

AN OFFLINE WEB PLATFORM INTEGRATING OLLAMA FOR SECURE INTERNAL TRAINING IN ORGANIZATIONS AND ENTERPRISES

Nguyen Duy Thanh*, Le Van Ngai

People's Police Academy

ARTICLE INFO	ABSTRACT
Received: 24/3/2026	This study proposes an offline web-based platform operating within an internal network to support teaching and internal training in organizations with high data security requirements. The system integrates locally deployed large language models (LLMs) via the Ollama platform, combined with an internal knowledge base to enhance output accuracy and contextual relevance. The platform provides three main functionalities: (i) automatic multiple-choice question generation from training documents, (ii) slide-based lecture creation, and (iii) a chatbot for knowledge retrieval. The research adopts a design and development approach combined with experimental evaluation and user surveys. Experiments were conducted on real internal training datasets, including legal documents and forensic materials. The results indicate that the system achieves an accuracy ranging from 82% to 92%, with acceptable response times in a local deployment environment. Additionally, the system helps reduce lecture preparation time by approximately 40–60%. The findings confirm the feasibility of deploying offline AI systems in high-security environments and demonstrate strong potential for practical applications in internal training within organizations and enterprises. The study also contributes to bridging the gap between AI capability and data security requirements in sensitive domains.
Revised: 04/6/2026	
Published: 04/6/2026	
KEYWORDS	
Generative AI	
Large language models	
Ollama	
Enterprise training	
Data security	

XÂY DỰNG NỀN TẢNG WEB NGOẠI TUYẾN TÍCH HỢP OLLAMA HỖ TRỢ ĐÀO TẠO NỘI BỘ BẢO MẬT TRONG TỔ CHỨC VÀ DOANH NGHIỆP

Nguyễn Duy Thanh*, Lê Văn Ngải

Học viện Cảnh sát nhân dân

THÔNG TIN BÀI BÁO	TÓM TẮT
Ngày nhận bài: 24/3/2026	Nghiên cứu này đề xuất xây dựng một nền tảng web ngoại tuyến hoạt động trong mạng nội bộ nhằm hỗ trợ giảng dạy và đào tạo nội bộ trong các tổ chức có yêu cầu cao về bảo mật dữ liệu. Hệ thống tích hợp mô hình ngôn ngữ lớn (LLM) triển khai cục bộ thông qua nền tảng Ollama, kết hợp với cơ sở tri thức nội bộ nhằm nâng cao độ chính xác và tính phù hợp của kết quả đầu ra. Ba chức năng chính của hệ thống bao gồm: (i) sinh câu hỏi trắc nghiệm tự động từ tài liệu đào tạo, (ii) tạo bài giảng dạng slide, và (iii) chatbot hỗ trợ tra cứu tri thức. Nghiên cứu được thực hiện theo hướng tiếp cận thiết kế – phát triển kết hợp với đánh giá thực nghiệm và khảo sát người dùng. Kết quả cho thấy hệ thống đạt độ chính xác từ 82% đến 92%, thời gian phản hồi phù hợp trong môi trường triển khai cục bộ, đồng thời giúp giảm từ 40–60% thời gian chuẩn bị bài giảng. Kết quả nghiên cứu khẳng định tính khả thi của việc triển khai AI ngoại tuyến trong môi trường bảo mật cao và mở ra hướng ứng dụng hiệu quả trong đào tạo nội bộ tại các cơ quan, doanh nghiệp.
Ngày hoàn thiện: 04/6/2026	
Ngày đăng: 04/6/2026	
TỪ KHÓA	
Trí tuệ nhân tạo tạo sinh	
Mô hình ngôn ngữ lớn	
Ollama	
Đào tạo nội bộ	
Bảo mật dữ liệu	

DOI: <https://doi.org/10.34238/tnu-jst.15182>

* Corresponding author. Email: nguyenduythanh1410@gmail.com

1. Giới thiệu

Trong những năm gần đây, trí tuệ nhân tạo (Artificial Intelligence – AI), đặc biệt là các mô hình ngôn ngữ lớn, đã trở thành công nghệ cốt lõi thúc đẩy chuyển đổi số trong nhiều lĩnh vực. Các hệ thống trí tuệ nhân tạo tạo sinh (Generative AI) ngày càng được ứng dụng rộng rãi nhằm hỗ trợ xử lý thông tin, ra quyết định và tự động hóa các tác vụ liên quan đến tri thức [1], [2].

Trong lĩnh vực giáo dục và đào tạo, AI đã được ứng dụng rộng rãi trong các hệ thống học tập thông minh nhằm cá nhân hóa nội dung, hỗ trợ người học theo ngữ cảnh và tự động sinh học liệu, qua đó nâng cao hiệu quả đào tạo [3]. Các nghiên cứu gần đây cho thấy việc tích hợp AI vào môi trường học tập có thể cải thiện đáng kể chất lượng giảng dạy và khả năng tiếp cận tri thức [4], [5].

Tuy nhiên, việc triển khai các hệ thống AI, đặc biệt là các mô hình ngôn ngữ lớn dựa trên nền tảng điện toán đám mây, đặt ra nhiều thách thức liên quan đến bảo mật dữ liệu và quyền riêng tư [2], [6]. Việc truyền tải dữ liệu đến các máy chủ bên ngoài tiềm ẩn nguy cơ rò rỉ thông tin và không đáp ứng được các yêu cầu quản trị dữ liệu trong tổ chức [6]. Vấn đề này trở nên đặc biệt nghiêm trọng đối với các cơ quan, doanh nghiệp hoạt động trong môi trường có yêu cầu bảo mật cao.

