Nghiên cứu ứng dụng SLM và RAG trong xây dựng trợ lý học tập nguyên lý kế toán

Lê Trung Thực[1*], Trần Hương Giang[2], Phan Thị Yến[3]
[1]Khoa Công nghệ thông tin, Trường Đại học Công nghệ Đông Á
[2]Khoa Tài chính Kế toán, Trường Đại học Công nghệ Đông Á
[3]Phòng Nghiên cứu Khoa học, Trường Đại học Công nghệ Đông Á
*Email: thuclt@eaut.edu.vn

TÓM TẮT

Hiện nay bối cảnh việc học tập dưới sự hỗ trợ của ứng dụng trí tuệ nhân tạo ngày càng phát triển, nhu cầu hỗ trợ cá nhân hóa tri thức tăng cao. Trong bài báo này, nhóm nghiên cứu tập trung vào mô hình ngôn ngữ nhỏ (Small Language Model – SLM), kỹ thuật Retrieval-Augmented Generation (RAG), các kiến trúc RAG, và đề xuất một hệ thống trợ lý học tập đáng tin cậy, hữu ích cho học phần Nguyên lý kế toán. Mục tiêu của hệ thống là hỗ trợ nhu cầu học tập của sinh viên, đồng thời bảo đảm tính phù hợp với điều kiện về tài nguyên tại Trường Đại học Công nghệ Đông Á. Kiến trúc hệ thống bao gồm ba thành phần chính: kho học liệu được số hóa; bộ phận truy xuất tri thức ứng dụng mô hình nhúng; và mô hình SLM đảm nhiệm vai trò tạo sinh phản hồi cho truy vấn của người học. Sau đó, nhóm thực hiện thử nghiệm, đối sánh 2 mô hình SLM đối với kết quả tạo sinh phản hồi là mô hình mistral và mô hình tinyLLaMA. Kết quả thu được 80% phản hồi đúng trọng tâm nội dung môn học với mô hình tinyLLaMA, 93% đối với mistral, 97% truy vấn có truy xuất đúng học liệu. Từ kết quả thử nghiệm bước đầu, nhóm nghiên cứu nhận thấy SLM và RAG có tiềm năng ứng dụng cao và mở ra khả năng phát triển cho các môn học khác thuộc chương trình đào tạo kế toán, cũng như tích hợp kiến trúc Modular RAG nhằm nâng cao hiệu quả hệ thống trong tương lai.

Từ khóa: SLM, RAG, Naive RAG, trợ lý ảo thông minh, nguyên lý kế toán.

ABSTRACT

In the current landscape of education increasingly supported by artificial intelligence applications, the demand for personalized knowledge assistance is rapidly growing. This paper focuses on the use of Small Language Models (SLMs), the Retrieval-Augmented Generation (RAG) technique, and various RAG architectures to propose a reliable and practical learning assistant system for the Principles of Accounting course. The objective is to support students' learning needs while ensuring compatibility with the resource constraints of East Asia University of Technology. The proposed system architecture consists of three main components: a digitized learning resource repository, a knowledge retrieval module utilizing embedding models, and an SLM responsible for generating responses to student

queries. The study conducted experimental comparisons between two SLMs—Mistral and TinyLLaMA—in response generation. Results show that 80% of responses generated by TinyLLaMA were relevant to the course content, while Mistral achieved 93%, and 97% of all queries retrieved the correct learning materials. These initial results suggest that SLMs and RAG offer significant potential for educational applications and may be extended to other accounting subjects. The study also outlines future directions for integrating a Modular RAG architecture to enhance system performance.

Keywords: SLM, RAG, Naive RAG, intelligent virtual assistant, principles of accounting.

1. MỞ ĐẦU

Sự phát triển mạnh mẽ của các công nghệ trí tuệ nhân tạo (AI) trong những năm gần đây đang đặt nền móng cho những thay đổi trong phương pháp dạy và học, đặc biệt ở bậc đại học. Học tập kết hợp trở thành xu hướng phổ biến, kết hợp giữa học truyền thống với các hình thức học tập số hóa nhằm hỗ trợ, nâng cao hiệu quả tiếp thu kiến thức trong quá trình học tập của sinh viên. Tuy nhiên, trong thực tiễn triển khai sinh viên vẫn gặp khó khăn trong việc chủ động nắm bắt tri thức, đặc biệt với các môn lý thuyết nền tảng như Nguyên lý kế toán – nơi đòi hỏi người học phải hiểu rõ các khái niệm trừu tượng, nguyên tắc ghi sổ và khả năng vận dụng vào tình huống cụ thể.

