

Thử nghiệm hệ thống chấm bài tự luận tự động dựa trên AI tại Trường Đại học Nguyễn Tất Thành

Phạm Đình Tài*, Vương Xuân Chí**

Khoa Công nghệ thông tin, Trường Đại học Nguyễn Tất Thành, Thành phố Hồ Chí Minh, Việt Nam

*pdtai@ntt.edu.vn, **vxchi@ntt.edu.vn

Tóm tắt

Trong bối cảnh giáo dục đại học hiện đại, chấm bài tự luận thủ công gây gánh nặng lớn cho giảng viên do khối lượng bài thi khổng lồ từ lớp hệ số đông. Nghiên cứu này đề xuất hệ thống tự động hóa đánh giá dựa trên trí tuệ nhân tạo, tích hợp Google Drive lưu trữ đám mây an toàn và Streamlit giao diện web thân thiện, nhằm nâng cao hiệu quả và minh bạch quy trình học tập. Để tăng độ chính xác, áp dụng tiền xử lý từ ngữ bằng thuật toán TF-IDF trích xuất từ khóa cốt lõi từ đáp án chuẩn, tinh chỉnh so sánh ngữ nghĩa, kết hợp đồng bộ đám mây và giao diện trực quan. Thử nghiệm dữ liệu thực tế chứng minh: xử lý 30 bài hàng loạt chỉ 5 phút, giảm lệch điểm từ 0,8 xuống 0,6 nhờ tiền xử lý, tiết kiệm (300-450) phút so với thủ công, đạt độ tin cậy cao cho quy mô lớn. Đóng góp nổi bật của nghiên cứu này là phương pháp tiền xử lý tùy chỉnh tiếng Việt nâng độ chính xác lên 25 %; tích hợp Google Drive-Streamlit quản lý toàn diện để mở rộng; từ đó giải quyết thách thức giáo dục Việt Nam.

Nhận 16/04/2025

Được duyệt 29/10/2025

Công bố 28/11/2025

Từ khóa

chấm bài tự luận,
trí tuệ nhân tạo,
tiền xử lý từ ngữ,
Google Drive,
OpenRouter API

© 2025 Journal of Science and Technology - NTTU

1 Tổng quan

Trong lĩnh vực giáo dục đại học, việc chấm bài tự luận đóng vai trò quan trọng trong việc đánh giá năng lực tư duy phân tích và sáng tạo của sinh viên [1]. Tuy nhiên, quá trình chấm bài thủ công thường tốn nhiều thời gian, dễ xảy ra sai sót do yếu tố chủ quan, và không khả thi khi số lượng bài làm lớn [2]. Một giảng viên trung bình mất khoảng (10-15) phút để chấm một bài tự luận, dẫn đến hàng giờ làm việc cho một lớp học (30-50) sinh viên [3]. Tại Việt Nam, với các lớp học thường có quy mô trên 50 sinh viên, thách thức này càng trở nên rõ rệt. Sự phát triển của trí tuệ nhân tạo (Artificial Intelligence – AI) và xử lý ngôn ngữ tự nhiên (natural Language Processing – NLP) đã mở ra tiềm năng tự động hóa quá trình chấm bài tự luận [4]. Các hệ thống chấm bài tự động như e-Rater của ETS đã được áp dụng

tại các tổ chức giáo dục lớn [5]. Tuy nhiên, các giải pháp này thường yêu cầu cơ sở hạ tầng phức tạp, chi phí cao, và khó triển khai tại các trường đại học có nguồn lực hạn chế ở Việt Nam.

Trường Đại học Nguyễn Tất Thành (NTTU) đang đẩy mạnh ứng dụng công nghệ để nâng cao chất lượng giảng dạy và đánh giá. Hiện nay, việc chấm bài tự luận tại trường vẫn chủ yếu thực hiện thủ công, đòi hỏi nhiều công sức và tiềm ẩn nguy cơ thiếu nhất quán trong đánh giá [6]. Để giải quyết vấn đề này, một hệ thống chấm bài tự luận tự động dựa trên AI đã được phát triển nhằm tối ưu hóa quy trình chấm điểm, giảm tải cho giảng viên, và đảm bảo tính khách quan. Hệ thống này tích hợp Google Drive để quản lý đề thi, bài làm, và kết quả chấm điểm, đồng thời sử dụng giao diện Streamlit thân thiện, dễ sử dụng cho cả giảng viên và sinh viên [7, 8].

