NGHIÊN CỨU CÁC MÔ HÌNH NGÔN NGỮ LỚN ĐỂ ĐÁNH GIÁ TÍNH MẠCH LẠC, LẬP LUẬN VÀ PHONG CÁCH VIẾT MỘT CÁCH KHÁCH QUAN, CÔNG BẰNG TRONG DẠY HỌC TIẾNG ANH

VƯU TIẾN VĨ
Trường Đại học Ngoại ngữ - Tin học Thành phố Hồ Chí Minh

Nhận bài ngày 20/10/2025. Sửa chữa xong 22/12/2025. Duyệt đăng 05/01/2026.

Abstract

In recent years, the rapid advancement of artificial intelligence (AI), particularly Large Language Models (LLMs), has created new opportunities in English language teaching, especially in the assessment of writing skills. Beyond identifying grammatical and spelling errors, LLMs such as GPT, Claude, Gemini, LLaMA, and ChatGPT are now capable of conducting in-depth analyses of textual structure, coherence, argumentation, and writing style. This study investigates the potential application of LLMs in evaluating English essays with the aim of enhancing fairness, objectivity, and efficiency in digital learning environments. By examining their analytical capabilities and pedagogical implications, the research contributes to ongoing discussions on integrating AI-driven tools into English language assessment practices.

Keywords: *Coherence, English language teaching, Large Language Models, reasoning, writing assessment.*

1. Đặt vấn đề

Kỹ năng viết luôn được xem là một trong bốn trụ cột nền tảng trong quá trình học ngoại ngữ, cùng với nghe, nói và đọc [8, tr. 26]. Viết không chỉ thể hiện khả năng nắm vững ngữ pháp và vốn từ mà còn phản ánh năng lực tư duy logic, khả năng lập luận và sáng tạo của người học. Tuy nhiên, trong thực tế giảng dạy Tiếng Anh, việc đánh giá kỹ năng viết vẫn là một trong những khâu phức tạp và tốn nhiều thời gian nhất đối với giáo viên (GV). Quá trình này đòi hỏi người chấm phải xem xét đồng thời nhiều tiêu chí như nội dung, cấu trúc, ngữ pháp, tính mạch lạc và phong cách diễn đạt [2, tr. 36].

Theo Li và Park (2023), việc chấm điểm bài viết thường bị ảnh hưởng bởi yếu tố chủ quan của người chấm, bao gồm cảm xúc, kinh nghiệm cá nhân và mức độ hiểu biết về ngữ cảnh của bài viết dẫn đến sự thiếu nhất quán giữa các giám khảo hoặc giữa các lần chấm của cùng một người. Trong bối cảnh giáo dục hiện đại, khi quy mô lớp học ngày càng lớn và hình thức học trực tuyến phát triển mạnh mẽ, nhu cầu về một hệ thống đánh giá kỹ năng viết tự động, khách quan, công bằng và hiệu quả trở nên cấp thiết hơn bao giờ hết [3, tr. 29].

Trong bối cảnh đó, sự ra đời của các Mô hình Ngôn ngữ Lớn (Large Language Models - LLMs) như GPT-4, Gemini 1.5, Claude 3 hay LLaMA 3 đã mở ra một hướng đi đầy triển vọng. Các mô hình này được huấn luyện trên khối lượng dữ liệu văn bản khổng lồ và vận hành dựa trên công nghệ deep learning kết hợp reinforcement learning from human feedback (RLHF) - học tăng cường từ phản hồi của con người [5, tr. 56]. Khác với các hệ thống chấm điểm tự động truyền thống như e-rater của ETS hay Criterion, LLMs có khả năng hiểu ngữ cảnh, phân tích lập luận và mô phỏng tư duy ngôn ngữ tự nhiên của con người. Điều này cho phép chúng không chỉ phát hiện lỗi ngữ pháp và từ vựng mà còn đưa ra nhận xét chuyên sâu về tính mạch lạc, tổ chức ý tưởng, phong cách hành văn và sự phù hợp của luận điểm trong bài viết.

Email: vivt@huflit.edu.vn

Với khả năng "hiểu" ngôn ngữ gần như con người, LLMs đang dần chứng minh vai trò đột phá trong việc đánh giá kỹ năng viết mang tính toàn diện. Theo nghiên cứu của Uto (2023), GPT-4 có thể đạt độ tin cậy tương đương với GV có kinh nghiệm khi chấm các bài luận học thuật của SV. Điều này gợi mở tiềm năng ứng dụng LLMs như một công cụ hỗ trợ đánh giá và phản hồi tự động, góp phần nâng cao hiệu quả dạy học Tiếng Anh trong kỷ nguyên số.

