khi so sánh hiệu suất của các phương pháp khác nhau, chẳng hạn như phương pháp được đề xuất trong khóa luận và các phương pháp truyền thống.

Exact Match (EM) đo lường tỷ lệ phần trăm các dự đoán chính xác khớp với các câu trả lời thực tế. Nó được định nghĩa như sau:

$$\text{EM}(\%) = \frac{|\{i \mid \hat{a}_i = a_i\}|}{N} \times 100,$$

trong đó \hat{a}_i là câu trả lời dự đoán, a_i là câu trả lời thực tế, và N là số mẫu.

F1 Score xem xét cả độ chính xác (precision) và độ bao phủ (recall) của các câu trả lời dự đoán, cung cấp một sự cân bằng giữa hai yếu tố này. Điểm F1 cho mỗi trường hợp được tính như sau:

$$\text{F1}(\%) = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100.$$

Độ chính xác (precision) và độ bao phủ (recall) được tính dựa trên các từ chung giữa câu trả lời dự đoán và câu trả lời đúng. Điều này đặc biệt hữu ích trong các trường hợp câu trả lời là các cụm từ hoặc câu.

Cohen's Kappa được dùng để đánh giá mức độ đồng thuận giữa hai người đánh giá hoặc giữa hai bộ dữ liệu phân loại. Trong phạm vi khóa luận này, Cohen's Kappa được sử dụng để đo lường mức độ đồng thuận giữa đáp án của mô hình và đáp án của con người. Công thức tính Cohen's Kappa được định nghĩa như sau:

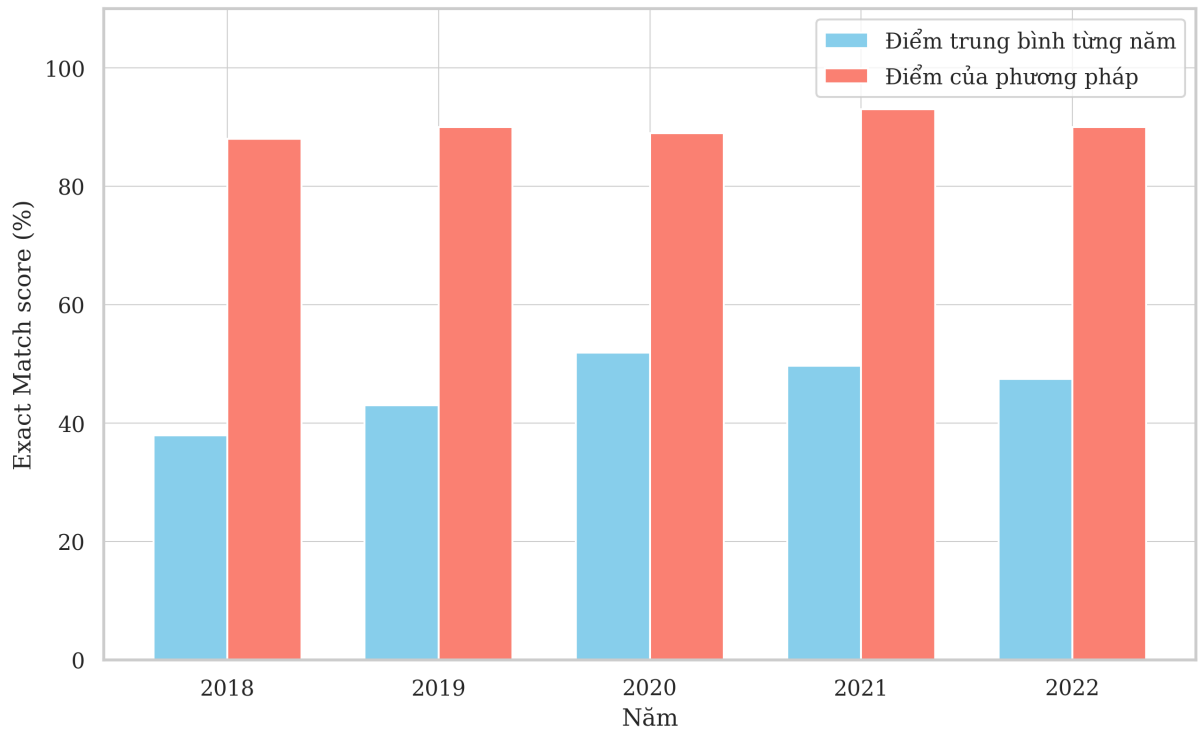
$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Trong đó p_o là tỷ lệ sự thỏa thuận giữa hai bộ phân loại, và p_e là tỷ lệ sự thỏa thuận ngẫu nhiên giữa hai bộ phân loại.

4.3.2 Kết quả so sánh với điểm trung bình của học sinh từng năm

Để đánh giá hiệu quả của hệ thống RAG được xây dựng, tôi đã so sánh điểm trung bình của các câu trả lời do hệ thống tạo ra với điểm trung bình thực tế của học sinh từng năm trong kỳ thi THPT Quốc gia môn Lịch sử. Dữ liệu tham khảo từ Bộ Giáo dục và Đào tạo cho thấy điểm trung bình của học sinh dao động trong khoảng từ 4.5 đến 6.0 trên thang điểm 10 trong giai đoạn 2018-2022. Kết quả trung bình do hệ thống RAG trả lời được thể hiện như hình 4.5.

Nhận xét: Điểm trung bình từ hệ thống RAG luôn đạt từ 8.8 đến 9 điểm, mức điểm đòi hỏi sự hiểu biết sâu rộng và kỹ năng suy luận tốt. Điểm số của hệ thống cũng cao tương đối hơn điểm trung bình của học sinh với sự chênh lệch từ 4 đến 5



Hình 4.5: Kết quả so sánh với kết quả trung bình của từng năm

điểm, cho thấy khả năng hiểu và trả lời câu hỏi tốt hơn phần lớn các thí sinh tham gia kỳ thi. Tổng quan, có thể thấy hệ thống RAG có khả năng trả lời câu hỏi trắc nghiệm môn Lịch sử với độ chính xác cao và đồng thuận tốt, giúp nâng cao chất lượng phản hồi của chatbot.

4.3.3 So sánh với các phương pháp RAG khác

Tôi đã thiết lập thực nghiệm so sánh với các phương pháp cơ sở khác để cung cấp cái nhìn tổng quan, bao gồm: phương pháp chỉ sử dụng LLM mà không có tri thức bên ngoài, RAG truyền thống dựa trên truy xuất văn bản trực tiếp từ tài liệu để trả lời câu hỏi. Bảng 4.3 trình bày kết quả so sánh:

Bảng 4.3: So sánh hiệu suất của các phương pháp khác nhau với GPT-4o-mini.

Baseline type	Metric		
	Exact Match	F1 score	Cohen's Kappa
LLM directly	89.41	89.34	85.67
LLM with CoT	90.04	89.95	85.89
Text-based RAG	82.07	82.13	75.89
Proposed	90.93	90.01	86.42

Nhận xét: kết quả trong Bảng 4.3 cho thấy phương pháp được đề xuất có hiệu suất vượt trội hơn so với các phương pháp khác ở cả ba chỉ số đánh giá: Exact Match, F1 score, và Cohen’s Kappa. Cụ thể khi so với việc để LLM trực tiếp trả lời, phương pháp đề xuất đã cải thiện 1.52% (Exact Match), 0.67% (F1 score) và 0.75 (Cohen’s Kappa). So với LLM sử dụng phương pháp CoT, hiệu suất tăng nhẹ với 0.89% (Exact Match), 0.06% (F1 score) và 0.53 (Cohen’s Kappa). Đặc biệt, so với RAG dựa trên truy xuất văn bản, sự cải thiện rõ rệt: 8.86% (Exact Match), 7.88% (F1 score), và 10.53 (Cohen’s Kappa), điều này có thể do với các câu hỏi trắc nghiệm chứa nhiều câu hỏi nhiễu, khiến cho việc truy xuất thông tin từ tri thức trở nên khó khăn và RAG truyền thống không thể đáp ứng được.

