ỨNG DỤNG MÔ HÌNH NGÔN NGỮ LỚN VÀ TÌM KIẾM NGỮ NGHĨA XÂY DỰNG TRỢ LÝ ẢO HỖ TRỢ TRA CỨU VĂN BẢN QUY ĐỊNH NỘI BỘ TẠI TRƯỜNG ĐẠI HỌC

Tô Đức Nhuận*, Nguyễn Thị Thu Thủy, Phùng Thị Thu Hiền, Bùi Thu Hải, Trần Thị Yến
1Trường Đại học Sư phạm Kỹ thuật Nam Định, Việt Nam
***Email: tdnhuan@nute.edu.vn**

Tóm tắt: *Trong bối cảnh chuyển đổi số trong giáo dục, việc cung cấp cho sinh viên khả năng truy cập nhanh chóng và chính xác các quy định học vụ ngày càng trở nên cấp thiết. Bài báo này trình bày quá trình xây dựng và triển khai eTutor - một hệ thống hỏi đáp thông minh chuyên biệt dành cho sinh viên Trường Đại học Sư phạm Kỹ thuật Nam Định. Khác với các chatbot trí tuệ nhân tạo thông thường, eTutor ưu tiên tính chính xác và hợp lệ về pháp lý bằng cách chỉ truy xuất câu trả lời từ cơ sở dữ liệu đã được kiểm chứng. Điểm đổi mới cốt lõi của hệ thống nằm ở quy trình vận hành lai, kết hợp giữa việc sử dụng mô hình ngôn ngữ lớn triển khai cục bộ thông qua nền tảng Ollama để tự động sinh các biến thể câu hỏi, và việc áp dụng các thuật toán tìm kiếm ngữ nghĩa nhằm nhận diện ý định người dùng. Kết quả triển khai thực tế cho thấy hệ thống không chỉ nâng cao đáng kể độ chính xác truy vấn nhờ quy trình tiền xử lý chuyên biệt, mà còn hỗ trợ cơ chế tự cải thiện liên tục thông qua vòng phản hồi của sinh viên.*

Từ khóa: *Hệ thống hỏi đáp thông minh; Tìm kiếm ngữ nghĩa; Mô hình ngôn ngữ lớn; Quy định nội bộ; Cơ chế phản hồi.*

APPLICATION OF LARGE LANGUAGE MODELS AND SEMANTIC SEARCH IN DEVELOPING A VIRTUAL ASSISTANT FOR RETRIEVING INTERNAL REGULATIONS IN UNIVERSITIES

Abstract*: In the era of digital transformation in education, providing students with rapid and precise access to academic rules and regulations has become a critical necessity. This paper discusses the development and operation of eTutor, a specialized smart question-and-answering system tailored for students at Nam Dinh University of Technology and Education. Unlike conventional AI chatbots, eTutor prioritizes legal accuracy by strictly retrieving answers from a pre-validated database. The system's core innovation lies in its hybrid operational workflow: leveraging locally deployed Large Language Models via the Ollama platform to automate the generation of question variants, while simultaneously employing semantic search algorithms for intent recognition. Practical implementation results demonstrate that the system not only significantly enhances query accuracy through specialized preprocessing but also facilitates continuous self-improvement via a robust student feedback loop.*

Keywords: *Smart question-answering system; Semantic search; Large Language Models; Internal regulations; Feedback mechanism.*

Nhận bài: 03/03/2026 Phản biện: 27/03/2026 Duyệt đăng: 31/03/2026

I. ĐẶT VẤN ĐỀ

Trong kỷ nguyên chuyển đổi số của giáo dục đại học, các trường đại học ngày càng chú trọng hiện đại hóa công tác quản lý và dịch vụ hỗ trợ sinh viên. Việc cung cấp cho sinh viên khả năng tiếp cận nhanh chóng, chính xác các quy định nội bộ như quy chế đào tạo, quy định công tác sinh viên hay hướng dẫn nghiên cứu khoa học đã trở thành một yêu cầu thiết yếu. Tại Trường Đại học Sư phạm Kỹ thuật Nam Định, mặc dù các văn bản này được công bố công khai, sinh viên vẫn gặp nhiều khó khăn trong việc tra cứu thông tin do tài liệu thường có dung lượng lớn và được lưu trữ dưới dạng PDF hoặc Word. Điều này dẫn đến tình trạng các phòng ban chức năng phải xử lý khối lượng lớn các câu hỏi lặp lại.