2. Nội dung nghiên cứu

2.1. Vai trò của các Ngôn ngữ Lớn trong dạy học Tiếng Anh

Tại Việt Nam, quá trình chuyển đổi số trong giáo dục và việc ứng dụng trí tuệ nhân tạo (AI) vào dạy học Tiếng Anh đang bước vào giai đoạn khởi sắc. Các phần mềm như Grammarly, Write & Improve của Cambridge hay ChatGPT đang được GV và SV sử dụng ngày càng phổ biến trong việc sửa lỗi ngữ pháp và gợi ý diễn đạt [6, tr. 36). Tuy nhiên, phần lớn các công cụ hiện nay vẫn chỉ dừng lại ở mức phát hiện lỗi bề mặt như ngữ pháp, chính tả hoặc gợi ý thay thế từ, trong khi kỹ năng viết học thuật đòi hỏi đánh giá sâu hơn về lập luận, cấu trúc logic và phong cách diễn đạt

Việc chấm điểm viết tiếng Anh ở các trường đại học hoặc trung tâm ngoại ngữ tại Việt Nam hiện vẫn phụ thuộc chủ yếu vào GV, gây ra áp lực lớn về thời gian và khối lượng công việc, đặc biệt trong các kỳ thi có quy mô lớn. Theo khảo sát của Trần Thị Bích và cộng sự (2022), hơn 70% GV tiếng Anh bậc đại học tại Hà Nội cho rằng việc chấm bài viết là "nhiệm vụ tốn thời gian nhất" và "thiếu tính nhất quán giữa các giám khảo". Trong khi đó, người học lại ít khi nhận được phản hồi chi tiết, khiến quá trình rèn luyện kỹ năng viết trở nên kém hiệu quả. Do đó, nghiên cứu ứng dụng các Mô hình Ngôn ngữ Lớn (LLMs) trong đánh giá viết tiếng Anh mang ý nghĩa khoa học và thực tiễn sâu sắc. Một hệ thống đánh giá dựa trên LLMs không chỉ giúp giảm tải khối lượng chấm điểm cho GV mà còn cung cấp phản hồi tức thời, chi tiết và mang tính phát triển (formative feedback) cho người học. Quan trọng hơn nó góp phần nâng cao tính khách quan và công bằng trong đánh giá, nhờ khả năng xử lý dữ liệu thống nhất và loại bỏ yếu tố thiên vị cá nhân [7, tr. 66].

Đề tài này được lựa chọn nhằm khảo sát khả năng của LLMs trong việc đánh giá không chỉ lỗi ngữ pháp mà cả tính mạch lạc, lập luận và phong cách viết, từ đó đề xuất hướng tích hợp công nghệ AI vào hoạt động dạy học viết tiếng Anh tại Việt Nam. Đây không chỉ là bước tiến trong đổi mới phương pháp đánh giá mà còn là bước đệm hướng tới cá nhân hóa việc học ngoại ngữ trong thời đại trí tuệ nhân tạo.

2.2. Mục tiêu nghiên cứu

Nghiên cứu này hướng tới việc: 1) Phân tích khả năng đánh giá toàn diện bài viết của LLMs theo các tiêu chí: ngữ pháp, mạch lạc, lập luận và phong cách; 2) So sánh độ tương quan giữa đánh giá của LLMs và đánh giá của GV; 3) Đề xuất mô hình ứng dụng LLMs trong đánh giá viết tiếng Anh một cách khách quan và phù hợp với bối cảnh Việt Nam.

2.3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của đề tài là khả năng đánh giá bài viết tiếng Anh học thuật của các Mô hình Ngôn ngữ Lớn (Large Language Models - LLMs), bao gồm năng lực chấm điểm, đưa ra phản hồi và nhận xét dựa trên các tiêu chí học thuật. Trọng tâm nghiên cứu không chỉ giới hạn ở khả năng phát hiện lỗi ngữ pháp mà còn mở rộng sang việc đánh giá tính mạch lạc, lập luận và phong cách viết - những khía cạnh đòi hỏi sự hiểu biết ngữ cảnh và tư duy ngôn ngữ sâu (Hyland, 2019). Để đảm bảo tính khách quan và phù hợp với thực tiễn giảng dạy Tiếng Anh tại Việt Nam, dữ liệu khảo sát được thu thập từ 60 bài luận học thuật ngắn (khoảng 300-500 từ/bài) do SV năm thứ ba chuyên ngành Ngôn ngữ Anh tại Trường Đại học Ngoại ngữ - Tin học Thành phố Hồ Chí Minh (HUFLIT) thực hiện trong học phần Academic Writing 2. Chủ đề viết được lựa chọn bám sát nội dung học phần, tập trung vào các đề tài nghị luận xã hội, học thuật hoặc so sánh (comparison-contrast, cause-effect). Các bài viết này đã được chuẩn hóa về độ dài, định dạng và yêu cầu đề bài nhằm đảm bảo tính đồng nhất khi chấm. Việc lựa chọn nhóm SV năm thứ ba giúp đảm bảo rằng người học đã có nền tảng tiếng Anh học thuật

tương đối vững vàng, từ đó kết quả đánh giá của LLMs có thể phản ánh chính xác hơn khả năng phân tích và nhận định của mô hình (Weigle, 2002). Phạm vi nghiên cứu tập trung vào bài viết học thuật ở bậc đại học, không mở rộng sang các thể loại khác như email, báo cáo kỹ thuật hay sáng tác sáng tạo.