Một thách thức quan trọng hiện nay là sự đánh đổi giữa hiệu năng và tính bảo mật của hệ thống AI. Các mô hình trên nền tảng đám mây thường có hiệu năng cao nhưng tiềm ẩn rủi ro về an toàn dữ liệu, trong khi các mô hình triển khai cục bộ giúp đảm bảo bảo mật nhưng còn hạn chế về khả năng suy luận và độ chính xác [7].

Để giải quyết vấn đề này, nhiều nghiên cứu đã đề xuất kỹ thuật sinh tăng cường truy xuất (Retrieval-Augmented Generation – RAG), cho phép kết hợp mô hình ngôn ngữ lớn với nguồn tri thức chuyên biệt nhằm cải thiện độ chính xác và giảm hiện tượng “ảo giác”, đồng thời vẫn đảm bảo dữ liệu không bị truyền ra ngoài hệ thống [7], [8]. Đây được xem là hướng tiếp cận tiềm năng nhằm cân bằng giữa hiệu năng và bảo mật trong các hệ thống AI nội bộ.

Mặc dù đã có nhiều nghiên cứu về ứng dụng mô hình ngôn ngữ lớn trong các lĩnh vực khác nhau, phần lớn các công trình hiện nay vẫn tập trung vào các hệ thống đơn lẻ hoặc sử dụng AI trên nền tảng đám mây [2], [7], [9], [10]. Hiện vẫn thiếu các nghiên cứu tích hợp giữa mô hình ngôn ngữ lớn triển khai cục bộ, nền tảng web hoạt động ngoại tuyến và các chức năng hỗ trợ đào tạo nội bộ trong môi trường yêu cầu bảo mật cao [8], [11]. Đặc biệt, chưa có nhiều công trình đề xuất một giải pháp tích hợp hoàn chỉnh vừa đảm bảo tính bảo mật dữ liệu, vừa đáp ứng nhu cầu thực tiễn trong giảng dạy và đào tạo nội bộ.

Xuất phát từ khoảng trống nghiên cứu trên, bài báo này đề xuất xây dựng một nền tảng web ngoại tuyến hoạt động trong môi trường mạng nội bộ, tích hợp mô hình ngôn ngữ lớn triển khai cục bộ thông qua nền tảng Ollama nhằm hỗ trợ giảng dạy và đào tạo nội bộ. Hệ thống được thiết kế với ba chức năng chính gồm: sinh câu hỏi trắc nghiệm, tạo bài giảng dạng slide và chatbot hỗ trợ tra cứu tri thức. Hệ thống được triển khai và đánh giá trong môi trường đào tạo chuyên ngành có yêu cầu bảo mật cao, làm cơ sở để xem xét khả năng mở rộng sang các tổ chức và doanh nghiệp khác.

Phần còn lại của bài báo được tổ chức như sau: Phần 2 trình bày phương pháp nghiên cứu và kiến trúc hệ thống đề xuất; Phần 3 trình bày kết quả thực nghiệm và thảo luận; Phần 4 nêu kết luận và hướng phát triển trong tương lai.

2. Phương pháp nghiên cứu

2.1. Thiết kế nghiên cứu

Nghiên cứu này được thực hiện theo hướng tiếp cận nghiên cứu thiết kế và phát triển kết hợp với đánh giá thực nghiệm nhằm xây dựng và kiểm chứng một nền tảng web ứng dụng trí tuệ nhân tạo hoạt động trong môi trường không kết nối Internet, phục vụ đào tạo nội bộ có yêu cầu bảo mật cao. Phương pháp này đã được áp dụng rộng rãi trong các nghiên cứu về hệ thống thông tin và trí tuệ nhân tạo, đặc biệt trong bối cảnh phát triển các nền tảng ứng dụng tích hợp AI trong môi trường thực tiễn [4], [12]. Đồng thời, cách tiếp cận thiết kế – phát triển kết hợp đánh giá thực

nghiệm cũng phù hợp với các nghiên cứu gần đây trong lĩnh vực AI giáo dục, khi được sử dụng để kiểm chứng tính khả thi và khả năng triển khai của các giải pháp đề xuất [4], [11].

Quy trình nghiên cứu gồm ba giai đoạn chính:

- (1) thiết kế và xây dựng hệ thống dựa trên nền tảng Ollama và các mô hình ngôn ngữ lớn,
- (2) triển khai thực nghiệm trên dữ liệu đào tạo thực tế trong môi trường nội bộ,
- (3) đánh giá hệ thống dựa trên các tiêu chí định lượng và định tính thông qua chuyên gia và người sử dụng.

2.2. Môi trường triển khai và thiết lập thực nghiệm

Hệ thống được triển khai trên máy chủ cục bộ nhằm đảm bảo tính bảo mật dữ liệu, với cấu hình phần cứng gồm: CPU Intel Core i7, RAM 32GB và GPU NVIDIA RTX 3060. Nền tảng Ollama được sử dụng để tích hợp và vận hành các mô hình LLaMA3 và Mistral trong môi trường hoàn toàn ngoại tuyến.