Để hỗ trợ quá trình học tập trở nên chủ động và cá nhân hóa hơn, nghiên cứu này đề xuất một hệ thống trợ lý học tập thông minh được xây dựng dựa trên mô hình ngôn ngữ nhỏ (SLM) kết hợp kỹ thuật tăng cường truy xuất dữ liệu nội bộ (RAG). SLM được lựa chọn vì khả năng triển khai nhẹ, phù hợp với điều kiện hạ tầng tại nhiều cơ sở giáo dục đại học Việt Nam nói chung và tại Trường Đại học Công nghệ Đông Á nói riêng. RAG giúp bổ sung ngữ cảnh tri thức chuyên ngành cho mô hình tạo sinh phản hồi, hạn chế hiện tượng ảo giác thông tin và nâng cao độ chính xác của phản hồi. Đặc biệt, trong các môn học có hàm lượng lý thuyết cao như Nguyên lý kế toán, sinh viên thường gặp khó khăn trong việc nắm bắt khái niệm, nguyên tắc định khoản và vận dụng bài tập. Việc xây dựng một hệ thống trợ lý học tập dựa trên công nghệ AI – cụ thể là mô hình ngôn ngữ nhỏ (SLM) kết hợp với kỹ thuật RAG – có thể giúp sinh viên truy xuất tri thức một cách tự nhiên, chủ động và hiệu quả hơn.

Trên cơ sở đó, nhóm nghiên cứu xây dựng hệ thống gồm ba thành phần chính: Kho học liệu số hóa (Đề cương chi tiết, chuẩn đầu ra, lịch trình đào tạo, giáo trình, bài giảng, câu hỏi ôn tập, đề cương ôn tập, các đề thi/đáp án tại các năm trước…), bộ truy xuất tri thức sử dụng mô hình nhúng và mô hình sinh phản hồi dựa trên SLM. Hai mô hình ngôn ngữ đã được huấn luyện và tích hợp là *tinyLLaMA* và *mistral* – đều là các mô hình mã nguồn mở dung lượng nhỏ, đã được tinh chỉnh cho tiếng Việt.

Thử nghiệm được tiến hành trên hơn 100 truy vấn, nhằm so sánh hiệu quả của hai mô hình SLM trong việc sinh phản hồi phù hợp với nội dung môn học. Kết quả sơ bộ cho thấy mô hình mistral đạt độ chính xác nội dung cao hơn (93%) so với tinyllama (80%), đồng thời hệ thống có khả năng truy xuất học liệu phù hợp trong 97% truy vấn. Những kết quả này đặt tiền đề cho việc mở rộng ứng dụng mô hình SLM – RAG trong trợ lý học tập các môn chuyên ngành khác, hướng đến sự hỗ trợ linh hoạt, tiết kiệm và cá nhân hóa hơn.

2. CƠ SỞ LÝ THUYẾT

2.1. Mô hình ngôn ngữ nhỏ (Small Language Model – SLM):

SLM là thuật ngữ dùng để chỉ các mô hình ngôn ngữ có quy mô nhỏ hơn đáng kể so với các mô hình ngôn ngữ lớn (Large Language Model – LLM) nhưng vẫn duy trì khả năng sinh ngôn ngữ tự nhiên và thực hiện các tác vụ xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP) thông dụng. Khái niệm này phổ biến từ năm 2023 – 2024, gắn liền với xu hướng phát triển các mô hình mã nguồn mở nhẹ, dễ triển khai, tiêu tốn ít tài nguyên và phù hợp cho các ứng dụng nội bộ. Một số nghiên cứu tiêu biểu như đã chỉ ra rằng, khi được tinh chỉnh đúng cách, SLM có thể đạt hiệu quả gần tương đương LLM trong một số tác vụ cụ thể, đồng thời giúp giảm đáng kể chi phí tính toán [5].

Ưu điểm nổi bật của SLM [5] bao gồm:

- Yêu cầu tài nguyên tính toán thấp hơn nhiều so với LLM;
- Thời gian phản hồi nhanh hơn, đặc biệt khi triển khai tại chỗ;
- Dễ dàng tùy biến và kiểm soát hành vi mô hình;
- Chi phí vận hành thấp – phù hợp với môi trường giáo dục đại học, đặc biệt tại các cơ sở đào tạo không có hạ tầng AI mạnh.