Nhằm giải quyết vấn đề trên, các hệ thống chatbot dựa trên trí tuệ nhân tạo (AI) đã được phát triển như một giải pháp hiệu quả, cho phép cung cấp dịch vụ tư vấn tự động 24/7. Nhiều nghiên cứu cho thấy chatbot trong môi trường đại học đang phát triển từ các mô hình dựa trên luật đơn giản sang các hệ thống AI phức tạp hơn, ứng dụng xử lý ngôn ngữ tự nhiên và quản lý hội thoại nhằm tối ưu hóa tương tác với sinh viên (Ahmad et al., 2024). Về mặt kỹ thuật, các chatbot này thường được xây dựng trên các nền tảng mã nguồn mở hoặc thương mại như Coze hay Rasa, cho phép tích hợp các cơ sở tri thức chuyên biệt để tự động trả lời các câu hỏi liên quan đến quy chế đào tạo, học phí hay công tác sinh viên (Chau & Vo, 2024). Tại Việt Nam, việc ứng dụng bước đầu của các hệ thống này đã góp phần nâng cao kỹ năng ngoại ngữ của sinh viên, tuy nhiên vẫn tồn tại những thách thức liên quan đến tính trung thực và chuẩn mực học thuật (Hop et al., 2025), (Hiền, 2025).

Tuy nhiên, các chatbot truyền thống dựa trên đối sánh từ khóa thường không nắm bắt được đầy đủ ngữ cảnh và ý định ẩn sau câu hỏi của người

dùng. Ngược lại, các mô hình AI sinh nội dung hiện đại như ChatGPT hay Gemini, dù có tính linh hoạt cao, lại tiềm ẩn nguy cơ "ảo giác", tức là tạo ra các thông tin nghe có vẻ hợp lý nhưng không chính xác về mặt pháp lý hoặc không phù hợp với quy định cụ thể của từng trường đại học. Bên cạnh đó, việc triển khai các mô hình này thông qua API đám mây còn làm phát sinh các vấn đề liên quan đến bảo mật dữ liệu và chi phí vận hành định kỳ.

Xuất phát từ những vấn đề nêu trên, bài báo này trình bày quá trình nghiên cứu, thiết kế và triển khai eTutor – một trợ lý ảo thông minh được phát triển riêng cho sinh viên Trường Đại học Sư phạm Kỹ thuật Nam Định. Hệ thống được xây dựng trên kiến trúc lai, kết hợp tìm kiếm ngữ nghĩa dựa trên biểu diễn vector với mô hình ngôn ngữ lớn (LLM) được triển khai cục bộ thông qua nền tảng Ollama. Mục tiêu cốt lõi của eTutor là đảm bảo tính chính xác và hợp lệ về mặt pháp lý bằng cách chỉ cung cấp câu trả lời từ cơ sở dữ liệu đã được xác thực, đồng thời tận dụng AI để nâng cao khả năng hiểu ngôn ngữ tự nhiên. Các phần tiếp theo của bài báo sẽ trình bày chi tiết kiến trúc hệ thống, quy trình xây dựng ngân hàng cặp câu hỏi - câu trả lời tự động, cũng như đánh giá sơ bộ hiệu quả của hệ thống thông qua triển khai thực tế.

