

MÔ HÌNH NGÔN NGỮ LỚN TRONG ĐÁNH GIÁ BÀI LUẬN TỰ ĐỘNG: TỔNG QUAN LÝ THUYẾT VÀ ĐỀ XUẤT KHUNG ỨNG DỤNG CHO GIÁO DỤC ĐẠI HỌC VIỆT NAM

Mai Thị Thúy Hà^{1*}, Nguyễn Thùy Linh², Trịnh Thị Xuân², Lê Thị Thanh Thùy²

*Tác giả liên hệ, Email: ha.mtt@phuongdong.edu.vn. ORCID: 0009-0002-1647-0237

Ngày tòa soạn nhận được bài báo: 15/06/2026

Ngày phản biện đánh giá: 22/06/2026

Ngày bài báo được duyệt đăng: 14/09/2026

DOI: 10.59266/houjs.2026.1435

Tóm tắt: Bài báo tổng quan đánh giá bài luận tự động (Automated Essay Scoring - AES) bằng mô hình ngôn ngữ lớn (LLMs) và đề xuất khung ứng dụng cho giáo dục đại học Việt Nam. Tổng quan tường thuật có cấu trúc phân tích 19 nguồn cốt lõi, trong đó có 6 công trình năm 2025-2026. Bằng chứng trực tiếp cho thấy mức đồng thuận LLM-người chấm thay đổi theo dữ liệu, mô hình, rubric và prompt; độ tin cậy không đồng nghĩa với độ giá trị. Khung LLM-Rubric-LMS-Human-in-the-Loop được đề xuất như một quy trình kỹ thuật-sư phạm chưa kiểm chứng trên bài luận tiếng Việt: LLM đề xuất điểm và phản hồi, giảng viên quyết định điểm cuối. Khung kèm rubric, prompt minh họa, cơ chế giảm thiểu rủi ro và kế hoạch kiểm định.

Từ khóa: chấm luận tự động, đánh giá theo rubric, giáo dục đại học, mô hình ngôn ngữ lớn, tiếng Việt

I. Đặt vấn đề

LLMs có thể hỗ trợ phản hồi tức thời và nâng cao hiệu quả đánh giá, nhưng giáo dục phải duy trì vai trò quyết định của con người và bảo đảm sử dụng AI có trách nhiệm (Kasneci và cộng sự, 2023; UNESCO, 2023).

Bài luận đo lường lập luận, tư duy phản biện và diễn đạt học thuật nhưng tạo khối lượng chấm lớn và nguy cơ thiếu nhất quán. AES đã chuyển từ dự đoán

điểm sang đánh giá theo tiêu chí và sinh phản hồi.

Bằng chứng trực tiếp còn chưa nhất quán: mức đồng thuận với người chấm phụ thuộc bộ dữ liệu, thang điểm, mô hình, rubric và prompt (Mizumoto và Eguchi, 2023; Mansour và cộng sự, 2024; Pack và cộng sự, 2024; Huang và Wilson, 2025; Yoshida, 2025). Các nghiên cứu mới cho thấy khó khăn ở đặc trưng diễn ngôn, công bằng và chấm điểm tuyệt đối, đồng thời

¹ Trường Đại học Phương Đông, Hà Nội, Việt Nam

² Trường Đại học Mở Hà Nội, Hà Nội, Việt Nam

đề xuất lập luận theo tiêu chí hoặc so sánh theo cặp (Chu và cộng sự, 2025; Shibata và Miyamura, 2025; Su và cộng sự, 2025). Trong phạm vi tìm kiếm, nhóm tác giả chưa xác định được bộ dữ liệu công khai hoặc chuẩn đối sánh (benchmark) chuẩn hóa dành riêng cho bài luận học thuật tiếng Việt bằng LLM; ViGLUE và ViLLM-Eval chỉ đánh giá năng lực ngôn ngữ nói chung (Nguyen và cộng sự, 2024; Tran và cộng sự, 2024).