Nhìn chung, phương pháp đề xuất không chỉ kết hợp hiệu quả các ưu điểm của LLM và RAG mà còn cải thiện độ chính xác và sự đồng thuận trong việc trả lời các câu hỏi nói chung và câu hỏi trắc nghiệm nói riêng, làm nổi bật tiềm năng của nó trong việc áp dụng vào các bài toán thực tế.

4.3.4 Đánh giá sự tác động của LLMs

Thêm vào đó để đánh giá tác động của LLM với khả năng của mô hình, tôi cũng tiến hành thực nghiệm với các LLM khác bao gồm: GPT-4o, GPT-4o-mini, Llama3.3-70B, Llama3.1-8B. Kết quả được thể hiện trong bảng 4.4.

Bảng 4.4: So sánh hiệu suất giữa câu trả lời trực tiếp và phương pháp đề xuất với các mô hình khác nhau. Các giá trị trong ngoặc đơn biểu thị phần trăm cải thiện.

	GPT-4o		GPT-4o-mini		Llama3.3-70B		Llama3.1-8B	
	Direct	Proposed	Direct	Proposed	Direct	Proposed	Direct	Proposed
Exact Match	89.65	91.37 (↑ 1.9%)	89.41	90.93 (↑ 1.7%)	77.53	79.13 (↑ 2.1%)	60.58	68.43 (↑ 12.9%)
F1 Score	89.62	90.92 (↑ 1.5%)	89.34	90.01 (↑ 0.8%)	78.44	79.45 (↑ 1.3%)	60.21	67.28 (↑ 11.7%)
Cohen’s Kappa	85.99	86.42 (↑ 0.5%)	85.67	86.14 (↑ 0.6%)	69.81	72.78 (↑ 4.3%)	47.26	56.86 (↑ 20.3%)

Nhận xét: bảng 4.4 cho thấy tác động tích cực của phương pháp đề xuất đối với hiệu suất của các mô hình ngôn ngữ lớn (LLMs). Cụ thể:

- **Tăng cường hiệu suất tổng thể** khi các chỉ số đánh giá Exact Match, F1 Score, và Cohen’s Kappa đều cải thiện dù là cải thiện nhẹ khi sử dụng phương pháp đề xuất so với câu trả lời trực tiếp trên tất cả các mô hình. Cụ thể khi sử dụng mô hình lớn là GPT-4o, GPT-4o-mini kết quả tăng khoảng 1.4%,

trong khi đó với các mô hình nhỏ hơn mang lại kết quả tốt hơn khoảng 7-9%.

- **Khả năng cải thiện khác biệt giữa các mô hình:** Các mô hình lớn hơn như GPT-4o, GPT-4o-mini, Llama3.3-70B cho thấy sự cải thiện khiêm tốn hơn so với ,điều này có thể đo mức độ phức tạp và khả năng biểu diễn vốn đã rất cao của các mô hình lớn. Nhưng với Llama3.1-8B, mức độ cải thiện rõ rệt hơn (Exact Match tăng từ 60.58% lên 68.43% và F1 Score từ 60.21% lên 67.28%). Điều này chỉ ra rằng các mô hình nhỏ hơn có tiềm năng được cải thiện đáng kể nhờ phương pháp đề xuất.

Từ kết quả trên có thể thấy phương pháp đề xuất đã thể hiện khả năng cải thiện đáng kể hiệu suất của các LLMs, đặc biệt đối với các mô hình nhỏ hoặc trung bình.

4.3.5 Đánh giá khả năng trả lời câu hỏi và đưa ra giải thích

Ngoài các đánh giá dựa trên sự chính xác của đáp án trắc nghiệm, tôi cũng đã thực hiện đánh giá khả năng của hệ thống trong việc trả lời câu hỏi mở và đưa ra giải thích cho câu trả lời. Kết quả các thực nghiệm được thể hiện trong các bảng 4.5, 4.6, và 4.7.

Bảng 4.5: Thực nghiệm trả lời câu hỏi mở.

Câu hỏi	Quốc gia nào phát động Chiến tranh lạnh chống Liên Xô?
Câu trả lời	Câu trả lời: Quốc gia phát động Chiến tranh lạnh chống Liên Xô là Mỹ. Giải thích: Căn cứ vào các đoạn văn đã cung cấp, đặc biệt từ đoạn 1, có thể thấy rõ rằng sau Chiến tranh thế giới thứ hai, mối quan hệ giữa Liên Xô và Mỹ đã nhanh chóng chuyển từ liên minh chống phát xít sang đối đầu. Trong đoạn 1, thông điệp của Tổng thống Truman vào ngày 12-3-1947 được nhấn mạnh như là khởi đầu cho chính sách chống Liên Xô của Mỹ, được gọi là "Học thuyết Truman". Điều này cho thấy Mỹ đã chủ động phát động chính sách chiến tranh lạnh để chống lại Liên Xô, với ý định ngăn chặn sự lan rộng của chủ nghĩa xã hội và tăng cường ảnh hưởng của mình trên toàn thế giới. Bên cạnh đó, trong đoạn 5, có đề cập đến việc các nước Tây phương do Mỹ lãnh đạo đã theo đuổi chính sách chống cộng và tiến hành Chiến tranh lạnh, điều này càng củng cố cho luận điểm rằng Mỹ là quốc gia phát động chiến tranh lạnh chống Liên Xô. Do đó, câu trả lời là Mỹ là quốc gia phát động Chiến tranh lạnh chống Liên Xô.
Đáp án	Mỹ

Nhận xét: Từ kết quả của các thực nghiệm trên, có thể thấy khả năng trả lời các câu hỏi mở và đưa ra giải thích của hệ thống đề xuất là khá tốt. Cụ thể, hệ

thống đã trả lời chính xác và cung cấp các đoạn văn liên quan đối với các câu hỏi tự luận với độ chính xác của quá trình trích xuất thông tin từ tri thức đạt trên 90%. Đối với các câu hỏi trắc nghiệm, hệ thống cũng đã trả lời chính xác và cung cấp giải thích cho đáp án đúng và lý do các đáp án còn lại sai. Điều này cho thấy hệ thống không chỉ có khả năng trả lời câu hỏi mà còn có khả năng giải thích câu trả lời, giúp người dùng hiểu rõ hơn và tin tưởng hơn vào câu trả lời của hệ thống.

Bảng 4.6: Thực nghiệm trả lời câu hỏi mở.

