SMART DIGITAL LIBRARY MODEL: INTEGRATING AI LIBRARY AND ACADEMIC FORUMS TO SUPPORT UNIVERSITY LEARNING

Ta Khoa Anh Dung*
Trung Vuong University
ROR ID: https://ror.org/05xzsm645
Email: takhoaanhdung2005@gmail.com
ORCID iD: https://orcid.org/0009-0002-6994-6737
Do Tien Dat
Trung Vuong University
ROR ID: https://ror.org/05xzsm645
Email: dovoe456789@gmail.com
ORCID iD: https://orcid.org/0009-0006-0293-5215
Do Van Binh
Trung Vuong University
ROR ID: https://ror.org/05xzsm645
Email: dovanbinh868@gmail.com
ORCID iD: https://orcid.org/0009-0009-4882-0992
Nguyen Thi My Hoa
Trung Vuong University
ROR ID: https://ror.org/05xzsm645
Email: ngmyhoa0@gmail.com
ORCID iD: https://orcid.org/0009-0007-8875-8906
Nguyen Duc An
Trung Vuong University
ROR ID: https://ror.org/05xzsm645
Email: nguyenducan.hilix26@gmail.com
ORCID iD: https://orcid.org/0009-0004-8889-413X

Article History
Received: 10/02/2026
Reviewed: 20/3/2026
Revised: 30/4/2026
Acceped: 28/5/2026
Released: 30/6/2026

DOI: https://doi.org/10.64223/tvj.e2026.v2.i6.a92

Abstract:

In the context of digital transformation in education and the rapid development of artificial intelligence (AI), the need to build learning support systems capable of providing accurate, transparent, and personalized academic information is becoming increasingly urgent. This paper proposes a smart digital library model integrating AI librarians and academic forums to support learners in searching, accessing, managing, and verifying academic resources in higher education.

The system is developed based on a combination of semantic search, augmented reality, vector storage, and local language models to build a multifunctional digital academic environment. The platform not only supports document management and academic querying based on document content but also provides functions such as text summarization, multiple-choice question generation, and academic discussion organization through interactive forums. AI librarians act as intelligent learning aids, helping learners access knowledge quickly while enhancing information verification through transparent source citation mechanisms.

Initial evaluation results from 261 valid surveys show high user satisfaction with the system's usefulness, self-learning support capabilities, and convenience. These results initially confirm the feasibility of a proposed model in improving learning efficiency, promoting personalized learning, and developing a digital academic environment in higher education.

Keywords: *Smart Digital Library; AI Librarian; Academic Discussion Forum; Higher Education Learning; Educational Digital Transformation.*

JEL: A20, I21, I23, K20, K40, M53
ISCED-F: 0421, 0111, 0413
OECD: 5.05, 5.09
FoR: 480307, 390303

1. Giới thiệu

Sự phát triển nhanh chóng của trí tuệ nhân tạo (Artificial Intelligence - AI) đang đặt ra những yêu cầu mới đối với hệ thống thư viện số trong giáo dục đại học. Thay vì chỉ thực hiện chức năng lưu trữ và tìm kiếm tài liệu theo từ khóa, thư viện số hiện nay cần hỗ trợ người học khai thác tri thức học thuật theo hướng nhanh chóng, chính xác và có khả năng kiểm chứng nguồn thông tin. Trong thực tiễn, sinh viên thường gặp khó khăn trong việc lựa chọn tài liệu phù hợp, xử lý khối lượng văn bản học thuật lớn và đánh giá độ tin cậy của thông tin. Mặc dù các Chatbot (phản hồi tự động giữa con người với máy tính) hỗ trợ thư viện đã góp phần tăng cường khả năng tương tác, nhiều hệ thống vẫn tồn tại hạn chế như trả lời thiếu căn cứ, diễn giải sai nội dung hoặc không cung cấp được nguồn trích dẫn rõ ràng.

Trong bối cảnh đó, mô hình truy hồi tăng cường sinh (Retrieval-Augmented Generation - RAG) được xem là một hướng tiếp cận phù hợp để phát triển thư viện số thông minh có khả năng hỏi - đáp dựa trên tài liệu. Khác với các mô hình sinh ngôn ngữ truyền thống,

RAG kết hợp giữa cơ chế truy hồi dữ liệu và mô hình ngôn ngữ lớn (Large Language Model - LLM), cho phép hệ thống tìm kiếm các đoạn nội dung liên quan trước khi tạo phản hồi. Cách tiếp cận này giúp giới hạn câu trả lời trong phạm vi dữ liệu học thuật đã được cung cấp, đồng thời nâng cao khả năng đối chiếu và kiểm chứng nguồn thông tin.

Từ cơ sở đó, nghiên cứu đề xuất một nền tảng thư viện số thông minh tích hợp AI thủ thư và diễn đàn học thuật nhằm hỗ trợ người học trong quá trình tìm kiếm, khai thác và trao đổi tri thức. Hệ thống sử dụng kỹ thuật tìm kiếm ngữ nghĩa (Semantic Search), trong đó tài liệu PDF được phân tách thành các đoạn nội dung (Chunks), mã hóa dưới dạng Vector và lập chỉ mục bằng Facebook AI Similarity Search (FAISS). Các dữ liệu truy hồi theo cơ chế Top-k (kết quả phù hợp nhất) được sử dụng làm ngữ cảnh đầu vào cho mô hình ngôn ngữ cục bộ vận hành thông qua nền tảng Ollama (nền tảng chạy mô hình ngôn ngữ lớn), từ đó tạo ra các phản hồi có căn cứ học thuật và gắn với nguồn tài liệu cụ thể.

Điểm mới của nghiên cứu nằm ở việc tích hợp đồng thời cơ chế truy hồi ngữ nghĩa, mô hình RAG cục bộ, chức năng trích dẫn nguồn và không gian thảo luận học thuật trong cùng một hệ thống thư viện số. Mô hình này không chỉ hỗ trợ người học đặt câu hỏi trực tiếp trên tài liệu mà còn tạo điều kiện để trao đổi, phản biện và kiểm chứng thông tin trong môi trường học thuật mở. Nghiên cứu được đánh giá thông qua kiểm thử chức năng hệ thống, đo lường chất lượng phản hồi của AI và khảo sát 261 người dùng, qua đó bước đầu khẳng định tính khả thi của giải pháp trong hỗ trợ tự học, nâng cao khả năng tiếp cận học liệu số và tăng cường tính minh bạch học thuật trong giáo dục đại học.