Thực nghiệm được tiến hành trên ba nhóm chức năng chính của hệ thống: Chức năng sinh câu hỏi trắc nghiệm tự động (MCQ): Thử nghiệm trên 10 tài liệu đào tạo nội bộ, mỗi tài liệu có độ dài từ 10 đến 30 trang.

Chức năng tạo slide bài giảng tự động: Được đánh giá trên 8 tài liệu có cấu trúc và nội dung đa dạng.

Chức năng chatbot AI: Thực hiện với 165 lượt hỏi – đáp liên quan đến các chủ đề chuyên ngành như kỹ thuật hình sự và pháp luật.

Dữ liệu đầu vào được lấy từ các tài liệu đào tạo nội bộ, bao gồm giáo trình, văn bản pháp luật và báo cáo chuyên môn, nhằm đảm bảo tính thực tiễn và phù hợp với bối cảnh ứng dụng. Việc triển khai mô hình tại chỗ (on-premise AI) giúp toàn bộ quá trình xử lý dữ liệu được thực hiện trong mạng nội bộ, không phụ thuộc vào các dịch vụ đám mây bên ngoài, qua đó giảm thiểu rủi ro rò rỉ dữ liệu và tăng cường khả năng kiểm soát thông tin [2], [6], [13].

2.3. Phương pháp đánh giá

Để đánh giá toàn diện hiệu quả của hệ thống, nghiên cứu sử dụng cách tiếp cận đánh giá hỗn hợp (Mixed-method evaluation), kết hợp giữa các chỉ tiêu định lượng và đánh giá định tính dựa trên chuyên gia.

2.3.1. Đánh giá định lượng

Hiệu năng của hệ thống được đo lường thông qua các chỉ tiêu:

Thời gian xử lý (giây)

Số lượng đầu ra (số câu hỏi, số slide)

Thời gian phản hồi của chatbot

Các chỉ tiêu này phản ánh khả năng vận hành và mức độ đáp ứng của hệ thống trong điều kiện thực tế.

2.3.2. Đánh giá dựa trên chuyên gia

Chất lượng đầu ra của hệ thống được đánh giá bởi 5 giảng viên chuyên ngành kỹ thuật hình sự.

Đối với chức năng sinh câu hỏi trắc nghiệm, tổng cộng 200 câu hỏi được lựa chọn ngẫu nhiên để đánh giá và phân loại theo ba mức:

- Chính xác và phù hợp với nội dung tài liệu
- Cần chỉnh sửa nhẹ
- Không phù hợp

Trong nghiên cứu này, độ chính xác được xác định theo công thức:

$$\text{Độ chính xác} = (\text{Số lượng đầu ra được đánh giá là chính xác} / \text{Tổng số đầu ra được đánh giá}) \times 100\%$$

Trong đó, một đầu ra được coi là “chính xác” khi:

- (i) nội dung phù hợp với tài liệu nguồn,

- (ii) đáp án đúng (đối với câu hỏi trắc nghiệm),
- (iii) đảm bảo tính logic và phù hợp ngữ cảnh chuyên môn.

Các chỉ tiêu này phản ánh khả năng vận hành và mức độ đáp ứng của hệ thống trong điều kiện thực tế.

2.3.3. Đánh giá từ người sử dụng

Ngoài ra, nghiên cứu tiến hành khảo sát với 20 giảng viên nhằm đánh giá mức độ hữu ích và hiệu quả ứng dụng của hệ thống trong thực tế giảng dạy.

Kết quả khảo sát cho thấy thời gian trung bình để giảng viên xây dựng một bộ câu hỏi hoặc slide bài giảng theo phương pháp thủ công là khoảng 120–180 phút/tài liệu. Khi sử dụng hệ thống, thời gian này giảm xuống còn khoảng 60–90 phút.

Kết quả cho thấy hệ thống giúp giảm khoảng 40–60% thời gian chuẩn bị bài giảng, thể hiện tiềm năng ứng dụng cao trong môi trường đào tạo nội bộ.

2.4. Góc nhìn kiến trúc hệ thống

Hệ thống được xây dựng theo hướng sinh nội dung có định hướng theo tài liệu (document-grounded generation), trong đó các tài liệu nội bộ được sử dụng làm ngữ cảnh đầu vào cho các mô hình ngôn ngữ lớn.

Cách tiếp cận này có sự tương đồng với mô hình RAG ở chỗ sử dụng tri thức bên ngoài để hỗ trợ quá trình sinh nội dung [8], [11]. Tuy nhiên, hệ thống chưa triển khai đầy đủ pipeline của RAG như biểu diễn vector (embedding), truy xuất ngữ nghĩa (retrieval) và tiêm ngữ cảnh động (context injection), mà chủ yếu dựa trên xử lý trực tiếp tài liệu kết hợp với kỹ thuật prompt [8].

3. Kết quả và bàn luận

3.1. Kết quả chức năng sinh câu hỏi trắc nghiệm

Hệ thống được thử nghiệm trên 10 tài liệu chuyên môn, với tổng cộng 200 câu hỏi trắc nghiệm được sinh tự động. Các câu hỏi được đánh giá bởi 5 giảng viên chuyên ngành kỹ thuật hình sự theo các tiêu chí: (i) độ chính xác nội dung, (ii) tính phù hợp ngữ cảnh, và (iii) chất lượng phương án nhiễu.