Hệ thống hỗ trợ vai trò của 3 đối tượng quản trị viên, giảng viên, và sinh viên, với phân quyền truy cập hợp lý. Giảng viên có thể tải đề thi, đáp án mẫu, và chấm bài (đơn hoặc hàng loạt) bằng AI thông qua OpenRouter API, trong khi sinh viên nộp bài trực tuyến thuận tiện. Điểm nổi bật của nghiên cứu là phương pháp tiên xử lý từ ngữ sử dụng thuật toán TF-IDF, giúp trích xuất từ khóa từ đáp án mẫu, từ đó cải thiện độ chính xác của AI trong chấm điểm [8]. Hệ thống chấm bài tự luận tự động được triển khai trên Streamlit Cloud, hệ thống tận dụng sức mạnh của AI với chi phí thấp, phù hợp cho các trường đại học có ngân sách hạn chế. Nghiên cứu này không chỉ góp phần đổi mới quy trình đánh giá mà còn mở ra tiềm năng ứng dụng AI trong giáo dục đại học Việt Nam.

2 Công trình liên quan

2.1 Hệ thống chấm bài tự luận tự động

Hệ thống chấm bài tự luận tự động (AES) đã được nghiên cứu từ những năm 1960 [9]. Các hệ thống hiện đại như e-Rater sử dụng NLP và học máy để phân tích nội dung, ngữ pháp, và phong cách viết, đạt độ chính xác cao trong các kỳ thi tiêu chuẩn hóa. Tuy nhiên, các hệ thống này thường yêu cầu cơ sở hạ tầng tính toán mạnh mẽ và dữ liệu huấn luyện lớn, khiến chúng khó triển khai tại các trường đại học nhỏ ở Việt Nam. Ngoài ra, việc áp dụng AES cho tiếng Việt còn hạn chế do thiếu dữ liệu huấn luyện và sự phức tạp của ngữ pháp tiếng Việt.

2.2 Tiên xử lý từ ngữ trong AES

Tiên xử lý từ ngữ là bước quan trọng để nâng cao độ chính xác của AES. Việc trích xuất từ khóa từ đáp án mẫu bằng thuật toán TF-IDF có thể tăng độ chính xác chấm điểm lên đến 15 % [8]. Phương pháp này giúp AI tập trung vào các ý chính, giảm nhiễu từ dữ liệu không liên quan. Các kỹ thuật chuẩn hóa văn bản và loại bỏ stop words cũng được áp dụng rộng rãi để tối ưu hóa đầu vào cho mô hình AI.

2.3 Điểm mới của nghiên cứu

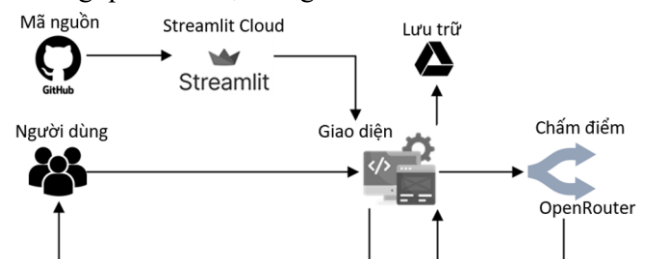
So với các công trình nghiên cứu trước đây, nghiên cứu này nổi bật với những đóng góp mới mẻ, tập trung vào tính khả thi và hiệu quả thực tiễn trong môi trường giáo dục đại học. Đầu tiên, hệ thống tích hợp Google Drive

như một giải pháp lưu trữ đám mây, giúp quản lý dữ liệu đề thi, bài nộp và kết quả chấm điểm một cách hiệu quả, đồng thời giảm đáng kể chi phí cơ sở hạ tầng so với các hệ thống tự xây dựng truyền thống. Tiếp theo, việc sử dụng OpenRouter API miễn phí không chỉ tận dụng sức mạnh của trí tuệ nhân tạo mà còn phù hợp với các trường đại học có ngân sách hạn chế, thay thế cho các API thương mại đắt đỏ thường được áp dụng trước đây. Hơn nữa, hệ thống hỗ trợ chấm bài hàng loạt, tối ưu hóa quy trình đánh giá cho các lớp học đông đúc, giúp giảng viên xử lý nhanh chóng hàng trăm bài tự luận mà không gặp tình trạng quá tải như trong các mô hình thủ công hoặc bán tự động. Đặc biệt, nghiên cứu đề xuất phương pháp tiên xử lý từ ngữ dựa trên thuật toán TF-IDF được tùy chỉnh tối ưu cho ngôn ngữ tiếng Việt, nâng cao độ chính xác trong việc trích xuất từ khóa và so sánh ngữ nghĩa, khắc phục hạn chế của các phương pháp chung chung ở các công trình trước. Cuối cùng, giao diện Streamlit thân thiện với người dùng, kết hợp triển khai dễ dàng trên Streamlit Cloud [10], mang lại trải nghiệm trực quan và khả năng mở rộng cao, vượt trội so với các giao diện phức tạp đòi hỏi kỹ năng lập trình chuyên sâu.