Tuy nhiên, SLM cũng có hạn chế về khả năng tổng quát hóa và dễ bị suy giảm chất lượng khi xử lý các truy vấn phức tạp hoặc yêu cầu suy luận dài [5]. Một số mô hình SLM tiêu biểu được công bố những năm gần đây (từ 2023 – 2024) như:

- Mistral-7B: Mô hình hiệu quả cao với thiết kế attention sliding window, đạt hiệu suất gần tương đương LLaMA-13B nhưng nhẹ hơn.
- TinyLLaMA (1.1B): Mô hình cực nhỏ, huấn luyện hiệu quả và phù hợp cho các thiết bị tài nguyên thấp.
- Phi-2 (2.7B): Do Microsoft phát triển, tập trung vào hiệu suất huấn luyện từ tập dữ liệu chất lượng cao thay vì số lượng lớn.
- Qwen-1.8B: Mô hình tiếng Trung – Anh lai, nổi bật nhờ tối ưu đa ngữ trong quy mô nhỏ.

- OpenHermes-2.5: Mô hình SLM tinh chỉnh từ Mistral, huấn luyện trên tập dữ liệu hội thoại hướng dẫn, được đánh giá cao về khả năng trả lời tự nhiên.
- Nous-Hermes-2 Yi-6B: Mô hình mở rộng từ SLM với khả năng đối thoại tốt, huấn luyện từ dữ liệu web và hỏi đáp chất lượng cao.

Các mô hình trên minh chứng cho xu thế dịch chuyển từ *"bigger is better"* sang *"efficient is better",* trong đó SLM đóng vai trò quan trọng trong việc cá nhân hóa ứng dụng AI cho giáo dục, doanh nghiệp và các tổ chức nghiên cứu quy mô nhỏ.

2.2. Truy xuất tri thức tăng cường (Retrieval-Augmented Generation – RAG):

Retrieval-Augmented Generation (RAG) là kiến trúc kết hợp giữa truy xuất thông tin và mô hình sinh ngôn ngữ, được giới thiệu bởi [4]. Cơ chế hoạt động chính của RAG bao gồm hai giai đoạn:

- Truy xuất các đoạn văn bản liên quan từ một kho tri thức đã lập chỉ mục;
- Sử dụng những đoạn văn bản trên làm ngữ cảnh để mô hình ngôn ngữ tạo sinh phản hồi đầu ra.

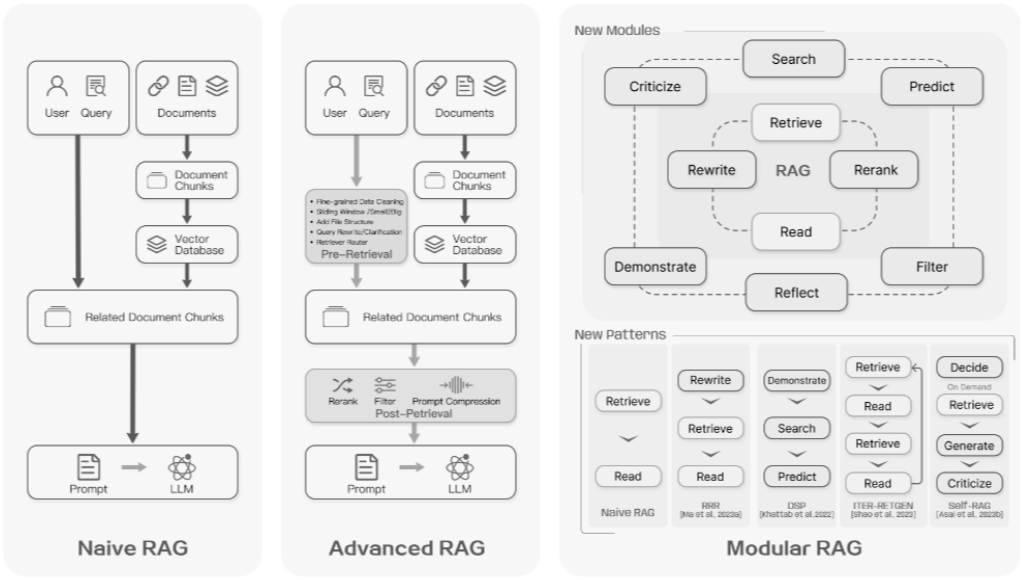
Nhờ vào khả năng cập nhật tri thức ngoại sinh một cách linh hoạt, RAG đã chứng minh hiệu quả vượt trội trong các tác vụ NLP đòi hỏi độ chính xác cao và tri thức cập nhật. Hiện nay, có ba biến thể kiến trúc RAG tiêu biểu đang được sử dụng rộng rãi trong nghiên cứu và ứng dụng: RAG cơ bản (Naive RAG), RAG nâng cao (Advance RAG) và Modular RAG, mỗi biến thể tương ứng với một mức độ kiểm soát tối ưu cho từng nhu cầu nghiệp vụ và hệ thống khác nhau.

