

NÂNG CAO HIỆU QUẢ ĐÀO TẠO THÔNG QUA GIẢI PHÁP CHATBOT AI: NGHIÊN CỨU THỰC NGHIỆM TRÊN HỆ THỐNG LMS MOODLE

● **TRƯƠNG TẤN LỘC^{1,2} - NGUYỄN DUY THANH^{2*}**

¹ Trường Cao đẳng Kinh tế Thành phố Hồ Chí Minh

² Trường Đại học Ngân hàng Thành phố Hồ Chí Minh

*Email: thanhnd@hub.edu.vn

TÓM TẮT:

Nghiên cứu này đề xuất giải pháp chatbot AI sử dụng công nghệ Tạo sinh tăng cường bằng truy xuất (Retrieval-Augmented Generation - RAG) chạy trên máy chủ cục bộ (Local LLM), tích hợp trực tiếp vào hệ thống LMS Moodle. Điểm đột phá của nghiên cứu nằm ở thuật toán Bộ nhớ đệm ngữ nghĩa lai (Hybrid Semantic Cache - HSC) giúp giảm 99,7% thời gian phản hồi và đưa chi phí biên cho mỗi truy vấn về mức bằng không. Kết quả thực nghiệm trên 320 sinh viên cho thấy, hệ thống đạt độ chính xác trung bình 92%, không ghi nhận hiện tượng ảo giác AI trong bộ câu hỏi kiểm thử, đồng thời tăng cường khả năng kiểm soát dữ liệu do xử lý cục bộ cho cơ sở giáo dục. Nghiên cứu cung cấp một mô hình khả thi về mặt kinh tế và kỹ thuật để các trường đại học, cao đẳng triển khai trợ giảng ảo quy mô lớn với chi phí vận hành tối thiểu.

Từ khóa: chuyển đổi số, quản trị LMS, RAG chatbot, hiệu quả kinh tế.

1. Đặt vấn đề

Chuyển đổi số trong giáo dục đang là ưu tiên chiến lược của Việt Nam, được thể hiện rõ qua Quyết định số 131/2022/QĐ-TTg phê duyệt Đề án “Tăng cường ứng dụng công nghệ thông tin và chuyển đổi số trong giáo dục và đào tạo giai đoạn 2022-2025, định hướng đến năm 2030”. Chuyển đổi số trong giáo dục không chỉ đơn thuần là việc số hóa tài liệu mà còn là sự thay đổi căn bản trong phương thức quản lý và tương tác dạy - học. Tại Việt Nam, hệ thống quản lý học tập (LMS Moodle) đã trở thành nền tảng phổ biến tại nhiều cơ sở đào tạo. Tuy nhiên, thực trạng

hiện nay cho thấy các hệ thống này chủ yếu đóng vai trò là kho chứa học liệu tĩnh, thiếu khả năng hỗ trợ tương tác tức thời cho sinh viên. Từ góc độ quản lý giáo dục, sự thiếu hụt công cụ hỗ trợ tương tác tự động đặt ra 2 bài toán nan giải: (1) sinh viên không được giải đáp kịp thời các vướng mắc ngoài giờ học, ảnh hưởng tiêu cực đến chất lượng tự học và động lực học tập; (2) giảng viên phải đối mặt với khối lượng công việc hỗ trợ ngày càng gia tăng, ảnh hưởng đến năng suất giảng dạy và nghiên cứu. Đây là khoảng trống quản lý mà công nghệ chatbot AI có thể lấp đầy một cách hiệu quả.