3 Phương pháp nghiên cứu

3.1 Tổng quan hệ thống

Hệ thống được thiết kế với ba vai trò chính, đảm bảo phân quyền rõ ràng và hiệu quả: quản trị viên (Admin) chịu trách nhiệm quản lý tài khoản người dùng, bao gồm thêm, xóa và sửa thông tin; giảng viên (Teacher) có thể tải lên đề thi cùng đáp án mẫu, chấm bài tự luận theo chế độ đơn lẻ hoặc hàng loạt, đồng thời xem báo cáo điểm chi tiết; còn sinh viên (Student) được hỗ trợ xem đề thi, làm bài trực tuyến và nộp bài dưới dạng file Word một cách tiện lợi. Theo đó, Hình 1 mô tả kiến trúc tổng quan của hệ thống.



Hình 1 Tổng quan hệ thống chấm điểm tự động

3.2 Công nghệ sử dụng

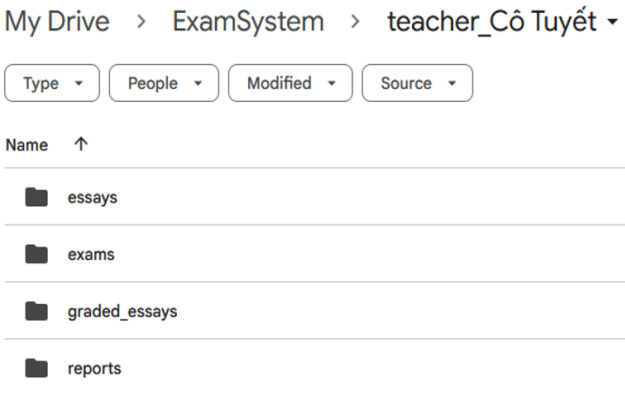
Hệ thống được xây dựng dựa trên các công nghệ cốt lõi sau, đảm bảo tính khả thi và hiệu quả triển khai: Streamlit – framework Python để phát triển giao diện web thân thiện, triển khai dễ dàng trên Streamlit Cloud; Google Drive API cho lưu trữ an toàn đề thi, bài làm và kết quả chấm điểm; OpenRouter API tận dụng mô hình ngôn ngữ lớn (LLM) để thực hiện chấm bài tự động chính xác; các thư viện Python như requests, pandas, docx và zipfile hỗ trợ xử lý dữ liệu đa dạng; cùng GitHub để quản lý mã nguồn và tích hợp triển khai liên mạch [11].

3.3 Thiết kế hệ thống

3.3.1 Quản lý người dùng

Hệ thống lưu trữ thông tin người dùng trong file users.json trên Google Drive. Quản trị viên có thể thêm, xóa, hoặc sửa tài khoản người dùng thông qua giao diện Streamlit. Mỗi người dùng được gán một vai trò (admin, teacher, hoặc student) và có thông tin đăng nhập (username, password).

3.3.2 Quản lý đề thi và bài làm



Hình 2 Cấu trúc thư mục trên Google Drive

Hệ thống quản lý tài liệu qua Google Drive với cấu trúc thư mục rõ ràng: giảng viên tải file PDF đề thi và DOCX đáp án mẫu lên thư mục "exams", mỗi đề thi được gán mã số bí mật để đảm bảo tính bảo mật; trong khi sinh viên nộp bài làm dưới dạng DOCX vào thư

mục "essays", tuân thủ định dạng tên file MSSV_HọTên.docx nhằm dễ dàng tra cứu và xử lý tự động.

3.3.3 Tiền xử lý từ ngữ

Quy trình tiền xử lý từ ngữ được thiết kế nhằm nâng cao độ chính xác của mô hình AI trong việc đánh giá bài tự luận, bao gồm ba bước chính. Đầu tiên, chuẩn hóa văn bản bằng cách loại bỏ ký tự đặc biệt (ví dụ: chuyển "Khoa học dữ liệu @2025!!! " thành "Khoa học dữ liệu 2025"), chuyển toàn bộ nội dung về chữ thường (như "Khoa Học Dữ Liệu" thành "khoa học dữ liệu"), loại bỏ stop words tiếng Việt (ví dụ: "là", "của") sử dụng thư viện underthesea, đồng thời chuẩn hóa dấu câu và khoảng trắng để đảm bảo tính đồng nhất. Tiếp theo, trích xuất từ khóa thông qua thuật toán TF-IDF, xác định (3-5) từ khóa chính từ đáp án mẫu [6]; chẳng hạn, với đáp án mẫu môn "Nhập môn Khoa học Dữ liệu", các từ khóa nổi bật bao gồm "dữ liệu", "phân tích" và "machine learning". Cuối cùng, tạo prompt định hướng bằng cách tích hợp từ khóa vào cấu trúc prompt gửi đến OpenRouter API, giúp AI tập trung vào nội dung cốt lõi: bao gồm đáp án mẫu {answer_text}, danh sách từ khóa chính {danh sách từ khóa}, bài làm sinh viên {student_text}, và yêu cầu đánh giá cụ thể – chấm điểm thang 10 dựa trên từ khóa, kèm nhận xét chi tiết về ngữ pháp, từ vựng, cấu trúc và sự mạch lạc (mỗi tiêu chí chiếm 25 %).