2. Tổng quan nghiên cứu vấn đề

Các nghiên cứu về thư viện số cho thấy xu hướng phát triển hiện nay không còn dừng ở chức năng lưu trữ và tra cứu tài liệu, mà chuyển sang hỗ trợ người học tương tác trực tiếp với tri thức học thuật. Theo định nghĩa của Digital Library Federation và IFLA/UNESCO, thư viện số cần bảo đảm việc tổ chức, quản lý và cung cấp khả năng truy cập tài nguyên thông tin trong môi trường số. Trong giáo dục đại học, yêu cầu này tiếp tục được mở rộng theo hướng tìm kiếm thông minh, gợi ý học liệu, tóm tắt nội dung và hỗ trợ tự học. Tuy nhiên, các hệ thống tìm kiếm truyền thống vẫn phụ thuộc chủ yếu vào từ khóa và siêu dữ liệu (Metadata), nên gặp hạn chế khi xử lý các truy vấn học thuật được diễn đạt bằng ngôn ngữ tự nhiên.

Sự phát triển của tìm kiếm ngữ nghĩa và biểu diễn Vector (Vector Representation) đã tạo nền tảng kỹ thuật cho thư viện số thông minh. Các nghiên cứu về biểu diễn nhúng (Embedding), mô hình BERT biểu diễn ngữ nghĩa câu (Sentence-BERT), truy xuất đoạn văn mật độ cao (Dense Passage Retrieval - DPR) và tìm kiếm Vecto (Vector Search) cho thấy truy vấn và văn bản có thể được mã hóa thành các Vector ngữ nghĩa nhằm đo lường mức độ tương đồng về ý nghĩa thay vì chỉ đối sánh từ khóa. Trong đó, Facebook AI Similarity Search (FAISS) được sử dụng phổ biến nhờ khả năng lập chỉ mục và truy hồi Vector hiệu quả trên quy mô lớn. Cách tiếp cận này cho phép hệ thống xác định chính xác các đoạn nội dung liên quan trong tài liệu, thay vì chỉ trả về danh mục tài liệu dựa trên nhan đề hoặc từ khóa.

Bên cạnh đó, truy hồi tăng cường sinh RAG được xem là hướng tiếp cận phù hợp nhằm khắc phục hạn chế của các Chatbot thư viện truyền thống. Một số nghiên cứu cho thấy Chatbot thư viện có thể nâng cao khả năng tương tác với người dùng, nhưng vẫn gặp khó khăn về độ chính xác, độ tin cậy và khả năng xử lý các truy vấn học thuật phức tạp. RAG giải quyết vấn đề này bằng cách truy xuất các đoạn tài liệu liên quan trước khi mô hình ngôn ngữ lớn LLM tạo phản hồi. Nhờ đó, câu trả lời được xây dựng dựa trên ngữ cảnh tài liệu truy hồi thay vì chỉ phụ thuộc vào tri thức tiềm ẩn của mô hình, qua đó tăng khả năng kiểm chứng và hạn chế hiện tượng "ảo giác AI" (AI Hallucination).

Mặc dù đã có nhiều nghiên cứu về thư viện số, Chatbot học thuật, tìm kiếm ngữ nghĩa và RAG, phần lớn các tiếp cận vẫn được triển khai riêng lẻ. Hiện chưa có nhiều nghiên cứu xây dựng mô hình tích hợp giữa AI thủ thư, cơ chế truy hồi có trích dẫn nguồn và diễn đàn học thuật trong cùng một

nền tảng. Đây là khoảng trống đáng chú ý trong giáo dục đại học, nơi người học không chỉ cần câu trả lời nhanh mà còn cần khả năng đối chiếu và xác thực nguồn tài liệu. Việc kết hợp RAG với cơ chế trích dẫn nguồn giúp giới hạn phản hồi trong phạm vi dữ liệu đã nạp, trong khi diễn đàn học thuật tạo môi trường để người dùng trao đổi, phản biện và đánh giá lại kết quả do AI tạo ra.

Từ khoảng trống đó, nghiên cứu này đề xuất mô hình thư viện số thông minh tích hợp AI thủ thư và diễn đàn học thuật dựa trên kiến trúc RAG cục bộ. Hệ thống xử lý tài liệu PDF bằng kỹ thuật phân đoạn nội dung Chunking, tạo Embedding, lưu trữ Vector trong FAISS và truy hồi Top-k đoạn văn liên quan để cung cấp ngữ cảnh cho mô hình ngôn ngữ vận hành cục bộ thông qua Ollama. Điểm khác biệt của mô hình nằm ở việc chuyển thư viện số từ công cụ tra cứu tài liệu truyền thống thành môi trường hỏi - đáp học thuật có căn cứ, trong đó mỗi phản hồi của AI đều gắn với nguồn tham chiếu cụ thể và có thể tiếp tục được kiểm chứng thông qua diễn đàn học thuật.

3. Cách tiếp cận và phương pháp nghiên cứu

Nghiên cứu sử dụng phương pháp thực nghiệm hệ thống nhằm đánh giá khả năng ứng dụng mô hình truy hồi tăng cường sinh RAG trong thư viện số thông minh. Kiến trúc hệ thống gồm bốn lớp chính: giao diện người dùng, máy chủ xử lý, lớp trí tuệ nhân tạo AI và mô hình ngôn ngữ lớn LLM vận hành cục bộ thông qua Ollama. Trọng tâm nghiên cứu không hướng đến mô tả các chức năng phần mềm, mà tập trung kiểm chứng khả năng truy hồi đúng nội dung tài liệu, sinh phản hồi bám sát ngữ cảnh và cung cấp trích dẫn để người học đối chiếu với nguồn gốc học liệu.

Nguồn dữ liệu thực nghiệm bao gồm các tài liệu học thuật định dạng PDF được nạp vào hệ thống để xây dựng kho tri thức. Văn bản từ mỗi tài liệu được trích xuất, phân đoạn theo chiến lược Chunking và gắn Metadata như tên tài liệu, số trang, vị trí đoạn và thông tin mô tả. Sau đó, hệ thống tạo Embedding cho từng đoạn văn bản và lưu trữ dưới dạng Vector trong Facebook AI Similarity Search (FAISS) nhằm phục vụ truy hồi ngữ nghĩa. Khi người dùng đặt câu hỏi, truy vấn được mã hóa thành Vector, đối sánh với cơ sở dữ liệu Vector và truy hồi Top-k = 12 đoạn có độ tương đồng cao nhất. Các đoạn truy hồi này được kết hợp thành ngữ cảnh đầu vào cho mô hình ngôn ngữ chạy cục bộ qua Ollama để tạo câu trả lời kèm theo trích dẫn nguồn.