II. NỘI DUNG NGHIÊN CỨU

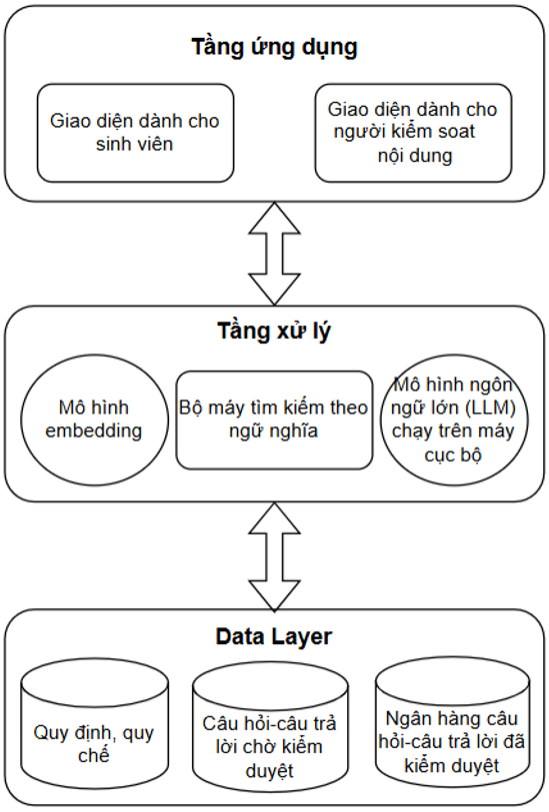
2.1. Kiến trúc hệ thống và phương pháp triển khai

2.1.1. Kiến trúc tổng thể

Hệ thống eTutor được thiết kế dựa trên kiến trúc lai, kết hợp giữa **cơ sở dữ liệu vector từ ngân hàng câu hỏi-câu trả lời đã được kiểm duyệt và mô hình ngôn ngữ lớn (LLM)** được triển khai cục bộ. Kiến trúc hệ thống gồm ba lớp chính: lớp dữ liệu, lớp xử lý và lớp ứng dụng.

Lớp dữ liệu lưu trữ các văn bản quy định chính thức của Nhà trường (quy định, quy chế) dưới dạng PDF và Word, cùng với ngân hàng câu hỏi–trả lời đã được chuẩn hóa (đã phê duyệt) và biểu diễn dưới dạng vector. Lớp xử lý đảm nhận nhiệm vụ xử lý ngôn ngữ tự nhiên, trong đó mô hình Sentence-BERT (Reimers & Gurevych, 2019) được sử dụng để mã hóa ngữ nghĩa văn bản, còn nền tảng Ollama đóng vai trò triển khai cục bộ các mô hình ngôn ngữ lớn nhằm hỗ trợ sinh và mở rộng câu hỏi. Lớp ứng dụng cung cấp giao diện truy cập cho sinh viên, đồng thời tích hợp cổng quản trị dành cho cán bộ phụ trách nhằm thẩm định và cập nhật nội dung.

Kiến trúc này cho phép hệ thống vừa đảm bảo tính linh hoạt trong xử lý ngôn ngữ tự nhiên, vừa duy trì quyền kiểm soát chặt chẽ đối với nội dung trả lời, phù hợp với yêu cầu quản lý trong môi trường giáo dục đại học.



Hình 1. Kiến trúc tổng thể hệ thống eTutor

2.1.2. Quy trình xây dựng cơ sở tri thức

Quy trình xây dựng cơ sở tri thức của eTutor tập trung vào việc chuyển đổi các tài liệu pháp quy không có cấu trúc thành một hệ thống tri thức có tổ chức, phục vụ hiệu quả cho hoạt động hỏi đáp tự động. Quy trình này bao gồm ba bước chính.

Thứ nhất, **trích xuất dữ liệu**, trong đó các văn bản như quy chế đào tạo, quy chế công tác sinh viên hay các chính sách hỗ trợ được thu thập và xử lý từ các tệp PDF và Word. Nội dung văn bản được phân tách theo điều khoản, mục và ngữ cảnh sử dụng.

Thứ hai, **mở rộng dữ liệu**, hệ thống sử dụng mô hình ngôn ngữ lớn triển khai cục bộ thông qua Ollama để tự động sinh các biến thể câu hỏi và câu trả lời tiềm năng. Việc sinh nội dung này không nhằm mục đích trả lời trực tiếp cho sinh viên mà nhằm mở rộng không gian biểu diễn ngữ nghĩa, giúp hệ thống nhận diện đa dạng cách đặt câu hỏi khác nhau của người học.