Question	Sự phát triển thịnh đạt của các quốc gia phong kiến ở Đông Nam Á thế kỷ X-XVIII được biểu hiện như thế nào?
Câu trả lời	<p>Câu trả lời: Sự phát triển thịnh đạt của các quốc gia phong kiến ở Đông Nam Á từ thế kỷ X đến XVIII được biểu hiện qua việc hình thành và phát triển của nhiều quốc gia phong kiến mạnh mẽ, sự tiến bộ của nền kinh tế và văn hóa, cũng như việc xây dựng những vùng kinh tế quan trọng.</p> <p>Giải thích:</p> <ol style="list-style-type: none"> 1. Hình thành quốc gia phong kiến mạnh mẽ: Trong khoảng thời gian này, nhiều quốc gia phong kiến lớn đã hình thành và củng cố quyền lực, như Vương quốc Cam-pu-chia, Đại Việt, Champa, và Vương quốc Pa-gan (Mi-an-ma). Các quốc gia này đã thực hiện các cuộc chinh phục và thống nhất các lãnh thổ nhỏ hơn, gia tăng quyền lực và sự giàu có. 2. Phát triển kinh tế: Kinh tế khu vực phát triển thịnh vượng với sản xuất nông nghiệp lớn, bao gồm lúa gạo và các sản phẩm thủ công như vải, đồ sứ... Sự phát triển này tạo ra khả năng cung cấp sản phẩm quý giá cho thị trường thế giới, thu hút thương nhân từ nhiều quốc gia đến giao thương. 3. Xây dựng văn hóa dân tộc: Đi cùng sự phát triển kinh tế là sự hình thành và phát triển văn hóa của các dân tộc Đông Nam Á, với việc tiếp thu và sáng tạo trên nền tảng văn hóa Ấn Độ. Điều này cho thấy sự phát triển không chỉ ở mặt kinh tế mà còn cả ở mặt văn hóa, hình thành nền văn hóa riêng của từng quốc gia. <p>Tổng hợp lại, sự phát triển thịnh đạt của các quốc gia phong kiến ở Đông Nam Á từ thế kỷ X đến XVIII thể hiện qua cả yếu tố chính trị, kinh tế và văn hóa, với những quốc gia lớn mạnh và các hoạt động giao thương sôi nổi.</p>
Đáp án	<p>Thế kỷ X-XVIII là thời kỳ thịnh vượng của các quốc gia phong kiến Đông Nam Á, thể hiện qua ba lĩnh vực chính:</p> <ul style="list-style-type: none"> • Chính trị: Sự hình thành và phát triển của nhiều quốc gia phong kiến, như Vương triều Mô-giô-pa-hit ở In-đô-nê-xi-a, Đại Việt dưới các triều Lý, Trần, Lê sơ, Vương quốc Ăng-co của Campuchia, quốc gia Pa-gan ở Mi-an-ma, Vương quốc Su-khô-thay ở lưu vực sông Mê Nam và Vương quốc Lan Xang ở sông Mê Công. • Kinh tế: Sự phát triển mạnh mẽ của các nền kinh tế, hình thành các vùng kinh tế quan trọng cung cấp lúa gạo, cá, sản phẩm thủ công và tài nguyên thiên nhiên. Đông Nam Á trở thành điểm giao thương của các lái buôn quốc tế. • Văn hóa: Nền văn hóa dân tộc được hình thành, đóng góp vào kho tàng văn hóa chung của nhân loại.

Bảng 4.7: Thực nghiệm lý giải câu hỏi trắc nghiệm.

Câu hỏi	<p>Thế kỉ X – XII, ở khu vực Đông Nam Á, Campuchia được gọi là</p> <p>A. vương quốc phát triển nhất.</p> <p>B. vương quốc hùng mạnh nhất.</p> <p>C. vương quốc mạnh và ham chiến trận nhất.</p> <p>D. vương quốc chịu ảnh hưởng sâu sắc nhất của văn hóa Ấn Độ.</p>
Câu trả lời	<p>Câu trả lời: Thế kỉ X – XII, ở khu vực Đông Nam Á, Campuchia được gọi là C. vương quốc mạnh và ham chiến trận nhất.</p> <p>Giải thích: Dựa trên đoạn văn thứ nhất, thông tin cho thấy "trong các thế kỉ X – XII, Cam-pu-chia trở thành một trong những vương quốc mạnh và ham chiến trận nhất Đông Nam Á." Từ đây, ta có thể suy ra rằng trong thời kỳ này, Campuchia không chỉ mạnh mẽ mà còn có sự tham gia tích cực trong các hoạt động chiến tranh, bao gồm cả việc chinh phục lãnh thổ của các quốc gia khác.</p> <p>Phân tích các lựa chọn còn lại:</p> <ul style="list-style-type: none"> • A. vương quốc phát triển nhất. - Mặc dù Campuchia phát triển nhưng không thể khẳng định nó là "phát triển nhất" trong toàn khu vực. • B. vương quốc hùng mạnh nhất. - Mặc dầu Campuchia mạnh mẽ, từ thông tin đề cập, từ "hùng mạnh nhất" có thể hiểu là thuộc về lĩnh vực quân sự, không nhất thiết là mạnh nhất tổng thể trong mọi khía cạnh. • D. vương quốc chịu ảnh hưởng sâu sắc nhất của văn hóa Ấn Độ. - Mặc dù văn hóa Ấn Độ có ảnh hưởng lớn đến các vương quốc Đông Nam Á, không có thông tin rõ ràng chứng minh Campuchia là vương quốc chịu ảnh hưởng sâu sắc nhất so với các quốc gia khác trong khu vực.
Đáp án	C. vương quốc mạnh và ham chiến trận nhất.

Chương 5

Kết luận

5.1 Kết luận

Trong bối cảnh mà mô hình ngôn ngữ lớn càng ngày càng chứng tỏ được tiềm năng lớn, việc ứng dụng vào các ứng dụng vào các bài toán thực tế trở thành một chủ đề được chú trọng phát triển và nghiên cứu, đặc biệt là trong lĩnh vực hỏi đáp khi mà các Chatbot hỗ trợ như GPT, Gemini,... ngày càng trở nên phổ biến. Việc đảm bảo chất lượng phản hồi của Chatbot trở thành yếu tố cốt lõi để nâng cao hiệu quả và tính ứng dụng của chúng. Khóa luận này nghiên cứu về phương pháp cải thiện chất lượng phản hồi của Chatbot thông qua việc sử dụng phương pháp tạo sinh tăng cường truy xuất kết hợp với đồ thị tri thức.

Cụ thể, khóa luận đã nghiên cứu và đề xuất một phương pháp tích hợp dựa trên mô hình tăng cường truy xuất tạo sinh (Retrieval-Augmented Generation) với đồ thị tri thức (Knowledge Graph) nhằm nâng cao hiệu quả phản hồi của Chatbot trong các miền tri thức cụ thể với các tài liệu đã được chuẩn hóa thuộc miền tri thức đó. Phương pháp được đề xuất gồm có các phần xây dựng cơ sở dữ liệu đồ thị, và tạo sinh phản hồi từ thông tin truy xuất thông qua suy luận trên đồ thị tri thức.

Để kiểm chứng phương pháp đề xuất, tôi đã phát triển hệ thống thực nghiệm và thực nghiệm với dữ liệu là các đề thi THPT Lịch sử các năm. Kết quả từ các thử nghiệm đã cho thấy rằng hệ thống RAG được đề xuất đã khắc phục một số hạn chế hiện tại của các mô hình ngôn ngữ lớn và phương pháp RAG truyền thống, như hiện tượng ảo giác thông tin, độ chính xác trong xử lý truy vấn đa bước. Kết quả đạt được không chỉ nâng cao chất lượng phản hồi của Chatbot mà còn mở ra các ứng dụng tiềm năng trong các lĩnh vực thực tế như giáo dục, luật, và tài chính.

Tổng quan, khóa luận đã đạt được các mục tiêu đề ra, đồng thời đóng góp thêm vào lĩnh vực nghiên cứu về Chatbot và các ứng dụng của mô hình ngôn ngữ lớn trong thực tế.

