VẬN DỤNG MÔ HÌNH CHATBOT TRONG XỬ LÝ NGÔN NGỮ VÀO HỆ THỐNG TRUY VẤN THÔNG MINH HỖ TRỢ TÌM KIẾM TÀI LIỆU THAM KHẢO

ThS Nguyễn Tâm Thanh Tùng, ThS Trần Đình Anh Huy, Huỳnh Phương Vi
Khoa Thư viện - Thông tin học, Trường Đại học KHXH&NV-ĐHQG Tp. Hồ Chí Minh

Tóm tắt: Trong bối cảnh dữ liệu ngày càng lớn, đa dạng và thay đổi liên tục, nhu cầu tìm kiếm thông tin chính xác, đáng tin cậy trở thành một yêu cầu cấp thiết. Tuy nhiên, các hệ thống tìm kiếm truyền thống chủ yếu dựa trên từ khóa và siêu dữ liệu thường gặp phải nhiều hạn chế như: không hiểu được ngữ cảnh của truy vấn, kết quả trả về thiếu chính xác và không có nguồn tham khảo rõ ràng. Điều này làm giảm độ tin cậy của thông tin và ảnh hưởng tiêu cực đến quá trình phân tích và ra quyết định của người dùng. Để khắc phục, hệ thống DataChatBot được xây dựng dựa trên nền tảng Mô hình ngôn ngữ lớn kết hợp với các kỹ thuật hiện đại như truy hồi thông tin từ nguồn ngoài và sử dụng cơ sở dữ liệu vector. Kết quả là hệ thống cho phép lưu trữ và tìm kiếm thông tin dưới dạng vector nhúng, hỗ trợ truy vấn ngữ nghĩa, suy luận logic và đưa ra câu trả lời chính xác kèm theo trích dẫn nguồn. Điều này giúp người dùng kiểm chứng dễ dàng, tiếp cận thông tin phù hợp và nâng cao hiệu quả trong công tác tìm kiếm, phân tích và ra quyết định.

Từ khóa: Truy vấn tài liệu tham khảo; mô hình ngôn ngữ lớn; xử lý ngôn ngữ tự nhiên; truy vấn ngữ nghĩa; cơ sở dữ liệu vector; mô hình xếp hạng lại.

APPLYING CHATBOT MODELS IN NATURAL LANGUAGE PROCESSING TO AN INTELLIGENT QUERY SYSTEM FOR REFERENCE DOCUMENT RETRIEVAL

Abstract: In the era of increasingly vast, diverse, and rapidly evolving data, the demand for accurate and reliable information retrieval has become more critical than ever. Traditional search systems, which primarily rely on keywords and metadata, often fail to understand contextual nuances, leading to imprecise results and a lack of credible references. These limitations reduce the reliability of retrieved information and hinder effective analysis and decision-making. To overcome these challenges, the DataChatBot system was developed, leveraging Large language models (LLMs) in combination with advanced techniques such as external information retrieval and vector databases. As a result, the system enables information to be stored and searched as embedded vectors, enhancing semantic search capabilities, supporting logical reasoning, and delivering accurate, source-cited responses. This approach not only improves the relevance and trustworthiness of search results but also empowers users to verify information with ease, thereby optimizing the processes of information discovery, analysis, and decision-making.

Keywords: Reference retrieval; large language model; natural language processing; semantic query; vector database; re-ranking model.

1. CÁCH VẬN HÀNH CỦA HỆ THỐNG TÌM KIẾM TRUYỀN THỐNG

Hệ thống tìm kiếm truyền thống hoạt động gồm nhiều bước để thu thập, lưu trữ và xử lý dữ liệu. Các công cụ tìm kiếm sử dụng bot để quét các trang web, thu thập dữ liệu, lập chỉ mục thông tin theo từ khóa, tiêu đề và các siêu dữ liệu khác. Khi người dùng truy vấn, hệ thống sẽ xử lý yêu cầu và

trả về kết quả dựa trên độ phù hợp và các yếu tố khác như độ tin cậy của nguồn thông tin theo các bước như hình sau:

Hình 1. *Bước vận hành hệ thống tìm tin truyền thống*

Bước 1- Thu thập dữ liệu: Các công cụ tìm kiếm sử dụng bot hoặc "crawler" để quét và thu thập dữ liệu từ các trang web và tài nguyên số hóa. Ví dụ: Googlebot quét các trang web tin tức như: vnexpress.net, thesaigontimes.vn,… Khi vnexpress.net xuất bản một bài viết với tiêu đề *"Xuất khẩu gạo Việt Nam tăng trưởng mạnh năm 2024"*, Googlebot sẽ quét và thu thập nội dung của bài viết này.

Bước 2- Lập chỉ mục dữ liệu: Thông tin thu thập được sắp xếp và lưu trữ trong cơ sở dữ liệu chỉ mục dùng để chứa thông tin như từ khóa, tiêu đề, tác giả và mô tả. Ví dụ: Thông tin từ bài viết *"Xuất khẩu gạo Việt Nam tăng trưởng mạnh năm 2024"* được lưu trữ trong chỉ mục của Google, trong đó các siêu dữ liệu như tiêu đề, từ khóa, tác giả và nguồn được ghi lại để hỗ trợ quá trình tìm kiếm nhanh chóng và hiệu quả. Bài viết do tác giả Nguyễn An thực hiện, được đăng tải trên trang vnexpress.net, tập trung vào các chủ đề liên quan đến xuất khẩu gạo, tình hình kinh tế - nông nghiệp Việt Nam trong năm 2024, với các từ khóa chính như "xuất khẩu gạo", "Việt Nam", "2024", "nông nghiệp" và "kinh tế".

Bước 3- Xử lý truy vấn: Khi người dùng nhập từ khóa, hệ thống sẽ tìm kiếm các mục trong chỉ mục phù hợp với từ khóa đó. Ví dụ: Người truy vấn nhập từ khóa: *"xuất khẩu gạo Việt Nam 2024"* vào ô tìm kiếm của Google và tìm kiếm trong chỉ mục để tìm các tài liệu chứa từ khóa liên quan như: "xuất khẩu gạo", "Việt Nam", "2024".