Bài báo có ba đóng góp: phân loại bằng chứng trực tiếp và gián tiếp; tổng hợp các yếu tố ảnh hưởng độ tin cậy; và đề xuất khung tích hợp có giám sát, quản trị dữ liệu và kế hoạch kiểm định. Ba câu hỏi nghiên cứu tập trung vào sự phát triển của AES, vai trò/độ tin cậy của LLM và khung phù hợp cho giáo dục đại học Việt Nam.

II. Cơ sở lý thuyết

2.1. Sự phát triển của hệ thống chấm luận tự động

AES phát triển từ phương pháp dựa trên luật và đặc trưng thủ công, qua học sâu, đến LLM. Các thế hệ đầu chủ yếu dự đoán điểm; LLM có thể đối chiếu rubric, giải thích và sinh phản hồi nhưng cũng làm tăng rủi ro thiếu ổn định, thiên lệch và nhận xét không có căn cứ. AES tập trung vào điểm; Automated Essay Evaluation (đánh giá bài luận tự động) bao gồm điểm và phân tích theo tiêu chí; Automated Writing Evaluation rộng hơn, gồm phản hồi và hỗ trợ chỉnh sửa. Bài báo dùng AES làm thuật ngữ trung tâm nhưng khung đề xuất có thêm chức năng phản hồi.

2.2. LLM và các cơ chế liên quan đến đánh giá bài luận

LLM có thể thích ứng nhiệm vụ chấm qua zero-shot, few-shot hoặc tinh chỉnh. Với AES, prompt phải chứa đề bài,

rubric, quy tắc chấm và cấu trúc đầu ra; kết quả vẫn cần hiệu chỉnh theo người chấm và bối cảnh sử dụng.

2.3. Các hướng tiếp cận và cơ sở bằng chứng

Có hai nhóm bằng chứng. G-EVAL, PROMETHEUS và LLM-RUBRIC cung cấp nguyên lý đánh giá theo tiêu chí nhưng không trực tiếp chứng minh hiệu quả chấm bài luận đại học (Hashemi và cộng sự, 2024; Kim và cộng sự, 2024; Liu và cộng sự, 2023). Nhóm AES trực tiếp cho thấy QWK và độ ổn định thay đổi theo mô hình, prompt và dữ liệu (Mansour và cộng sự, 2024; Mizumoto và Eguchi, 2023; Pack và cộng sự, 2024). Nghiên cứu năm 2025-2026 bổ sung bằng chứng về công bằng, độ chi tiết rubric, lập luận đa tiêu chí, đặc trưng diễn ngôn, so sánh theo cặp và hiệu năng trên benchmark (Chu và cộng sự, 2025; Huang và Wilson, 2025; Mughal và cộng sự, 2026; Shibata và Miyamura, 2025; Su và cộng sự, 2025; Yoshida, 2025).

2.4. Các mô hình đánh giá bài luận

Chấm tổng thể cho một điểm chung; chấm phân tích tách các thành phần; đánh giá theo rubric mô tả mức đạt của từng tiêu chí. Rubric tăng tính minh bạch và định hướng phản hồi nhưng không tự loại bỏ thiên lệch, vì vậy vẫn cần kiểm định với giảng viên và người học.

III. Phương pháp nghiên cứu

Nghiên cứu sử dụng tổng quan tường thuật có cấu trúc kết hợp phân tích khái niệm. Phương pháp nhằm hệ thống hóa lý thuyết và xây dựng khung, không phải tổng quan hệ thống hay phân tích gộp; luồng sàng lọc chỉ tham chiếu cấu trúc PRISMA.

Tìm kiếm ban đầu được thực hiện từ 15/01-28/02/2026 trong Scopus, Web of Science, IEEE Xplore, ACM Digital Library và SpringerLink, bổ sung Google Scholar; cập nhật ngày 13/08/2026 tại ACL Anthology và SpringerLink. Chuỗi lõi kết hợp các nhóm từ “automated essay scoring/essay evaluation/writing assessment”, “large language model/LLM/GPT/generative AI” và “rubric/feedback/higher education”, tìm trong tiêu đề, tóm tắt và từ khóa. Phạm vi gồm tiếng Anh, tiếng Việt, giai đoạn 2023-13/08/2026 và một số nguồn nền tảng cần thiết.