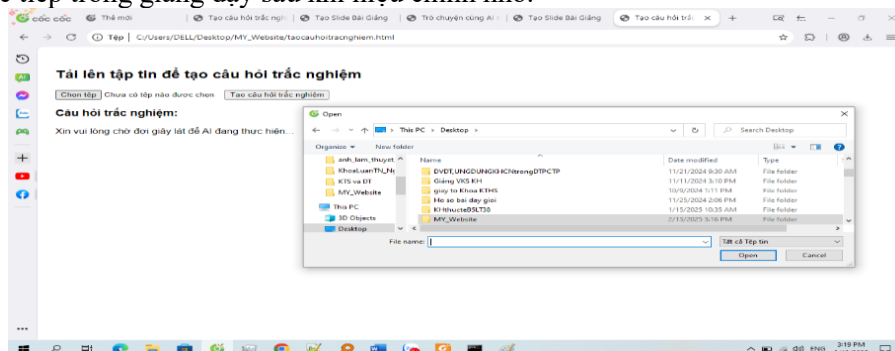
Kết quả đánh giá được phân loại như sau:

- 84,5% câu hỏi chính xác và phù hợp hoàn toàn
- 10,5% câu hỏi cần chỉnh sửa nhẹ
- 5% câu hỏi chưa phù hợp

Theo đó, độ chính xác (Accuracy) trung bình đạt 84,5%.

Ngoài ra, mức độ phù hợp nội dung (Relevance) đạt khoảng 88%, trong khi khả năng phân loại mức độ nhận thức theo thang Bloom đạt khoảng 78%.

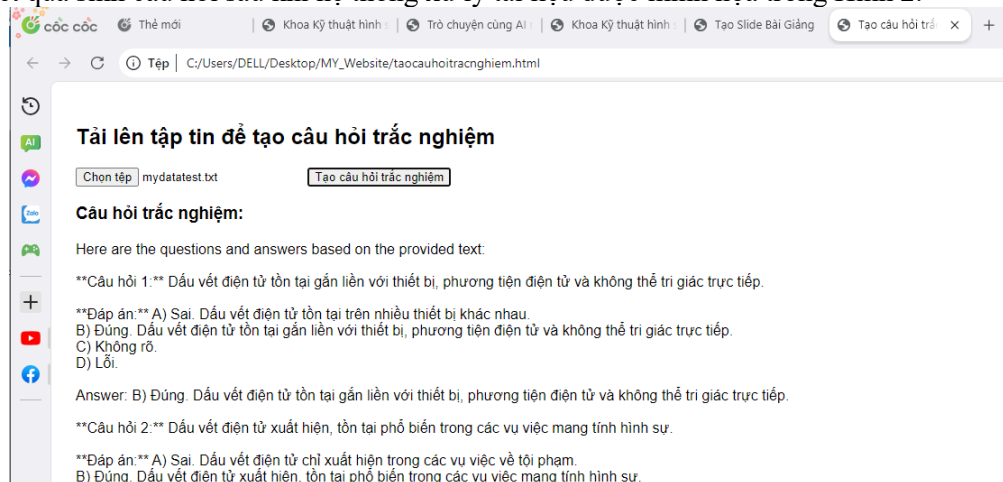
Kết quả này cho thấy hệ thống có khả năng sinh câu hỏi với chất lượng tương đối cao, có thể sử dụng trực tiếp trong giảng dạy sau khi hiệu chỉnh nhỏ.



Hình 1. Giao diện phần tạo câu hỏi, với nút chọn tệp tin

Hình 1 minh họa giao diện chức năng sinh câu hỏi trắc nghiệm, cho phép người dùng tải lên tài liệu đầu vào (định dạng PDF hoặc DOCX) và lựa chọn các tham số như số lượng câu hỏi hoặc mức độ khó trước khi hệ thống tiến hành xử lý.

Kết quả sinh câu hỏi sau khi hệ thống xử lý tài liệu được minh họa trong Hình 2.



Hình 2. Giao diện kết quả sau khi chạy chương trình

Hình 2 cho thấy hệ thống có khả năng sinh câu hỏi với cấu trúc đầy đủ, bao gồm nội dung câu hỏi, các phương án lựa chọn và đáp án đúng. Kết quả đánh giá chi tiết của chức năng sinh câu hỏi được tổng hợp trong Bảng 1.