5.2 Định hướng phát triển trong tương lai

Mặc dù hệ thống được đề xuất trong khóa luận đã cho thấy những kết quả đầy hứa hẹn, tuy nhiên vẫn còn một số điểm cần cải thiện và mở rộng trong tương lai để tăng cường hơn nữa các khả năng của nó:

1. **Mở rộng phạm vi tài liệu:** Hiện tại, hệ thống chỉ sử dụng dữ liệu từ các tài liệu được chuẩn hóa với nhau, cụ thể như là sách giáo khoa. Trong tương lai, việc mở rộng phạm vi tài liệu từ nhiều nguồn khác nhau như báo chí, sách báo, tạp chí,... sẽ giúp hệ thống trở nên phong phú hơn và đa dạng hơn.
2. **Mở rộng cơ sở dữ liệu đồ thị:** Cơ sở dữ liệu đồ thị mới chỉ được xây dựng từ các tài liệu nên vẫn còn hạn chế về số lượng thông tin, tri thức. Việc tích hợp với các nền tảng cơ sở dữ liệu đồ thị tri thức với lượng thông tin phong phú và cập nhật liên tục như Wikidata, DBpedia,... sẽ giúp hiệu quả truy xuất, suy luận được nâng cao.
3. **Mở rộng kiến trúc hệ thống:** Trong khóa luận, hệ thống chỉ tập trung vào việc cải thiện khả năng trả lời câu hỏi. Trong tương lai, việc mở rộng kiến trúc hệ thống để hỗ trợ nhiều mô hình ngôn ngữ lớn khác nhau, cũng như cung cấp các dịch vụ khác như tìm kiếm, gợi ý,... sẽ giúp hệ thống trở nên linh hoạt và đa dạng hơn.

Tài liệu tham khảo

- [1] A. Turing, “Computing machinery and intelligence,” *Mind*, vol. 59, no. October, pp. 433–60, 1950.
- [2] J. Weizenbaum, “Eliza a computer program for the study of natural language communication between man and machine,” *Commun. ACM*, vol. 9, pp. 36–45, Jan. 1966.
- [3] H.-Y. Shum, X. He, and D. Li, “From eliza to xiaoice: Challenges and opportunities with social chatbots,” 2018.
- [4] T. Zemčík, “A brief history of chatbots,” *DEStech Transactions on Computer Science and Engineering*, 2019.
- [5] J. Cahn, *CHATBOT: Architecture, Design, & Development*. Ph.d. thesis, University of Pennsylvania, School of Engineering and Applied Science, 2017.
- [6] L. Bradeško and D. Mladenec, “A survey of chatbot systems through a loebner prize competition,” in *Proceedings of Slovenian Language Technologies Society Eighth Conference of Language Technologies*, (Ljubljana, Slovenia), pp. 34–37, Institut Jožef Stefan, 2012.
- [7] B. Wilcox, “Winning the loebner’s,” 2014. Accessed on: 26/11/2020.
- [8] B. AbuShawar and E. Atwell, “Alice chatbot: Trials and outputs,” *Comput. Y Sist.*, vol. 19, p. 2326, 2015.
- [9] L. Dormehl, “Microsoft’s friendly xiaoice a.i can figure out what you want—before you ask,” 2018. Accessed on: 05/05/2021.
- [10] G. Spencer, “Much more than a chatbot: China’s xiaoice mixes ai with emotions and wins over millions of fans.” Microsoft Asia News Center, 2018. Accessed on: 29/06/2021.

- [11] L. Zhou, J. Gao, D. Li, and H. Shum, “The design and implementation of xiaoice, an empathetic social chatbot,” *arXiv*, 2019.
- [12] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, and et al., “Improving language understanding by generative pre-training,” *arXiv preprint*, 2018. arXiv:1801.06146.
- [13] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, “Retrieval-augmented generation for large language models: A survey,” *arXiv preprint arXiv:2312.10997*, 2023.
- [14] X. Li, J. Jin, Y. Zhou, Y. Zhang, P. Zhang, Y. Zhu, and Z. Dou, “From matching to generation: A survey on generative information retrieval,” *arXiv preprint arXiv:2404.14851*, 2024.
- [15] V. Karpukhin, B. O’guz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. t. Yih, “Dense passage retrieval for open-domain question answering,” *arXiv preprint arXiv:2004.04906*, 2020.
- [16] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, “Learning to rank using gradient descent,” in *Proceedings of the 22nd international conference on Machine learning*, pp. 89–96, 2005.
- [17] S. Robertson and H. Zaragoza, “The probabilistic relevance framework: Bm25 and beyond,” *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.
- [18] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, and et al., “A survey of large language models,” *arXiv preprint arXiv:2303.18223*, 2023.
- [19] R. Rosenfeld, “Two decades of statistical language modeling: Where do we go from here?,” *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1270–1278, 2000.
- [20] Y. Bengio, R. Ducharme, and P. Vincent, “A neural probabilistic language model,” in *Advances in neural information processing systems 13*, 2000.
- [21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.

- [22] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.
- [23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems 30*, 2017.
- [24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [25] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, and et al., “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [26] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, and et al., “Palm: Scaling language modeling with pathways,” *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.
- [27] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” *arXiv preprint arXiv:1910.13461*, 2019.
- [28] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.
- [29] G. Salton, E. A. Fox, and H. Wu, “Extended boolean information retrieval,” *Communications of the ACM*, vol. 26, no. 11, pp. 1022–1036, 1983.
- [30] H. P. Luhn, “A statistical approach to mechanized encoding and searching of literary information,” *IBM Journal of Research and Development*, vol. 1, no. 4, pp. 309–317, 1957.
- [31] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, “Learning to rank: from pairwise approach to listwise approach,” in *Proceedings of the 24th international conference on Machine learning*, pp. 129–136, 2007.

- [32] R. Nogueira and K. Cho, “Passage re-ranking with bert,” *arXiv preprint arXiv:1901.04085*, 2019.
- [33] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, “Realm: Retrieval augmented language model pre-training,” in *International conference on machine learning*, pp. 3929–3938, 2020.
- [34] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. t. Yih, T. Rocktäschel, and et al., “Retrieval-augmented generation for knowledge-intensive nlp tasks,” in *Advances in Neural Information Processing Systems 33*, pp. 9459–9474, 2020.
- [35] G. Izacard and E. Grave, “Leveraging passage retrieval with generative models for open domain question answering,” *arXiv preprint arXiv:2007.01282*, 2020.
- [36] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. V. D. Driessche, J.-B. Lespiau, B. Damoc, A. Clark, and et al., “Improving language models by retrieving from trillions of tokens,” in *International conference on machine learning*, pp. 2206–224, 2022.
- [37] B. Oguz, X. Chen, V. Karpukhin, S. Peshterliev, D. Okhonko, M. Schlichtkrull, S. Gupta, Y. Mehdad, and S. Yih, “Unik-qa: Unified representations of structured and unstructured knowledge for open-domain question answering,” *arXiv preprint arXiv:2012.14610*, 2020.
- [38] Y. Li, B. Peng, Y. Shen, Y. Mao, L. Liden, Z. Yu, and J. Gao, “Knowledge-grounded dialogue generation with a unified knowledge representation,” *arXiv preprint arXiv:2112.07924*, 2021.
- [39] S. Weijia, M. Sewon, Y. Michihiro, S. Minjoon, J. Rich, L. Mike, and et al., “Replug: Retrieval-augmented black-box language models,” *arXiv preprint arXiv:2301.12652*, 2023.
- [40] J. Berant, A. Chou, R. Frostig, and P. Liang, “Semantic parsing on freebase from question-answer pairs,” in *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1533–1544, 2013.
- [41] Z. Wang, P. Ng, R. Nallapati, and B. Xiang, “Retrieval, re-ranking and multi-task learning for knowledge-base question answering,” in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 347–357, 2021.