Hiệu quả của mô hình RAG được đánh giá thông qua bộ câu hỏi kiểm thử xây dựng từ chính nội dung tài liệu đã nạp vào hệ thống. Mỗi phản hồi được đối chiếu theo ba tiêu chí: mức độ trả lời đúng trọng tâm câu hỏi, mức độ bám sát tài liệu và độ chính xác của trích dẫn nguồn. Kết quả cho thấy độ chính xác đạt 85% đối với nhóm câu hỏi có thông tin xuất hiện trực tiếp trong tài liệu. Phương pháp đánh giá này phù hợp với đặc thù của thư viện số học thuật, nơi phản hồi do AI tạo ra không chỉ cần chính xác mà còn phải có căn cứ và khả năng kiểm chứng.

Bên cạnh đánh giá kỹ thuật, nghiên cứu còn kết hợp khảo sát người dùng nhằm phản ánh mức độ hiệu quả và khả năng chấp nhận của hệ thống trong môi trường giáo dục đại học. Phần đánh giá kỹ thuật tập trung vào các tiêu chí như độ chính xác câu trả lời, mức độ liên quan với tài liệu, tính đúng của trích dẫn và thời gian phản hồi. Song song với đó, khảo sát người dùng được thực hiện bằng thang đo Likert 5 mức với 261 phiếu hợp lệ, tập trung vào các khía cạnh gồm giao diện hệ thống, khả năng tìm kiếm, bộ lọc tài liệu, mức độ liên quan của kết quả, chức năng tóm tắt, hỏi-đáp AI, trích dẫn tự động và độ ổn định khi vận hành. Cách tiếp cận kết hợp này cho phép đánh giá đồng thời hiệu quả kỹ thuật của mô hình RAG và mức độ phù hợp của hệ thống đối với nhu cầu học tập thực tế của người dùng.

Để làm rõ nguyên nhân khác biệt giữa các nhóm kết quả, nghiên cứu phân tích riêng hai thành phần chính của hệ thống: quy trình truy hồi Vector bằng FAISS và quy trình sinh phản hồi thông qua Ollama. Hiệu quả truy hồi phụ thuộc chủ yếu vào chất lượng Embedding, chiến lược phân đoạn văn bản và tham số Top-k, trong khi chất lượng sinh phản hồi còn chịu tác động bởi độ dài ngữ cảnh, năng lực suy luận của mô hình và tài nguyên phần cứng khi vận hành cục bộ. Cách phân tích này giúp xác định rõ các điểm nghẽn kỹ thuật của hệ thống, đặc biệt đối với các chức năng AI sinh nội dung và yêu cầu ổn định trong môi trường triển khai thực tế.

4. Kết quả nghiên cứu

4.1. Kiến trúc hệ thống thư viện số thông minh dựa trên RAG

Kiến trúc hệ thống được thiết kế theo mô hình RAG cục bộ nhằm hỗ trợ người học hỏi-đáp trực tiếp trên tài liệu số có trích dẫn nguồn. Hệ thống gồm bốn thành phần chính: giao diện người dùng, dịch vụ máy chủ, dịch vụ AI và mô hình ngôn ngữ chạy cục bộ qua Ollama. Giao diện người dùng tiếp nhận yêu cầu tra cứu, đọc tài liệu và đặt câu hỏi. Dịch vụ máy chủ quản lý tài liệu, Metadata, tài khoản và điều phối các yêu cầu giữa người dùng với dịch vụ AI. Dịch vụ AI đảm nhiệm xử lý tài liệu, tạo Embedding, truy hồi ngữ nghĩa bằng FAISS và xây dựng ngữ cảnh cho mô hình ngôn ngữ sinh câu trả lời.

Điểm cốt lõi của kiến trúc nằm ở quy trình truy hồi trước khi sinh câu trả lời. Tài liệu PDF sau khi nạp vào hệ thống được trích xuất văn bản, chia thành các Chunk, gắn Metadata và mã hóa thành Vector

Bảng 1. Kết quả khảo sát minh họa về năng lực nghề của HSSV

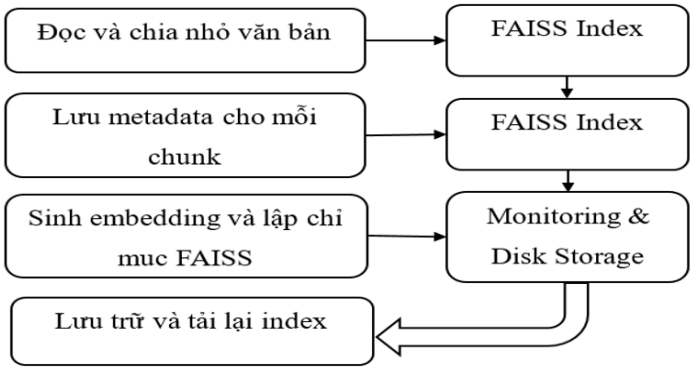
Thành phần	Vai trò chính	Công nghệ / Nền tảng	Đầu vào	Đầu ra
Dịch vụ máy khách	Giao diện người dùng; hiển thị tài liệu, Chat, tóm tắt, Thẻ ghi nhớ/ Câu hỏi luyện tập	Web UI	Tương tác người dùng; dữ liệu từ dịch vụ máy chủ/ dịch vụ AI	Trang hiển thị; kết quả Chat/tóm tắt/ Câu hỏi luyện tập/ Thẻ ghi nhớ
Dịch vụ máy chủ	Quản lý nghiệp vụ thư viện; quản lý Metadata tài liệu; cung cấp API trung tâm; tích hợp với dịch vụ AI	Node.js dịch vụ máy chủ	Request từ dịch vụ máy khách; yêu cầu từ dịch vụ AI	Response dữ liệu/ Metadata; kết quả tích hợp trả về dịch vụ máy khách/ dịch vụ AI
Dịch vụ AI	Xử lý NLP/RAG: lấy vào tài liệu, truy hồi, sinh câu trả lời; sinh tóm tắt/Thẻ ghi nhớ/ Câu hỏi luyện tập	Python FastAPI + LangChain + FAISS	Câu hỏi hoặc File-Path; dữ liệu truy hồi từ Vector DB Metadata từ dịch vụ máy chủ	Định dạng JSON cho Thẻ ghi nhớ/ Câu hỏi luyện tập
Ollama	Cung cấp môi trường chạy LLM cục bộ cho các tác vụ sinh nội dung	Ollama + model	Prompt từ dịch vụ AI	Text Output

ngữ nghĩa. Các Vector này được lưu trong FAISS để phục vụ tìm kiếm tương đồng. Khi người dùng đặt câu hỏi, hệ thống chuyển câu hỏi thành Vector, truy hồi các Chunk liên quan và đưa các đoạn này vào mô hình ngôn ngữ Qwen2.5:7b chạy qua Ollama. Cách tổ chức này giúp câu trả lời không được sinh độc lập từ mô hình, mà dựa trên nội dung tài liệu đã truy hồi.