Việc sử dụng các giải pháp AI thương mại qua API bên thứ ba như OpenAI bộc lộ nhiều rào cản về quản trị: chi phí duy trì cao theo lưu lượng sử dụng, nguy cơ lộ lọt dữ liệu nội bộ và hiện tượng “ảo giác AI” gây sai lệch kiến thức chuyên môn (OpenAI, 2025). Từ góc độ quản trị tài chính và an ninh thông tin, các cơ sở giáo dục cần một giải pháp AI tự chủ, an toàn và tối ưu về chi phí. Nghiên cứu này được thực hiện nhằm giải quyết nút thắt đó thông qua việc thiết kế và thử nghiệm hệ thống chatbot AI tích hợp RAG chạy cục bộ, giúp nâng cao hiệu quả tự học của sinh viên và giảm tải áp lực hỗ trợ cho giảng viên.

2. Cơ sở lý thuyết

2.1. Vai trò của LMS và chatbot trong quản trị giáo dục số

Chuyển đổi số trong giáo dục không đơn thuần là việc số hóa tài liệu hay dạy học trực tuyến, mà là quá trình tái cấu trúc toàn diện mô hình quản lý và vận hành của nhà trường thông qua ứng dụng công nghệ kỹ thuật số (UNESCO, 2023). Trong quản lý nhà trường hiện đại, LMS đóng vai trò là hạ tầng trung tâm, tích hợp các chức năng quản lý khóa học, phân phối học liệu, đánh giá học tập và theo dõi tiến độ của người học. Moodle với cộng đồng hơn 300 triệu người dùng tại hơn 240 quốc gia, là nền tảng LMS mã nguồn mở phổ biến nhất thế giới (Moodle, 2025), mang lại lợi thế lớn về tính linh hoạt, khả năng tùy chỉnh và đặc biệt là không phát sinh chi phí bản quyền phần mềm.

Chatbot giáo dục là hệ thống phần mềm tương tác bằng ngôn ngữ tự nhiên, có khả năng đóng vai trò trợ giảng ảo, hỗ trợ hành chính và cung cấp phản hồi cá nhân hóa cho người học (Winkler & Söllner, 2018). Từ góc độ quản lý, chatbot không thay thế giảng viên mà đóng vai trò xử lý các câu hỏi thường gặp và hỗ trợ học tập ngoài giờ, qua đó giải phóng thời gian của giảng viên cho các hoạt động giảng dạy và nghiên cứu có giá trị gia tăng cao hơn (Winkler & Söllner, 2018). LMS được xác định là thành phần cốt lõi của hệ sinh thái giáo dục số, hỗ trợ theo dõi tiến trình và cải thiện hiệu quả đào tạo. Việc tích hợp chatbot vào LMS giúp cá nhân hóa lộ trình học tập, cung cấp phản hồi khách quan và nhanh chóng cho người học. Các

nghiên cứu trước đây chỉ ra, chatbot giúp sinh viên tiếp cận tài nguyên giáo dục hiệu quả hơn, đồng thời giải phóng giảng viên khỏi các tác vụ hành chính lặp lại (UNESCO, 2023).

2.2. Công nghệ RAG và xu hướng LLM trong giáo dục

Tạo sinh tăng cường bằng truy xuất (Retrieval-Augmented Generation - RAG) là kiến trúc kết hợp giữa hệ thống truy xuất thông tin ngữ nghĩa và mô hình ngôn ngữ tạo sinh, được Lewis et al. (2020) đề xuất. Trong giáo dục, RAG có ý nghĩa đặc biệt quan trọng vì giải quyết được vấn đề “ảo giác AI” - hiện tượng mô hình ngôn ngữ sinh ra nội dung sai lệch không có căn cứ - một rào cản nghiêm trọng khi ứng dụng AI trong môi trường đòi hỏi độ chính xác cao về kiến thức chuyên môn. Từ góc độ quản lý chất lượng giáo dục, RAG cho phép chatbot chỉ trả lời dựa trên tài liệu học liệu đã được kiểm duyệt bởi giảng viên, đảm bảo tính nhất quán và chuẩn xác của thông tin cung cấp cho người học. Đây là yếu tố then chốt để các nhà quản lý giáo dục có thể tự tin triển khai AI trong nhà trường.