Thứ ba, **cơ chế kiểm duyệt có sự tham gia của con người được áp dụng.** Toàn bộ các cặp câu hỏi- câu trả lời được sinh tự động đều phải trải qua quá trình rà soát, chỉnh sửa và phê duyệt bởi cán bộ chuyên trách trước khi chính thức được cập nhật vào hệ thống. Cách tiếp cận này giúp đảm bảo tính chính xác, hợp lệ và phù hợp với các quy định hiện hành của nhà trường.

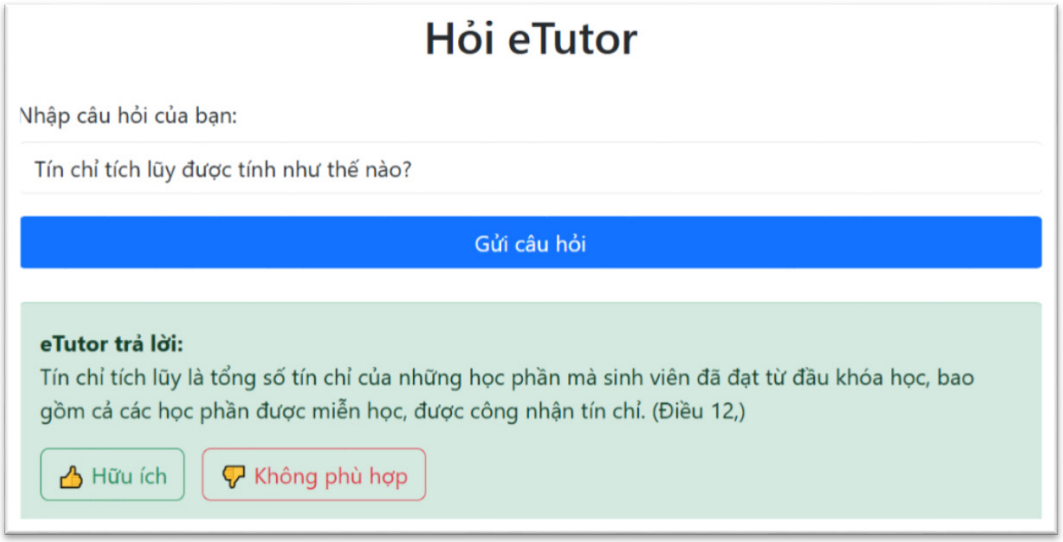
2.1.3. Xử lý truy vấn thời gian thực và truy hồi ngữ nghĩa

Khi sinh viên gửi một câu hỏi đến hệ thống, eTutor thực hiện chuỗi xử lý kỹ thuật nhằm đảm bảo độ chính xác của câu trả lời.

Trước hết, câu hỏi được tiền xử lý, bao gồm chuẩn hóa văn bản, mở rộng các từ viết tắt trong ngữ cảnh học vụ và loại bỏ các từ dừng không mang giá trị ngữ nghĩa. Tiếp theo, câu hỏi sau xử lý được vector hóa bằng mô hình paraphrase-multilingual-mpnet-base-v2 để chuyển thành một vector có số chiều cao, phản ánh ý nghĩa ngữ nghĩa của truy vấn.

Sau đó, hệ thống thực hiện đối sánh ngữ nghĩa bằng cách tính toán độ tương đồng Cosine giữa vector truy vấn và các vector đã được xây dựng sẵn trong ngân hàng câu hỏi - câu trả lời. Nếu độ tương đồng vượt qua một ngưỡng xác định trước (0,75), câu trả lời chính thức tương ứng sẽ được truy xuất và cung cấp ngay cho sinh viên.

Cơ chế này cho phép hệ thống trả lời gần như theo thời gian thực, đồng thời hạn chế tối đa các sai lệch nội dung do diễn giải không chính xác.