Kiến trúc RAG cục bộ tạo ra khác biệt so với Chatbot thư viện thông thường ở khả năng kiểm soát nguồn tri thức và trích dẫn kết quả. Mỗi câu trả lời được gắn với Metadata của Chunk truy hồi như tên tài liệu, trang hoặc vị trí đoạn, cho phép người dùng đối chiếu lại với tài liệu gốc. Cơ chế này phù hợp với yêu cầu học thuật, vì người học có thể kiểm chứng thông tin thay vì tiếp nhận câu trả lời AI như một nội dung không có nguồn. Việc chạy mô hình cục bộ qua Ollama cũng giúp hệ thống chủ động hơn về dữ liệu và hạ tầng, nhưng đồng thời đặt ra yêu cầu tối ưu phần cứng, độ dài ngữ cảnh và thời gian phản hồi khi triển khai ở quy mô lớn.

4.2. Kết quả xây dựng quy trình Ingest tài liệu và kho tri thức

Quy trình nạp tài liệu (Ingest) được xây dựng nhằm chuyển tài liệu PDF thành kho tri thức có thể truy hồi bằng ngữ nghĩa. Tài liệu sau khi nạp vào hệ thống được trích xuất văn bản, chuẩn hóa nội dung, chia thành các Chunk và gắn Metadata gồm tên tài liệu, số trang, vị trí đoạn và thông tin mô tả. Mỗi Chunk được xem là một đơn vị tri thức độc lập, phục vụ trực tiếp cho quá trình truy hồi khi người dùng đặt câu hỏi. Hệ thống tạo Embedding cho từng Chunk bằng mô hình biểu diễn ngữ nghĩa và lưu



Hình 1. Quy trình nạp tài liệu và xây dựng kho tri thức

các Vector này vào FAISS index. FAISS cho phép hệ thống tìm kiếm tương đồng giữa Vector câu hỏi và Vector nội dung tài liệu, nhờ đó truy hồi đúng các đoạn liên quan thay vì chỉ trả về toàn bộ tài liệu theo từ khóa. Cách tổ chức kho tri thức theo Chunk giúp mô hình RAG sử dụng đúng phần nội dung cần thiết, giảm nhiễu ngữ cảnh và tạo điều kiện gắn trích dẫn nguồn cho từng câu trả lời.

Kết quả xây dựng quy trình Ingest cho thấy chất lượng chia Chunk và Metadata ảnh hưởng trực tiếp đến độ chính xác của chức năng hỏi-đáp. Chunk quá ngắn dễ làm mất ngữ cảnh, trong khi Chunk quá dài làm tăng nhiễu và kéo dài thời gian xử lý của mô hình ngôn ngữ. Vì vậy, quy trình Ingest không chỉ là bước nạp dữ liệu kỹ thuật, mà là khâu quyết định chất lượng truy hồi, độ đúng của trích dẫn và khả năng giảm ảo giác AI trong hệ thống thư viện số thông minh.

4.3. Kết quả triển khai cơ chế RAG Chat có trích dẫn

Cơ chế RAG Chat được triển khai theo quy trình truy hồi nội dung trước khi sinh câu trả lời. Khi người dùng đặt câu hỏi, hệ thống chuyển câu hỏi thành Vector ngữ nghĩa, so khớp với kho FAISS và truy hồi Top-k = 12 Chunk có mức tương đồng cao nhất. Các Chunk này được ghép thành ngữ cảnh đầu vào cho mô hình ngôn ngữ qwen2.5:7b chạy cục bộ qua Ollama. Câu trả lời cuối cùng được sinh ra dựa trên phần ngữ cảnh đã truy hồi, thay vì để mô hình tự trả lời độc lập từ tri thức sẵn có.

Bảng 2. Thiết kế chức năng hỏi-đáp theo tài liệu

Hạng mục	Nội dung thiết kế	Tham số / Quy ước	Mục đích / Ghi chú
Quy trình truy hồi	Thực hiện truy hồi các đoạn liên quan từ Vector DB bằng tìm kiếm tương tự	top-k = 12	Lấy đủ ngữ cảnh để trả lời, tránh thiếu dữ kiện
Similarity Search	Tạo Embedding cho câu hỏi → tìm các Chunk gần nhất	k = 12	Cân bằng giữa độ phủ và nhiễu ngữ cảnh
Tổ chức ngữ cảnh	Ghép các đoạn truy hồi thành bối cảnh đưa vào LLM	Định dạng: [Giáo trình – Trang] + đoạn văn	Tạo điều kiện để LLM trích dẫn rõ ràng, có thể kiểm chứng
Thiết kế Prompt	Prompt ràng buộc mô hình chỉ trả lời theo bối cảnh	Quy ước: "chỉ dùng thông tin trong bối cảnh"	Giảm Hallucination, tăng tính học thuật
Cơ chế trích dẫn	Bắt buộc câu trả lời có trích dẫn nguồn	Trích dẫn gồm: tên giáo trình + số trang	Minh bạch nguồn, hỗ trợ đối chiếu lại tài liệu
Resolve Title qua dịch vụ máy chủ	dịch vụ AI gọi dịch vụ máy chủ để ánh xạ Filename → Title chuẩn	/api/documents/resolve-titles	Trích dẫn "đúng tên" giáo trình , không hiển thị Filename thô
Kiểm soát ngữ cảnh	Giới hạn độ dài bối cảnh trước khi gửi sang LLM	Hiện tại: Truncate theo ký tự	Tránh vượt giới hạn ngữ cảnh của LLM, giảm Latency
Đề xuất cải tiến	Kiểm soát bối cảnh theo hướng thông minh hơn	Giới hạn theo mã lọc theo ngưỡng điểm	Tối ưu chất lượng bối cảnh và giảm nhiễu; tăng độ chính xác câu trả lời & trích dẫn

Cơ chế trích dẫn được tích hợp vào đầu ra của RAG Chat nhằm tăng khả năng kiểm chứng học thuật. Mỗi Chunk được gắn Metadata như tên tài liệu, số trang và vị trí đoạn; khi Chunk được sử dụng làm căn cứ trả lời, hệ thống đưa thông tin nguồn vào kết quả hiển thị. Cách triển khai này cho phép người học đối chiếu câu trả lời với tài liệu gốc, phát hiện trường hợp diễn giải chưa sát nguồn và hạn chế rủi ro tiếp nhận thông tin thiếu căn cứ.