Đưa vào các bài báo/hội nghị phản biện liên quan AES hoặc đánh giá văn bản bằng LLM; preprint, báo cáo kỹ thuật và

chính sách chỉ dùng khi cần cho nền tảng hoặc quản trị. Loại nguồn không có đánh giá theo tiêu chí/phản hồi, không thuộc giáo dục hoặc chỉ khai thác đặc trưng bề mặt. Đối chiếu DOI, tiêu đề và tác giả để loại trùng. Một thành viên sàng lọc lần đầu và nhóm tác giả kiểm tra trường hợp chưa rõ; chưa có hai người sàng lọc độc lập là một hạn chế.

Đợt đầu xác định 52 bản ghi, còn 41 sau loại trùng, 28 sau sàng lọc và 23 sau đọc toàn văn. Đợt cập nhật bổ sung 6 công trình; sau khi ưu tiên bằng chứng AES trực tiếp, quản trị và bối cảnh tiếng Việt, 10 nguồn nền tảng/gián tiếp được lược bỏ, còn 19 nguồn cốt lõi. Bảng 1 trình bày luồng lựa chọn; Bảng 2 tóm tắt mức độ bằng chứng.

Bảng 1. Quy trình nhận diện, sàng lọc và cập nhật tài liệu

Bước sàng lọc	Số lượng
Bản ghi xác định ở đợt ban đầu	52
Bản ghi sau loại trùng	41
Bản ghi sau sàng lọc tiêu đề và tóm tắt	28
Toàn văn bị loại: không benchmark có rubric/phản hồi (3); không thuộc giáo dục (2)	5
Nguồn đưa vào ở đợt ban đầu	23
Nguồn bổ sung qua cập nhật ngày 13/08/2026	6
Nguồn sau cập nhật trước khi rút gọn phạm vi	29
Nguồn nền tảng/gián tiếp được lược bỏ	10
Nguồn cốt lõi giữ lại trong bài rút gọn	19

Nguồn: Nhóm tác giả đề xuất

Bảng 2. Tổng hợp các nguồn bằng chứng trực tiếp và cốt lõi tiêu biểu

Nguồn	Loại (mức bằng chứng)	Nhiệm vụ và dữ liệu	Kết quả / vai trò chính
Mizumoto và Eguchi (2023)	Bài báo phản biện - AES trực tiếp	Chấm 12.100 bài TOEFL, GPT-3.5	QWK ≈ 0,39; AES bổ trợ, không thay người
Mansour và cộng sự (2024)	Hội nghị phản biện - AES trực tiếp	Chấm bài, ChatGPT & Llama	QWK ≈ 0,30-0,31; đồng thuận thấp-trung bình
Yoshida (2024)	Hội nghị phản biện - AES trực tiếp	Few-shot AES với GPT	Cách chọn ví dụ ảnh hưởng điểm và thiên lệch
Pack và cộng sự (2024)	Bài báo phản biện - AES trực tiếp	AES cho người học tiếng Anh	Mạnh bề mặt, yếu diễn ngôn; tin cậy ≠ giá trị