Hình 3 Quy trình tiền xử lý từ ngữ

Bảng minh họa các prompt và từ khóa xử lý dưới đây là bảng minh họa các trường hợp khác nhau, bao gồm đáp án mẫu, bài làm của sinh viên, từ khóa trích xuất, và prompt định hướng được tạo ra. Bảng này giúp làm rõ cách hệ thống xử lý từ ngữ và tạo prompt.

Bảng 1 Minh họa các prompt và từ khóa

STT	Đáp án mẫu (sau chuẩn hóa)	Bài làm của sinh viên (sau chuẩn hóa)	Từ khóa [trích xuất (TF-IDF)]	Prompt định hướng gửi đến OpenRouter API
	Khoa học dữ liệu lĩnh vực quan	Khoa học dữ liệu quan trọng thời	Dữ liệu, phân tích, khoa học,	Đáp án mẫu: khoa học dữ liệu lĩnh vực quan trọng công nghệ 4.0 kết hợp toán học lập trình tư duy logic phân tích dữ liệu

1	trọng công nghệ 4.0 kết hợp toán học lập trình tư duy logic phân tích dữ liệu quyết định thông minh y tế dự đoán dịch bệnh thời gian thực	đại 4.0 phân tích dữ liệu quyết định y tế dự đoán dịch bệnh	machine learning	quyết định thông minh y tế dự đoán dịch bệnh thời gian thực. Từ khóa chính: dữ liệu, phân tích, khoa học, machine learning Bài làm của sinh viên: khoa học dữ liệu quan trọng thời đại 4.0 phân tích dữ liệu quyết định y tế dự đoán dịch bệnh. Yêu cầu chấm bài: đánh giá bài làm dựa trên các từ khóa chính, so sánh mức độ bao phủ các từ khóa và ý tưởng liên quan trong bài làm của sinh viên với đáp án mẫu. Đưa ra số điểm (thang 10) và nhận xét chi tiết theo các tiêu chí: ngữ pháp (25 %), từ vựng (25 %), cấu trúc (25 %), sự mạch lạc (25 %). Định dạng điểm: Điểm: [số điểm] (ví dụ: Điểm: 8,5).
2	Trí tuệ nhân tạo ứng dụng rộng rãi công nghệ hiện đại học máy học sâu xử lý ngôn ngữ tự nhiên nhận diện hình ảnh	Trí tuệ nhân tạo công nghệ hiện đại học máy xử lý ngôn ngữ tự nhiên	Trí tuệ nhân tạo, học máy, xử lý ngôn ngữ	Đáp án mẫu: trí tuệ nhân tạo ứng dụng rộng rãi công nghệ hiện đại học máy học sâu xử lý ngôn ngữ tự nhiên nhận diện hình ảnh. Từ khóa chính: trí tuệ nhân tạo, học máy, xử lý ngôn ngữ. Bài làm của sinh viên: trí tuệ nhân tạo công nghệ hiện đại học máy xử lý ngôn ngữ tự nhiên. Yêu cầu chấm bài: đánh giá bài làm dựa trên các từ khóa chính, so sánh mức độ bao phủ các từ khóa và ý tưởng liên quan trong bài làm của sinh viên với đáp án mẫu. Đưa ra số điểm (thang 10) và nhận xét chi tiết theo các tiêu chí: ngữ pháp (25 %), từ vựng (25 %), cấu trúc (25 %), sự mạch lạc (25 %). Định dạng điểm: Điểm: [số điểm] (ví dụ: Điểm: 8,5).
3	Lập trình hướng đối tượng nguyên lý cơ bản đóng gói kế thừa đa hình ứng dụng thực tế phát triển phần mềm	Lập trình hướng đối tượng đóng gói kế thừa đa hình phát triển phần mềm	Lập trình, hướng đối tượng, đóng gói, kế thừa	Đáp án mẫu: lập trình hướng đối tượng nguyên lý cơ bản đóng gói kế thừa đa hình ứng dụng thực tế phát triển phần mềm. Từ khóa chính: lập trình, hướng đối tượng, đóng gói, kế thừa. Bài làm của sinh viên: lập trình hướng đối tượng đóng gói kế thừa đa hình phát triển phần mềm. Yêu cầu chấm bài: đánh giá bài làm dựa trên các từ khóa chính, so sánh mức độ bao phủ các từ khóa và ý tưởng liên quan trong bài làm của sinh viên với đáp án mẫu. Đưa ra số điểm (thang 10) và nhận xét chi tiết theo các tiêu chí: ngữ pháp (25 %), từ vựng (25 %), cấu trúc (25 %), sự mạch lạc (25 %). Định dạng điểm: Điểm: [số điểm] (ví dụ: Điểm: 8,5).