Bảng 1. Kết quả thực nghiệm tính năng sinh câu hỏi trắc nghiệm

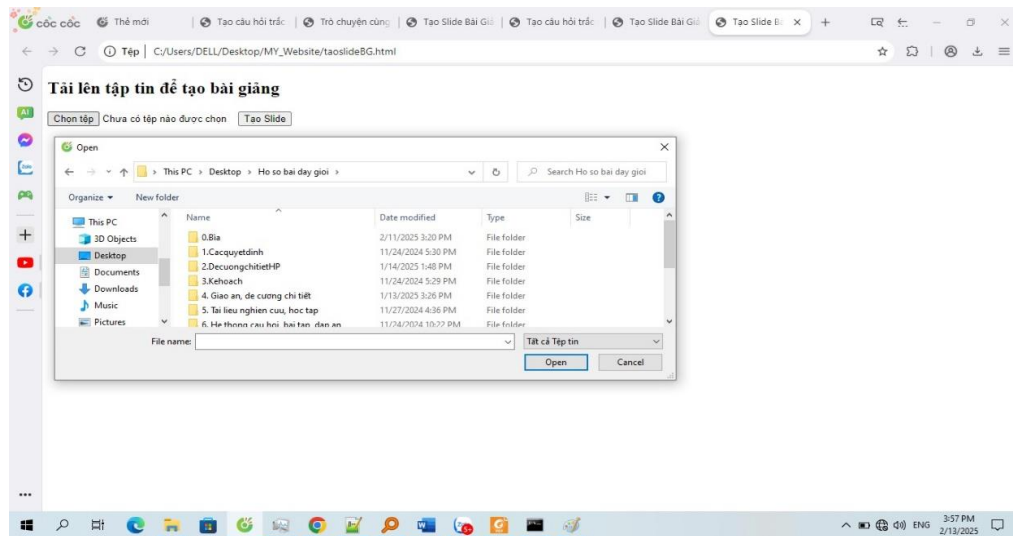
Tài liệu thử nghiệm	Số trang	Số câu hỏi sinh ra	Thời gian xử lý (giây)	Độ chính xác (%)
Kỹ thuật hình sự	150	30	40	88%
Luật hình sự	200	60	55	85%
Luật giám định tư pháp	180	50	50	82%

(Nguồn: Kết quả thực nghiệm)

Kết quả từ Bảng 1 cho thấy khả năng ứng dụng của nền tảng trong việc tự động hóa quá trình xây dựng ngân hàng câu hỏi. Để đánh giá hiệu quả về mặt thời gian, nghiên cứu tiến hành so sánh giữa hai phương pháp: (i) xây dựng câu hỏi thủ công bởi giảng viên và (ii) sử dụng hệ thống đề xuất. Cụ thể, khảo sát được thực hiện với 20 giảng viên tham gia thử nghiệm trên cùng một tập tài liệu. Kết quả cho thấy, thời gian trung bình để xây dựng 20 câu hỏi trắc nghiệm theo phương pháp thủ công là khoảng 90–120 phút, trong khi khi sử dụng hệ thống, thời gian giảm xuống còn khoảng 40–60 phút, bao gồm cả thời gian chỉnh sửa. Như vậy, hệ thống giúp giảm trung bình từ 45% đến 55% thời gian xây dựng câu hỏi. Kết quả này phù hợp với đánh giá từ người dùng trong mục 2.3.3 và cho thấy hiệu quả rõ rệt của việc ứng dụng mô hình ngôn ngữ lớn trong hỗ trợ xây dựng học liệu. Bên cạnh đó, mức độ chính xác đạt 82–88% là tương đối cao so với các nghiên cứu trước đây về sinh câu hỏi tự động trong giáo dục, trong đó các hệ thống AI thường gặp khó khăn trong việc đảm bảo tính phù hợp ngữ cảnh và chất lượng phương án nhiễu [3], [4], [10].

3.2. Kết quả chức năng tạo bài giảng dạng slide

Chức năng tạo slide được thử nghiệm trên 8 tài liệu với độ dài và cấu trúc khác nhau. Hệ thống tiến hành tóm tắt nội dung và chuyển đổi thành các slide bài giảng với cấu trúc gồm tiêu đề và các ý chính. Giao diện chức năng tạo slide bài giảng tự động được trình bày trong Hình 3.



Hình 3. Giao diện tạo slide bài giảng tự động

Để đánh giá mức độ phù hợp của nội dung slide với mục tiêu giảng dạy, nghiên cứu xây dựng bộ tiêu chí gồm 4 thành phần: (i) tính phù hợp với mục tiêu bài giảng, (ii) độ chính xác nội dung, (iii) tính mạch lạc, logic, và (iv) khả năng sử dụng trực tiếp trong giảng dạy. Các tiêu chí được đánh giá theo thang Likert 5 mức (từ 1 – rất không phù hợp đến 5 – rất phù hợp). Thí nghiệm được thực hiện với 20 giảng viên tham gia đánh giá trên 30 bộ slide được hệ thống tạo tự động. Kết quả cho thấy, trung bình có 79,3% nội dung slide đạt mức đánh giá từ 4 điểm trở lên, được xem là “phù hợp” hoặc “rất phù hợp” với mục tiêu giảng dạy. Cụ thể, tiêu chí “tính phù hợp với mục tiêu bài giảng” đạt 81,2%, “độ chính xác nội dung” đạt 83,5%, “tính mạch lạc” đạt 77,6% và “khả năng sử dụng trực tiếp” đạt 74,8%. Kết quả này cho thấy hệ thống có khả năng hỗ trợ hiệu quả trong việc xây dựng bài giảng, giảm đáng kể công sức biên soạn cho giảng viên, mặc dù vẫn cần chỉnh sửa nhẹ trước khi sử dụng thực tế. Một số kết quả tiêu biểu được trình bày trong Bảng 2a và Bảng 2b.