Naive RAG là kiến trúc cơ bản nhất, trong đó truy vấn của người dùng được chuyển thành vector nhúng và dùng để truy xuất top-k đoạn văn bản từ cơ sở dữ liệu. Sau đó, tất cả các đoạn này được đưa trực tiếp cùng với truy vấn vào mô hình tạo sinh phản hồi. Ưu điểm lớn của Naive RAG là dễ cài đặt, phù hợp với các ứng dụng nhỏ hoặc không yêu cầu kiểm soát gắt gao nguồn tri thức. Tuy nhiên, do không có bước lọc hoặc đánh giá lại chất lượng thông tin truy xuất, mô hình tạo sinh có thể dựa vào các đoạn không liên quan hoặc chứa thông tin sai lệch, làm giảm độ tin cậy của đầu ra [4]. Hơn nữa, khi số lượng tài liệu lớn hoặc đa dạng về chủ đề, mô hình dễ bị nhiễu bởi các đoạn không đúng ngữ cảnh. Việc gộp toàn bộ các đoạn vào mô hình sinh cũng có thể vượt quá giới hạn từ, dẫn đến mất mát thông tin đầu vào hoặc làm giảm hiệu suất xử lý. Do đó, Naive RAG tuy hiệu quả trong môi trường nhỏ và dữ liệu đồng nhất, nhưng không phù hợp với các hệ thống đòi hỏi độ chính xác cao và kiểm soát nguồn tri thức chuyên sâu, đa nguồn.

Advance RAG khắc phục nhược điểm trên bằng cách thêm một bước đánh giá lại (re-ranking) sau truy xuất. Sau khi hệ thống tìm top-k đoạn văn bản liên quan, một mô hình reranker – thường là mô hình huấn luyện học sâu – được sử dụng để xếp

hạng lại các đoạn theo độ liên quan đến truy vấn. Chỉ các đoạn có điểm số cao nhất mới được chọn làm đầu vào cho mô hình tạo sinh. Cách tiếp cận này giúp giảm đáng kể nhiễu, tăng độ chính xác và tính mạch lạc của phản hồi [1]. Ưu điểm lớn nhất của kiến trúc RAG nâng cao là khả năng kiểm soát chất lượng thông tin và đảm bảo rằng mô hình sinh chỉ sử dụng các nguồn đáng tin cậy. Tuy nhiên, chi phí tính toán cao hơn do bổ sung bước đánh giá lại, đồng thời hệ thống trở nên phức tạp hơn về mặt kiến trúc. Kiến trúc này đòi hỏi dữ liệu huấn luyện reranker chất lượng cao, khả năng xử lý đồng thời nhiều mô hình, nên phù hợp hơn với các hệ thống có yêu cầu chính xác cao và hạ tầng mạnh.

Modular RAG là biến thể hiện đại của RAG với đặc điểm nổi bật là kiến trúc tách biệt hoàn toàn giữa ba thành phần: truy xuất (retriever), đánh giá lại (reranker) và tạo sinh phản hồi (generator). Cách thiết kế này cho phép huấn luyện, điều chỉnh và tối ưu riêng biệt từng mô-đun theo mục tiêu hoặc dữ liệu cụ thể [2]. Ưu điểm lớn nhất của Modular RAG là tính linh hoạt, khả năng mở rộng và dễ tích hợp với các pipeline có sẵn. Ngoài ra, nó còn tạo điều kiện thuận lợi cho việc thử nghiệm các thuật toán khác nhau tại từng bước xử lý. Nhược điểm chính là yêu cầu hạ tầng triển khai cao hơn và cần nhiều công đoạn huấn luyện mô-đun riêng, đặc biệt khi triển khai trong môi trường hạn chế tài nguyên hoặc thời gian phản hồi yêu cầu nhanh [2].