Bước 4- Xếp hạng kết quả: Kết quả xếp hạng dựa trên mức độ phù hợp, tần suất của từ khóa và các yếu tố như độ phổ biến của trang. Ví dụ: Google xếp hạng kết quả tìm kiếm dựa trên nhiều yếu tố, trong đó bao gồm mức độ phù hợp giữa nội dung trang và từ khóa mà người dùng nhập vào, độ tin cậy của nguồn thông tin (chẳng hạn như trang vnexpress.net thường được đánh giá cao về độ tin cậy) và thời gian cập nhật của nội dung, bởi các bài viết mới thường có khả năng được xếp hạng cao hơn.

Các hệ thống tìm kiếm truyền thống gặp phải nhiều khó khăn và bất cập đáng kể, một trong những vấn đề lớn là sự hiểu sai mục đích tìm kiếm. Các công cụ tìm kiếm truyền thống không thể hiểu được ngữ cảnh của người dùng, dẫn đến việc cung cấp kết quả không phù hợp với yêu cầu tìm kiếm thực tế. Quá tải thông tin là một vấn đề nghiêm trọng khác, người dùng phải lọc rất nhiều

kết quả không liên quan hoặc không hữu ích, dẫn đến mất thời gian và công sức trong tìm kiếm thông tin đúng. Thêm vào đó, các hệ thống tìm kiếm truyền thống vẫn gặp phải hạn chế trong xử lý dữ liệu phi văn bản như hình ảnh, video, khiến cho những tài liệu dạng này khó được truy xuất hiệu quả. Ngoài ra, các hệ thống tìm kiếm truyền thống đang đối mặt với nhiều hạn chế như hiểu sai mục đích tìm kiếm do thiếu khả năng nắm bắt ngữ cảnh người dùng và phụ thuộc vào từ khóa chính xác gây khó khăn cho người sử dụng [6].

2. VẬN DỤNG MÔ HÌNH XỬ LÝ NGÔN NGỮ VÀO HỆ THỐNG TRUY VẤN THÔNG MINH ĐỂ TÌM KIẾM TÀI LIỆU

2.1. Chuyển đổi từ tìm kiếm từ khóa sang tìm kiếm ngữ nghĩa

Hệ thống sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) để phân tích và hiểu rõ ngữ cảnh cũng như ý nghĩa ẩn sau các từ khóa. Nhờ đó, hệ thống có thể hiểu được ý định tìm kiếm của người dùng thay vì chỉ dựa vào so khớp từ khóa, hệ thống nhận diện được các yếu tố như thời gian, lĩnh vực hoặc mục đích tìm kiếm. Điều này giúp truy xuất thông tin chính xác hơn, hạn chế tình trạng trả về kết quả không liên quan và đáp ứng đúng nhu cầu tìm kiếm [11]. Hơn nữa, hệ thống có khả năng xử lý tốt các truy vấn phức tạp như tìm kiếm liên quan đến xu hướng thị trường, báo cáo nghiên cứu hoặc phân tích dữ liệu, giúp người dùng tiếp cận thông tin cần thiết hiệu quả hơn.

2.2. Ứng dụng trí tuệ nhân tạo vào hệ thống tìm kiếm thông minh

Trí tuệ nhân tạo - (Artificial Intelligence - AI) đang trở thành công cụ quan trọng trong việc phát triển các hệ thống tìm kiếm thông minh, các mô hình ngôn ngữ lớn (Large Language Models-LLMs) như GPT [7] và BERT [6] giúp phân tích và hiểu ngữ cảnh sâu sắc của các truy vấn dài, cho phép phân tích và nhận diện ý định tìm kiếm phức tạp để cung cấp câu trả lời chính xác và có ý nghĩa hơn. Trong lĩnh vực kinh tế, AI có thể hỗ trợ tìm kiếm và tổng hợp các báo cáo tài chính, xu hướng đầu tư, hay dữ liệu vĩ mô từ nhiều nguồn khác nhau. Hệ thống tìm kiếm thông minh ứng dụng AI sở hữu nhiều tính năng nổi bật, giúp cải thiện đáng kể trải nghiệm tìm kiếm thông tin.

2.3. Hỗ trợ ra quyết định và tối ưu hóa thông tin

Ứng dụng AI trong hệ thống tìm kiếm với khối lượng dữ liệu lớn, hỗ trợ tìm và phân tích thông tin từ các kho dữ liệu đồ sộ mà các hệ thống tìm kiếm truyền thống thường gặp khó khăn khi xử lý. Trong các lĩnh vực như nghiên cứu, giáo dục, kinh doanh và kinh tế, khả năng tìm kiếm thông tin chính xác và kịp thời mang lại lợi thế đáng kể. AI giúp tìm thông tin nhanh hơn, hỗ trợ người dùng ra quyết định hiệu quả hơn giúp tiết kiệm thời gian nhờ khả năng xử lý ngữ nghĩa, lọc thông tin không cần thiết và cung cấp thông tin minh bạch, đáng tin cậy kèm nguồn tham khảo rõ ràng, giúp người dùng dễ dàng kiểm chứng độ chính xác.