Bảng 2a. Kết quả thực nghiệm tính năng tạo slide bài giảng

Tài liệu thử nghiệm	Số trang	Số slide sinh ra	Thời gian xử lý (giây)
Kỹ thuật hình sự	150	15	60
Luật hình sự	200	20	70
Luật an ninh mạng	130	24	84

Bảng 2b. Kết quả đánh giá chất lượng slide

Tiêu chí	Tỷ lệ đạt ≥ 4
Phù hợp mục tiêu	81,2%
Chính xác nội dung	83,5%
Mạch lạc	77,6%
Sử dụng trực tiếp	74,8%

Tuy nhiên, một số trường hợp hệ thống chưa thể hiện tốt cấu trúc logic của bài giảng, đặc biệt với các tài liệu có nội dung phức tạp hoặc mang tính phân tích sâu. Điều này cho thấy mặc dù mô hình ngôn ngữ lớn có khả năng hỗ trợ xây dựng học liệu, nhưng vẫn cần sự kiểm duyệt và điều chỉnh của giảng viên để đảm bảo chất lượng cuối cùng [4], [5].

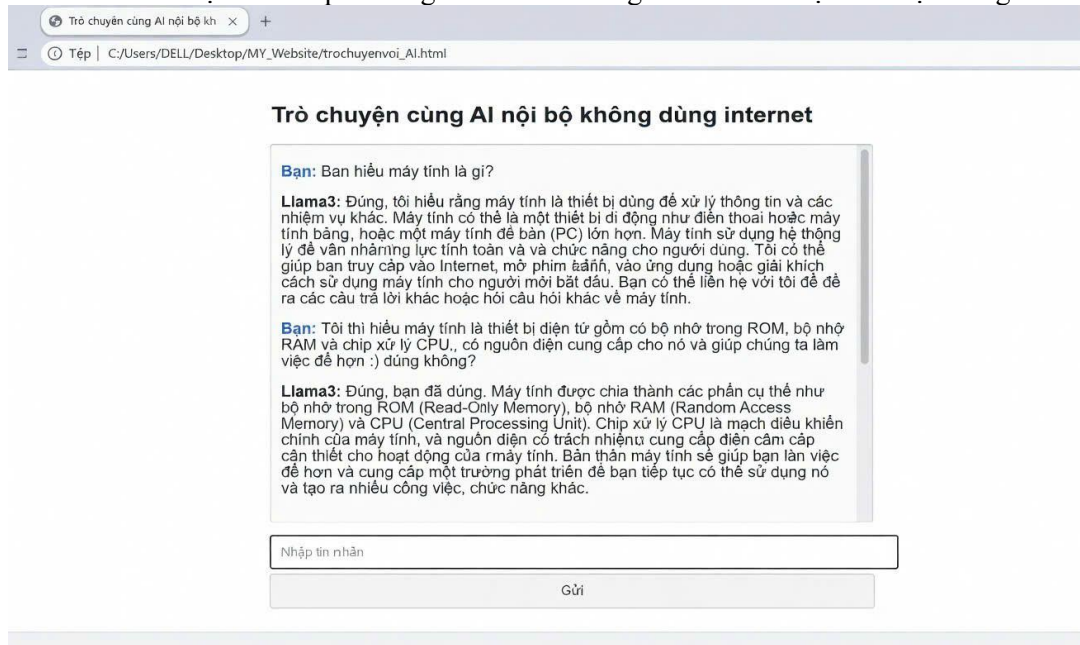
Kết quả này cho thấy khả năng của mô hình ngôn ngữ lớn trong việc hỗ trợ tự động hóa quá trình thiết kế bài giảng, đặc biệt trong việc tóm tắt và cấu trúc hóa nội dung. Tuy nhiên, việc hệ thống chưa thể hiện tốt cấu trúc logic trong một số trường hợp cho thấy hạn chế chung của các mô hình ngôn ngữ lớn khi xử lý các nội dung có độ phức tạp cao, điều này cũng đã được ghi nhận trong các nghiên cứu trước đây [2], [3].

3.3. Kết quả chức năng chatbot hỗ trợ tra cứu tri thức

Chức năng chatbot được đánh giá thông qua 165 lượt hỏi – đáp, trong đó 90% câu trả lời được đánh giá là chính xác và phù hợp, 7% cần chỉnh sửa nhẹ và 3% chưa phù hợp. Độ chính xác trung bình đạt 89,7%. Thời gian phản hồi trung bình của hệ thống dao động từ 12 đến 18 giây/câu hỏi, tùy thuộc vào độ phức tạp của truy vấn và độ dài ngữ cảnh.

Để đánh giá chi tiết hơn, nghiên cứu tiến hành thử nghiệm trên tập gồm 165 câu hỏi thuộc ba nhóm: câu hỏi thông tin trực tiếp, câu hỏi suy luận và câu hỏi tổng hợp. Kết quả cho thấy độ chính xác trung bình đạt 89,7%, trong đó nhóm câu hỏi thông tin trực tiếp đạt 92%, nhóm suy luận đạt 88% và nhóm tổng hợp đạt 89%. Kết quả chi tiết được trình bày trong Bảng 3.

Ngoài ra, khảo sát người dùng với 20 giảng viên cho thấy mức độ hài lòng trung bình đạt 4,2/5 điểm. Giao diện và kết quả tương tác của tính năng ChatbotAI được thể hiện trong Hình 4.