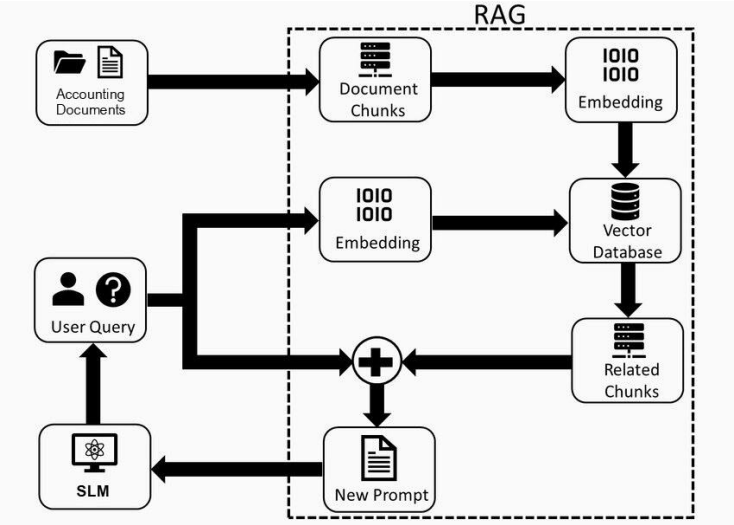
Hình 1. Các kiến trúc RAG [2]

3. MÔ HÌNH ĐỀ XUẤT

Dựa trên phân tích các kiến trúc RAG và khả năng triển khai hiệu quả trong môi trường giáo dục đại học có tài nguyên giới hạn, nhóm nghiên cứu đề xuất mô hình hệ thống trợ lý học tập học phần Nguyên lý kế toán sử dụng kiến trúc Naive RAG kết

hợp với mô hình ngôn ngữ nhỏ (SLM). Mục tiêu của hệ thống là hỗ trợ sinh viên truy xuất nhanh chóng, chính xác các nội dung học tập cốt lõi, đồng thời tạo điều kiện học tập cá nhân hóa theo nhu cầu.

Cấu trúc tổng thể của hệ thống bao gồm ba thành phần chính: kho học liệu nguyên lý kế toán được số hóa, bộ truy xuất tri thức sử dụng mô hình nhúng và mô hình tạo sinh phản hồi SLM như "bộ não" của trợ lý này. Trong đó, kho học liệu đầu vào bao gồm các tài liệu học thuật quan trọng của học phần Nguyên lý kế toán đang được áp dụng tại Khoa Tài chính Kế toán, Trường Đại học Công nghệ Đông Á như: đề cương chi tiết môn học, chuẩn đầu ra của học phần, lịch trình học tập, bài giảng, tài liệu tham khảo, bộ câu hỏi ôn tập, đề thi và đáp án các năm học trước... Các tài liệu này được tiền xử lý, phân đoạn và mã hóa thành vector nhúng bằng mô hình nhúng phù hợp với tiếng Việt chuyên ngành (trong bài báo này nhóm sử dụng mô hình sentence-transformer). Các vector thu được sẽ được lưu trữ trong cơ sở dữ liệu vector để phục vụ truy xuất.



Hình 2. Mô hình RAG, SLM do nhóm đề xuất

Khi sinh viên nhập một truy vấn dưới dạng câu hỏi tự nhiên, hệ thống sẽ ánh xạ câu hỏi thành vector và truy xuất top-k đoạn học liệu có mức độ tương đồng cao nhất từ cơ sở dữ liệu. Sau đó, toàn bộ các đoạn này cùng với truy vấn gốc sẽ được đưa vào mô hình SLM để sinh ra phản hồi. Trong kiến trúc đề xuất, mô hình SLM đóng vai trò là bộ não của trợ lý học tập, có nhiệm vụ tổng hợp, diễn giải và trình bày thông tin từ các đoạn truy xuất một cách rõ ràng, mạch lạc và phù hợp với ngữ cảnh người học.