Hình 2. Giao diện của eTutor dành cho sinh viên

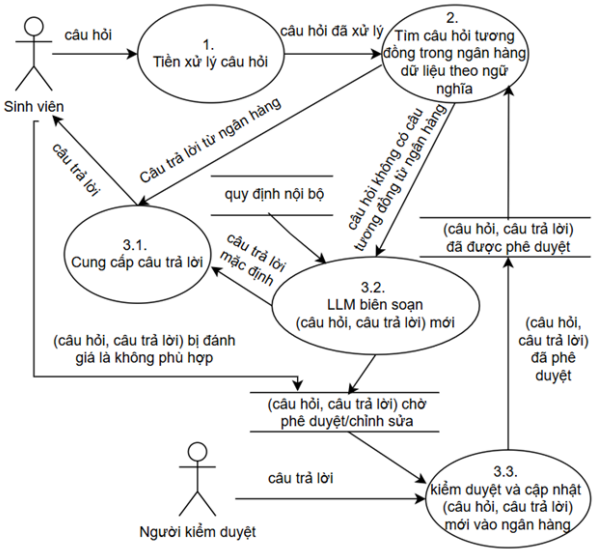
2.1.4. Xử lý truy vấn chưa khớp và cơ chế phản hồi

Để đảm bảo tính mở rộng và khả năng học hỏi liên tục, eTutor được thiết kế với cơ chế xử lý các truy vấn chưa có trong cơ sở dữ liệu.

Đối với các câu hỏi có độ tương đồng thấp, mô hình ngôn ngữ lớn sẽ đề xuất bản nháp câu trả lời dựa trên ngữ cảnh phù hợp trong các tài liệu gốc. Tuy nhiên, các câu trả lời này không được cung cấp trực tiếp cho sinh viên mà được đưa vào hàng đợi kiểm duyệt để cán bộ phụ trách rà soát và phê duyệt.

Sau khi được chấp thuận, các cặp câu hỏi - câu trả lời mới sẽ được vector hóa và bổ sung vào ngân hàng dữ liệu chính thức, qua đó giúp hệ thống ngày càng hoàn thiện. Đồng thời, hệ thống cho phép sinh viên đánh giá chất lượng câu trả lời thông qua cơ chế phản hồi tích cực hoặc tiêu cực. Ngoài ra, cơ chế đánh giá câu trả lời của hệ thống

là một cách kiểm duyệt lần hai. Câu trả lời của eTutor mà sinh viên thấy không hợp lý thì đánh giá là "Không phù hợp", khi đó câu hỏi và câu trả lời tương ứng được gửi vào hàng đợi để người kiểm duyệt xem xét lại, đảm bảo chất lượng thông tin luôn được cập nhật và duy trì ở mức cao.



Hình 3. Quy trình xử lý truy vấn thời gian thực của hệ thống eTutor

2.2. Kết quả thực nghiệm và thảo luận

2.2.1. Đánh giá kỹ thuật và đặc điểm tập dữ liệu

Hệ thống eTutor đã xây dựng thành công một cơ sở tri thức toàn diện gồm 1.146 cặp câu hỏi - câu trả lời. Tập dữ liệu này được trích xuất và tổng hợp từ bốn văn bản quy định cốt lõi của Trường Đại học Sư phạm Kỹ thuật Nam Định, bao gồm: Quy chế đào tạo trình độ đại học, Quy chế công tác sinh viên, Quy định quản lý hoạt động nghiên cứu khoa học của sinh viên và Quy định thực hiện các chính sách đối với sinh viên (Trường Đại học Sư phạm Kỹ thuật Nam Định, 2024a; 2018; 2024b; 2014).

Kết quả thực nghiệm cho thấy mô hình paraphrase-multilingual-mpnet-base-v2 thể hiện hiệu quả vượt trội so với các phương pháp tìm kiếm truyền thống dựa trên từ khóa. Cụ thể, việc áp dụng các kỹ thuật tiền xử lý chuyên biệt cho ngữ cảnh học vụ, chẳng hạn như ánh xạ các từ viết tắt (ví dụ: "đk tín chỉ" sang "đăng ký tín chỉ"), giúp hệ thống nhận diện chính xác hơn các truy vấn được diễn đạt bằng ngôn ngữ tự nhiên của sinh viên.