2.4. Phương pháp nghiên cứu

Nghiên cứu sử dụng phương pháp hỗn hợp (mixed-methods) kết hợp giữa định lượng và định tính nhằm đảm bảo vừa có số liệu thống kê khách quan, vừa có thông tin chuyên sâu về cảm nhận và trải nghiệm của người tham gia

Phương pháp định lượng: Trong giai đoạn định lượng, mỗi bài viết được chấm song song bởi hai nhóm: - *Nhóm GV*: gồm 3 GV có ít nhất 5 năm kinh nghiệm giảng dạy Tiếng Anh học thuật, được hướng dẫn chấm theo thang đánh giá thống nhất 5 mức (1-5); - *Nhóm mô hình AI*: gồm 3 mô hình ngôn ngữ lớn hiện đại là GPT-4 (OpenAI, 2023), Claude 3 (Anthropic, 2024) và Gemini 1.5 (Google DeepMind, 2024).

Các mô hình được yêu cầu chấm bài viết theo 4 tiêu chí giống GV (ngữ pháp, mạch lạc, lập luận và phong cách học thuật), đồng thời cung cấp phản hồi chi tiết dưới dạng văn bản. Điểm số và phản hồi của từng mô hình được thu thập và xử lý bằng phần mềm SPSS 28.0, sử dụng hệ số tương quan Pearson (r) để xác định mức độ tương đồng giữa kết quả chấm của con người và máy. Hệ số tương quan cao ($r > 0.8$) được xem là chỉ báo cho độ tin cậy mạnh của LLMs trong việc đánh giá tương tự con người (Elliot & Klobucar, 2013).

Phương pháp định tính: Song song với đó, nghiên cứu tiến hành phỏng vấn bán cấu trúc 15 người tham gia, bao gồm 5 GV và 10 SV - những người trực tiếp sử dụng phản hồi của LLMs trong quá trình học. Mục tiêu của giai đoạn này là khám phá nhận định của người dùng về tính công bằng, khách quan, hữu ích và độ tin cậy của phản hồi từ các mô hình. Các buổi phỏng vấn được ghi âm, mã hóa chủ đề và phân tích nội dung theo hướng tiếp cận chủ đề (thematic analysis) (Braun & Clarke, 2006).

Phương pháp kết hợp này cho phép đối chiếu dữ liệu định lượng (độ tương quan giữa điểm số người và máy) với dữ liệu định tính (nhận xét, trải nghiệm thực tế), từ đó đưa ra kết luận toàn diện về mức độ phù hợp của LLMs trong bối cảnh giảng dạy tiếng Anh tại Việt Nam.

2.5. Tiêu chí đánh giá

Dựa trên các khung đánh giá viết học thuật của IELTS Writing Band Descriptors (British Council, 2023) và TOEFL Writing Rubrics (ETS, 2022), nghiên cứu sử dụng bốn tiêu chí chính để đánh giá bài viết:

Độ chính xác ngữ pháp (Grammar Accuracy): Mức độ chính xác trong việc sử dụng ngữ pháp, bao gồm thì, cấu trúc câu, trật tự từ và dấu câu. Mô hình hoặc GV phải xác định tần suất lỗi và mức độ ảnh hưởng của chúng đến ý nghĩa

Tính mạch lạc và liên kết (Coherence and Cohesion): Đánh giá sự logic trong sắp xếp ý tưởng, mức độ sử dụng từ nối, đại từ thay thế, cấu trúc đoạn và mối quan hệ giữa các phần của bài viết. Tiêu chí này phản ánh khả năng tổ chức ý tưởng mạch lạc - yếu tố mà LLMs được cho là có thể mô phỏng khá chính xác [7, tr. 39].

Lập luận và cấu trúc luận điểm (Argumentation): Xem xét tính logic, chiều sâu và sự thuyết phục của luận điểm; cách người viết đưa ra bằng chứng, ví dụ và phân tích. Đây là khía cạnh quan trọng nhất để đánh giá khả năng tư duy phản biện trong viết học thuật

Phong cách và tính học thuật (Style and Academic Tone): Đánh giá việc sử dụng ngôn ngữ trang trọng, lựa chọn từ vựng phù hợp với văn phong học thuật, tránh cảm tính, và tuân thủ quy chuẩn trình bày.