Ưu điểm nổi bật của hệ thống là đơn giản về mặt kiến trúc, phù hợp với điều kiện triển khai tại các cơ sở giáo dục đại học có hạ tầng giới hạn. Nhờ sử dụng Naive RAG, hệ thống tránh được yêu cầu về mô hình re-ranking trong khi vẫn đảm bảo khả năng khai thác tri thức từ kho học liệu. Mặt khác, việc lựa chọn mô hình SLM giúp tiết kiệm chi phí vận hành mà vẫn duy trì chất lượng phản hồi nếu được tinh chỉnh

đúng cách. Tuy nhiên, để giảm thiểu các hạn chế của Naive RAG như nguy cơ nhiễu thông tin, hệ thống cần được xây dựng trên một kho học liệu được biên soạn kỹ lưỡng, có cấu trúc rõ ràng và nhất quán.

Hệ thống đề xuất có thể tích hợp dưới dạng ứng dụng web, ứng dụng mobile, giao diện chatbot hoặc tích hợp vào nền tảng học tập điện tử sử dụng tại trường. Sinh viên có thể truy vấn kiến thức theo từng chủ đề như nguyên tắc kế toán, định khoản, hệ thống tài khoản hoặc cách làm bài tập thực hành. Nhờ khả năng tương tác tự nhiên và truy xuất dựa trên nội dung thực tế từ học liệu, trợ lý học tập này không chỉ giúp tiết kiệm thời gian tra cứu mà còn nâng cao khả năng học tập chủ động và chính xác của người học.

4. THỬ NGHIỆM VÀ ĐÁNH GIÁ

Để kiểm chứng tính khả thi của mô hình đề xuất, nhóm nghiên cứu đã triển khai một hệ thống thử nghiệm đầy đủ với các thành phần: giao diện ứng dụng, mô hình nhúng, cơ sở dữ liệu vector và mô hình ngôn ngữ nhỏ tạo sinh phản hồi.

Ở tầng ứng dụng, UI/UX dành cho người dùng nhóm cài đặt Streamlit – một framework mã nguồn mở nhẹ, phù hợp cho việc phát triển giao diện web tương tác đơn giản trong nghiên cứu khoa học. Streamlit hỗ trợ tích hợp nhanh mô hình AI, đồng thời cung cấp trải nghiệm tương tác trực quan cho người dùng cuối – trong trường hợp này là sinh viên chuyên ngành kế toán.

Về truy xuất tri thức, hệ thống sử dụng mô hình nhúng từ thư viện sentence-transformers [3] để chuyển đổi các câu hỏi của người học và các đoạn tài liệu thành biểu diễn vector. Trong bối cảnh ngôn ngữ tiếng Việt, nhóm sử dụng mô hình *distiluse-base-multilingual-cased-v1* – một mô hình đa ngôn ngữ hiệu quả và nhẹ, tương thích tốt với dữ liệu kế toán mang tính liệt kê, định nghĩa, quy tắc.

Các vector sinh ra được lưu trữ trong ChromaDB – một cơ sở dữ liệu vector hiện đại, hỗ trợ truy vấn nhanh trên các không gian vector có kích thước lớn. Việc sử dụng ChromaDB giúp đảm bảo tốc độ truy xuất tốt ngay cả khi mở rộng quy mô học liệu trong tương lai.

Ở tầng tạo phản hồi, nhóm thử nghiệm với hai mô hình ngôn ngữ nhỏ là mistral-7B và tinyLLaMA (1.1B) nhằm đối sánh năng lực. Các mô hình này đều được triển khai cục bộ thông qua API, có thể nhận đầu vào gồm truy vấn và các đoạn văn bản đã truy xuất từ ChromaDB. Kết quả phản hồi được hiển thị ngay trên giao diện streamlit để phục vụ so sánh chất lượng giữa hai mô hình và thu thập phản hồi người dùng.