- [42] J. Baek, A. F. Aji, J. Lehmann, and S. J. Hwang, “Direct fact retrieval from knowledge graphs without entity linking,” *arXiv preprint arXiv:2305.12416*, 2023.
- [43] Y. Wang, N. Lipka, R. A. Rossi, A. Siu, R. Zhang, and T. Derr, “Knowledge graph prompting for multi-document question answering,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 19206–19214, 2024.
- [44] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson, “From local to global: A graph rag approach to query-focused summarization,” *Microsoft Research*, 2024.
- [45] B. Sarmah, B. Hall, R. Rao, S. Patel, S. Pasquali, and D. Mehta, “Hybridrag: Integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction,” *arXiv preprint arXiv:2408.04948*, 2024.
- [46] S. Ma, C. Xu, X. Jiang, M. Li, H. Qu, C. Yang, J. Mao, and J. Guo, “Think-on-graph 2.0: Deep and faithful large language model reasoning with knowledge-guided retrieval augmented generation,” *IDEA Research, International Digital Economy Academy*, 2024.
- [47] J. Sun, C. Xu, L. Tang, S. Wang, C. Lin, Y. Gong, L. M. Ni, H.-Y. Shum, and J. Guo, “Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph,” *arXiv preprint arXiv:2407.10805*, 2024.
- [48] R. Rao, B. Hall, S. Pasquali, S. Patel, B. Sarmah, and D. Mehta, “Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation,” *arXiv preprint arXiv:2402.03216*, 2024.
- [49] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, “Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers,” 2020.
- [50] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “SQuAD: 100,000+ questions for machine comprehension of text,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (J. Su, K. Duh, and X. Carreras, eds.), (Austin, Texas), pp. 2383–2392, Association for Computational Linguistics, Nov. 2016.

- [51] M. McHugh, “Interrater reliability: The kappa statistic,” *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, vol. 22, pp. 276–82, 10 2012.

Phụ lục A

PSEUDOCODE

Algorithm 1: Phương pháp RAG tích hợp truy xuất, suy luận đồ thị tri thức với truy xuất cơ sở dữ liệu vector

Input: Question q

Output: Answer Ans

1 **Initialization**

2 $Clues \leftarrow \text{Null};$

3 Topic entities $\mathcal{E}_{\text{topic}}^0 \leftarrow$ Using LLM to extract topic entities;

4 Reasoning paths $\mathcal{P} \leftarrow$ Start with each entity in $\mathcal{E}_{\text{topic}}^0$;

5 $Ctx^0 \leftarrow$ Retrieve top- k documents related to $\mathcal{E}_{\text{topic}}^0$ using DRM;

6 Prompt LLM evaluates the sufficiency of knowledge to answer q :

$PROMPT_{rs}(q, Ctx^0, Clues^0);$

7 **if** *Information is sufficient* **then**

8 Output Ans ;

9 **return**;

10 **else**

11 Output $Clues^0$

 (Using LLM to rewrite q for retrieving potential evidence);

Algorithm 2: Phương pháp RAG tích hợp truy xuất, suy luận đồ thị tri thức với truy xuất cơ sở dữ liệu vector (tiếp tục)

```

1 for  $i = 1$  to  $I$  do
2   Knowledge-guided Graph Search
3   foreach  $p_n \in \mathcal{P}$  with  $e_n^{i-1}$  is the tail entity of  $p_n$  do
4     Find relations  $\text{Edge}(e_n^{i-1}) = \{(r_{n,m}^i, h_m)\}$ ;
5     Use LLM to select and score relations then add to corresponding  $p$ 
         $\mathcal{P} \leftarrow \text{PROMPT}_{RP\_all}(q, \{p_n, \text{Edge}(e_n^{i-1})\}_{n=1}^N)$ ;
6   foreach  $(r_{n,m}^i \in R^i)$  do
7     Find connected entities  $c_{n,m}^i = \text{Tail}(e_n^i, (r_{n,m}^i, h_m))$ ;
8     Get all related chunks  $\text{Chunk}_{\text{related}}^I = \{\text{LinkedDoc}(c_{n,m}^i)\}_{n=1}^N$ ;
9     foreach  $c_{n,m}^i \in \text{Chunk}_{\text{related}}^I$  do
10      Compute relevance score
11       $\text{Ctx}^i \leftarrow$  Select top- $k$  chunks with highest relevance score;
12       $\mathcal{P} \leftarrow$  Select top- $N$  reasoning paths  $\text{PROMPT}_{EP}(\text{Ctx}, q, p_n, \mathcal{C}_n^{I-1})$ ;
13   Reasoning Knowledge
14   Prompt LLM evaluates sufficiency of knowledge to answer  $q$ :
         $\text{PROMPT}_{RS}(q, \mathcal{P}, \text{Clue}^{I-1}, \text{Ctx})$  if Information is sufficient
        then
15     Output  $\text{Ans}$ ;
16     return;
17   else
18     Output  $\text{Clues}^0$ 
        (Using LLM to rewrite  $q$  for retrieving potential evidence);

```

Phụ lục B

PROMPT

Bảng B.1: Prompt cho tác vụ nhận diện thực thể

Prompt Nhận diện thực thể

Bạn là một người phân tích, cải thiện danh sách thuật ngữ, người phân tích các thuật ngữ từ đoạn văn bản được cung cấp, sử dụng lý thuyết phạm trù (category theory). Bạn sẽ được cung cấp một đoạn VĂN BẢN (được phân cách bởi ```) và danh sách các thuật ngữ (được phân cách bởi ```). Nhiệm vụ của bạn là phân tích và xem xét các thuật ngữ một cách khắt khe trong ngữ cảnh VĂN BẢN đó thì có phải thuật ngữ về 'concept.name' dựa theo mô tả sau về thuật ngữ: concept.description để cải thiện danh sách thuật ngữ trở nên tổng quát.

Nhiệm vụ:

Suy nghĩ 1: Khi duyệt qua từng khái niệm, những khái niệm cùng chỉ về một ý nghĩa thì hãy gộp chúng lại thành một khái niệm suy nhất. Ví dụ:

Suy nghĩ 2: Khi duyệt qua từng câu, hãy suy nghĩ về ngữ cảnh trong câu để xác định các thuật ngữ được cung cấp thuộc phạm trù.

Suy nghĩ 3: Thuật ngữ nên chi tiết nhất có thể. Các thuật ngữ phải càng nguyên tử càng tốt.

Bảng B.2: Prompt cho tác vụ cắt giảm các quan hệ trong quá trình khám phá tri thức

Prompt Cắt giảm các quan hệ trong quá trình khám phá tri thức
<p>Nhiệm vụ:</p> <ol style="list-style-type: none">1. Carefully examine the question provided.2. From the list of relationships, select the one that you think is most relevant to the question to help answer it.3. For each relationship selected, provide a score from 0 to 10 reflecting its usefulness in answering the question, with 10 being the most useful.4. Do not add or remove relationships from the list provided.5. Always use <code>function_call</code> provided. <p>Đầu vào theo định dạng dưới đây:</p> <p>Câu hỏi: [Nội dung câu hỏi]</p> <p>Thực thể chính: [Tên của thực thể]</p> <p>Danh sách mối quan hệ: [Danh sách các quan hệ của thực thể để lựa chọn.]</p>

Bảng B.3: Prompt cho tác vụ cắt giảm tất cả các quan hệ cùng 1 lúc trong quá trình khám phá tri thức

Prompt Cắt giảm tất cả các quan hệ cùng 1 lúc trong quá trình khám phá tri thức
<p>Nhiệm vụ:</p> <ol style="list-style-type: none">1. Carefully examine the question provided.2. From the list of relationships, select the one that you think is most relevant to the question to help answer it.3. For each relationship selected, provide a score from 0 to 10 reflecting its usefulness in answering the question, with 10 being the most useful.4. Do not add or remove relationships from the list provided.5. Always use <code>function_call</code> provided. <p>Đầu vào theo định dạng dưới đây:</p> <p>Câu hỏi: [Nội dung câu hỏi]</p> <p>Danh sách thực thể và mối quan hệ [Tên của thực thể]</p> <p>Tên thực thể: [Danh sách các quan hệ của thực thể để lựa chọn.]</p> <p>Danh sách mối quan hệ tương ứng: [Danh sách các quan hệ của thực thể]</p> <p>...</p>