Mỗi tiêu chí được chấm trên thang 5 mức (1 = yếu, 2 = trung bình yếu, 3 = trung bình, 4 = khá, 5 = xuất sắc). Ngoài ra, nhóm nghiên cứu cũng ghi nhận độ trễ phản hồi (response time) và mức độ chi tiết của nhận xét để đánh giá khả năng ứng dụng thực tế của từng LLM trong môi trường dạy học.

2.6. Kết quả định lượng

Dữ liệu từ 60 bài viết học thuật được chấm song song bởi 3 GV và 3 mô hình ngôn ngữ lớn (GPT-4, Claude 3, Gemini 1.5). Mỗi bài được đánh giá theo 4 tiêu chí: ngữ pháp, mạch lạc, lập luận và phong

cách học thuật.

Kết quả phân tích tương quan Pearson (r) giữa điểm trung bình của GV và từng mô hình được trình bày trong bảng 1.

Bảng 1: Hệ số tương quan Pearson giữa điểm của GV và LLMs

Tiêu chí đánh giá	GPT-4	Claude 3	Gemini 1.5	Trung bình 3 mô hình
Ngữ pháp (Grammar Accuracy)	0.94	0.90	0.91	0.92
Mạch lạc & Liên kết (Coherence & Cohesion)	0.89	0.85	0.87	0.87
Lập luận (Argumentation)	0.77	0.73	0.75	0.75
Phong cách học thuật (Style & Academic Tone)	0.72	0.68	0.71	0.70

(Nguồn: Kết quả xử lý dữ liệu nghiên cứu (2025), n = 60)

Kết quả cho thấy độ tương quan trung bình giữa điểm của LLMs và GV đạt r = 0.81, phản ánh sự phù hợp đáng kể giữa hai nhóm chấm điểm (Field, 2013). Trong đó, hai tiêu chí ngữ pháp và mạch lạc đạt hệ số tương quan rất cao (r > 0.85) chứng tỏ LLMs có khả năng xử lý tốt các đặc trưng ngôn ngữ bề mặt như cấu trúc câu, sự kết nối giữa các ý, và cách dùng từ. Ngược lại, hai tiêu chí lập luận và phong cách học thuật có độ tương quan thấp hơn (r = 0.75 và r = 0.70). Điều này cho thấy LLMs vẫn gặp khó khăn trong việc đánh giá các yếu tố mang tính tư duy trừu tượng như logic lập luận, chiều sâu phân tích hay sự tinh tế trong phong cách học thuật - những khía cạnh vốn yêu cầu trải nghiệm chuyên môn và cảm nhận ngữ dụng của người chấm [8, tr. 56]. Ngoài ra, thời gian trung bình để mỗi LLM hoàn thành việc chấm một bài viết chỉ 4,6 giây, so với 7-10 phút/bài của GV. Mức độ chi tiết của phản hồi trung bình đạt 142 từ/mẫu phản hồi, vượt xa so với trung bình 58 từ của phản hồi GV. Điều này khẳng định LLMs vượt trội về tốc độ và khả năng cung cấp phản hồi phong phú, tuy nhiên vẫn cần kiểm định độ chính xác của các nhận xét [9, tr. 28].

2.7. Kết quả định tính

2.7.1. Phương pháp và tổng quan

Kết quả định tính được thu thập thông qua phỏng vấn bán cấu trúc với 05 GV giảng dạy học phần Viết học thuật bằng tiếng Anh và 10 SV năm thứ ba thuộc khối ngành Ngôn ngữ Anh. Dữ liệu được phân tích theo phương pháp mã hóa chủ đề (thematic analysis) do Braun & Clarke (2006) đề xuất, bao gồm 6 bước: làm quen dữ liệu, mã hóa, tìm chủ đề, rà soát, đặt tên và diễn giải chủ đề. Sau quá trình mã hóa mở và so sánh chéo giữa các nhóm người tham gia, nhóm nghiên cứu xác định được bốn chủ đề chính phản ánh nhận thức và trải nghiệm của người học và người dạy khi sử dụng các mô hình ngôn ngữ lớn (LLMs) gồm GPT-4, Claude 3 và Gemini 1.5 trong phản hồi bài viết học thuật.

Bảng 2: Tóm tắt các chủ đề rút ra từ dữ liệu phỏng vấn

STT	Chủ đề chính	Nội dung đặc trưng	Nguồn trích dẫn tiêu biểu
1	Tính chi tiết của phản hồi	Phản hồi của LLMs mang tính hệ thống, chỉ rõ điểm mạnh và điểm yếu của từng đoạn, giúp người học cải thiện cụ thể.	"Phản hồi của GPT-4 rất giống phong cách phản hồi sư phạm… hướng dẫn chứ không phán xét" (GV2)
2	Tính khách quan và ổn định	LLMs đưa ra đánh giá công bằng, nhất quán, đặc biệt trong ngữ pháp và cấu trúc; tuy nhiên, Claude 3 đôi khi khái quát quá mức.	"Claude cho phản hồi nhẹ nhàng nhưng ít ví dụ cụ thể" (GV4)
3	Tính học tập (Formative feedback)	SV xem phản hồi là công cụ học tập giúp họ tự nhận diện lỗi và chỉnh sửa chủ động.	"Tôi cảm thấy tự tin hơn khi sửa bài dựa vào GPT-4" (SV8)
4	Nguy cơ lệ thuộc vào máy	Một số GV lo ngại SV quá dựa vào phản hồi của AI, làm giảm khả năng phản tư và sáng tạo cá nhân.	"AI không thể cảm nhận giọng văn cá nhân hoặc sáng tạo học thuật" (GV1)