Nguồn	Loại (mức bằng chứng)	Nhiệm vụ và dữ liệu	Kết quả / vai trò chính
Liu và cộng sự (2023) - G-EVAL	Hội nghị phản biện - phương pháp chung	Chăm chất lượng văn bản, GPT-4	Nguyên lý chuỗi suy luận + điền biểu mẫu
Kim và cộng sự (2024) - PROMETHEUS	Hội nghị phản biện - phương pháp chung	Đánh giá theo rubric	Mô hình đánh giá chuyên biệt
Hashemi và cộng sự (2024) - LLM-RUBRIC	Hội nghị phản biện - phương pháp chung	Đánh giá đa chiều	Hiệu chỉnh theo người chấm
Huang và Wilson (2025)	Hội nghị - AES trực tiếp	541 bài lớp 3-4; nhiều LLM/prompt	Prompt ảnh hưởng độ chính xác; kiểm tra công bằng
Yoshida (2025)	Hội nghị - AES trực tiếp	TOEFL11; 4 LLM; 3 mức rubric	Rubric giảm lược có thể tương đương, giảm token
Chu và cộng sự (2025)	Hội nghị - AES trực tiếp	ASAP/ASAP++/Feedback Prize; đa tiêu chí	Lập luận theo tiêu chí hỗ trợ chấm đa đặc trưng
Su và cộng sự (2025)	Hội nghị - benchmark AES	EssayJudge; 18 MLLM	Khoảng cách lớn ở đặc trưng diễn ngôn
Shibata và Miyamura (2025)	Hội nghị - AES trực tiếp	So sánh theo cặp + RankNet	Cải thiện chấm zero-shot so với chấm tuyệt đối
Mughal và cộng sự (2026)	Bài báo - AES trực tiếp	ASAP-AES, LA-AES; GPT/Gemini	QWK trung bình khoảng 0,43-0,45

Nguồn: Nhóm tác giả đề xuất

IV. Kết quả nghiên cứu và thảo luận

4.1. Xu hướng phát triển AES trong bối cảnh LLM (Câu hỏi 1)

AES đang chuyển từ dự đoán điểm sang đánh giá theo rubric và phản hồi. Bảng chứng trực tiếp báo cáo kết quả khác nhau: một số cấu hình trước 2025 đạt QWK khoảng 0,30-0,39, trong khi Mughal và cộng sự (2026) báo cáo khoảng 0,43-0,45 trên hai benchmark; các giá trị không thể so sánh trực tiếp khi dữ liệu và quy trình chấm khác nhau. Các nghiên cứu năm 2025 cho thấy prompt, độ chi tiết rubric, công bằng và cách biểu diễn nhiệm vụ đều ảnh hưởng kết quả (Huang và Wilson, 2025; Shibata và Miyamura, 2025; Yoshida, 2025).

4.2. Vai trò và độ tin cậy của LLM trong đánh giá theo rubric (Câu hỏi 2)

LLM có thể đề xuất điểm, sinh phản hồi, hỗ trợ chỉnh sửa và cá nhân hóa học

tập. Mô hình thường mạnh hơn ở ngữ pháp/từ vựng nhưng yếu hơn ở diễn ngôn và phát triển ý; lập luận theo tiêu chí hỗ trợ truy vết nhưng không phản ánh đầy đủ cơ chế nội tại (Chu và cộng sự, 2025; Pack và cộng sự, 2024; Su và cộng sự, 2025). Vì vậy, LLM phù hợp vai trò đề xuất hơn là quyết định.

4.3. So sánh một số mô hình ứng viên cho AES tiếng Việt

Việc chọn LLM cho tiếng Việt phải dựa trên cùng dữ liệu, rubric, phản cứng và quy trình chấm. Báo cáo kỹ thuật về khả năng đa ngữ không chứng minh độ tin cậy trong AES tiếng Việt; do đó không xếp hạng mô hình khi chưa có benchmark trực tiếp.

4.4. Thách thức và khoảng trống tại Việt Nam

Các thách thức chính gồm thiên lệch, ảo giác, khả năng giải thích hạn chế,

công bằng, liên chính học thuật và quyền riêng tư (Bender và cộng sự, 2021; Tlili và cộng sự, 2023). Ở Việt Nam, thiếu dữ liệu công khai, benchmark và rubric tham chiếu trong phạm vi tổng quan làm tăng khó khăn kiểm định.