2.3. Bộ nhớ đệm ngữ nghĩa cho hệ thống AI

Trong hệ thống AI dựa trên LLM, mỗi truy vấn đều phát sinh chi phí tính toán đáng kể do mô hình phải xử lý toàn bộ ngữ cảnh đầu vào. Bộ nhớ đệm ngữ nghĩa là kỹ thuật lưu trữ các cặp câu hỏi - câu trả lời đã xử lý, sau đó tái sử dụng kết quả khi gặp truy vấn tương tự về mặt ý nghĩa (Bang, 2023). Khác với bộ nhớ đệm truyền thống chỉ khớp chính xác từng ký tự, bộ nhớ đệm ngữ nghĩa sử dụng các vector nhúng để đo độ tương đồng cosine giữa câu hỏi mới và câu hỏi đã lưu, cho phép nhận diện các biến thể ngôn ngữ cùng ý nghĩa. Trong môi trường giáo dục - nơi sinh viên thường đặt nhiều câu hỏi tương tự nhau xung quanh các chủ đề chính - bộ nhớ đệm ngữ nghĩa được kỳ vọng mang lại hiệu quả tiết kiệm tài nguyên đáng kể.

3. Thiết kế nghiên cứu

3.1. Phương pháp nghiên cứu

Nghiên cứu sử dụng phương pháp Nghiên cứu Thiết kế và Thực nghiệm (Design Science Research - DSR) kết hợp với phân tích chi phí - lợi ích. DSR phù

hợp cho các nghiên cứu nhằm tạo ra và đánh giá các giải pháp công nghệ trong bối cảnh tổ chức cụ thể (Hevner et al., 2004). Quy trình nghiên cứu gồm 4 giai đoạn: (1) phân tích nhu cầu tổ chức và thiết kế giải pháp; (2) triển khai hệ thống trong môi trường thực tế; (3) thực nghiệm và thu thập dữ liệu đa chiều; (4) phân tích hiệu quả quản lý, kinh tế và kỹ thuật.

Nghiên cứu được triển khai tại Trường Cao đẳng Kinh tế TP. Hồ Chí Minh (HCE), một cơ sở giáo dục công lập điển hình tại Việt Nam với khoảng 9.000 sinh viên đang theo học. HCE đã vận hành LMS Moodle từ năm 2021 nhưng chưa tích hợp bất kỳ công cụ AI tương tác nào. Đối tượng thực nghiệm là 320 sinh viên, được chọn do tính phổ biến của học phần và sự đa dạng trong nhu cầu câu hỏi của sinh viên. Thời gian thực nghiệm kéo dài 8 tuần trong học kỳ. Dữ liệu được thu thập gồm:

(i) nhật ký truy vấn hệ thống gồm thời gian phản hồi, trạng thái bộ đệm, độ dài câu trả lời; (ii) bộ 200 câu hỏi kiểm thử do giảng viên xây dựng để đánh giá độ chính xác phản hồi; (iii) khảo sát mức độ hài lòng của 320 sinh viên trước và sau triển khai theo thang đo Likert 5 điểm; (iv) báo cáo khối lượng công việc của giảng viên phụ trách trước và sau khi triển khai hệ thống.