Bảng 1. Kết quả phản hồi của các mô hình mistral, tinyLLaMA

Câu hỏi sinh viên	Phản hồi Mistral-7B	Phản hồi TinyLLaMA (1.1B)
"Tài khoản 111 dùng để làm gì?"	Tài khoản 111 dùng để phản ánh tiền mặt hiện có và các giao dịch liên quan đến tiền mặt tại doanh nghiệp.	TK 111 là tài khoản trong kế toán dùng cho các khoản liên quan đến tiền.
"Khi mua hàng hóa chưa trả tiền thì ghi vào đâu?"	Khi mua hàng hóa chưa trả tiền, kế toán ghi Nợ TK Hàng tồn kho, Có TK Phải trả người bán (TK 331).	Ghi vào khoản mua chưa trả, trong tài khoản nợ và có, cụ thể là khoản phải trả.
"Kế toán kép là gì?"	Kế toán kép là nguyên tắc ghi nhận một nghiệp vụ vào ít nhất hai tài khoản, đảm bảo cân đối giữa bên Nợ và Có.	Là một khái niệm về kế toán, có nhiều tài khoản liên quan cùng ghi nhận.
"Phân biệt tài sản ngắn hạn và dài hạn?"	Tài sản ngắn hạn là tài sản có thời gian sử dụng hoặc thu hồi dưới 12 tháng, dài hạn thì trên 12 tháng.	Ngắn hạn là ngắn, dài hạn là dài, phụ thuộc vào thời gian.
"Doanh thu ghi nhận khi nào?"	Doanh thu được ghi nhận khi doanh nghiệp đã chuyển giao hàng hóa hoặc hoàn thành dịch vụ cho khách hàng.	Khi bán hàng hoặc có tiền thì ghi doanh thu.

Trong nội dung thử nghiệm và đánh giá, nhóm phối hợp cùng các giảng viên phụ trách học phần, giảng viên giảng dạy học phần Nguyên lý kế toán xây dựng mẫu 100 câu hỏi chuyên môn. Sau đó, nhóm nghiên cứu cùng nhóm giảng viên kiểm thử nhằm đánh giá tính đúng đắn chuyên môn, chất lượng phù hợp của phản hồi từ hệ thống.

Bảng 2. Đối sánh tỷ lệ phản hồi hiệu quả của hai mô hình SLM trong hệ thống trợ lý học tập nguyên lý kế toán (Dựa trên 100 mẫu câu hỏi đầu vào)

TT	Tiêu chí đánh giá	Mistral-7B	TinyLLaMA (1.1B)
1	Tỷ lệ phản hồi đúng trọng tâm *(%)*	93	80
2	Tỷ lệ sử dụng đúng thuật ngữ kế toán *(%)*	92	76
3	Độ mạch lạc và dễ hiểu của câu trả lời *(thang điểm từ 1-5)*	4.6	3.8
4	Tốc độ phản hồi trung bình *(giây)*	1.8	1.4
5	Số lỗi ngữ nghĩa nghiêm trọng *(%)*	2	9

5. TRAO ĐỔI VÀ KẾT LUẬN

Với các kết quả thực nghiệm đạt được, nghiên cứu khẳng định tính khả thi và hiệu quả bước đầu của việc ứng dụng mô hình SLM kết hợp Naive RAG trong xây dựng hệ thống trợ lý học tập cho học phần Nguyên lý kế toán. Tuy nhiên, mô hình hiện tại còn hạn chế ở khả năng kiểm soát chất lượng truy xuất tri thức và độ sâu phản hồi trong các truy vấn phức tạp. Ngoài ra, việc đánh giá vẫn còn dựa trên dữ liệu giả lập và số lượng mẫu chưa lớn. Trong các nghiên cứu tiếp theo, nhóm sẽ mở rộng thử nghiệm trên nhiều môn học kế toán khác, tích hợp kiến trúc Advance/Modular RAG nhằm nâng cao tính chính xác và khả năng thích ứng của hệ thống. Đồng thời, việc phát triển bộ đánh giá chuẩn hóa phản hồi AI trong giáo dục đại học cũng được đề xuất như một hướng nghiên cứu độc lập có giá trị.

TÀI LIỆU THAM KHẢO

[1]. G. Izacard and E. Grave (2021). Leveraging passage retrieval with generative models for open domain question answering, *arXiv preprint*, arXiv:2007, 01282, Website: https://arxiv.org/abs/2007.01282.

[2]. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A Surve, *arXiv preprint*, arXiv:2312, 10997, Website: https://arxiv.org/abs/2312.10997.

[3]. Nils Reimers, Iryna Gurevych (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, *EMNLP-IJCNLP*, Hong Kong, pp. 3982–3992

[4]. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, et al (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks, *arXiv preprint*, arXiv:2005, 11401, Website: https://arxiv.org/abs/2005.11401.

[5]. Phan Thị Yến, Lê Văn Chung, Lê Trung Thực (2024). Nghiên cứu mô hình ngôn ngữ nhỏ (SLM) trong công tác giảng dạy lập trình: cơ hội và thách thức, *Kỷ yếu Hội thảo khoa học quốc gia lần thứ 3*, Hà Nội, Tập 1, tr. 319–329.