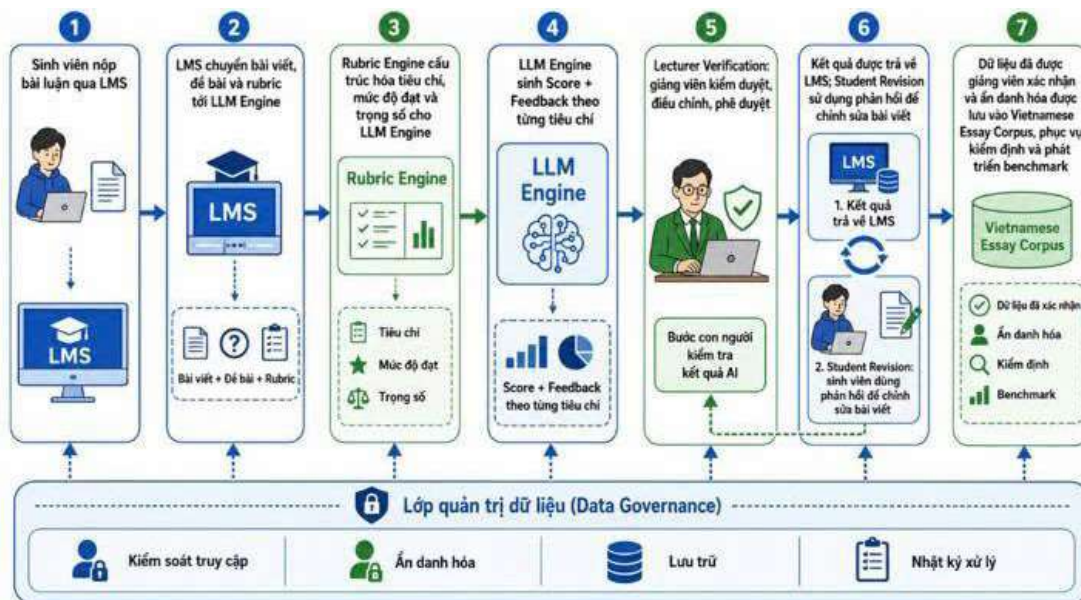
V. Đề xuất khung ứng dụng (Câu hỏi 3)

5.1. Phạm vi, bản chất và nguyên tắc vận hành của khung đề xuất

Khung LLM-Rubric-LMS-Human-in-the-Loop là khung khái niệm và quy trình kỹ thuật-sư phạm, chưa phải mô hình LLM mới, phần mềm đã triển khai hay hệ thống đã kiểm chứng. Bốn trụ cột lần lượt đảm nhiệm đánh giá, kiểm soát tiêu chí, quản lý

luồng dữ liệu và giám sát của con người; bảy cấu phần trong Hình 1 là khai triển vận hành. Đóng góp nằm ở cách tích hợp cho bối cảnh giáo dục đại học Việt Nam, không phải ở từng thành phần riêng lẻ.

LLM chỉ đề xuất; giảng viên kiểm duyệt và quyết định điểm cuối. Trong giai đoạn chưa kiểm định, mọi bài đều phải được duyệt. Ngưỡng gắn cờ phải xác lập trên tập hiệu chỉnh từ độ lệch qua nhiều lần chạy, chênh lệch LLM-giảng viên và vị trí gần ngưỡng đạt. Khi bất đồng, lưu điểm LLM, điểm giảng viên và lý do điều chỉnh. Sinh viên phải được thông báo AI tham gia, được tiếp cận căn cứ theo rubric và được yêu cầu xem xét lại.



Hình 1. Sơ đồ khung LLM-Rubric-LMS-Human-in-the-Loop

Nguồn: Nhóm tác giả đề xuất.

5.2. Rubric minh họa (đầy đủ năm mức)

Bảng 3 minh họa năm tiêu chí, năm mức và trọng số. Mô tả mức phải

quan sát được; rubric phải được chuyên gia thẩm định và thử nghiệm đồng thuận giữa giảng viên trước khi dùng làm chuẩn cho LLM.