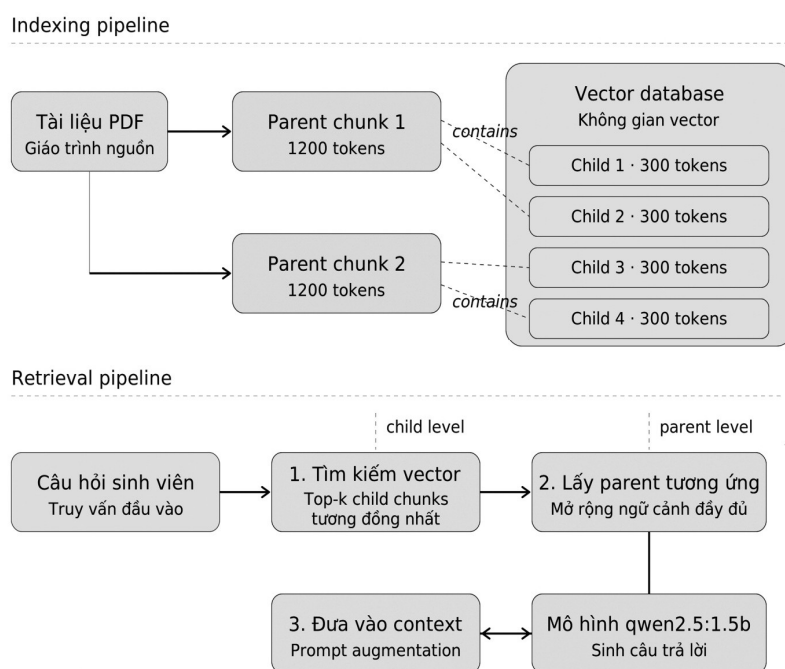
3.2. Kiến trúc hệ thống RAG phân cấp

Để giải quyết mâu thuẫn giữa độ chính xác tìm kiếm và tính toàn vẹn của ngữ cảnh, nghiên cứu áp dụng kiến trúc của bộ truy xuất phân cấp Cha-Con. Đoạn con (Child Chunks) với kích thước 300 token, được dùng để so sánh vector với câu hỏi người dùng nhằm định vị thông tin chính xác nhất. Đoạn cha (Parent Chunks) với kích thước 1.200 token, bao trùm đoạn con để cung cấp đầy đủ ngữ cảnh cho mô hình AI sinh câu trả lời logic. Thực nghiệm trên bộ

kiểm thử 200 câu hỏi cho thấy bộ tham số 1.200/300 token đạt độ chính xác trung bình 92%, cao hơn rõ rệt so với các cấu hình chunk đơn cấp truyền thống.

Hình 1.

Hình 1: Sơ đồ chiến lược phân cấp Cha-Con



3.3. Thuật toán Bộ nhớ đệm ngữ nghĩa lai (HSC)

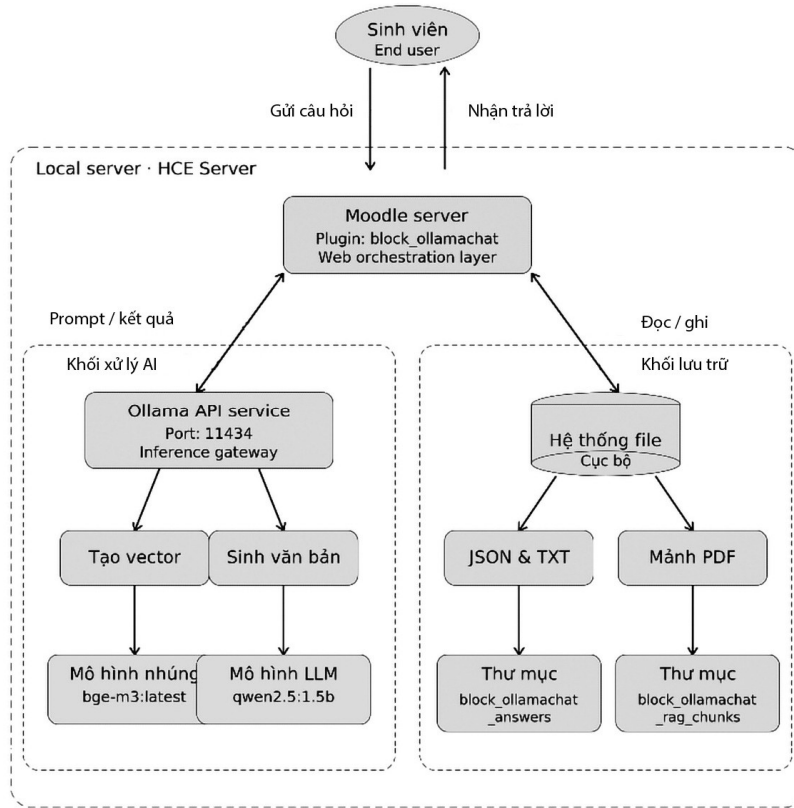
Đây là đóng góp trọng yếu về mặt kỹ thuật và kinh tế của nghiên cứu. Thuật toán kết hợp 2 lớp tra cứu: (1) Exact Match - tra cứu nhanh dựa trên hàm băm MD5 cho các câu hỏi giống hệt; (2) Semantic Match - sử dụng mô hình nhúng bge-m3 để đo lường độ tương đồng cosine với ngưỡng khắt khe 0,92. Cơ chế này cho phép hệ thống nhận diện các câu hỏi biến thể có cùng ý nghĩa nhưng khác từ ngữ để trả lời ngay lập tức từ bộ nhớ đệm, giúp tiết kiệm tài nguyên máy chủ đáng kể. Toàn bộ kiến trúc tổng thể của hệ thống được mô tả trong Hình 2.