3. ĐỀ XUẤT THIẾT KẾ HỆ THỐNG TRUY VẤN THÔNG MINH

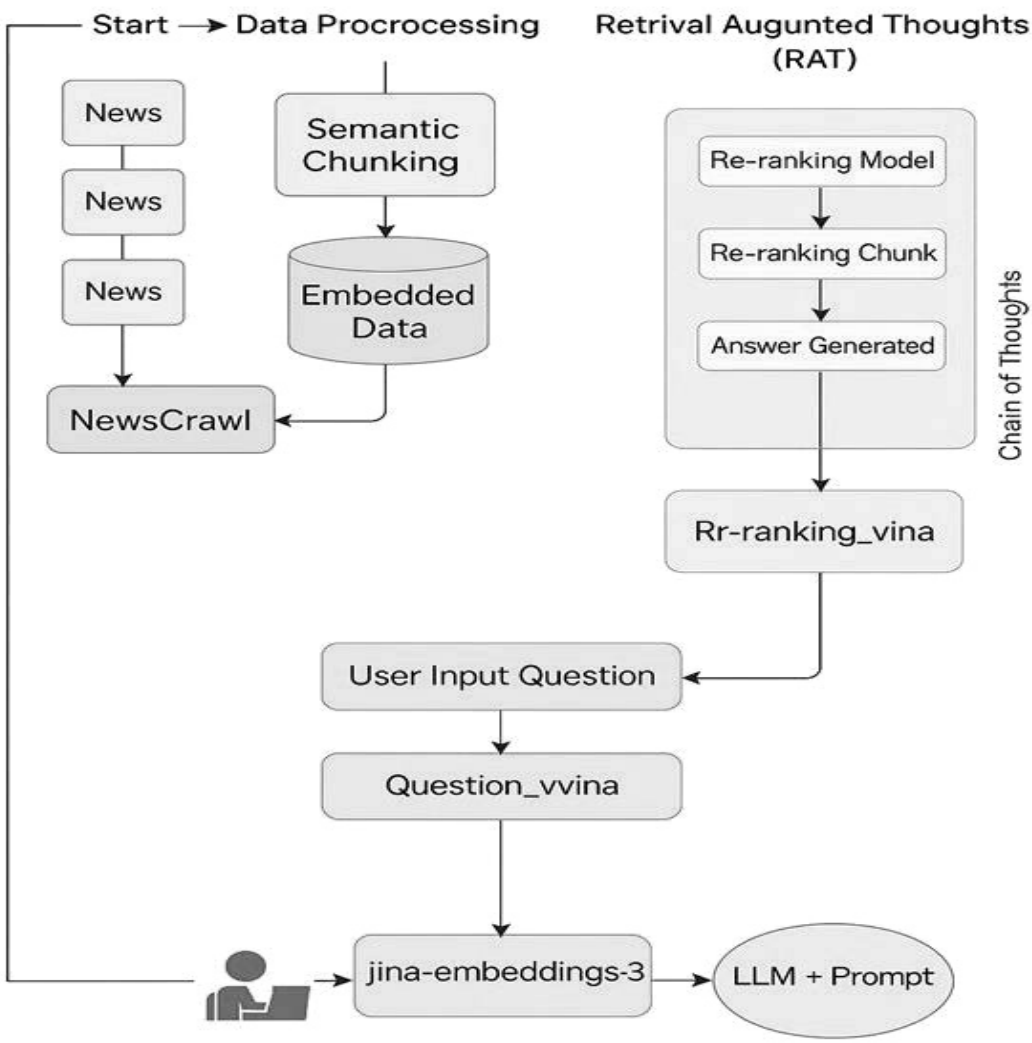
3.1. Sơ đồ kiến trúc hệ thống

Hệ thống truy vấn thông minh hoạt động thông qua ba bước chính của sơ đồ (Hình 2) như sau:

Bước 1: Dữ liệu được thu thập từ các trang báo kinh tế uy tín tại Việt Nam như: VnExpress, CafeF, CafeBiz, VietnamNet thông qua các công cụ hỗ trợ.

Bước 2: Các bài viết sau khi thu thập sẽ được lưu trữ tập trung phục vụ cho quá trình xử lý tiếp theo.

Bước 3: Sau khi được thu thập, nội dung được chia nhỏ và xử lý bằng mô hình nhúng văn bản (jina-embeddings-v3) [7] [9] để chuyển đổi nội dung và tạo ra các vector mang ý nghĩa ngữ cảnh. Các vector này được lưu trong cơ sở dữ liệu vector để phục vụ tìm kiếm nhanh chóng. Cuối cùng, khi có truy vấn từ người dùng, hệ thống tìm kiếm các đoạn nội dung phù hợp nhất, sau đó phân tích, tổng hợp thông tin thông qua mô-đun CoT [10] để tạo ra câu trả lời đầy đủ, mạch lạc và có trích dẫn nguồn tham khảo nhằm đảm bảo tính tin cậy và minh bạch.



Hình 2. *Sơ đồ kiến trúc hệ thống*

3.2. Mô hình tích hợp của hệ thống DataChatBot

LLM - GPT-3.5-turbo

Trong hệ thống DataChatBot, LLM đóng vai trò là thành phần chính chịu trách nhiệm xử lý và diễn giải các truy vấn ngôn ngữ tự nhiên từ người dùng. Phiên bản GPT-3.5-turbo mà DataChatBot sử dụng được thiết kế để đạt hiệu suất cao trong việc hiểu và phân tích ngữ cảnh của câu hỏi [6]. Nó giúp đảm bảo rằng người dùng nhận được thông tin chính xác và dễ hiểu ngay cả khi đặt các câu hỏi phức tạp hoặc không rõ ràng. Cụ thể, LLM tiếp nhận các truy vấn ngôn ngữ tự nhiên, phân tích mục đích và bối cảnh truy vấn, từ đó xây dựng các phản hồi phù hợp đảm bảo các câu trả lời không chỉ mang tính chính xác mà còn có tính liên kết và khả năng giải thích sâu sắc. Bên cạnh đó, LLM không chỉ trả lời trực tiếp các câu hỏi mà còn cung cấp các phản hồi chi tiết, hỗ trợ giải thích và phân tích các vấn đề phức tạp.

Vector Database

Cơ sở dữ liệu vector (Vector Database) đóng vai trò là thành phần lưu trữ và quản lý dữ liệu dưới dạng vector ngữ nghĩa. Mỗi vector được thiết kế để biểu diễn ý nghĩa của một đoạn văn bản, thay vì chỉ lưu trữ văn bản thô. Khi người dùng đặt một truy vấn, câu hỏi sẽ được chuyển đổi thành vector

ngữ nghĩa và so sánh với các vector đã lưu trong cơ sở dữ liệu để tìm ra các nội dung có mức độ tương đồng cao nhất. Công nghệ Vector Database không chỉ giúp tăng tốc độ xử lý mà còn giảm tải cho hệ thống, đảm bảo rằng các kết quả truy vấn luôn phản ánh đúng ý nghĩa của câu hỏi, ngay cả khi cơ sở dữ liệu mở rộng đến hàng triệu mục tin tức.