Bên cạnh đó, việc lưu trữ các vector ngữ nghĩa dưới dạng tensor PyTorch (.pt) cho phép hệ thống tính toán độ tương đồng Cosine gần như tức thời, đáp ứng yêu cầu tương tác thời gian thực, đồng thời tránh được hiện tượng "ảo giác", vốn thường xuất hiện trong các hệ thống AI sinh nội dung tổng quát (tăng precision). Đặc biệt, việc sử dụng mô hình ngôn ngữ lớn để sinh các biến thể câu hỏi giúp mở rộng ngân hàng câu hỏi giúp nâng cao khả năng trả lời các câu hỏi đồng nghĩa (tăng recall).

2.2.2. Phân tích so sánh các phương pháp truy khớp

Để thực hiện sơ khớp giữa câu hỏi (câu truy vấn) của sinh viên và các câu hỏi đã kiểm duyệt trong ngân hàng, nhóm tác giả đã thử nghiệm nhiều phương pháp khác nhau:

Trước hết, phương pháp đối sánh mờ (fuzzy matching) cho kết quả không cao do không xử lý hiệu quả các biến thể ngôn ngữ tự nhiên và các cách diễn đạt đa dạng của sinh viên. Ngược lại, tìm kiếm ngữ nghĩa dựa trên vector thể hiện ưu thế rõ rệt nhờ khả năng nắm bắt ý định bên trong câu hỏi, mặc dù phương pháp này vẫn gặp khó khăn khi truy vấn được diễn đạt quá khác biệt so với dữ liệu đã được chuẩn bị ban đầu.

Trong hệ thống eTutor, mô hình ngôn ngữ lớn không đóng vai trò tạo sinh câu trả lời độc lập. Thay vào đó, vai trò chính của LLM là hỗ trợ tìm kiếm và xác định câu hỏi phù hợp nhất trong ngân hàng hiện có. Đối với các truy vấn chưa được ánh xạ thành công, LLM được sử dụng để soạn câu trả lời nháp dựa trên dữ liệu văn bản gốc, sau đó chuyển vào giao diện để người kiểm duyệt xem xét, điều chỉnh trước khi đưa vào sử dụng chính thức. Cách tiếp cận này giúp tận dụng được khả năng xử lý ngôn ngữ mạnh mẽ của LLM, đồng thời hạn chế rủi ro sai lệch thông tin, đặc biệt trong bối cảnh các quy định học vụ yêu cầu độ chính xác cao.

2.2.3. Hạn chế và định hướng phát triển

Mặc dù có nhiều ưu điểm về tính chính xác, khả năng kiểm soát nội dung, và đặc biệt là tiết kiệm chi phí vận hành (do sử dụng mô hình ngôn ngữ lớn miễn phí (open source)máy chủ hiện có của Nhà trường thay vì sử dụng API của các nhà cung cấp), hệ thống eTutor hiện tại vẫn tồn tại một số hạn chế khi so sánh với các chatbot AI đa năng như ChatGPT.

Trước hết, hệ thống chưa hỗ trợ ghi nhớ hội thoại đa lượt, chưa cung cấp các câu trả lời cá nhân hóa dựa trên lịch sử người dùng, và chưa thực hiện được các suy luận phức tạp trên nhiều nguồn thông tin khác nhau. Tuy nhiên, các hạn chế này một phần nhằm đảm bảo tính chính xác nội dung, tuân thủ quy định và duy trì khả năng kiểm soát thông tin trong môi trường giáo dục.

Trong các giai đoạn phát triển tiếp theo, nhóm nghiên cứu dự kiến tập trung vào một số hướng mở rộng chính. Thứ nhất, triển khai bộ nhớ hội thoại theo phiên, cho phép hệ thống duy trì ngữ cảnh trong chuỗi các câu hỏi liên quan. Thứ hai, cá nhân hóa người dùng, thông qua xác thực sinh viên để điều chỉnh nội dung trả lời phù hợp với ngành học hoặc tình trạng học tập.

III. KẾT LUẬN

Hệ thống eTutor tại Trường Đại học Sư phạm Kỹ thuật Nam Định đã được nghiên cứu, xây dựng và triển khai với mục tiêu trọng tâm là hỗ trợ sinh viên tiếp cận thông tin học vụ một cách nhanh chóng, chính xác và được kiểm soát chặt chẽ về nội dung. Thông qua việc tích hợp các kỹ thuật trí tuệ nhân tạo hiện đại, đặc biệt là mô hình ngôn ngữ lớn triển khai cục bộ kết hợp với cơ chế kiểm duyệt có sự tham gia của con người hệ thống bảo đảm tính chính thống và độ tin cậy cao cho mọi câu trả lời được cung cấp.