Kết quả triển khai cho thấy RAG Chat hoạt động hiệu quả nhất với các câu hỏi có thông tin xuất hiện trực tiếp trong tài liệu đã nạp. Khi FAISS truy hồi đúng các Chunk liên quan, mô hình tạo được câu trả lời mạch lạc, đúng trọng tâm và có nguồn đối chiếu. Một số hạn chế vẫn xuất hiện ở các câu hỏi khái quát hoặc yêu cầu tổng hợp nhiều đoạn rời nhau, do chất lượng câu trả lời phụ thuộc vào chiến lược Chunking, tham số Top-k, độ dài ngữ cảnh và năng lực suy luận của mô hình chạy cục bộ.

4.4. Kết quả khảo sát người dùng

Kết quả khảo sát 261 phiếu hợp lệ cho thấy người dùng đánh giá tích cực đối với nền tảng thư viện số thông minh, nhưng mức hài lòng có sự khác biệt rõ giữa nhóm chức năng tìm kiếm và nhóm chức năng AI. Nhóm tìm kiếm đạt kết quả cao nhất: tìm kiếm tài liệu theo từ khóa có 218/261 lựa chọn ở mức 4-5, trong đó 93/261 lựa chọn mức 5; độ liên quan của kết quả tìm kiếm có 214/261 lựa chọn ở mức 4-5, trong đó 96/261 lựa chọn mức 5. Kết quả này phản ánh các chức năng tra cứu truyền thống vận hành ổn định hơn, do phụ thuộc chủ yếu vào dữ liệu mô tả, chỉ mục tìm kiếm và thao tác truy hồi ít biến động.

Bảng 3: Kết quả khảo sát mức độ hài lòng của người dùng đối với các chức năng của hệ thống

Tiêu chí	Mức độ đánh giá				
	1	2	3	4	5
Mức độ hài lòng về các chức năng giao diện tổng thể	3	8	40	120	90
Mức độ hài lòng về tìm kiếm tài liệu theo từ khóa	2	6	35	125	93
Mức độ hài lòng về các chức năng bộ lọc kết quả	3	10	45	120	83
Mức độ hài lòng về độ liên quan của kết quả tìm kiếm	2	7	38	118	96
Mức độ hài lòng về gợi ý tài liệu liên quan	3	9	50	115	84
Mức độ hài lòng về chức năng tóm tắt tài liệu bằng AI	4	12	55	110	80
Mức độ hài lòng về chức năng hỏi-đáp AI theo nội dung tài liệu	5	15	60	105	76
Mức độ hài lòng về chức năng trích dẫn tự động hỏi-đáp AI theo nội dung tài liệu	6	18	65	102	70
Mức độ hài lòng của bạn về độ ổn định	8	22	80	95	56
Mức độ hài lòng chung	3	9	45	120	84

Nhóm chức năng AI nhận được mức đánh giá tích cực nhưng thấp hơn nhóm tìm kiếm. Chức năng tóm tắt tài liệu bằng AI có 190/261 lựa chọn ở mức 4-5; hỏi-đáp AI theo nội dung tài liệu có 181/261 lựa chọn ở mức 4-5, trong đó 76/261 lựa chọn mức 5; trích dẫn tự động trong hỏi-đáp AI có 172/261 lựa chọn ở mức 4-5, trong đó 70/261 lựa chọn mức 5. Khoảng chênh lệch này cho thấy trải nghiệm với AI phụ thuộc vào nhiều yếu tố kỹ thuật hơn, gồm chất lượng Chunking, độ phù hợp của các đoạn truy hồi, tham số Top-k, độ dài ngữ cảnh và khả năng sinh câu trả lời của mô hình ngôn ngữ.

Độ ổn định là tiêu chí có mức đánh giá thấp nhất, với 151/261 lựa chọn ở mức 4-5 và chỉ 56/261 lựa chọn mức 5. Kết quả này phù hợp với đặc điểm vận hành của mô hình RAG cục bộ, trong đó FAISS xử lý truy hồi tương đồng tương đối nhanh, còn bước sinh câu trả lời qua Ollama phụ thuộc nhiều vào tài nguyên phần cứng và độ dài ngữ cảnh đầu vào. Vì vậy, điểm nghẽn chính của hệ thống không nằm ở chức năng tìm kiếm, mà nằm ở giai đoạn suy luận của mô hình ngôn ngữ và kiểm soát ngữ cảnh khi triển khai chức năng AI ở quy mô lớn hơn.

4.5. Phân tích nguyên nhân kỹ thuật của kết quả khảo sát

Kết quả khảo sát phản ánh sự khác biệt rõ giữa nhóm chức năng truy hồi ổn định và nhóm chức năng AI sinh nội dung. Các chức năng tìm kiếm theo từ khóa và đánh giá độ liên quan kết quả có mức hài lòng cao hơn vì quy trình xử lý ngắn, phụ thuộc chủ yếu vào dữ liệu mô tả, chỉ mục tìm kiếm và thao tác truy hồi. Trong khi đó, các chức năng hỏi-đáp AI và trích dẫn tự động phải đi qua nhiều bước kỹ thuật hơn, gồm chia Chunk, tạo Embedding, truy hồi FAISS, ghép ngữ cảnh, gọi mô hình ngôn ngữ qua Ollama và chuẩn hóa trích dẫn. Mỗi bước đều có thể ảnh hưởng đến chất lượng đầu ra, nên mức đánh giá của người dùng thấp hơn nhóm tìm kiếm là kết quả phù hợp với đặc điểm vận hành của hệ thống.

Độ ổn định được đánh giá thấp nhất chủ yếu do điểm nghẽn ở bước suy luận của mô hình ngôn ngữ cục bộ. FAISS thực hiện tìm kiếm tương đồng trên Vector với tốc độ tương đối ổn định, nhưng mô hình qwen2.5:7b chạy qua Ollama cần xử lý ngữ cảnh gồm nhiều Chunk trước khi sinh câu trả lời. Khi câu hỏi dài, tài liệu có cấu trúc phức tạp hoặc Top-k truy hồi nhiều đoạn, độ dài Prompt tăng lên và thời gian phản hồi cũng tăng theo. Vì vậy, hiện tượng chậm phản hồi hoặc đầu ra chưa nhất quán có khả năng xuất phát từ tài nguyên phần cứng, độ dài ngữ cảnh và quá trình sinh văn bản của LLM, hơn là từ bước tìm kiếm FAISS.