Giao diện người dùng

Giao diện người dùng được thiết kế mang lại trải nghiệm trực quan, thân thiện và dễ sử dụng. Đây được coi là cầu nối giữa người dùng với các thành phần của hệ thống, cho phép người dùng nhập câu hỏi, hiển thị kết quả truy vấn và cung cấp các tùy chọn tìm kiếm nâng cao. Giao diện được tích hợp chặt chẽ với LLM và Vector Database giúp đảm bảo câu hỏi từ người dùng được xử lý mượt mà và kết quả được hiển thị một cách rõ ràng. Người dùng có thể nhận được câu trả lời có kèm thông tin chi tiết như nguồn tin gốc hoặc các gợi ý liên quan.

Re-ranking Model

Mô hình xếp hạng lại (Re-ranking Model) đóng vai trò tối ưu hóa thứ tự kết quả truy vấn được trả về từ Vector Database. Re-ranking sử dụng độ đo tương đồng cosine [8] cùng với các kỹ thuật suy luận hiện đại như Chain of Thought Prompting [10], để phân tích và xếp hạng các kết quả một cách chính xác. Sau khi nhận danh sách các nội dung tương tự từ cơ sở dữ liệu, mô hình Re-ranking sẽ thực hiện đánh giá lại mức độ phù hợp từng kết quả dựa trên các tiêu chí ngữ nghĩa và ngữ cảnh.

Retrieval-Augmented Thoughts (RAT)

RAT là một phương pháp tiên tiến được thiết kế để cải thiện khả năng suy luận và tạo nội dung của các LLM trong các nhiệm vụ phức tạp. Bằng cách kết hợp giữa suy luận từng bước (Chain of Thought - CoT) và truy xuất thông tin bổ sung (Retrieval Augmented Generation - RAG) [7], RAT giảm thiểu đáng kể các lỗi suy đoán không chính xác (hallucination). Điểm nổi bật của phương pháp này là khả năng sửa đổi từng bước suy luận một cách liên tục và có hệ thống dựa trên thông tin được truy xuất từ nguồn dữ liệu bên ngoài, giúp các mô hình mô phỏng quy trình suy nghĩ như con người. RAT đã chứng minh hiệu quả vượt trội trong các lĩnh vực như lập trình, lập kế hoạch nhiệm vụ, suy luận toán học và sáng tạo nội dung, mang đến một giải pháp đột phá để khai thác tối đa tiềm năng của các mô hình AI.

Kết nối LLM với cơ sở dữ liệu và giao diện

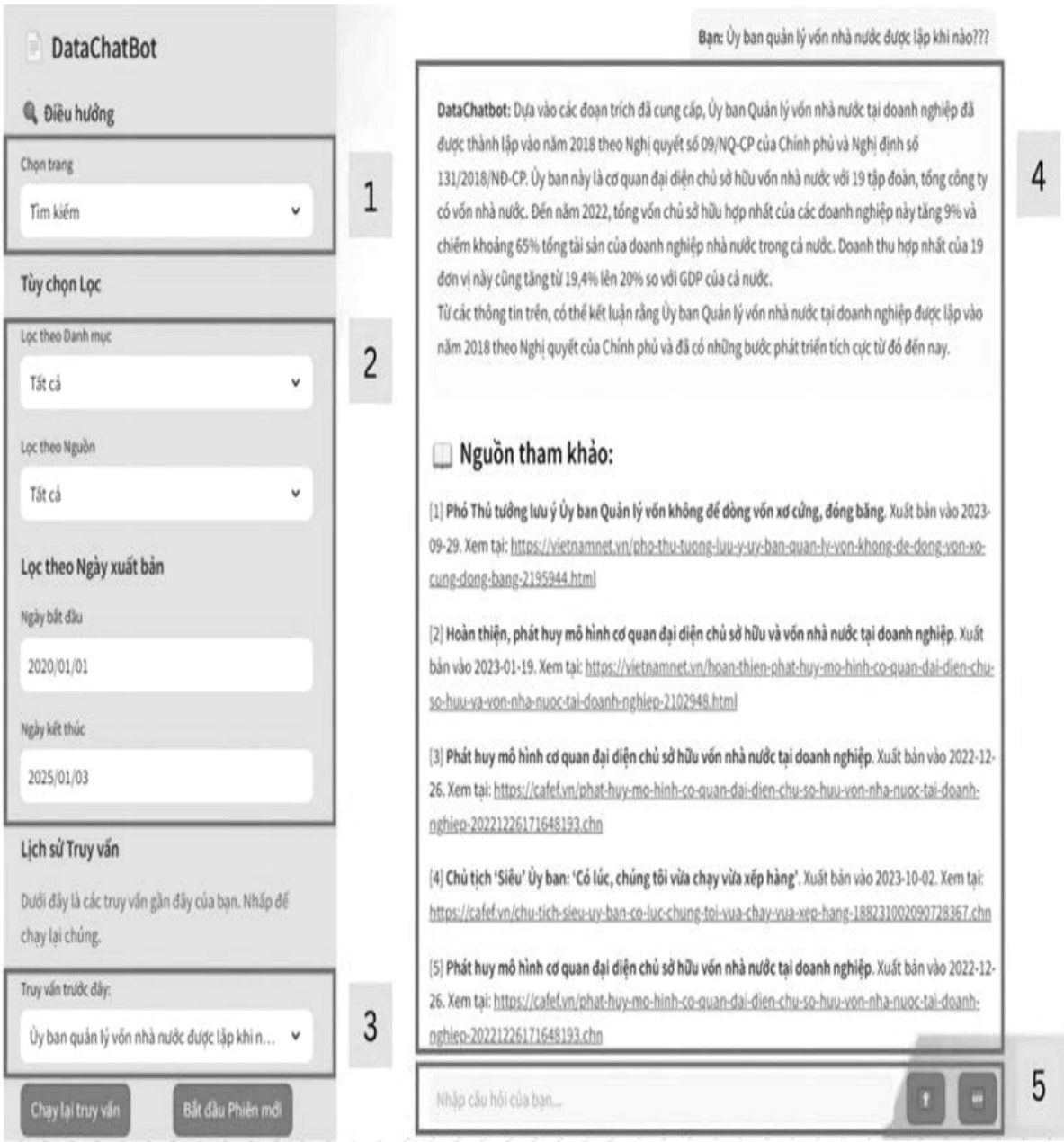
Trong bối cảnh phát triển công nghệ trí tuệ nhân tạo hiện đại, việc tích hợp các LLM như GPT-3.5-turbo [6] vào hệ thống quản lý và truy vấn thông tin đã trở thành một xu hướng đột phá. Kiến trúc hệ thống DataChatBot minh họa rõ nét quá trình chuyển đổi từ các hệ thống tìm kiếm truyền thống sang các giải pháp xử lý ngôn ngữ tự nhiên tiên tiến.