Thông qua hai quy trình cốt lõi gồm xây dựng ngân hàng câu hỏi – câu trả lời từ các văn bản quy định chính thức và xử lý truy vấn của sinh viên theo thời gian thực, eTutor đã đạt được độ chính xác vượt trội so với các phương pháp tra cứu truyền thống và khắc phục hiệu quả khó khăn trong việc hiểu ngôn ngữ tự nhiên của người dùng. Mặc dù hiện tại hệ thống còn tồn tại một số hạn chế so với các chatbot AI đa năng, cơ chế kiểm soát nội dung hai lớp bảo đảm tuân thủ quy định, kiểm soát nội dung và việc triển khai mô hình ngôn ngữ lớn miễn phí trên máy chủ cục bộ của Nhà trường giúp tiết kiệm chi phí vận hành thực tế.

Trong thời gian tới, việc bổ sung bộ nhớ hội thoại theo phiên, cá nhân hóa theo hồ sơ sinh viên sẽ tiếp tục nâng cao trải nghiệm người dùng, đồng thời vẫn giữ vững nguyên tắc chính xác và an toàn thông tin. Kết quả nghiên cứu này cho thấy eTutor là một mô hình khả thi và hiệu quả để triển khai trợ lý ảo học vụ trong các cơ sở giáo dục đại học tại Việt Nam.

TÀI LIỆU THAM KHẢO

Ahmad, C. W. S. B. C. W., Ahmed, S. A., Rahman, K. A., Ahmad, S., & Basri, M. (2024). A systematic literature review on AI-powered chatbot for universities. Journal of Advanced Research in Applied Sciences and Engineering Technology, 62(2), 112–126. https://doi.org/10.37934/araset.62.4.111126

Châu, N. N., & Võ, H. N. (2024). Thử nghiệm ứng dụng AI Chatbot làm cố vấn học tập ảo hỗ trợ trực tuyến cho sinh viên tại Trường Đại học Kiên Giang. *Tạp chí Khoa học Quản lý Giáo dục*, 03(43), 34-40.

Hop, N. H., Cuc, N. T., Canh, P. T. T., & Cuong, D. H. (2025). AI-Chatbot applications in university learning: A survey on current status, effectiveness, and usage orientations of Technical Pedagogical University students in Vietnam. *Tạp chí Khoa học Giáo dục Việt Nam*, 21(09).

Hiền , P. T. (2025). Phản hồi của sinh viên về sử dụng chatbot trí tuệ nhân tạo cải thiện kỹ năng hội thoại tiếng Anh. *Tạp chí Khoa học Trường Đại học Mở Hà Nội*, 11(1), 264-271.

Reimers, N., & Gurevych, I. (2019, November). Sentence-bert: Sentence embeddings using siamese bert-networks. *In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3982-3992).

Trường Đại học Sư phạm Kỹ thuật Nam Định (2014). *Quy định thực hiện chế độ chính sách đối với sinh viên trường Đại học Sư phạm Kỹ thuật Nam Định (1234/QĐ-ĐHSPKTNĐ ngày 26 tháng 12 năm 2014).*

Trường Đại học Sư phạm Kỹ thuật Nam Định. (2018*). Quy chế công tác sinh viên hệ chính quy (78/ QĐ-ĐHSPKTNĐ ngày 16 tháng 01 năm 2018).*

Trường Đại học Sư phạm Kỹ thuật Nam Định (2024a). *Quy chế đào tạo trình độ đại học (414/QĐ-ĐHSPKTNĐ ngày 26 tháng 08 năm 2024).*

Trường Đại học Sư phạm Kỹ thuật Nam Định. (2024b). *Quy định về hoạt động NCKH của sinh viên (398/QĐ-ĐHSPKTNĐ ngày 30 tháng 08 năm 2024).*