Bảng 3. Rubric minh họa cho bài luận học thuật (thang 1-5)

Tiêu chí	Mức 1	Mức 2	Mức 3	Mức 4	Mức 5	Trọng số
Đáp ứng đề bài và nội dung	Lạc đề, thiếu nội dung	Bám đề một phần, sơ sài	Đáp ứng cơ bản, còn thiếu ý	Đáp ứng tốt, đầy đủ	Đáp ứng đầy đủ, sâu sắc	25%
Lập luận và dẫn chứng	Không luận điểm/dẫn chứng	Luận điểm mờ, dẫn chứng yếu	Có luận điểm, dẫn chứng cơ bản	Lập luận chặt, dẫn chứng phù hợp	Lập luận sắc bén, thuyết phục	25%
Độ chính xác kiến thức	Nhiều sai sót nghiêm trọng	Còn sai sót đáng kể	Cơ bản đúng, vài lỗi nhỏ	Chính xác, ít lỗi	Chính xác, hiểu sâu	15%
Cấu trúc và mạch lạc	Không có bố cục	Bố cục rời rạc	Bố cục cơ bản, chuyển ý chưa mượt	Bố cục rõ, mạch lạc	Bố cục chặt, liên kết tốt	20%
Ngôn ngữ và trích dẫn học thuật	Nhiều lỗi, không trích dẫn	Nhiều lỗi, trích dẫn sai chuẩn	Điển đạt ổn, trích dẫn cơ bản	Văn phong học thuật, trích dẫn đúng	Văn phong chuẩn, trích dẫn đầy đủ	15%

Nguồn: Nhóm tác giả xây dựng.

5.3. Prompt và đầu ra minh họa

Prompt gồm vai trò, đề bài, rubric, quy tắc không suy diễn và lược đồ JSON. Điểm quy đổi: $S_{10} = 2\Sigma(w_i s_i)$, $\Sigma w_i = 1$, làm tròn đến 0,5. Với điểm 4, 3, 4, 3, 4 trong ví dụ, kết quả là 7,1 và làm tròn thành 7,0. Nếu thiếu căn cứ, score và diem_tong

bằng null, chuyển cho giảng viên. “Độ tự tin” do LLM tự sinh không phải xác suất đã hiệu chỉnh; độ bất định phải ước lượng qua nhiều lần chạy hoặc trên tập giảng viên đã chấm. Cấu trúc prompt vận hành và lược đồ đầu ra được trình bày tại Phụ lục A; rubric sử dụng trực tiếp các mô tả mức tại Bảng 3.

PROMPT MINH HỌA	
[VAI TRÒ]	Trợ lý chấm bài; chỉ đề xuất, không quyết định điểm cuối.
[ĐẦU VÀO]	Đề bài; toàn bộ rubric 5 mức ở Bảng 4; bài luận.
[QUY TẮC]	Mỗi điểm phải có trích dẫn nguyên văn. Thiếu căn cứ: score=null, diem_tong=null, route_to_lecturer=true. Không suy đoán danh tính/ý định.
[ĐẦU RA]	JSON gồm score, evidence, comment theo từng tiêu chí; diem_tong; uncertainty_note; route_to_lecturer.
[TÍNH ĐIỂM]	$S_{10}=2\Sigma(w_i*s_i)$, làm tròn 0,5. Ví dụ 4-3-4-3-4 cho 7,0, không phải 7,5.

Hình 2. Prompt và đầu ra minh họa có cấu trúc của mô-đun LLM

Nguồn: Nhóm tác giả xây dựng.

5.4. Cơ chế giảm thiểu rủi ro

Các cơ chế dưới đây nhằm giảm thiểu (không phải loại bỏ) các rủi ro đã nêu ở mục 4.4:

Thiên lệch và công bằng: so sánh MAE/QWK kèm khoảng tin cậy giữa các nhóm được xác định trước; không phân tích nhóm quá nhỏ hoặc dùng

thuộc tính nhạy cảm ngoài mục đích được phê duyệt.

Ảo giác: yêu cầu trích dẫn căn cứ; giảng viên kiểm tra độ chính xác và mức phù hợp giữa căn cứ với nhận xét.

Giải thích: đầu ra theo tiêu chí hỗ trợ truy vết nhưng không được coi là phân ánh đầy đủ cơ chế nội tại của LLM.