Quy trình tương tác bắt đầu khi người dùng nhập câu hỏi vào giao diện. LLM sẽ đóng vai trò là bộ não trung tâm, thực hiện phân tích sâu về ngữ nghĩa và ngữ cảnh của truy vấn. Ngoài ra, mô hình này còn có khả năng hiểu được ý đồ và ngữ cảnh phức tạp của người dùng. Sau khi phân tích, LLM sẽ tương tác với Vector Database - một cơ sở dữ liệu tiên tiến sử dụng học máy để lưu trữ và tìm kiếm thông tin. Các vector được tạo ra từ các đoạn văn bản sẽ được so sánh với vector của câu hỏi, cho phép tìm kiếm thông tin có độ liên quan cao. Kế tiếp, một mô hình xếp hạng lại (Re-ranking Model) sẽ được sử dụng để tinh chỉnh và lọc ra những kết quả quan trọng nhất.

4. KẾT QUẢ ĐẠT ĐƯỢC

4.1. Kết quả

Màn hình tương tác chính (Hình 3) được chia thành hai phần rõ rệt với bố cục được sắp xếp hợp lý. Phần bên trái là thanh điều hướng và tìm kiếm, bên phải là khu vực hiển thị nội dung tương tác chính.



Hình 3. *Giao diện màn hình tương tác chính*

Phần điều hướng được đánh số (1), thể hiện đây là bước đầu tiên trong quá trình người dùng tương tác với hệ thống. Thanh điều hướng hiển thị các mục chính gồm: Tìm kiếm, Giới thiệu, Báo lỗi/Đóng góp và Đăng xuất. Cách sắp xếp này giúp người dùng dễ dàng chuyển đổi giữa các tính năng mà không cần cuộn trang hay tìm kiếm thủ công.

Tiếp theo là phần chọn lọc, cũng được đánh số (2), bao gồm hai danh mục con là lọc theo danh mục và lọc theo nguồn. Mỗi danh mục đều có tùy chọn "Tất cả" được chọn mặc định, nhằm tạo điều kiện thuận lợi cho việc tìm kiếm thông tin đa dạng và không bị giới hạn. Đặc biệt, phần lọc theo ngày xuất bản được thiết kế với hai ô nhập liệu riêng biệt cho ngày bắt đầu và ngày kết thúc. Mỗi ô đều được gắn nhãn rõ ràng nhằm giúp người dùng dễ dàng nhận biết và sử dụng. Khoảng thời gian này có thể được tùy chọn linh hoạt, cung cấp phạm vi tìm kiếm rộng và chính xác hơn.

Phần lịch sử truy vấn cũng được đánh số (3),

hiển thị các câu truy vấn gần đây của người dùng. Tính năng này cho phép người dùng dễ dàng quay lại với các tìm kiếm trước đó, góp phần tạo nên một trải nghiệm tương tác mượt mà và logic, theo trình tự từ trên xuống dưới.

Khu vực tương tác hiển thị kết quả chính được thiết kế chiếm phần lớn không gian màn hình được đánh số (4), với nền trắng sáng nhằm tối ưu khả năng đọc. Nội dung trong khu vực này được tổ chức theo một hệ thống phân cấp rõ ràng, sử dụng màu sắc và khoảng cách hợp lý để phân biệt các phần. Bên dưới mỗi câu trả lời là phần nguồn tham khảo, liệt kê các nguồn thông tin đánh số từ [1] đến [n], trong đó [n] phụ thuộc vào số lượng kết quả tìm được sau khi hệ thống thực hiện lọc theo yêu cầu người dùng. Mỗi nguồn tham khảo đều có đường dẫn màu xanh dương, cho phép truy cập nhanh đến nguồn gốc thông tin gốc.

Cuối cùng là phần ô nhập câu hỏi mới, cũng được đánh số (5). Giao diện tại đây được thiết kế tối giản nhưng vẫn đảm bảo đầy đủ chức năng cần thiết. Người dùng có thể nhập câu hỏi mới, đồng thời sử dụng nút "Bắt đầu phiên mới" để làm mới câu hỏi và tiếp tục các phiên truy vấn tiếp theo (Hình 3).

Qua các thử nghiệm thực tiễn, DataChatBot đã chứng minh khả năng hỗ trợ người dùng trong việc truy xuất thông tin kinh tế. Hệ thống cung cấp kết quả kèm theo tài liệu tham khảo minh bạch, giúp người dùng không chỉ tin tưởng vào độ chính xác của thông tin mà còn dễ dàng sử dụng cho các nghiên cứu chuyên sâu hoặc ra quyết định chiến lược. Bên cạnh đó, việc triển khai DataChatBot trong các bối cảnh thực tế đã cho thấy hệ thống này góp phần đáng kể trong việc tiết kiệm thời gian và nâng cao hiệu suất làm việc của người dùng.

Việc ứng dụng LLM đã mang lại khả năng hiểu ngữ nghĩa và ý định của người dùng vượt xa so với các hệ thống tìm kiếm truyền thống dựa trên từ khóa. LLM giúp xử lý các truy vấn phức tạp, nắm bắt ngữ cảnh của câu hỏi và cung cấp các câu trả lời có tính liên quan cao. Bên cạnh đó, kỹ thuật RAT đóng vai trò quan trọng trong việc kết hợp tìm kiếm thông tin từ cơ sở dữ liệu với việc tạo nội dung trả lời phù hợp dựa trên dữ liệu đã thu thập. Thông qua việc kết nối và bổ sung ngữ nghĩa cho dữ liệu đã truy xuất, RAT giúp đảm bảo rằng câu trả lời được tạo ra bởi LLM trở nên chính xác hơn, phù hợp hơn với truy vấn của người dùng. Tích hợp Vector Database vào hệ thống đã mang lại những cải tiến quan trọng trong việc xử lý và truy xuất thông tin. Công nghệ này cho phép hệ thống hiểu rõ hơn về mối quan hệ ngữ nghĩa giữa các truy vấn và thông tin trong cơ sở dữ liệu, ngay cả khi các từ khóa hoặc cụm từ không khớp chính xác.