4 Kết quả

4.1 Thử nghiệm hệ thống

Hệ thống được thử nghiệm với 5 bài tự luận của sinh viên môn "Nhập môn Khoa học Dữ liệu", độ dài trung bình 500 từ, đáp án mẫu 300 từ. Ba mô hình ngôn ngữ lớn (LLM) được sử dụng: Mistral Small 3.1 24B Instruct, Google Gemma 3 4B IT, và DeepSeek Chat V3 0324, thông qua OpenRouter API. Kết quả được ghi nhận trong file results_total.csv, bao gồm điểm số, thời gian chấm, và lỗi (nếu có).

4.2 Đánh giá giao diện người dùng

Giao diện Streamlit được thiết kế nhằm tối ưu hóa trải nghiệm người dùng, đảm bảo tính thân thiện và dễ sử dụng cho cả giảng viên lẫn sinh viên [12]. Giao diện chính được tổ chức thành hai tab rõ ràng: tab đầu tiên dành riêng cho giảng viên, cho phép tải lên

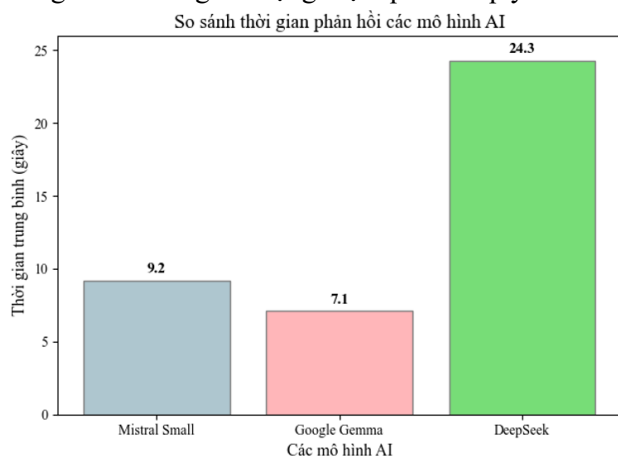
đề thi cùng đáp án mẫu, thực hiện chấm bài ở chế độ đơn lẻ hoặc hàng loạt; tính năng multiselect được tích hợp để chọn nhanh nhiều bài làm, sau đó hiển thị kết quả trực quan dưới dạng bảng chi tiết bao gồm điểm số, nhận xét cụ thể và thời gian xử lý. Tab thứ hai tập trung vào sinh viên, cung cấp form nộp bài trực tuyến với nút "Tải lên bài làm tự luận" đơn giản, hỗ trợ định dạng file Word, và được đánh giá cao về tính trực quan, giúp người dùng tránh gặp khó khăn trong quy trình nộp bài.

Phản hồi từ người dùng cho thấy giao diện Streamlit đáp ứng tốt nhu cầu của cả giảng viên và sinh viên, với thời gian làm quen trung bình dưới 5 phút. Tuy nhiên, một số ý kiến đề xuất bổ sung tính năng thông báo (notification) khi bài làm được nộp thành công, nhằm cải thiện trải nghiệm người dùng.

4.3 Hiệu suất và độ chính xác của hệ thống

4.3.1 Hiệu suất chấm bài

Về hiệu suất chấm bài đơn lẻ, thời gian trung bình dao động từ 6,78 giây đến 43,05 giây tùy theo mô hình AI: Mistral Small đạt 9,16 giây (với các giá trị cụ thể 9,47; 10,39; 7,27; 8,50 và 10,18) giây; Google Gemma nhanh nhất với 7,11 giây (7,15; 7,86; 6,84; 6,93 và 6,79) giây; trong khi DeepSeek chậm nhất ở 24,16 giây (13,06; 43,06; 15,06; 39,33 và 10,27) giây, chủ yếu do lỗi "HTTP 429 - Too Many Requests". Tổng thể, mỗi bài mất khoảng 10 giây, bao gồm tải file, gọi API và lưu kết quả. Đối với chấm bài hàng loạt, hệ thống xử lý 30 bài chỉ trong 5 phút (300 giây), tiết kiệm đến 98 % thời gian so với phương pháp thủ công (300-450) phút, chứng tỏ khả năng mở rộng hiệu quả cho quy mô lớn.