4. Kết quả thực nghiệm và thảo luận

4.1. Hiệu năng kỹ thuật

Thử nghiệm trên các kịch bản thực tế cho thấy, kết quả tích cực ở nhiều chỉ tiêu kỹ thuật then chốt. Bảng 1 tổng hợp các chỉ tiêu hiệu năng đo được

Hình 2: Sơ đồ kiến trúc hệ thống tổng thể



trong giai đoạn thực nghiệm. Kết quả cho thấy, độ chính xác trung bình 92% là mức chấp nhận được cho môi trường giáo dục, trong đó các câu hỏi nằm trọn trong tài liệu đạt 100%, các câu hỏi cần suy luận liên đoạn đạt khoảng 90%. Tỷ lệ không xảy ra ảo giác AI đạt 100% nhờ cơ chế Prompt Engineering ràng buộc mô hình chỉ trả lời dựa trên ngữ cảnh được cung cấp. Đặc biệt, khi truy vấn trùng bộ nhớ đệm, thời gian phản hồi giảm từ ~89

giây xuống dưới 0,5 giây, tương đương mức tiết kiệm 99,7% thời gian tính toán.

4.2. Hiệu quả quản lý đối với giảng viên và sinh viên

Kết quả khảo sát và đo lường khối lượng công việc trong 8 tuần thực nghiệm cho thấy tác động tích cực rõ rệt đến cả hai nhóm đối tượng. Đối với giảng viên, số lượng câu hỏi của sinh viên gửi qua email và tin nhắn ngoài giờ giảm 67% so với cùng kỳ học trước, từ trung bình 45 câu/tuần xuống còn 15 câu/tuần. Phần lớn câu hỏi còn lại là các vấn đề chuyên sâu cần tư vấn cá nhân hoặc các tình huống phức tạp ngoài phạm vi tài liệu. Đối với sinh viên, điểm hài lòng trung bình về dịch vụ hỗ trợ học tập - đo theo thang Likert 5 điểm, tăng từ 3,7 trước khi triển khai lên 4,47 sau khi

4.3. Phân tích chi phí sở hữu và hiệu quả kinh tế

Phân tích chi phí - lợi ích là công cụ quan trọng giúp các nhà quản lý giáo dục ra quyết định đầu tư công nghệ có căn cứ. Trong bối cảnh ngân sách công hạn chế tại các trường đại học và cao đẳng Việt Nam,