Bảng B.4: Prompt cho tác vụ cắt giảm thực thể trong quá trình khám phá

Prompt Cắt giảm thực thể trong quá trình khám phá
<p>Bạn được cung cấp câu hỏi và danh sách các đường dẫn suy luận chứa các BỘ 3 (triples) được truy xuất từ đồ thị tri thức, vui lòng chấm điểm mức độ liên quan của các BỘ 3 (triples) cho việc trả lời câu hỏi theo thang điểm 100.</p> <p>Thực hiện suy luận theo hướng dẫn:</p> <p>Bước 1: Đọc kỹ câu hỏi và xem xét cách trả lời câu hỏi.</p> <p>Bước 2: Phân tích mức độ liên quan giữa BỘ 3 và câu hỏi.</p> <p>Bước 3: Chấm điểm cho mức độ đóng góp thông tin để trả lời câu hỏi dựa trên BỘ 3 đó. Đánh giá điểm một cách công tâm và chính xác.</p> <p>Đầu vào theo định dạng dưới đây:</p> <p>Câu hỏi: [Nội dung câu hỏi]</p> <p>Danh sách các đường dẫn suy luận:</p> <p>Đường dẫn 1: Nội dung đường dẫn suy luận với các BỘ 3</p> <p>Đường dẫn 2: Nội dung đường dẫn suy luận với các BỘ 3</p> <p>...</p>

Bảng B.5: Prompt cho tác vụ suy luận câu trả lời

Prompt Suy luận câu trả lời
<p>Nhiệm vụ:</p> <ol style="list-style-type: none"> Đánh giá xem dựa vào các thông tin có trong đoạn văn bản có thể trả lời được câu hỏi hay không Câu trả lời của bạn phải bắt đầu bằng 'Có' hoặc 'Không' và không giải thích gì thêm. Nếu {Có}, lưu ý rằng thực thể câu trả lời đã phân tích phải được đặt trong dấu ngoặc nhọn {xxxxxxx}. Nếu {Không}, điều đó có nghĩa là các tài nguyên này vô dụng hoặc chỉ cung cấp các manh mối có ích nhưng không đủ để trả lời câu hỏi một cách chắc chắn, hãy xác định các khía cạnh còn thiếu và tinh chỉnh truy vấn tìm kiếm để nhắm cụ thể vào thông tin cần thiết để hoàn thiện câu trả lời. Truy vấn tìm kiếm được nhắm mục tiêu cũng phải được đặt trong dấu ngoặc nhọn {xxxxxxx}. <p>Đầu vào theo định dạng dưới đây:</p> <p>Câu hỏi: [Nội dung câu hỏi]</p> <p>Manh mối: [Nội dung manh mối để tìm kiếm thông tin]</p> <p>Các đoạn văn bản:</p> <ol style="list-style-type: none"> Nội dung đoạn văn bản 1 Nội dung đoạn văn bản 2 ...Tiếp tục theo cùng cách cho các đoạn văn bản <p>Ví dụ:</p> <p>Câu hỏi: Nam tước Yamaji Motoharu là một tướng lĩnh trong Quân đội Đế quốc Nhật Bản đầu thế kỷ, thuộc về Đế quốc nào?</p> <ol style="list-style-type: none"> Nội dung đoạn văn bản 1 Nội dung đoạn văn bản 2 <p>Trả lời: Có. Dựa trên manh mối, các đoạn văn ngữ cảnh thu thập được và kiến thức của tôi, Nam tước Yamaji Motoharu, người là một tướng lĩnh trong Quân đội Đế quốc Nhật Bản đầu thế kỷ, thuộc về Đế quốc Nhật Bản. Do đó, câu trả lời cho câu hỏi là {Đế quốc Nhật Bản}.</p>

Bảng B.6: Prompt cho tác vụ dự đoán các manh mối

Prompt để dự đoán các manh mối
<p>Dựa trên một câu hỏi và một số kiến thức đã thu thập được cho đến nay, hãy dự đoán bằng chứng bổ sung cần được tìm thấy để trả lời câu hỏi hiện tại, sau đó đưa ra một truy vấn phù hợp để tìm kiếm bằng chứng tiềm năng này. Lưu ý rằng truy vấn phải được đặt trong dấu ngoặc nhọn xxx.</p>

Phụ lục C

Đề thi THPT Lịch sử

Bảng C.1: Câu trả lời của phương pháp đề xuất cho đề thi THPT Lịch sử 2018, mã đề 302. Với các câu in đậm là câu trả lời sai

Câu	1	2	3	4	5	6	7	8	9	10
Đáp án	B	B	C	C	D	C	B	A	D	D
Câu	11	12	13	14	15	16	17	18	19	20
Đáp án	B	B	A	C	C	C	A	B	B	B
Câu	21	22	23	24	25	26	27	28	29	30
Đáp án	A	D	D	D	B	B	A	C	A	C
Câu	31	32	33	34	35	36	37	38	39	40
Đáp án	D	C	D	A	C	A	B	C	D	C

BỘ GIÁO DỤC VÀ ĐÀO TẠO**KỲ THI TRUNG HỌC PHỔ THÔNG QUỐC GIA NĂM 2018****ĐỀ THI CHÍNH THỨC****Bài thi: KHOA HỌC XÃ HỘI***(Đề thi có 04 trang)***Môn thi thành phần: LỊCH SỬ***Thời gian làm bài: 50 phút, không kể thời gian phát đề***Họ, tên thí sinh:****Mã đề thi 302****Số báo danh:****Câu 1:** Một trong những hệ quả tích cực của toàn cầu hóa là

- A. giải quyết triệt để những bất công xã hội. B. thúc đẩy sự phát triển của lực lượng sản xuất.
C. giải quyết căn bản sự phân hóa giàu nghèo. D. giúp các nước giữ nguyên cơ cấu kinh tế.

Câu 2: Giai cấp công nhân Việt Nam có mối quan hệ gắn bó mật thiết nhất với lực lượng xã hội nào?

- A. Thợ thủ công. B. Nông dân. C. Tiểu thương. D. Tiểu tư sản.

Câu 3: Năm 1953, thực dân Pháp gặp khó khăn nào trong cuộc chiến tranh xâm lược ở Việt Nam?

- A. Bước đầu đề mất quyền chủ động. B. Mỹ cắt giảm nguồn viện trợ.
C. Vùng chiếm đóng bị thu hẹp. D. Bị Mỹ ép kết thúc chiến tranh.

Câu 4: Nhiệm vụ cơ bản của cách mạng miền Nam được Hội nghị lần thứ 21 Ban Chấp hành Trung ương Đảng Lao động Việt Nam (tháng 7 - 1973) xác định là gì?

- A. Giải phóng miền Nam trong năm 1975.
B. Chỉ đấu tranh chính trị để thống nhất đất nước.
C. Tiếp tục cuộc cách mạng dân tộc dân chủ nhân dân.
D. Tiến hành cuộc cách mạng ruộng đất.

Câu 5: “Đánh đuổi giặc Pháp, khôi phục nước Việt Nam, thành lập nước Cộng hòa Dân quốc Việt Nam” là

- A. mục đích của phong trào Đông du. B. mục đích của phong trào Duy tân.
C. chủ trương của Hội Duy tân. D. chủ trương của Việt Nam Quang phục hội.

Câu 6: Năm 1949, Liên Xô đạt được thành tựu khoa học - kỹ thuật nào dưới đây?

- A. Phóng tàu vũ trụ Phương Đông. B. Phóng thành công vệ tinh nhân tạo.
C. Chế tạo thành công bom nguyên tử. D. Đưa con người lên thám hiểm Mặt Trăng.

Câu 7: Từ năm 1952 đến năm 1960, tình hình kinh tế Nhật Bản như thế nào?

- A. Phát triển xen lẫn suy thoái. B. Có bước phát triển nhanh.
C. Bước đầu suy thoái. D. Cơ bản được phục hồi.

Câu 8: Cuộc khởi nghĩa nào dưới đây thuộc phong trào Cần vương ở Việt Nam cuối thế kỷ XIX?