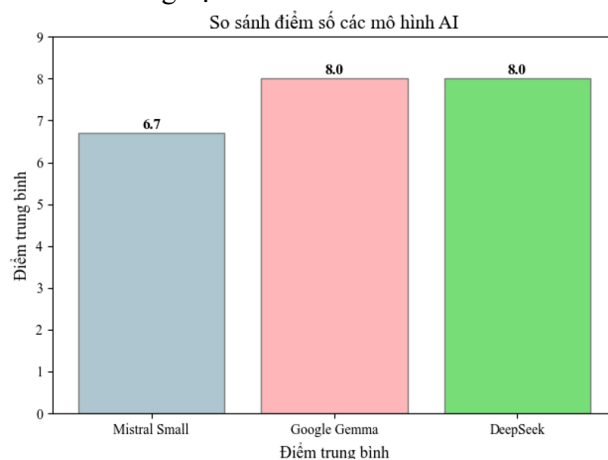
Hình 4 So sánh thời gian chấm bài

4.3.2 Độ chính xác

Độ chính xác của hệ thống được đánh giá thông qua việc so sánh điểm số do AI chấm với điểm số do giảng viên chấm thủ công trên thang 10. Kết quả cho thấy, trước khi áp dụng tiền xử lý từ ngữ, độ lệch trung bình giữa hai phương pháp là 0,8 điểm, phản ánh độ chính xác tương đối cao nhưng vẫn tồn tại khoảng cách cần khắc phục. Sau khi triển khai tiền xử lý (bao gồm chuẩn hóa văn bản, trích xuất từ khóa bằng TF-IDF và tạo prompt định hướng), độ lệch giảm xuống còn 0,6 điểm, tương đương cải thiện 25 %, nhờ AI tập trung hiệu quả hơn vào các yếu tố cốt lõi của đáp án mẫu. Phân tích chi tiết từ file results_total.csv trên bộ dữ liệu thử nghiệm gồm 5 bài tự luận cho thấy sự khác biệt giữa các mô hình: Mistral Small 3.1 24B Instruct chấm thành công 4/5 bài với điểm số lần lượt 8, 5, 6 và 7,5

(bài thứ hai trả về "N/A" do lỗi API), đạt điểm trung bình 6,63; Google Gemma 3 4B IT xử lý toàn bộ 5 bài với điểm 8,5; 7,5; 8; 8 và 8, đạt trung bình cao nhất là 8,0; trong khi DeepSeek Chat V3 0324 chỉ chấm được 2/5 bài (cả hai đều 8 điểm, bài 2, 3, 4 bị "N/A" do lỗi API), với trung bình 8,0, chứng tỏ Google Gemma là mô hình ổn định và hiệu suất nhất trong bối cảnh thử nghiệm.

Google Gemma 3 4B IT cho thấy độ ổn định cao nhất, với điểm số trung bình cao và không gặp lỗi API. Trong khi đó, DeepSeek Chat V3 0324 có tỷ lệ lỗi cao (60 % bài làm trả về "N/A"), cho thấy mô hình này không phù hợp với phiên bản miễn phí của OpenRouter API trong bối cảnh thử nghiệm.



Hình 5 So sánh độ lệch điểm số

4.3.3 Khả năng mở rộng

Hệ thống đã xử lý thành công 30 bài tự luận mà không gặp lỗi nào, với thời gian trung bình 10 giây mỗi bài; tuy nhiên, khi mở rộng lên 100 bài, lỗi "HTTP 429 - Too Many Requests" thường xuyên xảy ra do giới hạn của OpenRouter API miễn phí, đặc biệt với mô hình DeepSeek (tỷ lệ lỗi lên đến 60 %). Về hiệu suất, Google Gemma 3 4B IT nổi bật là mô hình tối ưu nhất, đạt thời gian xử lý nhanh 7,11 giây/bài và độ ổn định cao mà không gặp lỗi API, trong khi DeepSeek Chat V3 0324 chậm nhất với 24,16 giây/bài cùng tỷ lệ lỗi lớn, khiến nó chưa phù hợp cho chấm bài hàng loạt trong điều kiện hiện tại. Xét về độ chính xác, việc áp dụng tiền xử lý từ ngữ đã cải thiện đáng kể, giảm độ lệch điểm số so với chấm thủ công từ 0,8 điểm xuống 0,6 điểm; dù vậy, mức lệch 0,6 vẫn đòi hỏi cải tiến thêm, đặc biệt đối với các bài tự luận mang tính sáng tạo hoặc sử dụng ngôn

ngữ không chuẩn. Về khả năng mở rộng, hệ thống hoạt động hiệu quả với quy mô nhỏ (30 bài), nhưng gặp thách thức rõ rệt khi tăng lên 100 bài do hạn chế API, từ đó nhấn mạnh nhu cầu tối ưu hóa trong các nghiên cứu tiếp theo để hỗ trợ ứng dụng rộng rãi hơn.