Mặc dù hệ thống DataChatBot mang lại nhiều lợi ích nhưng cũng tồn tại một số hạn chế cần khắc phục. Để giúp người đọc có cái nhìn tổng quan và dễ dàng so sánh, bảng 1 tổng hợp các ưu điểm và nhược điểm chính của hệ thống DataChatBot.

Bảng 1. *Ưu điểm và nhược điểm của DataChatBot*

Ưu điểm	Nhược điểm
- Khả năng hiểu ngữ cảnh và ý định của người dùng nhờ ứng dụng các kỹ thuật xử lý ngôn ngữ tự nhiên hiện đại. - Tích hợp Vector Database giúp tăng tốc độ và độ chính xác trong truy xuất ngữ nghĩa. - Re-ranking Model cải thiện chất lượng tìm kiếm, đảm bảo kết quả được sắp xếp theo mức độ liên quan. - Giao diện thân thiện, dễ sử dụng, phù hợp cho cả người dùng có chuyên môn và không chuyên môn.	- Chi phí triển khai và vận hành hệ thống cao, đặc biệt khi xử lý dữ liệu lớn. - Hiệu quả của hệ thống phụ thuộc vào chất lượng và độ phong phú của dữ liệu đầu vào. - Cần thêm thời gian và nguồn lực để tinh chỉnh thuật toán và mở rộng ứng dụng sang các lĩnh vực mới. - Đòi hỏi hạ tầng kỹ thuật hiện đại, nguồn tài nguyên tính toán mạnh để vận hành hệ thống. - Phức tạp về công nghệ, cần hỗ trợ kỹ thuật và đào tạo để người dùng khai thác hiệu quả.

4.2. So sánh giữa DataChatBot và Chat GPT

Mô hình DataChatBot được đề xuất sử dụng các phương pháp xử lý ngôn ngữ tự nhiên (NLP) tiên tiến và các kỹ thuật AI để cải thiện khả năng truy vấn thông tin, đặc biệt trong các lĩnh vực chuyên sâu như kinh tế. So với Chat GPT, mô hình này có sự khác biệt rõ rệt về cách thức hoạt động và mục đích sử dụng.

Bảng 2. *So sánh DataChatBot và Chat GPT*

	DataChatBot	Chat GPT
Mục đích sử dụng	Được thiết kế chủ yếu để hỗ trợ tìm kiếm và truy xuất thông tin từ các cơ sở dữ liệu lớn, giúp người dùng (đặc biệt là các nhà phân tích dữ liệu và các chuyên gia) nhanh chóng truy xuất thông tin kinh tế một cách chính xác. Nó được tối ưu hóa cho việc trả lời các câu hỏi liên quan đến dữ liệu kinh tế và tài liệu tham khảo.	Ngược lại, đây là mô hình tổng quát, có thể thực hiện nhiều tác vụ từ tạo văn bản, trả lời câu hỏi, đến hỗ trợ sáng tạo nội dung và giải quyết các vấn đề hàng ngày.
Công nghệ sử dụng	Sử dụng LLM kết hợp với Vector Database, RAT và các kỹ thuật như Embedding để cải thiện khả năng hiểu và truy vấn thông tin trong các cơ sở dữ liệu chuyên sâu. Điều này giúp DataChatBot có thể hiểu ngữ cảnh và ý định của người dùng tốt hơn so với các hệ thống tìm kiếm truyền thống, trả về những kết quả ngữ nghĩa và chính xác hơn.	Chủ yếu dựa vào mô hình GPT với các kỹ thuật ngôn ngữ tự nhiên mạnh mẽ, nhưng không được tối ưu hóa cho việc truy xuất thông tin từ cơ sở dữ liệu bên ngoài hay thực hiện các tác vụ cụ thể như DataChatBot
Khả năng tối ưu hóa và cải thiện kết quả	Sử dụng kỹ thuật re-ranking và retrieval-augmented generation để lọc và tối ưu hóa kết quả truy vấn. Điều này đảm bảo rằng các câu trả lời không chỉ chính xác mà còn phù hợp với ngữ cảnh và yêu cầu cụ thể của người dùng.	Không có khả năng truy xuất thông tin từ cơ sở dữ liệu cụ thể mà chỉ dựa vào các thông tin đã được huấn luyện trước, nên đôi khi không thể cung cấp những câu trả lời chính xác khi cần dữ liệu cập nhật hoặc rất chuyên sâu.

DataChatBot còn đặc biệt cung cấp tài liệu tham khảo cho các câu trả lời, giúp người dùng có thể kiểm chứng thông tin, điều này cực kỳ quan trọng trong các lĩnh vực nghiên cứu và phân tích dữ liệu. Như vậy, mô hình DataChatBot vượt trội trong việc tìm kiếm thông tin và phân tích dữ liệu chuyên sâu, đặc biệt là trong các lĩnh vực như kinh tế, nhờ vào sự kết hợp giữa các LLM và cơ sở dữ liệu vector, trong khi ChatGPT chủ yếu hoạt động như một trợ lý ngôn ngữ tổng quát cho các tác vụ đa dạng hơn, nhưng không chuyên sâu vào việc truy xuất thông tin từ cơ sở dữ liệu ngoài.

4.3. So sánh giữa DataChatBot và phương pháp tìm kiếm truyền thống

Các hệ thống tìm kiếm truyền thống chủ yếu dựa vào việc tìm kiếm từ khóa và siêu dữ liệu. Tuy nhiên, các phương pháp này thường gặp phải một số hạn chế, đặc biệt trong việc xử lý các truy vấn phức tạp hoặc khi cần phải cung cấp các kết quả tìm kiếm chính xác với ngữ cảnh rõ ràng. DataChatBot, được xây dựng dựa trên nền tảng LLM và các công nghệ như vector database, mang lại một giải pháp cải tiến trong việc tìm kiếm thông tin chính xác hơn và hỗ trợ các quyết định dựa trên dữ liệu.