Chất lượng trích dẫn tự động thấp hơn chức năng hỏi-đáp AI do trích dẫn phụ thuộc trực tiếp vào Metadata và độ chính xác của chunk được truy hồi.

Khi Chunk chứa đúng nội dung nhưng Metadata chưa đầy đủ, hoặc khi câu trả lời tổng hợp từ nhiều đoạn rời nhau, hệ thống có thể tạo câu trả lời đúng ý nhưng phần nguồn chưa thật sự thuyết phục. Vấn đề này cho thấy cần tối ưu quy trình Ingest tài liệu, kiểm soát kích thước Chunk, bổ sung cơ chế xếp hạn lại (Reranking) và hiển thị đoạn trích nguồn kèm trích dẫn. Các cải tiến này có thể giảm nhiễu ngữ cảnh, tăng độ đúng của nguồn tham chiếu và nâng cao độ tin cậy của chức năng RAG Chat trong giai đoạn triển khai tiếp theo.

5. Thảo luận

Kết quả nghiên cứu cho thấy mô hình thư viện số thông minh dựa trên truy hồi tăng cường sinh RAG cục bộ có nhiều ưu thế so với các Chatbot thư viện truyền thống, đặc biệt ở khả năng trả lời dựa trên tài liệu và cung cấp nguồn tham chiếu kiểm chứng. Phần lớn các nghiên cứu về Chatbot thư viện trước đây chủ yếu tập trung vào hỗ trợ tra cứu nhanh, giải đáp dịch vụ và tăng cường tương tác với người dùng. Tuy nhiên, hạn chế phổ biến của các hệ thống này là phản hồi thường được tạo ra trực tiếp từ mô hình ngôn ngữ mà thiếu cơ chế đối chiếu với tài liệu nguồn, dẫn đến nguy cơ sai lệch hoặc khó kiểm chứng. Trong nghiên cứu này, mô hình ngôn ngữ không hoạt động độc lập mà được kết hợp với cơ chế truy hồi vector từ FAISS trước khi sinh phản hồi, qua đó bảo đảm rằng nội dung trả lời được hình thành trên cơ sở các đoạn tài liệu liên quan đã được truy xuất.

Việc tích hợp cơ chế RAG kèm trích dẫn nguồn góp phần nâng cao tính minh bạch của AI trong môi trường học thuật. Hệ thống sử dụng quy trình Chunking, tạo Embedding, truy hồi Top-k = 12 và mô hình ngôn ngữ Qwen2.5:7b vận hành cục bộ qua Ollama để sinh phản hồi dựa trên ngữ cảnh tài liệu. Mỗi câu trả lời đều được gắn với thông tin tham chiếu như tên tài liệu, số trang hoặc vị trí đoạn văn, cho phép người học kiểm chứng lại nội dung thay vì tiếp nhận phản hồi AI như một "hộp đen" không rõ căn cứ. Cách tiếp cận này đặc biệt phù hợp với môi trường giáo dục đại học, nơi các hoạt động học tập, viết tiểu luận và nghiên cứu khoa học đòi hỏi thông tin phải gắn với nguồn trích dẫn cụ thể và có khả năng đối chiếu học thuật.

Bên cạnh đó, diễn đàn học thuật được tích hợp trong hệ thống đóng vai trò như một lớp kiểm chứng cộng đồng đối với kết quả do AI tạo ra. Người học có thể đưa các phản hồi, đoạn trích hoặc vấn đề chưa rõ lên diễn đàn để trao đổi, phản biện và bổ sung tài liệu liên quan. Thông qua quá trình tương tác này, các trường hợp AI diễn giải chưa chính xác hoặc chưa sát với nội dung gốc có thể được phát hiện và điều chỉnh. Cơ chế kết hợp giữa truy hồi tài liệu và phản hồi cộng đồng giúp giảm nguy cơ "ảo giác AI" (AI Hallucination) theo hai hướng: giới hạn phạm vi phản hồi trong tập tài liệu đã truy hồi và mở rộng khả năng kiểm chứng thông qua trao đổi học thuật. Đây cũng là điểm khác biệt quan trọng của mô hình đề xuất so với các hệ thống chỉ cung cấp Chatbot hoặc kho học liệu số đơn thuần.

Tuy nhiên, kết quả khảo sát cũng cho thấy một số giới hạn kỹ thuật của mô hình RAG cục bộ. Nhóm chức năng tìm kiếm và truy hồi tài liệu được đánh giá cao hơn so với nhóm chức năng hỏi-đáp AI và trích dẫn tự động, trong khi độ ổn định hệ thống có mức đánh giá thấp nhất. Kết quả này cho thấy điểm nghẽn chủ yếu nằm ở giai đoạn suy luận của mô hình ngôn ngữ khi vận hành qua Ollama, đặc biệt trong các trường hợp ngữ cảnh đầu vào dài hoặc tài liệu có cấu trúc phức tạp. Do đó, các hướng cải tiến tiếp theo cần tập trung vào tối ưu chiến lược Chunking, bổ sung cơ chế Reranking nhằm nâng cao độ liên quan của kết quả truy hồi, kiểm soát độ dài Prompt, xây dựng bộ nhớ đệm (Cache) cho các truy vấn phổ biến và nâng cấp hạ tầng phần cứng phục vụ vận hành mô hình cục bộ.

6. Kết luận

Nghiên cứu đã đề xuất và triển khai mô hình thư viện số thông minh tích hợp AI thủ thư và diễn đàn học thuật dựa trên kiến trúc truy hồi tăng cường sinh RAG cục bộ. Hệ thống được xây dựng trên quy trình xử lý gồm nạp tài liệu PDF, phân đoạn văn bản Chunking, gắn Metadata, tạo Embedding, lưu trữ Vector bằng FAISS và truy hồi Top-k = 12 đoạn nội dung liên quan để cung cấp ngữ cảnh cho mô hình ngôn ngữ Qwen2.5:7b vận hành thông qua Ollama. Cách tiếp cận này giúp chuyển đổi thư viện số từ công cụ tra cứu tài liệu truyền thống thành môi trường hỏi–đáp học thuật có căn cứ, trong đó người học có thể tương tác trực tiếp với học liệu, nhận phản hồi gắn với nguồn tham chiếu cụ thể và kiểm chứng lại thông tin thông qua trích dẫn.