- A. Hương Khê. B. Yên Thế. C. Yên Bái. D. Thái Nguyên.

Câu 9: Tháng 3 - 1921, Đảng Bôn-sê-vích quyết định thực hiện Chính sách kinh tế mới (NEP) ở nước Nga trong bối cảnh

- A. đã hoàn thành nhiệm vụ công nghiệp hóa. B. quan hệ sản xuất phong kiến vẫn thống trị.
C. đã hoàn thành tập thể hóa nông nghiệp. D. nền kinh tế quốc dân bị tàn phá nghiêm trọng.

Câu 10: Ngày 8 - 9 - 1945, Chủ tịch Hồ Chí Minh ký Sắc lệnh thành lập Nha Bình dân học vụ để thực hiện nhiệm vụ gì?

- A. Khai giảng các bậc học. B. Cải cách giáo dục.
C. Bổ túc văn hóa. D. Chống giặc đói.

Câu 11: Nhiệm vụ cấp thiết của cách mạng Việt Nam sau Đại thắng mùa Xuân 1975 là gì?

- A. Thực hiện ngay công cuộc đổi mới đất nước.
B. Hoàn thành thống nhất đất nước về mặt nhà nước.
C. Hoàn thành thống nhất đất nước về mặt lãnh thổ.
D. Hoàn thành cải cách ruộng đất và triệt để giảm tô.

Câu 12: Căn cứ địa cách mạng đầu tiên của Việt Nam trong giai đoạn 1939 - 1945 là

- A. Bắc Kạn. B. Bắc Sơn - Võ Nhai.
C. Tân Trào - Tuyên Quang. D. Thái Nguyên.

Trang 1/4 - Mã đề thi 302

Hình C.1: Trang 1 đề thi THPT Lịch sử 2021

Câu 13: Thời cơ “ngàn năm có một” để nhân dân Việt Nam tổng khởi nghĩa giành chính quyền năm 1945 kết thúc khi

- A. quân Đồng minh vào Đông Dương giải giáp quân đội Nhật.
- B. thực dân Pháp bắt đầu nổ súng xâm lược trở lại Việt Nam.
- C. Nhật cùng thực dân Anh chống phá chính quyền cách mạng.
- D. Nhật giao Đông Dương cho quân Trung Hoa Dân quốc.

Câu 14: Chính phủ nước Việt Nam Dân chủ Cộng hòa ký với Chính phủ Pháp bản Hiệp định Sơ bộ (6 - 3 - 1946) nhằm

- A. buộc Pháp phải thừa nhận Việt Nam là một quốc gia độc lập.
- B. tạo điều kiện để quân Đồng minh vào giải giáp quân đội Nhật.
- C. tránh việc cùng lúc phải chống lại nhiều thế lực ngoại xâm.
- D. buộc Pháp phải thừa nhận Việt Nam là một quốc gia dân chủ.

Câu 15: Một trong những nguyên nhân thất bại của phong trào yêu nước theo khuynh hướng dân chủ tư sản ở Việt Nam những năm đầu thế kỷ XX là

- A. phong trào còn mang nặng tính tự phát.
- B. chưa sử dụng nhiều phương pháp đấu tranh.
- C. thiếu sự lãnh đạo của một giai cấp tiên tiến.
- D. không gắn cứu nước với canh tân đất nước.

Câu 16: Biểu hiện nào dưới đây **không** phản ánh đúng xu thế phát triển của thế giới từ khi Chiến tranh lạnh chấm dứt đến năm 2000?

- A. Hòa bình, hợp tác và phát triển là xu thế chủ đạo của thế giới.
- B. Các quốc gia đều tập trung lấy phát triển kinh tế làm trọng điểm.
- C. Trật tự thế giới đơn cực được xác lập trong quan hệ quốc tế.
- D. Các nước lớn điều chỉnh quan hệ theo hướng đối thoại, thỏa hiệp.

Câu 17: Nguyên nhân sâu xa dẫn đến cuộc Chiến tranh thế giới thứ hai (1939 - 1945) là gì?

- A. Quy luật phát triển không đồng đều của chủ nghĩa tư bản.
- B. Các thế lực phát xít lên cầm quyền ở một số nước Âu - Mỹ.
- C. Mâu thuẫn gay gắt giữa phe Đồng minh với phe phát xít.
- D. Chính sách theo “chủ nghĩa biệt lập” của Mỹ và Anh.

Câu 18: Nội dung nào **không** phản ánh đúng âm mưu của Mỹ ở Đông Dương trong những năm 1951 - 1954?

- A. Ngăn chặn ảnh hưởng của chủ nghĩa xã hội.
- B. Đưa quân đội tham chiến trực tiếp.
- C. Từng bước thay chân quân Pháp.
- D. Quốc tế hóa cuộc chiến tranh Đông Dương.

Câu 19: Xu hướng bạo động và xu hướng cải cách trong phong trào yêu nước và cách mạng ở Việt Nam những năm đầu thế kỷ XX đều

- A. không bị động trông chờ vào sự giúp đỡ từ bên ngoài.
- B. xuất phát từ truyền thống yêu nước của dân tộc.
- C. có sự kết hợp nhiệm vụ chống đế quốc và phong kiến.
- D. do giai cấp tư sản khởi xướng và lãnh đạo.

Câu 20: Trong chiến dịch Hồ Chí Minh (1975), địa bàn tác chiến của các lực lượng vũ trang nhân dân Việt Nam là

- A. rừng núi.
- B. đô thị.
- C. nông thôn.
- D. trung du.

Câu 21: Nét nổi bật về nghệ thuật chỉ đạo chiến tranh cách mạng của Đảng Lao động Việt Nam trong cuộc Tổng tiến công và nổi dậy Xuân 1975 là

- A. lựa chọn đúng địa bàn và chủ động tạo thời cơ tiến công.
- B. đi từ khởi nghĩa từng phần tiến lên chiến tranh cách mạng.
- C. kết hợp tiến công và khởi nghĩa của lực lượng vũ trang.
- D. kết hợp đánh thẳng nhanh và đánh chắc, tiến chắc.

Câu 22: Trong quá trình thực hiện chiến lược toàn cầu từ sau Chiến tranh thế giới thứ hai đến năm 2000, Mỹ đạt được kết quả nào dưới đây?

- A. Duy trì vị trí cường quốc số một thế giới trên tất cả các lĩnh vực.
- B. Trực tiếp xóa bỏ hoàn toàn chế độ phân biệt chủng tộc.
- C. Duy trì sự tồn tại và hoạt động của tất cả các tổ chức quân sự.
- D. Làm chậm quá trình giành thắng lợi của phong trào giải phóng dân tộc.

Trang 2/4 - Mã đề thi 302

Hình C.2: Trang 2 đề thi THPT Lịch sử 2021

Câu 23: Hiệp ước Patonôt (1884) được ký kết giữa triều đình nhà Nguyễn với thực dân Pháp là mốc đánh dấu

- A. các vua nhà Nguyễn hoàn toàn đầu hàng thực dân Pháp.
- B. thực dân Pháp căn bản hoàn thành công cuộc bình định Việt Nam.
- C. thực dân Pháp thiết lập xong bộ máy cai trị ở Việt Nam.
- D. thực dân Pháp căn bản hoàn thành công cuộc xâm lược Việt Nam.

Câu 24: Nội dung nào dưới đây là điểm khác nhau giữa Luận cương chính trị tháng 10 - 1930 của Đảng Cộng sản Đông Dương với Cương lĩnh chính trị đầu tiên của Đảng Cộng sản Việt Nam?

- A. Chủ trương tập hợp lực lượng cách mạng.
- B. Đề ra phương hướng chiến lược.
- C. Xác định phương pháp đấu tranh.
- D. Xác định giai cấp lãnh đạo.