Bảng 3. *So sánh giữa DataChatBot và phương pháp tìm kiếm truyền thống*

	DataChatBot	Phương pháp tìm kiếm truyền thống
Công nghệ sử dụng	Sử dụng các mô hình ngôn ngữ tự nhiên (NLP) để hiểu rõ hơn ngữ nghĩa của các truy vấn và các yếu tố ẩn sau chúng. Với sự kết hợp của LLM như GPT-3.5-turbo và vector database, DataChatBot có thể: *Hiểu ngữ nghĩa*: Hệ thống không chỉ dựa vào từ khóa mà còn có thể nhận diện ngữ cảnh, ý định và mục đích tìm kiếm của người dùng, giúp trả về kết quả chính xác và phù hợp hơn với nhu cầu thực tế. *Truy vấn ngữ nghĩa:* Thay vì so khớp từ khóa đơn giản, DataChatBot sử dụng vector ngữ nghĩa để tìm kiếm thông tin trong cơ sở dữ liệu. Các vector này biểu diễn ý nghĩa sâu sắc của văn bản, cho phép hệ thống hiểu mối quan hệ giữa các khái niệm và trả về kết quả có liên quan ngay cả khi từ khóa không khớp hoàn hảo. *Ứng dụng AI và Retrieval-Augmented Thoughts (RAT):* Sự kết hợp giữa *suy luận từng bước và truy xuất thông tin từ nguồn bên ngoài* giúp cải thiện khả năng suy luận của hệ thống và tránh được các lỗi suy đoán không chính xác (hallucinations).	Sử dụng các bot hoặc crawlers để thu thập thông tin từ các trang web và tài nguyên số hóa. Dữ liệu sau đó được lưu trữ trong các chỉ mục dựa trên các từ khóa và siêu dữ liệu như tiêu đề, tác giả và mô tả. Khi người dùng nhập truy vấn, hệ thống tìm kiếm so khớp các từ khóa trong chỉ mục và trả về kết quả phù hợp. Tuy nhiên, phương pháp này có một số hạn chế lớn: *Không hiểu được ngữ cảnh của truy vấn:* Hệ thống không thể nắm bắt được các yếu tố như mục đích tìm kiếm hoặc ngữ cảnh sử dụng, dẫn đến việc trả về các kết quả không liên quan hoặc thiếu chính xác. *Quá tải thông tin*: Hệ thống chỉ dựa vào từ khóa, dẫn đến việc người dùng phải lọc qua rất nhiều kết quả không hữu ích. *Khó xử lý dữ liệu phi văn bản:* Những tài liệu không phải văn bản như hình ảnh và video không thể được tìm kiếm hiệu quả bằng phương pháp này.
Khả năng tối ưu hóa và cải thiện kết quả	Mô hình xếp hạng lại (Re-ranking) trong DataChatBot sử dụng các kỹ thuật suy luận nâng cao để đánh giá lại và tối ưu hóa các kết quả truy vấn, đảm bảo rằng các kết quả cuối cùng đều có mức độ liên quan cao và phù hợp với ngữ cảnh truy vấn của người dùng.	Chỉ dựa vào tần suất từ khóa và độ phổ biến của trang web để xếp hạng kết quả, thiếu khả năng tối ưu hóa các kết quả theo ngữ nghĩa và ngữ cảnh cụ thể.
Khả năng cung cấp tài liệu tham khảo và đảm bảo tính minh bạch	Ưu điểm của DataChatBot là khả năng cung cấp tài liệu tham khảo cho các câu trả lời của mình, giúp người dùng có thể kiểm chứng thông tin dễ dàng hơn. Điều này đặc biệt quan trọng trong các lĩnh vực nghiên cứu và phân tích dữ liệu, nơi tính minh bạch và độ tin cậy của thông tin là yếu tố quyết định.	Các hệ thống tìm kiếm truyền thống thường không cung cấp tài liệu tham khảo, khiến người dùng phải tự tìm kiếm lại nguồn gốc của thông tin.

Như vậy, DataChatBot vượt trội so với hệ thống tìm kiếm truyền thống nhờ vào khả năng hiểu ngữ nghĩa và tối ưu hóa kết quả tìm kiếm thông qua các LLM và cơ sở dữ liệu vector. Sự kết hợp giữa AI và truy vấn ngữ nghĩa giúp DataChatBot xử lý các truy vấn phức tạp và cung cấp kết quả chính xác và đáng tin cậy hơn, điều mà các hệ thống tìm kiếm truyền thống không thể làm được. Hơn nữa, tính năng cung cấp tài liệu tham khảo trong DataChatBot cũng đảm bảo tính minh bạch và độ tin cậy của thông tin, mang lại trải nghiệm tìm kiếm vượt trội và hiệu quả hơn.

4.4. Đề xuất hướng phát triển

Hệ thống truy vấn thông minh sử dụng mô hình chatbot được đề xuất đã chứng minh tính hiệu quả trong việc hỗ trợ người dùng tìm kiếm tài liệu học thuật một cách tự nhiên và tiện lợi. Bằng cách kết hợp xử lý ngôn ngữ tự nhiên với kiến trúc dựa trên LLM, hệ thống không chỉ cải thiện độ chính xác của kết quả truy vấn mà còn nâng cao trải nghiệm tương tác người dùng. Trong tương lai, hệ thống có thể được mở rộng theo một số hướng phát triển sau:

Tận dụng khả năng xử lý đa ngôn ngữ: DataChatBot được trang bị các mô hình xử lý đa ngôn ngữ tiên tiến, hỗ trợ đến hơn 100 ngôn ngữ khác nhau nhờ tích hợp các mô hình dựa trên cơ chế Transformer. Nhưng hiện tại DataChatBot chủ yếu sử dụng dữ liệu tiếng Việt nên không tận dụng tối đa khả năng xử lý đa ngôn ngữ của hệ thống. Để tận dụng tốt khả năng của hệ thống, việc bổ sung dữ liệu từ nhiều ngôn ngữ khác sẽ là chìa khóa để khai thác tiềm năng này. Khi tận dụng được tối đa khả năng xử lý đa ngôn ngữ, DataChatBot sẽ hỗ trợ tốt hơn trong việc truy vấn thông tin kinh tế, tài liệu tham khảo và cả khi hệ thống mở rộng thêm nhiều lĩnh vực như y tế, giáo dục và pháp lý,...