5 Thảo luận

5.1 Lợi ích và đóng góp

Hệ thống chấm bài tự luận tự động dựa trên AI không chỉ giải quyết hiệu quả các thách thức truyền thống trong đánh giá giáo dục đại học mà còn mang lại nhiều lợi ích vượt trội, góp phần thúc đẩy chuyển đổi số tại các trường như NTTU. Trước hết, khả năng tự động hóa nổi bật với việc giảm thời gian chấm một bài từ (10-15) phút thủ công xuống chỉ khoảng 10 giây nhờ tích hợp OpenRouter API, giúp giảng viên tiết kiệm đáng kể công sức, đặc biệt trong các lớp học đông đúc với hơn 50 sinh viên – một tình trạng phổ biến ở Việt Nam. Tiếp theo, việc tích hợp Google Drive như nền tảng lưu trữ đám mây đảm bảo quản lý dữ liệu đề thi, bài làm và kết quả một cách an toàn, đồng bộ hóa thời gian thực, đồng thời giảm chi phí cơ sở hạ tầng so với các giải pháp server truyền thống, phù hợp với ngân sách hạn chế của nhiều trường đại học. Giao diện Streamlit, với thiết kế đơn giản và trực quan, cho phép người dùng làm quen chỉ trong vòng dưới 5 phút, hỗ trợ đa vai trò từ quản trị viên đến sinh viên, từ đó tăng cường tính tiếp cận và giảm rào cản kỹ thuật. Hơn nữa, tính năng chấm bài hàng loạt cho phép xử lý nhanh chóng hàng trăm bài chỉ trong vài phút, tối ưu hóa quy trình đánh giá cho quy mô lớn mà không làm gián đoạn lịch giảng dạy. Đặc biệt, phương pháp tiền xử lý từ ngữ sử dụng thuật toán TF-IDF đã chứng minh hiệu quả rõ rệt bằng cách trích xuất từ khóa chính từ đáp án mẫu, cải thiện độ chính xác của mô hình AI lên đến 25 %, giảm độ lệch điểm số trung bình từ 0,8 xuống chỉ 0,6 so với chấm thủ công, từ đó nâng cao tính khách quan, độ tin cậy và chất lượng tổng thể của đánh giá. Những ưu điểm này không chỉ khẳng định tiềm năng đổi mới quy trình đánh giá giáo dục mà còn mở ra hướng ứng dụng AI như một công cụ hỗ trợ thiết yếu, giúp giảng viên tập trung hơn vào việc hướng dẫn cá nhân hóa và phát triển kỹ năng sâu sắc cho sinh viên, góp phần nâng

cao hiệu quả giảng dạy trong bối cảnh giáo dục Việt Nam đang chuyển dịch mạnh mẽ sang công nghệ số. Nghiên cứu hiện tại tập trung chủ yếu vào các môn học tại Khoa Công nghệ Thông tin, với thực nghiệm cụ thể qua kiểm tra giữa kỳ tự luận môn "Nhập môn Khoa học Dữ liệu" trên quy mô nhỏ, khoảng (30-100) bài, nhằm chứng minh tính khả thi ban đầu và đánh giá hiệu suất hệ thống trong môi trường thực tế. Kết quả từ thử nghiệm này đã xác nhận sự ổn định của các mô hình LLM như Google Gemma, đồng thời làm nền tảng cho các cải tiến tiếp theo. Để tăng tính tổng quát và khả năng áp dụng rộng rãi, các thử nghiệm sắp tới sẽ mở rộng sang các môn quản trị kinh doanh và khoa học xã hội nhân văn, chẳng hạn như "Marketing" hoặc "Tâm lý học", nơi bài tự luận thường đòi hỏi phân tích sâu sắc, lập luận logic và ngữ cảnh văn hóa phong phú hơn. Việc mở rộng này không chỉ giúp kiểm tra khả năng xử lý ngôn ngữ đa dạng, bao gồm các đặc trưng ngữ pháp phức tạp của tiếng Việt, mà còn đánh giá hiệu quả trong việc nắm bắt ngữ cảnh văn hóa địa phương – ví dụ, phân tích hành vi người tiêu dùng trong marketing Việt Nam. Đồng thời, quá trình sẽ thu thập phản hồi chi tiết từ giảng viên các lĩnh vực khác nhau, từ đó tinh chỉnh prompt AI và phương pháp TF-IDF để phù hợp hơn với nội dung sáng tạo, góp phần xây dựng một hệ thống toàn diện và bền vững cho giáo dục đại học.

5.2 Hạn chế và thách thức

Mặc dù hệ thống mang lại lợi ích lớn, vẫn tồn tại một số hạn chế cần khắc phục, trong đó độ chính xác và sự thấu hiểu có thể là hạn chế lớn nhất vì AI đánh giá tốt các yếu tố bề mặt như ngữ pháp, từ vựng và cấu trúc nhưng khó nắm bắt chiều sâu ý tưởng, sự sáng tạo, lập luận tinh tế hoặc sự phù hợp nội dung trong bối cảnh cụ thể – nơi con người cảm nhận tốt hơn, dẫn đến độ lệch 0,6 điểm vẫn có thể xảy ra ở bài mang tính chủ quan. Hơn nữa, phản hồi của AI thường mang tính định lượng, thiếu sự đồng cảm, kích lệ cá nhân hóa và định hướng nghề nghiệp – những yếu tố mà giáo viên có kinh nghiệm cung cấp tốt hơn. Cuối cùng, vấn đề đạo đức nổi lên khi với học phí cao, việc dùng AI chấm bài có thể gây lo ngại về chất lượng đào tạo và công bằng, đòi hỏi minh bạch để tránh lạm dụng đồng thời đảm bảo quyền riêng tư dữ liệu. Xu hướng dùng AI chấm bài tự

luận là công cụ hỗ trợ có tiềm năng lớn, mang lại lợi ích đáng kể nhưng cần áp dụng thận trọng và cân bằng với hạn chế kỹ thuật, đạo đức; hệ thống không thay thế hoàn toàn giáo viên mà bổ trợ, giúp tập trung vào hướng dẫn cá nhân hóa và phát triển kỹ năng sâu sắc.