Quyền riêng tư: cần đồng thuận, phân quyền, mã hóa, thời hạn lưu, xóa dữ liệu, kiểm soát nhà cung cấp và đánh giá nguy cơ tái nhận diện.

Liên chính: LMS có thể lưu lịch sử chỉnh sửa và bản khai sử dụng AI của sinh viên để hỗ trợ xem xét quá trình viết. Lịch sử chỉnh sửa hoặc công cụ phát hiện văn bản AI không phải bằng chứng độc lập để kết luận vi phạm; cần chính sách công khai, quy trình giải trình và quyền khiếu nại.

5.5. Kế hoạch kiểm định thực nghiệm

Kiểm định cần bộ bài luận tiếng Việt do ít nhất hai giảng viên chấm độc lập. Trước khi so sánh LLM phải báo cáo QWK/ICC liên giám khảo, quy trình hòa giải và điểm chuẩn. Đánh giá gồm đồng thuận, sai số, ổn định, độ giá trị của rubric/điểm, chất lượng phản hồi, công bằng, tỷ lệ điều chỉnh, thời gian và chi phí.

Bảng 4. Tiêu chí và chỉ số kiểm định khung đề xuất

Tiêu chí	Chỉ số	Ý nghĩa
Đồng thuận LLM-giảng viên	QWK, Pearson, tỷ lệ trùng khớp chính xác/lần cận	Độ tin cậy so với người chấm
Sai số điểm	MAE, RMSE	Mức lệch điểm
Độ ổn định	ICC qua k lần chạy, phương sai	Tính lặp lại
Chất lượng phản hồi	Điểm rubric do chuyên gia đánh giá	Giá trị sư phạm
Mức phụ thuộc con người	Tỷ lệ bài giảng viên phải điều chỉnh điểm	Khả năng thay thế công sức
Kiểm soát ảo giác	Tỷ lệ phản hồi không có căn cứ	Độ tin cậy của nhận xét
Công bằng	Chênh lệch QWK/MAE theo nhóm	Bình đẳng đánh giá
Hiệu quả vận hành	Thời gian tiết kiệm, chi phí/bài	Tính khả thi triển khai
Đồng thuận liên giám khảo	QWK/ICC giữa ≥2 giảng viên	Xác lập độ tin cậy của chuẩn tham chiếu
Độ giá trị	Thẩm định rubric; phân tích cấu trúc; hệ quả	Đánh giá mô hình có đo đúng năng lực cần đo
Quyền người học	Tỷ lệ yêu cầu xem xét; kết quả giải quyết	Khả năng giải trình và khiếu nại

Nguồn: Nhóm tác giả đề xuất

5.6. Tính khả thi và lộ trình triển khai tại Việt Nam

Lộ trình gồm: ngắn hạn dùng AI hỗ trợ có giảng viên duyệt; trung hạn tích hợp LMS và tích lũy dữ liệu ẩn danh; dài hạn xây dựng benchmark, rubric tham chiếu và kiểm định độc lập. Cần đo chi phí API/GPU, gán nhãn, rubric, tích hợp, kiểm duyệt và bảo mật. Chỉ khẳng định giảm tải sau khi đo thời gian tiết kiệm ở đọc sơ bộ, trích căn cứ, gợi ý điểm và soạn phản hồi.

VI. Kết luận và kiến nghị

Bài báo cho thấy AES đang mở rộng từ dự đoán điểm sang đánh giá theo rubric và phản hồi, nhưng kết quả phụ thuộc dữ liệu, mô hình và prompt. Thách thức về diễn ngôn, công bằng, ổn định và độ giá trị khiến LLM chỉ nên đề xuất, còn giảng viên quyết định. Đóng góp chính là phân loại bằng chứng cập nhật đến 2026 và đề xuất khung tích hợp có giám sát, quản trị dữ liệu, xử lý bất đồng và kế hoạch kiểm

định; khung chưa được xác nhận ưu việt hơn phương án khác.