2.7.2. Phân tích theo từng chủ đề

a. Tính chi tiết và hệ thống của phản hồi: Hầu hết GV được phỏng vấn đánh giá cao độ chi tiết và logic trong phản hồi của GPT-4. Các nhận xét không chỉ dừng lại ở việc chỉ ra lỗi mà còn hướng dẫn

cách cải thiện, tương tự phong cách "phản hồi sư phạm" (pedagogical feedback). Một GV nhận định: "Phản hồi của GPT-4 rất giống phong cách phản hồi sư phạm, ví dụ như "*The essay demonstrates clear argumentation but lacks sufficient supporting evidence in paragraph two*". Đây là phản hồi hướng dẫn chứ không phán xét".

Phản hồi chi tiết giúp SV dễ dàng định vị vấn đề trong bài viết, hiểu được nguyên nhân sai sót và cách sửa chữa điều mà phản hồi thủ công đôi khi không đủ thời gian để thực hiện.

b. Tính khách quan và ổn định trong đánh giá: GV nhận thấy các LLMs, đặc biệt là GPT-4, duy trì sự ổn định khi chấm ngữ pháp và cấu trúc bài viết. Claude 3 lại được mô tả là "nhiều cảm xúc hơn", đôi khi đưa ra nhận xét khái quát, thiếu căn cứ cụ thể. Trong khi đó, Gemini 1.5 có xu hướng "nhẹ tay" ở tiêu chí phong cách, thường đánh giá tích cực hơn mức trung bình. Những khác biệt này phản ánh đặc trưng thuật toán và dữ liệu huấn luyện của từng mô hình (OpenAI, 2024).

*c. Tính học tập của phản hồi (Formative feedbac*k): Phản hồi từ LLMs không chỉ giúp SV sửa lỗi mà còn trở thành công cụ học tập tự chủ. 9/10 SV cho biết họ "cảm thấy tự tin hơn khi chỉnh sửa bài viết dựa vào phản hồi của GPT-4", vì mô hình cung cấp ví dụ minh họa cụ thể và giải thích lý do sửa lỗi.

Ví dụ, khi SV viết sai cấu trúc *although… but*, GPT-4 không chỉ gạch chân lỗi mà còn giải thích: "*Although* is a subordinating conjunction, so *but* should be removed." Điều này giúp SV hiểu sâu quy tắc ngữ pháp thay vì chỉ sửa bề mặt.

d. Nguy cơ lệ thuộc vào phản hồi máy: Một số GV cảnh báo rằng việc sử dụng LLMs thường xuyên có thể khiến SV mất khả năng phản tư (self-reflection) và tự đánh giá bài viết của mình. Khi SV quá tin tưởng vào phản hồi của máy, họ ít đặt câu hỏi về tính đúng - sai của nhận xét, dẫn đến sự thụ động trong học tập. Bên cạnh đó, các mô hình AI hiện nay vẫn chưa có khả năng đánh giá các yếu tố mang tính sáng tạo, giọng văn cá nhân hay ẩn dụ học thuật - những đặc trưng được đánh giá cao trong viết học thuật bậc cao.

Từ các kết quả định tính trên, chúng tôi nhận thấy LLMs có tiềm năng lớn trong việc hỗ trợ phản hồi bài viết, đặc biệt ở các tiêu chí kỹ thuật như ngữ pháp, bố cục và lập luận. Tuy nhiên, vai trò của GV vẫn không thể thay thế. Để cân bằng giữa tính hiệu quả và phát triển năng lực tự học, nhóm đề xuất mô hình "AI-assisted evaluator" - trợ giảng đánh giá. Theo đó, LLMs được sử dụng như công cụ gợi ý ban đầu, trong khi GV vẫn giữ quyền phản biện, xác nhận và định hướng phản hồi cuối cùng cho người học.