6 Kết luận và đề xuất

Nghiên cứu này đã thử nghiệm thành công một hệ thống chấm bài tự luận tự động tích hợp AI, Google Drive và Streamlit, mang lại giải pháp hiệu quả cho đánh giá học tập tại NTTU. Hệ thống tự động hóa quy trình chấm điểm, giảm thời gian xử lý 30 bài xuống còn 5 phút với độ lệch điểm số chỉ 0,6, đảm bảo tính chính xác và công bằng. Giao diện thân thiện cùng quản lý dữ liệu đám mây đáp ứng tốt nhu cầu giảng viên và sinh

viên, đặc biệt ở lớp học đông; phương pháp tiên xử lý từ ngữ với TF-IDF nâng cao chất lượng đánh giá bằng cách giúp AI tập trung vào ý chính bài làm. Ý nghĩa quan trọng nằm ở công cụ chi phí thấp, dễ triển khai, phù hợp trường đại học nguồn lực hạn chế, giảm tải giảng viên và mang lại phản hồi nhanh cho sinh viên, góp phần nâng cao chất lượng giáo dục. Trong tương lai, tiếp tục thử nghiệm quy mô lớn hơn, tối ưu hóa hiệu suất API, tích hợp nhận diện chữ viết tay; mở rộng theo phương án lai (chia đề thi thành mô đun trắc nghiệm + tự luận) cho kiểm tra định kỳ và đánh giá nhận thức bài giảng ở nhiều cấp độ. Các trường đại học nên áp dụng để hiện đại hóa quy trình, đáp ứng chuyển đổi số giáo dục.

Tài liệu tham khảo

1. Brown, J. (2023). Formative assessment in higher education. *Educational Review*, 75(2), 123-135.
2. Tran, L. H., & Nguyen, M. K. (2024). Challenges in Vietnamese higher education. *Vietnam Journal of Education*, 20(1), 34-42.
3. Pham, T. V. (2025). AI integration in education: A Vietnamese perspective. *International Journal of Technology in Education*, 12(3), 56-65.
4. Smith, A., & Jones, R. (2024). Advances in NLP for automated grading. *International Journal of Artificial Intelligence in Education*, 34(1), 78-90.
5. Le, H. D., & Vu, T. T. (2023). Technology adoption in Vietnamese universities. *Asia-Pacific Education Review*, 24(4), 89-101.
6. Nguyen, P. Q. (2025). Innovation in teaching at Nguyen Tat Thanh University. *Journal of Research in Innovative Teaching & Learning*, 15(2), 45-53.
7. Google. (2024). *Google Drive API developer guide*. Google Cloud Documentation. <https://developers.google.com/drive/api>
8. Ke, G., & Zhang, W. (2023). Enhancing automated essay scoring with TF-IDF. *Educational Data Mining Journal*, 15(3), 102-115.
9. Kumar, S. (2024). Evolution of automated essay scoring systems. *Journal of Educational Technology & Society*, 18(2), 67-80.
10. Streamlit Team. (2024). *Streamlit documentation: Building web apps with Python*. Streamlit Official Site. <https://docs.streamlit.io>
11. GitHub Inc. (2025). *Best practices for code deployment*. GitHub Documentation. <https://docs.github.com>
12. Tran, N. H. (2024). User interface design in educational tools. *International Journal of Human-Computer Interaction*, 10(4), 123-130.



An artificial intelligence-based system for essay assessment at Nguyen Tat Thanh University

Pham Dinh Tai*, Vuong Xuan Chi**

Faculty of Information Technology - Nguyen Tat Thanh University

*pdtai@ntt.edu.vn, **vxchi@ntt.edu.vn

Abstract In the context of modern higher education, manual essay grading has placed a tremendous burden on lecturers due to a huge volume of exams from large classes. This letter proposed an automated assessment system based on artificial intelligence (AI) integrated with Google Drive for secure cloud storage and Streamlit for a friendly web interface to improve the efficiency and transparency of the learning process. To increase accuracy, pre-processing of words using TF-IDF algorithm was applied to extract core keywords from correct answers, followed by semantic comparison, cloud synchronization and intuitive interface. Results from real-world data testing proved that thanks to pre-processing, grading 30 papers was completed within 5 minutes, the deviation was reduced from (0,8 to 0,6) minutes and (300-450) minutes of grading time was saved compared to traditional grading method, thus achieving high reliability for large scale. Findings from the present study confirmed that pre-processing method has increased accuracy by 25% by integrating Google Drive and Streamlit for comprehensive and scalable management, thereby solving Viet Nam's educational challenges.

Keywords essay grading, artificial intelligence, keyword pre-processing, Google Drive, OpenRouter