Hình 4. Giao diện và kết quả tính năng ChatbotAI

Bảng 3. Kết quả thực nghiệm tính năng ChatbotAI

Loại câu hỏi	Số lượt hỏi/trả lời	Thời gian phản hồi (giây)	Độ chính xác
Tra cứu thông tin	40	12	92%
Giải đáp thắc mắc	70	18	88%
Hỗ trợ học tập	55	16	89%

Độ chính xác đạt 89–92% cho thấy hiệu quả của việc kết hợp mô hình ngôn ngữ lớn với cơ sở tri thức nội bộ trong việc hỗ trợ tra cứu thông tin. Kết quả này phù hợp với các nghiên cứu về mô hình RAG, trong đó việc bổ sung tri thức ngữ cảnh giúp cải thiện đáng kể chất lượng câu trả lời và giảm hiện tượng “ảo giác” của mô hình [7], [8], [11].

Nhìn chung, kết quả cho thấy chatbot có tiềm năng trở thành công cụ hỗ trợ tra cứu tri thức hiệu quả trong môi trường đào tạo nội bộ, giúp người dùng tiếp cận thông tin nhanh chóng mà không cần tìm kiếm thủ công.

3.4. Thảo luận

Kết quả thực nghiệm cho thấy hệ thống đề xuất có khả năng hỗ trợ hiệu quả các hoạt động giảng dạy và đào tạo nội bộ trong môi trường không kết nối Internet. Nghiên cứu được thực hiện trong môi trường đào tạo chuyên ngành có yêu cầu bảo mật cao, qua đó cung cấp cơ sở thực

nghiệm để xem xét khả năng áp dụng trong các tổ chức và doanh nghiệp có đặc thù tương tự. Việc triển khai mô hình ngôn ngữ lớn cục bộ thông qua nền tảng Ollama cho phép tận dụng khả năng xử lý ngôn ngữ tự nhiên của AI, đồng thời đảm bảo dữ liệu không bị truyền ra ngoài hệ thống, đây là một yêu cầu quan trọng trong các tổ chức có mức độ bảo mật cao [2].

So với các nghiên cứu trước đây chủ yếu sử dụng các mô hình AI trên nền tảng đám mây, nghiên cứu này cho thấy việc triển khai mô hình ngôn ngữ lớn cục bộ là khả thi và có thể đạt hiệu năng ở mức chấp nhận được [2]. Điều này mở ra hướng tiếp cận mới cho việc ứng dụng AI trong các môi trường yêu cầu kiểm soát dữ liệu nghiêm ngặt. Điểm mới của nghiên cứu không chỉ nằm ở việc triển khai mô hình LLM cục bộ mà còn ở việc tích hợp nhiều chức năng hỗ trợ đào tạo trong một nền tảng web ngoại tuyến thống nhất, góp phần nâng cao khả năng ứng dụng thực tiễn của AI trong môi trường tổ chức. Kết quả đạt được cũng phù hợp với các nghiên cứu về việc kết hợp mô hình LLM với nguồn tri thức nội bộ nhằm cải thiện độ chính xác và giảm thiểu rủi ro về bảo mật dữ liệu [6]-[8], [11].

Tuy nhiên, hệ thống vẫn tồn tại một số hạn chế, tương tự như những thách thức thường gặp của các hệ thống mô hình ngôn ngữ lớn trong thực tiễn [2]. Thứ nhất, hiệu năng của mô hình ngôn ngữ lớn cục bộ còn phụ thuộc vào cấu hình phần cứng và kích thước mô hình, dẫn đến độ chính xác chưa cao trong một số trường hợp. Thứ hai, chất lượng kết quả phụ thuộc lớn vào dữ liệu đầu vào và cơ sở tri thức nội bộ. Thứ ba, hệ thống chưa tích hợp đầy đủ các cơ chế nâng cao như RAG hoặc fine-tuning để tối ưu hóa hiệu quả mô hình [8], [11].

Trong tương lai, nghiên cứu có thể mở rộng theo hướng tích hợp các kỹ thuật nâng cao như retrieval-augmented generation, tối ưu hóa mô hình theo miền tri thức cụ thể và cải thiện khả năng tương tác của hệ thống. Ngoài ra, việc mở rộng quy mô dữ liệu và đánh giá trên nhiều nhóm người dùng khác nhau cũng cần được thực hiện nhằm nâng cao tính tổng quát và độ tin cậy của kết quả nghiên cứu. Đóng góp của nghiên cứu không chỉ nằm ở việc triển khai mô hình ngôn ngữ lớn cục bộ mà còn ở việc tích hợp các chức năng hỗ trợ đào tạo trong một nền tảng web hoạt động ngoại tuyến, phù hợp với môi trường có yêu cầu cao về bảo mật dữ liệu. Bên cạnh đó, kết quả nghiên cứu cũng cho thấy tiềm năng ứng dụng rộng rãi của mô hình AI triển khai cục bộ trong các lĩnh vực khác ngoài giáo dục, như điều tra số, phân tích dữ liệu nghiệp vụ và hỗ trợ ra quyết định trong các cơ quan thực thi pháp luật. Đây là hướng nghiên cứu có ý nghĩa thực tiễn cao, đặc biệt trong bối cảnh chuyển đổi số gắn với yêu cầu bảo mật dữ liệu ngày càng khắt khe.