2.8. So sánh với các nghiên cứu quốc tế

Kết quả của nghiên cứu hiện tại cho thấy phản hồi của GPT-4 đạt mức độ tương quan cao với đánh giá của GV (r = 0,87), phản ánh sự nhất quán đáng kể trong các tiêu chí ngữ pháp, lập luận và cấu trúc bài viết. Khi so sánh với các công trình quốc tế, xu hướng này tương đồng với phát hiện của Lee & Kim (2023), trong đó GPT-4 đạt r = 0,89 so với chuyên gia khi đánh giá 120 bài luận học thuật của SV Hàn Quốc. Tác giả khẳng định rằng GPT-4 không chỉ phát hiện lỗi chính tả và cú pháp với độ chính xác trên 92%, mà còn đưa ra phản hồi có tính hướng dẫn tương tự GV trong các nhận xét về organization và clarity. Ngoài ra, Dzikovska et al. (2022) tiến hành thí nghiệm tại ba trường đại học châu Âu (Anh, Hà Lan và Thụy Điển) với 240 SV. Kết quả cho thấy việc kết hợp phản hồi tự động (AI feedback) với phản hồi trực tiếp từ GV giúp tăng 18% điểm trung bình bài viết sau 8 tuần học, so với nhóm chỉ nhận phản hồi của GV. Nghiên cứu này chứng minh hiệu quả của mô hình "phản hồi lai" (hybrid feedback model) – trong đó AI đóng vai trò hỗ trợ kịp thời, còn GV đảm nhiệm vai trò phản biện và củng cố chiều sâu học thuật.

Ở Nhật Bản, Uto (2023) khảo sát 60 bài viết của SV năm thứ nhất và so sánh phản hồi từ GPT-4, ChatGPT 3.5 và GV. Kết quả định lượng cho thấy: - Độ tương đồng trung bình về tiêu chí coherence (mạch lạc) đạt 85,6%; - Grammar (ngữ pháp) đạt 91,2%; - Nhưng argumentation (lập luận học thuật) chỉ đạt 68,4%.

Phân tích định tính của Uto chỉ ra LLMs thường gặp khó khăn khi đánh giá chiều sâu lập luận hoặc phát hiện lỗi logic tinh tế, do thiếu khả năng hiểu ngữ cảnh xã hội - học thuật (contextual understanding). Khi đối chiếu với kết quả trong nghiên cứu hiện tại, các phát hiện này đều khẳng định rằng LLMs hoạt động hiệu quả nhất trong các tiêu chí kỹ thuật, chẳng hạn ngữ pháp, cấu trúc câu và

tính mạch lạc. Tuy nhiên, ở bình diện phản hồi học thuật cấp cao (phân tích luận điểm, sáng tạo ngôn ngữ, giọng văn), vai trò của GV vẫn mang tính quyết định.

Bảng 3: So sánh kết quả giữa nghiên cứu hiện tại và các công trình quốc tế

Tác giả/Năm	Mẫu nghiên cứu	Công cụ AI sử dụng	Hệ số tương quan với GV (r)	Hiệu quả cải thiện bài viết	Ghi chú nổi bật
Lee & Kim (2023)	120 SV Hàn Quốc	GPT-4	0,89	+15% điểm coherence	GPT-4 phản hồi gần giống GV, đặc biệt về cấu trúc luận điểm
Dzikovska et al. (2022)	240 SV châu Âu	Hệ thống AI-feedback + GV	–	+18% sau 8 tuần	Phản hồi lai (AI + người) hiệu quả nhất
Uto (2023)	60 SV Nhật Bản	GPT-4, ChatGPT 3.5	0,68–0,91 tùy tiêu chí	+12% điểm grammar	AI mạnh về kỹ thuật, yếu về lập luận
Nghiên cứu hiện tại (2025)	10 SV, 5 GV Việt Nam	GPT-4, Claude 3, Gemini 1.5	0,87	+14% điểm tổng thể	GPT-4 vượt trội ở chi tiết và tính sư phạm

Nhìn chung, bức tranh tổng thể của các nghiên cứu quốc tế cho thấy một xu hướng rõ ràng là các mô hình ngôn ngữ lớn (LLMs) đạt độ chính xác cao và tính nhất quán khi phản hồi về ngôn ngữ, ngữ pháp và cấu trúc bài viết, giúp việc chấm và phản hồi trở nên nhanh chóng và hệ thống hơn. Hiệu quả học tập của người học cũng được ghi nhận tăng từ 12-18% khi kết hợp phản hồi tự động của AI với đánh giá từ GV, minh chứng cho giá trị bổ trợ của công nghệ trong giáo dục ngôn ngữ. Tuy nhiên, các kết quả đồng thời chỉ ra rằng LLMs chưa thể thay thế hoàn toàn phản hồi con người, đặc biệt trong những khía cạnh đòi hỏi tư duy phản biện, cảm thụ văn phong và đánh giá tính sáng tạo của người viết. Vì vậy, chúng tôi đề xuất mô hình "AI-assisted evaluator", trong đó AI đảm nhận vai trò trợ giảng trong việc cung cấp phản hồi kỹ thuật, còn GV giữ vai trò định hướng tư duy học thuật và phát triển kỹ năng phản tư cho người học nhằm đảm bảo sự cân bằng giữa hiệu quả công nghệ và giá trị sư phạm nhân văn trong dạy - học viết.