Bảng 1. Hiệu năng kỹ thuật của hệ thống

Chỉ tiêu	Kết quả đo lường	Ghi chú
Độ chính xác phản hồi	90 - 100%	Trung bình 92% trên 200 câu hỏi kiểm thử
Tỷ lệ không ảo giác AI	100%	Nhờ cơ chế lệnh ràng buộc nguồn
Thời gian phản hồi (LLM trực tiếp)	~89 giây	Đo trên máy chủ nội bộ của HCE
Thời gian phản hồi	< 0,5 giây	Tiết kiệm 99,7% thời gian tính toán
Tỷ lệ trùng bộ đệm cao điểm	> 80%	Đo trong giai đoạn ôn thi cuối kỳ

các quyết định đầu tư công nghệ cần được đánh giá dựa trên tổng chi phí sở hữu (TCO), lợi ích định lượng được như giảm chi phí vận hành và tăng năng suất, cùng các lợi ích định tính như cải thiện trải nghiệm học tập và nâng cao uy tín nhà trường. Mô hình LLM cục bộ (Local LLM) nổi lên như một lựa chọn chiến lược phù hợp với bài toán này khi kết hợp được ưu điểm về chi phí của phần mềm mã nguồn mở với khả năng AI tiên tiến.

Dưới góc độ quản lý kinh tế, nghiên cứu so sánh 2 phương án triển khai trong vòng 3 năm cho một trường có 9.000 sinh viên. Mỗi ngày trung bình có 1.000 câu hỏi, mỗi truy vấn gồm Prompt nội dung RAG và câu trả lời tốn khoảng 2.000 token. Bảng 2 trình bày kết quả so sánh chi tiết.

Bảng 2. So sánh hiệu quả kinh tế giữa Local LLM và Cloud API

Tiêu chí	Phương án Local LLM (Ollama)	Phương án Cloud API (Commercial LLM)
Đầu tư ban đầu	130.000.000 VNĐ	0 VNĐ
Chi phí vận hành/năm	~36.000.000 VNĐ (điện, bảo trì)	~100.000.000 VNĐ (phí API)
Tổng chi phí 3 năm	~238.000.000 VNĐ	~300.000.000 VNĐ
Bảo mật dữ liệu	Tuyệt đối do lưu trữ nội bộ	Rủi ro lộ lọt thông tin
Chi phí biên mỗi câu hỏi	0 VNĐ	Tăng theo số lượng sinh viên

Phân tích cho thấy dù Local LLM yêu cầu đầu tư ban đầu cao, nhưng về dài hạn, khi lượng sinh viên và tần suất truy vấn tăng lên, hệ thống nội bộ mang lại hiệu quả kinh tế vượt trội. Tổng tiết kiệm trong 3 năm đạt khoảng 62 triệu đồng so với phương án Cloud API, chưa kể phần tiết kiệm gián tiếp từ việc giảm tải khối lượng công việc của giảng viên - khoảng 100 triệu đồng/năm dựa trên đơn giá giờ giảng quy đổi. Đặc biệt, chi phí biên gần như bằng 0 ở góc độ phí API, chưa tính điện, bảo trì và khấu hao phần cứng, cho phép nhà trường cung cấp dịch vụ hỗ trợ 24/7 mà không lo ngại về ngân sách biến động.

4.4. Khả năng chịu tải và tính bền vững

Trong các giai đoạn cao điểm như đầu học kỳ khi đăng ký môn học hay giai đoạn ôn thi cuối kỳ,

tỷ lệ trùng bộ nhớ đệm dự kiến đạt trên 80% do sinh viên thường hỏi các câu tương tự nhau. Điều này giúp hệ thống duy trì thời gian phản hồi ổn định dưới 5 giây ngay cả khi có hàng chục yêu cầu đồng thời, tránh hiện tượng nghẽn cổ chai phần cứng và giữ được trải nghiệm người dùng trong điều kiện tải cao.