Câu 25: Nội dung đổi mới về kinh tế ở Việt Nam (từ tháng 12 - 1986) và Chính sách kinh tế mới (NEP, 1921) ở nước Nga có điểm tương đồng là

- A. ưu tiên phát triển công nghiệp nặng và giao thông vận tải.
- B. xây dựng nền kinh tế nhiều thành phần có sự quản lý của nhà nước.
- C. xây dựng kinh tế thị trường xã hội chủ nghĩa, kiểm soát bằng pháp luật.
- D. thay chế độ trung thu lương thực thừa bằng thuế lương thực.

Câu 26: Trong bối cảnh Hiệp ước Hoa - Pháp được ký kết (tháng 2 - 1946), Chính phủ nước Việt Nam Dân chủ Cộng hòa thực hiện chủ trương gì?

- A. Hòa với thực dân Pháp để đánh quân Trung Hoa Dân quốc.
- B. Hòa hoãn và nhân nhượng có nguyên tắc với thực dân Pháp.
- C. Phát động toàn quốc kháng chiến chống thực dân Pháp.
- D. Chiến đấu chống thực dân Pháp và quân Trung Hoa Dân quốc.

Câu 27: Những chuyển biến về kinh tế, xã hội ở Việt Nam trong cuộc khai thác thuộc địa lần thứ nhất (1897 - 1914) của thực dân Pháp đã

- A. tạo điều kiện cho sự hình thành khuynh hướng cứu nước mới.
- B. thúc đẩy phong trào công nhân bước đầu chuyển sang tự giác.
- C. làm cho tầng lớp tư sản Việt Nam trở thành một giai cấp.
- D. giúp các sĩ phu phong kiến chuyển hẳn sang lập trường tư sản.

Câu 28: Một trong những tác động của phong trào giải phóng dân tộc đối với quan hệ quốc tế sau Chiến tranh thế giới thứ hai là

- A. làm xuất hiện xu thế hòa hoãn Đông - Tây ở châu Âu.
- B. buộc Mỹ phải chấm dứt Chiến tranh lạnh với Liên Xô.
- C. góp phần làm đảo lộn chiến lược toàn cầu của Mỹ.
- D. tạo cơ sở hình thành các liên minh kinh tế - quân sự.

Câu 29: Trong giai đoạn sau Chiến tranh lạnh, để xây dựng sức mạnh thực sự, các quốc gia trên thế giới đều tập trung vào

- A. phát triển kinh tế.
- B. hội nhập quốc tế.
- C. phát triển quốc phòng.
- D. ổn định chính trị.

Câu 30: Trong thời kỳ 1954 - 1975, đâu là một trong những nguyên nhân trực tiếp làm cho Việt Nam trở thành nơi diễn ra “sự kiện có tầm quan trọng quốc tế to lớn và có tính thời đại sâu sắc”?

- A. Phe xã hội chủ nghĩa ủng hộ Việt Nam chống Mỹ.
- B. Hiệp định Giơnevơ chia cắt Việt Nam thành hai miền.
- C. Việt Nam chịu sự chi phối của cục diện hai cực, hai phe.
- D. Mỹ chuyển trọng tâm chiến lược toàn cầu sang Việt Nam.

Câu 31: Sự kiện có tính đột phá làm xói mòn trật tự hai cực Ianta là

- A. thắng lợi của cuộc kháng chiến chống Pháp ở Việt Nam (1954).
- B. cách mạng Cuba lật đổ được chế độ độc tài Batista (1959).
- C. ba nước Ấn Độ, Việt Nam, Lào tuyên bố độc lập (1945).
- D. cách mạng dân tộc dân chủ Trung Quốc thành công (1949).

Câu 32: Một trong những điểm chung của Cách mạng tháng Tám năm 1945 và hai cuộc kháng chiến chống đế quốc xâm lược (1945 - 1975) ở Việt Nam là có sự kết hợp

- A. đấu tranh đồng thời trên ba mặt trận: quân sự, chính trị và ngoại giao.
- B. phong trào đấu tranh ở nông thôn với phong trào đấu tranh ở thành thị.
- C. tác chiến trên cả ba vùng rừng núi, nông thôn đồng bằng và đô thị.
- D. hoạt động của bộ đội chủ lực, bộ đội địa phương và dân quân du kích.

Câu 33: Căn cứ địa cách mạng là nơi cần có những điều kiện thuận lợi, trong đó “nhân hòa” là yếu tố quan trọng nhất. Đây là yếu tố “nhân hòa” để Nguyễn Ái Quốc chọn Cao Bằng làm nơi xây dựng căn cứ địa vào năm 1941?

- A. Mọi người đều tham gia Việt Minh.
- B. Có lực lượng du kích phát triển sớm.
- C. Sớm hình thành các Hội Cứu quốc.
- D. Có phong trào quần chúng tốt từ trước.

Câu 34: Trong hai cuộc kháng chiến chống đế quốc xâm lược (1945 - 1975), nhân dân Việt Nam đã giành thắng lợi bằng việc

- A. dùng sức mạnh của nhiều nhân tố thắng sức mạnh quân sự và kinh tế.
- B. dùng sức mạnh của vật chất đánh thắng sức mạnh của ý chí và tinh thần.
- C. lấy số lượng quân đông thắng vũ khí chất lượng cao.
- D. lấy lực thắng thế, lấy ít thắng nhiều về quân số.

Câu 35: Sự ra đời tổ chức Hiệp ước Bắc Đại Tây Dương (NATO, 1949) và tổ chức Hiệp ước Vácava (1955) là hệ quả trực tiếp của

- A. các cuộc chiến tranh cục bộ trên thế giới.
- B. xung đột vũ trang giữa Tây Âu và Đông Âu.
- C. cuộc Chiến tranh lạnh do Mỹ phát động.
- D. chiến lược Ngăn đe thực tế của Mỹ.

Câu 36: Tính chất triệt để của phong trào cách mạng Việt Nam 1930 - 1931 được biểu hiện ở chỗ

- A. diễn ra trên quy mô rộng lớn chưa từng thấy.
- B. hình thức đấu tranh phong phú và quyết liệt.
- C. lần đầu tiên có sự lãnh đạo của một chính đảng.
- D. không ảo tưởng vào kẻ thù của dân tộc và giai cấp.

Câu 37: Nội dung nào dưới đây phản ánh đúng giá trị của chiến thắng Đường số 14 - Phước Long (tháng 1 - 1975) đối với việc củng cố quyết tâm và hoàn thành kế hoạch giải phóng miền Nam của Đảng Lao động Việt Nam?

- A. Trận mở màn chiến lược.
- B. Trận trinh sát chiến lược.
- C. Trận nghi binh chiến lược.
- D. Trận tập kích chiến lược.

Câu 38: Cách mạng tháng Tám năm 1945 ở Việt Nam là một cuộc cách mạng

- A. không mang tính bạo lực.
- B. có tính dân chủ điển hình.
- C. không mang tính cải lương.
- D. chỉ mang tính chất dân tộc.

Câu 39: Trong cuộc Tiến công chiến lược Đông - Xuân 1953 - 1954, quân đội và nhân dân Việt Nam thực hiện kế sách gì để đối phó với kế hoạch Nava?

- A. Lừa địch để đánh địch.
- B. Đánh điểm, diệt viện.
- C. Đánh vận động và công kiên.
- D. Điều địch để đánh địch.

Câu 40: Một trong những yếu tố tác động đến sự hình thành trật tự thế giới giai đoạn sau Chiến tranh lạnh là

- A. tư bản tài chính xuất hiện và chi phối nền kinh tế thế giới.
- B. các trung tâm kinh tế - tài chính Tây Âu và Nhật Bản ra đời.
- C. sự phát triển của các cường quốc và Liên minh châu Âu (EU).
- D. sự xuất hiện và ngày càng mở rộng của các công ty độc quyền.

----- HẾT -----