Kết quả thực nghiệm bước đầu cho thấy mô hình có khả năng hỗ trợ hiệu quả quá trình tự học và khai thác học liệu trong giáo dục đại học. Chức năng hỏi-đáp AI đạt độ chính xác 85% đối với nhóm câu hỏi có thông tin xuất hiện trực tiếp trong tài liệu, đồng thời khảo sát 261 phản hồi hợp lệ ghi nhận mức đánh giá tích cực đối với các chức năng tìm kiếm ngữ nghĩa, hỏi-đáp AI và trích dẫn tự động. Kết quả này cho thấy việc kết hợp giữa truy hồi tài liệu, mô hình ngôn ngữ cục bộ và cơ chế trích dẫn nguồn có thể nâng cao tính minh bạch và độ tin cậy của AI trong môi trường học thuật.

Bên cạnh những kết quả đạt được, nghiên cứu cũng chỉ ra một số giới hạn kỹ thuật khi triển khai mô hình RAG cục bộ trong thực tế. Các chức năng truy hồi và tìm kiếm tài liệu có mức ổn định cao hơn nhóm chức năng sinh phản hồi AI, trong khi độ ổn định toàn hệ thống vẫn chịu ảnh hưởng bởi tài

nguyên phần cứng, độ dài ngữ cảnh và chất lượng các đoạn văn bản được truy hồi. Điều này cho thấy hiệu quả của hệ thống không chỉ phụ thuộc vào mô hình ngôn ngữ mà còn gắn chặt với chiến lược chunking, chất lượng embedding và cơ chế lựa chọn ngữ cảnh đầu vào.

Trong giai đoạn tiếp theo, nghiên cứu cần tập trung tối ưu kích thước Chunk và Overlap, bổ sung cơ chế Reranking để nâng cao độ liên quan của kết quả truy hồi, kiểm soát độ dài Prompt, xây dựng bộ nhớ đệm Cache cho các truy vấn phổ biến và nâng cấp hạ tầng vận hành Ollama nhằm giảm độ trễ và tăng tính ổn định. Đồng thời, việc tích hợp hiển thị Snippet nguồn kèm trích dẫn trực tiếp sẽ góp phần nâng cao khả năng kiểm chứng và trải nghiệm người dùng. Những cải tiến này không chỉ mở rộng khả năng ứng dụng của mô hình trong thư viện số thông minh mà còn tạo tiền đề cho việc triển khai các hệ thống hỗ trợ học tập và nghiên cứu khoa học dựa trên AI tại các cơ sở giáo dục đại học.

Tài liệu tham khảo

Bạch, T. H., & Nguyễn, V. A. (2022). Chuyển đổi số trong thư viện đại học Việt Nam hiện nay. *Tạp chí Thông tin và Tư liệu, 4*, 15–24. https://doi.org/10.31276/VJST.64(4).15-24

Bai, J., Bai, S., Chu, Y., et al. (2023). *Qwen technical report*. arXiv preprint. https://doi.org/10.48550/arXiv.2309.16609

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*. https://doi.org/10.48550/arXiv.2005.14165

Cao, L. (2023). AI in libraries and information services. *Journal of Information Science, 49*(2), 145–162. https://doi.org/10.1177/01655515221076541

Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading Wikipedia to answer open-domain questions. *Proceedings of ACL 2017*. https://doi.org/10.18653/v1/P17-1171

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*. https://doi.org/10.18653/v1/N19-1423

Đặng, T. T., & Lê, H. M. (2021). Ứng dụng trí tuệ nhân tạo trong giáo dục đại học tại Việt Nam. Tạp chí Giáo dục, 498, 12–18. https://doi.org/10.54404/JER.498.2021.18

Gao, Y., Xiong, Y., Gao, X., et al. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint. https://doi.org/10.48550/arXiv.2312.10997

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. W. (2020). REALM: Retrieval-augmented language model pre-training. *Proceedings of ICML 2020*. https://doi.org/10.48550/arXiv.2002.08909

He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with disentangled attention. *Proceedings of ICLR 2021*. https://doi.org/10.48550/arXiv.2006.03654

Hinton, G., Deng, L., Yu, D., et al. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine, 29*(6), 82–97. https://doi.org/10.1109/MSP.2012.2205597

Hoàng, A. T., & Phạm, N. H. (2020). *Phát triển thư viện số trong bối cảnh chuyển đổi số quốc gia*. *Tạp chí Khoa học Đại học Quốc gia Hà Nội, 36*(2), 45–56. https://doi.org/10.25073/2588-1116/vnupam.4213

Johnson, J., Douze, M., & Jégou, H. (2017). Billion-scale similarity search with GPUs. IEEE *Transactions on Big Data*. https://doi.org/10.1109/TBDATA.2019.2921572

Karpukhin, V., Oguz, B., Min, S., et al. (2020). Dense passage retrieval for open-domain question answering. *Proceedings of EMNLP 2020*. https://doi.org/10.18653/v1/2020.emnlp-main.550

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of EMNLP 2014*. https://doi.org/10.3115/v1/D14-1181

Lê, T. H., & Nguyễn, Q. T. (2023). Chatbot AI hỗ trợ học tập trong môi trường đại học thông minh. *Tạp chí Giáo dục và Xã hội, 145*, 66–73. https://doi.org/10.56794/KHXH.2023.145.66

Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.2005.11401

Li, X., Xiong, C., & Callan, J. (2020). Pre-trained language models for information retrieval.

Proceedings of SIGIR 2020. https://doi.org/10.1145/3397271.3401328

Liu, Y., Ott, M., Goyal, N., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint. https://doi.org/10.48550/arXiv.1907.11692

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint. https://doi.org/10.48550/arXiv.1301.3781

Mikolov, T., Sutskever, I., Chen, K., et al. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1310.4546

Nguyễn, H. T., & Trần, P. V. (2021). Ứng dụng AI trong quản trị tri thức thư viện số. *Tạp chí Thư viện Việt Nam, 3*, 22–31. https://doi.org/10.54404/VNL.2021.31

Nguyễn, M. Q., & Đỗ, T. K. (2022). Phân tích dữ liệu lớn trong giáo dục thông minh. *Tạp chí Công nghệ Thông tin và Truyền thông, 5*, 55–63. https://doi.org/10.32913/mic-ict-research.v2022.n5.1023

OpenAI. (2023). *GPT-4 technical report*. arXiv preprint. https://doi.org/10.48550/arXiv.2303.08774

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. *Proceedings of EMNLP 2014*. https://doi.org/10.3115/v1/D14-1162

Phạm, Q. H., & Võ, M. T. (2022). Mô hình thư viện số thông minh trong giáo dục đại học. *Tạp chí Khoa học và Công nghệ Việt Nam, 64(*9), 48–56. https://doi.org/10.31276/VJST.64(9).48-56