4.5. Thảo luận kết quả

Kết quả nghiên cứu cung cấp bằng chứng thực nghiệm quan trọng cho các nhà quản lý giáo dục trong việc ra quyết định ứng dụng AI. Mức giảm 67% câu hỏi ngoài giờ không chỉ là con số thống kê đơn thuần mà phản ánh sự thay đổi thực sự trong mô hình vận hành hỗ trợ học tập. Khi giảng viên được giải phóng khỏi các câu hỏi lặp lại mang tính thủ

tục, họ có thể tập trung vào các hoạt động có giá trị học thuật cao hơn như thiết kế bài giảng, nghiên cứu khoa học và tư vấn học thuật chuyên sâu. Đây là sự dịch chuyển quan trọng trong phân bổ nguồn lực con người tại các cơ sở giáo dục.

Từ góc độ quản lý chất lượng, tỷ lệ không xảy ra ảo giác AI đạt 100% trong môi trường kiểm soát là kết quả đặc biệt có ý nghĩa. Điều này cho thấy, khi được thiết kế đúng cách với kiến trúc RAG, chatbot AI có thể đảm bảo tính chính xác của thông tin học thuật ở mức độ chấp nhận được cho môi trường giáo dục. Đây là điều kiện tiên quyết để các nhà quản lý giáo dục tự tin phê duyệt triển khai AI trong môi trường học thuật chính thức.

Từ góc độ kinh tế giáo dục, nghiên cứu đề xuất

một mô hình triển khai 3 giai đoạn phù hợp với bối cảnh các trường công lập Việt Nam. Giai đoạn 1 (năm đầu): đầu tư phần cứng tập trung và triển khai thí điểm trên 2-3 học phần có lượng sinh viên lớn để đạt hiệu quả cache cao nhất. Giai đoạn 2 (năm thứ hai): mở rộng sang toàn bộ các học phần trực tuyến, tận dụng hạ tầng sẵn có. Giai đoạn 3 (từ năm thứ ba): tích hợp thêm tính năng phân tích học tập và phát triển mô hình liên trường để chia sẻ hạ tầng, giảm thêm chi phí đầu tư bình quân.

Mô hình liên trường đặc biệt phù hợp với bối cảnh Việt Nam, khi nhiều trường cao đẳng và đại học địa phương có quy mô nhỏ, ngân sách hạn chế nhưng có thể cùng chia sẻ một máy chủ AI dùng chung và phân chia chi phí theo số lượng sinh viên. Đây là hướng đi có thể nhân rộng mô hình lên cấp độ tỉnh, thành phố hoặc theo cụm trường cùng ngành đào tạo.

5. Kết luận

Nghiên cứu này đã cung cấp bằng chứng toàn diện về hiệu quả quản lý, kinh tế và kỹ thuật của việc ứng dụng chatbot AI tích hợp LMS Moodle trong bối cảnh chuyển đổi số giáo dục tại Việt Nam. 3 kết quả cốt lõi gồm: (1) giảm 67% câu hỏi ngoài giờ cho giảng viên và cải thiện 20,8% điểm hài lòng của sinh viên; (2) tiết kiệm khoảng 62 triệu đồng TCO trong 3 năm so với giải pháp Cloud API, chưa kể giá trị tiết kiệm gián tiếp từ thời gian giảng viên; (3) đạt độ chính xác trung bình 92% và tỷ lệ không xảy ra ảo

giác AI 100%, với thời gian phản hồi dưới 0,5 giây khi trùng cache.

Điểm đặc biệt quan trọng trong bối cảnh chính sách giáo dục quốc gia là mô hình hoàn toàn cục bộ, không phụ thuộc vào nhà cung cấp nước ngoài, góp phần tăng cường khả năng kiểm soát dữ liệu giáo dục, giảm phụ thuộc vào nền tảng AI thương mại bên ngoài và phù hợp với định hướng tự chủ công nghệ trong chuyển đổi số giáo dục - trong chiến lược chuyển đổi số quốc gia. Nghiên cứu này có thể là tài liệu tham khảo hữu ích để các nhà quản lý giáo dục và hoạch định chính sách đưa ra quyết định đầu tư công nghệ AI có căn cứ khoa học, góp phần thúc đẩy chuyển đổi số giáo dục Việt Nam một cách bền vững và hiệu quả.