4. Kết luận

Nghiên cứu này đã xây dựng và triển khai một nền tảng web ngoại tuyến tích hợp mô hình ngôn ngữ lớn thông qua nền tảng Ollama nhằm hỗ trợ giảng dạy và đào tạo nội bộ trong môi trường yêu cầu bảo mật cao. Kết quả thực nghiệm cho thấy hệ thống có khả năng tự động sinh câu hỏi, tạo bài giảng và hỗ trợ tra cứu tri thức với độ chính xác từ 82% đến 92%, đáp ứng tốt yêu cầu sử dụng trong thực tế.

Về mặt khoa học, nghiên cứu góp phần làm rõ tính khả thi của việc triển khai mô hình ngôn ngữ lớn cục bộ trong môi trường không kết nối Internet, đồng thời đề xuất một kiến trúc tích hợp giữa AI và nền tảng đào tạo nội bộ. Kết quả này mở ra hướng tiếp cận mới cho việc ứng dụng AI trong các tổ chức có yêu cầu cao về bảo mật dữ liệu [4], [5].

Tuy nhiên, nghiên cứu vẫn còn một số hạn chế. Hiệu năng của hệ thống phụ thuộc vào cấu hình phần cứng và chất lượng dữ liệu đầu vào, dẫn đến kết quả chưa ổn định trong một số trường hợp. Ngoài ra, hệ thống chưa tích hợp các kỹ thuật nâng cao như retrieval-augmented generation hoặc tinh chỉnh mô hình theo miền tri thức cụ thể [8], [11].

Trong tương lai, nghiên cứu sẽ tập trung vào việc tối ưu hóa mô hình, tích hợp các kỹ thuật nâng cao và mở rộng quy mô dữ liệu nhằm nâng cao độ chính xác và khả năng ứng dụng của hệ thống trong các bối cảnh thực tiễn đa dạng hơn.

TÀI LIỆU THAM KHẢO/ REFERENCES

- [1] A. Agarwal, M. P. Sebastian, and S. Krishnan, "Understanding GenAI teammates in the workplace: A sensemaking and sensegiving analysis of user reviews," *Information Systems Frontiers*, 2026, doi: 10.1007/s10796-025-10672-5.
- [2] J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy, "Challenges and applications of large language models," *arXiv preprint arXiv:2307.10169*, 2023, doi: 10.48550/arXiv.2307.10169.
- [3] H. Pu and Y. Hua, "Adaptive course recommendation using federated learning and graph convolutional networks in IoT-enhanced e-learning," *Scientific Reports*, vol. 15, 2025, Art. no. 42040, doi: 10.1038/s41598-025-42040-0.
- [4] L. Yan, L. Sha, L. Zhao, Y. Li, R. Martinez-Maldonado, G. Chen, X. Li, Y. Jin, and D. Gašević, "Practical and ethical challenges of large language models in education: A systematic scoping review," *arXiv preprint arXiv:2303.13379*, 2023, doi: 10.48550/arXiv.2303.13379.
- [5] E. Kasneci *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education," *npj Science of Learning*, vol. 8, no. 1, 2023, doi: 10.1038/s41539-023-00172-3.
- [6] V. Koniakou, "From the 'rush to ethics' to the 'race for governance' in Artificial Intelligence," *Information Systems Frontiers*, vol. 25, pp. 71-102, 2023, doi: 10.1007/s10796-021-10164-5.
- [7] A. Wada, Y. Tanaka, M. Nishizawa, *et al.*, "Retrieval-augmented generation elevates local LLM quality in radiology contrast media consultation," *npj Digital Medicine*, vol. 8, 2025, Art. no. 395, doi: 10.1038/s41746-025-01802-z.
- [8] Z. Li, Z. Wang, W. Wang, K. Hung, H. Xie, and F. L. Wang, "Retrieval-augmented generation for educational application: A systematic survey," *Computers and Education: Artificial Intelligence*, vol. 8, 2025, Art. no. 100417, doi: 10.1016/j.caeai.2025.100417.
- [9] J. Xu, H. Zhang, X. Zhou, Y. Yang, R. Mao, X. Li, *et al.*, "Automated detection of affected libraries from vulnerability reports," *Automated Software Engineering*, vol. 32, 2025, Art. no. 73, doi: 10.1007/s10515-025-00473-6.
- [10] H. Wang, C. Gao, C. Dantona, B. Hull, and J. Sun, "DRG-LLaMA: Tuning LLaMA model to predict diagnosis-related group for hospitalized patients," *npj Digital Medicine*, vol. 7, 2024, Art. no. 16, doi: 10.1038/s41746-024-00916-5.
- [11] R. V. Barenji, N. Salimi, and S. Khoshgoftar, "An LLM-powered assessment system using retrieval-augmented generation for higher education," *arXiv preprint arXiv:2601.06141*, 2026, doi: 10.48550/arXiv.2601.06141.
- [12] I. Taj, S. W. Dadi, A. S. Wazan, R. Laborde, and A. Shoufan, "CheatGuard: A cybersecurity inspired anti-cheating platform for higher education," *Education and Information Technologies*, vol. 30, pp. 23729-23759, 2025, doi: 10.1007/s10639-024-12578-4.
- [13] A. R. Asadi, Y. Zhang, and H. Said, "What do personas say about privacy & security: A systematic literature review through human-AI collaboration," *Journal of Ambient Intelligence and Humanized Computing*, vol. 16, pp. 1175-1190, 2025, doi: 10.1007/s12652-024-04834-2.