Chi tiết hóa thông tin trả về và tốc độ phản hồi nhanh: DataChatBot cần được cải tiến về việc phản hồi một cách chi tiết hơn về các thông tin trả về với tốc độ nhanh chóng hơn hiện tại. Ở thời điểm hiện tại, mặc dù đã có tốc độ trả ra kết quả khá tốt nhưng vẫn còn cần nhiều thời gian để tổng hợp nội dung và gửi phản hồi cho người dùng hệ thống. Bên cạnh đó, tuy có điểm mạnh ở việc đưa ra nhận định rõ ràng cụ thể về mục tiêu câu truy vấn nhưng vẫn cần nhiều thông tin chi tiết chứng minh cho nhận định để tăng sức thuyết phục người dùng.

Lưu trữ toàn diện và học tăng cường từ phản hồi của người dùng: Thay vì lưu lại câu truy vấn cũ của người dùng như hiện tại thì hệ thống DataChatBot sẽ mở rộng khả năng lưu trữ toàn bộ nội dung các cuộc trò chuyện. Điều này giúp DataChatBot hiểu rõ hơn về ngữ cảnh, ý định của người dùng và đưa ra câu trả lời chính xác hơn. Hệ thống sẽ áp dụng phương pháp học tăng cường từ phản hồi của con người (Reinforcement Learning from Human Feedback - RLHF) để hệ thống có thể học từ phản hồi và dữ liệu thực tế [6]. Cùng với đó, Vector Database sẽ đóng vai trò quan trọng trong việc quản lý và cập nhật dữ liệu, đảm bảo rằng hệ thống luôn sẵn sàng đáp ứng các thay đổi nhanh chóng trong nhu cầu thông tin. Khả năng này kết hợp với RLHF giúp DataChatBot dễ dàng thích ứng với các thay đổi nhanh chóng trong nhu cầu thông tin.

Về việc tối ưu hóa chi phí và hiệu suất: DataChatBot có thể tận dụng công nghệ đám mây và phân tán để tối ưu hóa việc lưu trữ và xử lý dữ liệu. DataBot sẽ thay việc lưu trữ dữ liệu trên máy chủ vật lý bằng sử dụng dịch vụ đám mây để giảm chi phí hạ tầng đáng kể. Thêm vào đó, triển khai cơ chế lưu trữ dữ liệu thông minh như lưu trữ dựa trên mức độ ưu tiên sử dụng hoặc nén dữ liệu hiệu quả, giúp giảm dung lượng cần thiết mà vẫn đảm bảo khả năng truy cập nhanh chóng. Với các cải tiến này, DataChatBot không chỉ mang lại giá trị cao hơn cho người dùng mà còn trở thành một giải pháp bền vững và tiết kiệm chi phí dài hạn, đặc biệt khi triển khai trên quy mô lớn.

KẾT LUẬN

DataChatBot đã chứng minh được những khả năng hiểu câu hỏi và xử lý nhu cầu tin đa dạng liên quan đến lĩnh vực kinh tế một cách hiệu quả, phân tích và những nhận định có chiều sâu. Đặc biệt, DataChatBot có khả năng học hỏi và thích ứng với các loại dữ liệu mới linh hoạt, cung cấp thông tin chính xác và cập nhật cho người dùng. So với Chat GPT, DataChatBot có ưu thế trong xử lý câu hỏi có tính chuyên môn cao về kinh tế, tạo các văn bản có cấu trúc rõ ràng và logic. Đồng thời, DataChatBot có khả năng trình bày rõ ràng kèm các số liệu minh họa và nguồn tham khảo đáng tin cậy. Chat GPT lại có ưu thế ở sự cô đọng và dễ hiểu, phù hợp với tình huống cần câu trả lời nhanh chóng và người dùng phổ thông. DataChatBot vẫn cần cải thiện ở việc chi tiết hóa thông tin trả về, tốc độ phản hồi nhanh, trình bày kết quả hợp lý và dễ nhìn. Việc tích hợp sự súc tích của Chat GPT với độ chi tiết và minh bạch của DataChatBot, hệ thống sẽ trở thành công cụ toàn diện, đáp ứng mọi nhu cầu thông tin của người dùng.

TÀI LIỆU THAM KHẢO

1. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4171-4186. https://doi.org/10.18653/v1/N19-1423

2. Enterprise Knowledge. (2021, November 10). Expert analysis: Keyword search vs semantic search - Part one. Enterprise Knowledge. https://enterprise-knowledge.com/expert-analysis-keyword-search-vs-semantic-search-part-one

3. Jina AI. (2024). jina-embeddings-v3: Multilingual Embeddings With Task LoRA. arXiv preprint arXiv:2409.10173. https://huggingface.co/jinaai/jina-embeddings-v3

4. Kaufmann, T., Weng, P., Bengs, V., & Hüllermeier, E. (2023). A survey of reinforcement learning from human feedback. arXiv. https://doi.org/10.48550/arXiv.2312.14925

5. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N.,… & Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474. https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

6. OpenAI. (2023). GPT-4 Technical Report. ArXiv:2303.08774 [Cs]. https://doi.org/10.48550/arXiv.2303.08774

7. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI. https://openai.com/research/language-unsupervised

8. Rahutomo, F., Kitasuka, T., & Aritsugi, M. (2012). Semantic Cosine Similarity. http://www.researchgate.net/publication/262525676_Semantic_Cosine_Similarity

9. Sturua, S., Mohr, I., Akram, M. K., Günther, M., Wang, B., Krimmel, M., Wang, F., Mastrapas, G., Koukounas, A., Wang, N., & Xiao, H. (2024). jina-embeddings-v3: Multilingual Embeddings With Task LoRA. arXiv preprint arXiv:2409.10173. https://arxiv.org/abs/2409.10173

10. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. ArXiv:2201.11903 [Cs]. https://doi.org/10.48550/arXiv.2201.11903

11. Whitfield, S., & Hofmann, M. A. (2023). Elicit: AI literature review research assistant. Public Services Quarterly, 19(3), 201-207. https://doi.org/10.1080/15228959.2023.2224125

Ngày Tòa soạn nhận được bài: 02-8-2025;
Ngày phản biện đánh giá: 06-9-2025;
Ngày chấp nhận đăng: 15-10-2025