Nghiên cứu có một số hạn chế cần được thừa nhận. Thứ nhất, quy mô thực nghiệm còn giới hạn ở một học phần và một trường, cần mở rộng để có tính đại diện cao hơn. Thứ hai, dữ liệu hiệu quả quản lý được thu thập trong 8 tuần, chưa đủ để đánh giá tác động dài hạn đến kết quả học tập của sinh viên. Thứ ba, phân tích chi phí - lợi ích chưa tính đến chi phí đào tạo nhân lực kỹ thuật nội bộ để vận hành và bảo trì hệ thống. Về hướng nghiên cứu tiếp theo, cần có nghiên cứu dọc đánh giá tác động của chatbot AI đến kết quả học tập như điểm thi và tỷ lệ hoàn thành học phần, đồng thời thực hiện nghiên cứu so sánh giữa các trường triển khai và chưa triển khai để đánh giá lợi thế cạnh tranh ■

TÀI LIỆU THAM KHẢO:

Thủ tướng Chính phủ (2022). Quyết định số 131/QĐ-TTg ngày 25/01/2022 phê duyệt Đề án “Tăng cường ứng dụng công nghệ thông tin và chuyển đổi số trong giáo dục và đào tạo giai đoạn 2022-2025, định hướng đến năm 2030”. Hà Nội.

Bang, F. (2023). GPtCache: An open-source semantic cache for LLM applications enabling faster answers and cost savings. *Proceedings of NLP-OSS 2023*, 212-218.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75-105.

Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.

Moodle (2025). Moodle documentation. [Online] Available at <https://docs.moodle.org>

Ollama (2025). Ollama documentation. [Online] Available at <https://docs.ollama.com>

OpenAI (2025). API pricing. [Online] Available at <https://openai.com/api/pricing>

UNESCO (2023). Global Education Monitoring Report 2023: Technology in education: A tool on whose terms. UNESCO Publishing. [Online] Available at <https://unesdoc.unesco.org/ark:/48223/pf0000385723>.

Winkler, R., & Söllner, M. (2018). Unleashing the potential of chatbots in education: A state-of-the-art analysis. Academy of Management Proceedings, 2018(1), 15903.

Ngày nhận bài: 10/3/2026

Ngày phản biện đánh giá và sửa chữa: 25/3/2026

Ngày chấp nhận đăng bài: 15/4/2026

ENHANCING EDUCATIONAL EFFECTIVENESS THROUGH AI CHATBOT INTEGRATION: AN EXPERIMENTAL STUDY WITHIN THE MOODLE LEARNING MANAGEMENT SYSTEM (LMS)

● **TRUONG TAN LOC^{1,2}**

● **NGUYEN DUY THANH²**

¹ Ho Chi Minh City College of Economics

² Ho Chi Minh University of Banking

ABSTRACT:

This study proposes an AI chatbot solution based on Retrieval-Augmented Generation (RAG) technology deployed on a local server using a Local Large Language Model (Local LLM) and directly integrated into the Moodle Learning Management System (LMS). The primary contribution of the research is the development of the Hybrid Semantic Cache (HSC) algorithm, which reduces response time by 99.7% while effectively lowering the marginal cost per query to nearly zero. Experimental results involving 320 students demonstrate that the system achieved an average accuracy rate of 92%, with no instances of AI hallucination detected in the evaluation questions. In addition, the locally deployed architecture significantly enhances institutional data control and privacy protection. The findings provide a technically feasible and economically sustainable model for universities and colleges seeking to implement large-scale virtual teaching assistants with minimal operational costs.

Keywords: digital transformation, LMS management, RAG chatbot, economic efficiency.