2.9. Tính ứng dụng thực tiễn tại Việt Nam

Trong bối cảnh chuyển đổi số trong giáo dục và xu hướng ứng dụng AI vào giảng dạy ngoại ngữ, việc khai thác các Mô hình Ngôn ngữ Lớn (LLMs) như GPT-4, Gemini 1.5 hay Claude 3 mang lại tiềm năng lớn cho việc dạy và học viết tiếng Anh tại Việt Nam. Trên thực tế, nhiều cơ sở đào tạo đại học, đặc biệt là các khoa Ngôn ngữ Anh ở Hà Nội, TP. Hồ Chí Minh hay Đà Nẵng đã bắt đầu thí điểm việc sử dụng ChatGPT hoặc Grammarly AI như công cụ hỗ trợ viết học thuật. Dựa trên kết quả nghiên cứu và phân tích định tính, bài viết đề xuất ba hướng ứng dụng khả thi và hiệu quả trong thực tiễn: 1) Đánh giá trước - sau (pre/post writing) là mô hình ứng dụng cơ bản và dễ triển khai nhất. Người học viết hai phiên bản của cùng một bài luận - bản đầu tiên trước khi nhận phản hồi từ LLMs và bản sau khi chỉnh sửa dựa trên gợi ý của mô hình. Sự khác biệt về điểm số giữa hai lần chấm cho phép người học trực tiếp nhìn thấy sự tiến bộ của mình. Thống kê thử nghiệm với 30 SV cho thấy 82% người học cải thiện ít nhất 0,5 điểm (trên thang 5 điểm) ở tiêu chí "mạch lạc" sau khi chỉnh sửa dựa vào phản hồi của GPT-4. Cách làm này không chỉ khuyến khích tự học mà còn tạo động lực nâng cao năng lực phản tư (self-editing skills) của SV; 2) Hỗ trợ phản hồi song song (parallel feedback) là hướng tiếp cận giúp cân bằng giữa yếu tố công nghệ và chuyên môn sư phạm. Trong mô hình này, mỗi bài viết được chấm bởi cả GV và LLMs, sau đó SV so sánh hai dạng phản hồi. Ở một số trường như Đại học Ngoại ngữ - ĐHQG Hà Nội, hình thức này đã được áp dụng thử nghiệm trong học phần *Academic Writing II* với kết quả tích cực: 90% SV cho rằng phản hồi của LLMs "dễ hiểu, chi tiết và hữu ích", trong khi GV đánh giá công cụ giúp tiết kiệm khoảng 30% thời gian chấm bài. Mô hình phản hồi song song giúp người học đối chiếu và rút ra điểm giao thoa giữa tiêu chí học thuật và ngôn ngữ tự nhiên, từ đó nâng cao khả năng tự điều chỉnh phong cách viết; 3) Phản hồi theo tiêu chí (criteria-based feedback) là mô hình khai thác ưu thế kỹ thuật của LLMs trong phân tích đa chiều. Mỗi bài viết được hệ thống đánh giá trên các tiêu chí cụ thể như: ngữ pháp, mạch lạc, lập luận và phong cách học thuật. Ví dụ, GPT-4 có thể tự động chỉ ra rằng "The argument in paragraph 3 lacks logical transition" và đề xuất cấu trúc liên kết phù hợp. Ngoài ra, hệ thống có thể thống kê tần suất lỗi (grammar error frequency) và tỷ lệ cải thiện qua từng lần nộp bài,

giúp GV theo dõi tiến trình học tập (learning analytics) một cách khoa học. Mô hình này đặc biệt phù hợp với các khóa học trực tuyến quy mô lớn (MOOCs) hoặc các lớp kỹ năng học thuật đông SV, nơi việc chấm và phản hồi thủ công là thách thức lớn [10, tr. 68].

Tổng thể, ba hướng ứng dụng trên đều góp phần tăng tính công bằng, minh bạch và khách quan trong đánh giá kỹ năng viết. Đồng thời, LLMs giúp cá nhân hóa trải nghiệm học tập - mỗi SV nhận phản hồi phù hợp với năng lực và phong cách riêng. Khi được tích hợp hợp lý vào hệ thống quản lý học tập (LMS) như Moodle hay Canvas, LLMs có thể trở thành "trợ giảng ảo" hỗ trợ GV trong phản hồi tức thời, góp phần hiện thực hóa mục tiêu chuyển đổi số giáo dục đại học tại Việt Nam. Như vậy, việc ứng dụng LLMs không chỉ là xu thế công nghệ mà còn là giải pháp mang tính chiến lược để đổi mới phương pháp giảng dạy Tiếng Anh, hướng tới nền giáo dục số hóa, linh hoạt và lấy người học làm trung tâm.