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of EMNLP-IJCNLP 2019*. https://doi.org/10.18653/v1/D19-1410

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval, 3*(4), 333–389. https://doi.org/10.1561/1500000019

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology. *Transactions of the Association for Computational Linguistics, 8*, 842–866. https://doi.org/10.1162/tacl_a_00349

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613–620. https://doi.org/10.1145/361219.361220

Seyed Monir, S. M., Lau, I., Yang, S., & Zhao, D. (2024). VectorSearch: Enhancing document retrieval with semantic embeddings and optimized search. arXiv preprint. https://doi.org/10.48550/arXiv.2409.17383

Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin, 24(*4), 35–43. https://doi.org/10.1109/69.908985

Tạ, K. A. D., & Nguyễn, T. M. (2024). Mô hình AI thủ thư trong thư viện đại học số. *Tạp chí Thông tin và Tư liệu, 2,* 40–52. https://doi.org/10.31276/VJST.66(2).40-52

Trần, N. P., & Lưu, H. T. (2023). Chuyển đổi số và hệ sinh thái học tập đại học thông minh. *Tạp chí Giáo dục, 23*(4), 88–97. https://doi.org/10.54404/JER.2023.234.88

Touvron, H., Lavril, T., Izacard, G., et al. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint. https://doi.org/10.48550/arXiv.2302.13971

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1706.03762

Võ, T. M., & Nguyễn, H. P. (2021). Hệ thống metadata trong thư viện số hiện đại. *Tạp chí Khoa học Thông tin và Tư liệu, 1*, 14–23. https://doi.org/10.31276/VJST.63(1).14-23

Wang, S., & Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. *Proceedings of ACL 2012*. https://doi.org/10.3115/v1/P12-2018

Weibel, S., Kunze, J., Lagoze, C., & Wolf, M. (1998). Dublin core metadata for resource discovery. RFC 2413. https://doi.org/10.17487/RFC2413

Wolf, T., Debut, L., Sanh, V., et al. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of EMNLP 2020*. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Yan, R., Zhao, Y., & Li, X. (2023). Chatbots in libraries: A systematic literature review. *Education for Information, 39*(4), 431–449. https://doi.org/10.3233/EFI-230045

Yang, Z., Dai, Z., Yang, Y., et al. (2019). XLNet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1906.08237

Zhao, W. X., Zhou, K., Li, J., et al. (2023). A survey of large language models. arXiv preprint. https://doi.org/10.48550/arXiv.2303.18223

MÔ HÌNH THƯ VIỆN SỐ THÔNG MINH: *TÍCH HỢP AI THỦ THƯ VÀ DIỄN ĐÀN HỌC THUẬT HỖ TRỢ HỌC TẬP ĐẠI HỌC*

Tạ Khoa Anh Dũng*
Trường Đại học Trưng Vương
ROR ID: https://ror.org/05xzsm645
Email: takhoaanhdung2005@gmail.com
ORCID iD: https://orcid.org/0009-0002-6994-6737
Đỗ Tiến Đạt
Trường Đại học Trưng Vương
ROR ID: https://ror.org/05xzsm645
Email: dovoe456789@gmail.com
ORCID iD: https://orcid.org/0009-0006-0293-5215
Đỗ Văn Bình
Trường Đại học Trưng Vương
ROR ID: https://ror.org/05xzsm645
Email: dovanbinh868@gmail.com
ORCID iD: https://orcid.org/0009-0009-4882-0992
Nguyễn Thị Mỹ Hoa
Trường Đại học Trưng Vương
ROR ID: https://ror.org/05xzsm645
Email: ngmyhoa0@gmail.com
ORCID iD: https://orcid.org/0009-0007-8875-8906
Nguyễn Đức An
Trường Đại học Trưng Vương
ROR ID: https://ror.org/05xzsm645
Email: nguyenducan.hilix26@gmail.com
ORCID iD: https://orcid.org/0009-0004-8889-413X

Lịch sử bài báo

Ngày nhận bài:	10/02/2026
Ngày phản biện:	20/3/2026
Ngày tác giả sửa:	30/4/2026
Ngày duyệt đăng:	28/5/2026
Ngày phát hành:	30/6/2026

DOI: https://doi.org/10.64223/tvj.e2026.v2.i6.a92

Tóm tắt:

Trong bối cảnh chuyển đổi số giáo dục và sự phát triển nhanh của trí tuệ nhân tạo (AI), nhu cầu xây dựng các hệ thống hỗ trợ học tập có khả năng cung cấp thông tin học thuật chính xác, minh bạch và cá nhân hóa ngày càng trở nên cấp thiết. Bài báo đề xuất mô hình thư viện số thông minh tích hợp AI thủ thư và diễn đàn học thuật nhằm hỗ trợ người học trong quá trình tìm kiếm, khai thác, quản lý và kiểm chứng nguồn tài liệu học thuật trong giáo dục đại học.

Hệ thống được phát triển trên cơ sở kết hợp các công nghệ tìm kiếm ngữ nghĩa, tăng cường truy xuất, lưu trữ Vector và mô hình ngôn ngữ cục bộ để xây dựng một môi trường học thuật số đa chức năng. Nền tảng không chỉ hỗ trợ quản lý tài liệu và truy vấn học thuật dựa trên nội dung tài liệu mà còn cung cấp các chức năng như tóm tắt văn bản, ra câu hỏi trắc nghiệm và tổ chức thảo luận học thuật thông qua diễn đàn tương tác. AI thủ thư đóng vai trò như một tác nhân hỗ trợ học tập thông minh, giúp người học tiếp cận tri thức nhanh chóng, đồng thời tăng cường khả năng kiểm chứng thông tin thông qua cơ chế trích dẫn nguồn minh bạch.

Kết quả đánh giá ban đầu với 261 phiếu khảo sát hợp lệ cho thấy người dùng có mức độ hài lòng cao đối với tính hữu ích, khả năng hỗ trợ tự học và mức độ thuận tiện của hệ thống. Các kết quả này bước đầu khẳng định tính khả thi của một mô hình đề xuất trong việc nâng cao hiệu quả học tập, thúc đẩy học tập cá nhân hóa và phát triển môi trường học thuật số trong giáo dục đại học.

Từ khóa: *Thư viện số thông minh; AI thủ thư; Diễn đàn học thuật số; Học tập đại học; Chuyển đổi số giáo dục.*

JEL: A20, I21, I23, K20, K40, M53
ISCED-F: 0421, 0111, 0413
OECD: 5.05, 5.09
FoR: 480307, 390303