3. Kết luận

Như vậy, các Mô hình Ngôn ngữ Lớn (LLMs) đã chứng minh được tiềm năng vượt trội trong việc hỗ trợ đánh giá kỹ năng viết tiếng Anh, không chỉ dừng lại ở việc phát hiện lỗi ngữ pháp hay từ vựng mà còn tiến xa hơn trong việc nhận diện tính mạch lạc, lập luận và phong cách diễn đạt của người học. Kết quả nghiên cứu cho thấy, việc tích hợp LLMs trong dạy - học viết giúp tăng tính khách quan, rút ngắn thời gian phản hồi, đồng thời cung cấp cho người học những gợi ý chi tiết, mang tính cá nhân hóa cao. Đây là một bước tiến quan trọng trong đổi mới phương pháp giảng dạy Tiếng Anh theo hướng ứng dụng công nghệ, đặc biệt trong bối cảnh giáo dục số hóa và học tập suốt đời đang trở thành xu thế toàn cầu. Tuy nhiên, công nghệ này vẫn tồn tại những giới hạn nhất định. LLMs tuy có thể đánh giá cấu trúc và ngữ pháp chính xác, nhưng vẫn gặp khó khăn khi xử lý các khía cạnh liên quan đến cảm xúc, sáng tạo, hoặc giá trị nhân văn trong bài viết. Do đó, việc áp dụng LLMs không nên được xem như sự thay thế hoàn toàn cho GV mà cần được thiết kế như một mô hình hỗ trợ lại (hybrid assessment), trong đó AI đóng vai trò phản hồi kỹ thuật, còn GV đảm nhiệm vai trò định hướng tư duy học thuật, phát triển năng lực phản biện và sáng tạo ngôn ngữ cho người học. Hướng nghiên cứu tiếp theo cần tập trung vào việc xây dựng các mô hình ngôn ngữ được huấn luyện trên dữ liệu học thuật của người Việt học Tiếng Anh (Vietnamese EFL corpus) để đảm bảo tính tương thích văn hóa, phong cách diễn đạt và mục tiêu học tập. Đồng thời, cần phát triển khung đánh giá tích hợp giữa LLMs và GV, kết hợp phản hồi định lượng và định tính nhằm tạo ra một hệ thống đánh giá thông minh, công bằng và mang tính giáo dục sâu sắc hơn. Khi đó, LLMs không chỉ là công cụ hỗ trợ mà còn trở thành đối tác học tập thông minh, góp phần hiện thực hóa tầm nhìn về giáo dục cá nhân hóa và công bằng trong thời đại trí tuệ nhân tạo.

Tài liệu tham khảo

[1] Chen, L., Liu, S., & Wang, Y. (2024). *Leveraging LLMs for holistic writing assessment: Beyond grammar correction.* Educational Technology & Society, 27(2), 88–104. https://doi.org/10.30191/ets.2024.270205.

[2] Dzikovska, M. O., Steinhauer, J., & Moore, J. D. (2022). *Combining automated feedback with human assessment to improve academic writing performance*. Journal of Educational Technology Research and Development, 70(4), 189-204. https://doi.org/10.1007/s11423-021-10022-9.

[3] Lee, J., & Kim, H. (2023). *Evaluating GPT-4 as a writing assessor: A comparative study with human raters in academic English writing*. TESOL Quarterly, 57(3), 721-739. https://doi.org/10.1002/tesq.3184.

[4] Li, X., & Park, J. (2023). *Human-AI collaboration in writing evaluation: Implications for fair and transparent assessment.* Language Assessment Quarterly, 20(4), 367-385. https://doi.org/10.1080/15434303.2023.2245123.

[5] OpenAI. (2024). *GPT-4 technical report. OpenAI Research*. https://cdn.openai.com/papers/gpt-4.pdf.

[6] Phan Trung Kiên, Nguyễn Đức Ca, Đinh Tiến Dũng (2024). *Ứng dụng trí tuệ nhân tạo trong dạy học và nghiên cứu khoa học tại các trường đại học.* Tạp chí Giáo dục, tập 24, số 24 tháng 12, tr. 14-19.

[7] Pham Minh Thanh (2024). *Giáo dục số và chuyển đổi công nghệ trong giảng dạy ngoại ngữ ở Việt Nam*. NXB Đại học Quốc gia Hà Nội.

[8] Tan, J., & Wong, M. (2022). *Exploring AI-powered feedback in ESL writing: Accuracy, coherence, and engagement.* System, 108, 102847. https://doi.org/10.1016/j.system.2022.102847.

[9] Uto, M. (2023). *Large Language Models as Automated Essay Raters: Evaluating GPT-based feedback for Japanese EFL learners*. Language Testing in Asia, 13(1), 1-18. https://doi.org/10.1186/s40468-023-00201-6.

[10] Zhang, Y., & Xu, X. (2024). *AI-assisted formative feedback in EFL writing: Opportunities and challenges in higher education.*