Hạn chế gồm phạm vi của tổng quan tường thuật, sàng lọc chưa độc lập bởi hai người, bằng chứng trực tiếp chủ yếu ngoài tiếng Việt và khung chưa kiểm định. Nghiên cứu tiếp theo cần xây dựng dữ liệu tiếng Việt có đồng thuận giám khảo, so sánh baseline và đánh giá độ tin cậy, độ giá trị, công bằng, chi phí và tác động sự phạm.

Lời cảm ơn. Nghiên cứu này được tài trợ bởi đề tài cấp Trường Đại học Mở Hà Nội, mã số MHN2026-02-10.

Phụ lục A. Prompt vận hành và lược đồ đầu ra

Vai trò và giới hạn. Bạn là trợ lý hỗ trợ chấm bài. Bạn chỉ được đối chiếu bài làm với đề bài và rubric tại Bảng 3 để đề xuất điểm và phản hồi. Không được thay đổi tiêu chí hoặc trọng số; suy đoán thông tin không có trong bài; kết luận về đạo văn, vi phạm liên chính hoặc đặc điểm cá nhân. Điểm cuối cùng do giảng viên quyết định.

Đầu vào. Đề bài; rubric đầy đủ tại Bảng 3; bài luận đã loại thông tin định danh.

Quy tắc đánh giá. Mỗi tiêu chí phải có score, evidence và comment. Evidence phải là câu hoặc đoạn trích nguyên văn từ bài làm. Nếu không đủ căn cứ, đặt score = null, evidence = null, nêu lý do trong comment, đặt diem_tong = null và route_to_lecturer = true. Không tự loại tiêu chí thiếu căn cứ để tính lại điểm.

Cách tính. $S_{10} = 2\sum(w_i S_i)$, với $\sum w_i = 1$; kết quả làm tròn đến 0,5.

Điều kiện chuyển giảng viên. Đặt route_to_lecturer = true khi thiếu căn cứ; kết quả giữa các lần chạy không ổn định;

bài nằm gần ngưỡng đạt/không đạt; xuất hiện nội dung cần con người xem xét; hoặc trong toàn bộ giai đoạn hệ thống chưa được kiểm định.

Bảo mật. Không ghi lại họ tên, mã sinh viên, lớp hoặc thông tin liên hệ trong đầu ra. Nếu phát hiện thông tin định danh, đặt canh_bao_thong_tin_nhay_cam = true.

Lược đồ đầu ra:

```
{
  "tieu_chi": {
    "score": null,
    "evidence": "Trích dẫn nguyên văn
                  hoặc null",
    "comment": "Nhận xét theo rubric"
  },
  "diem_tong": null,
  "uncertainty_note": "Nguyên nhân bất
                       định",
  "route_to_lecturer": true,
  "tieu_chi_thieu_can_cu": [],
  "canh_bao_thong_tin_nhay_cam": false
}
```

Ghi chú. *uncertainty_note chỉ mô tả nguyên nhân bất định, không phải xác suất tin cậy đã hiệu chỉnh. Trong giai đoạn chưa kiểm định, route_to_lecturer mặc định là true cho mọi bài.*

Nguồn: Nhóm tác giả xây dựng.

Tài liệu tham khảo

- 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>
- Chu, S. Y., Kim, J. W., Wong, B., & Yi, M. Y. (2025). Rationale behind essay scores: Enhancing S-LLM's multi-trait essay scoring with rationale generated by LLMs. *Findings of the Association for Computational Linguistics: NAACL 2025*, 5811-5829. <https://doi.org/10.18653/v1/2025.findings-naacl.322>

- Hashemi, H., Eisner, J., Rosset, C., Van Durme, B., & Kedzie, C. (2024). LLM-RUBRIC: A multidimensional, calibrated approach to automated evaluation of natural language texts. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 13806-13834.
- Huang, Y., & Wilson, J. (2025). Evaluating LLM-based automated essay scoring: Accuracy, fairness, and validity. *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress*, 71-83. <https://aclanthology.org/2025.aimecon-wip.9/>
- Kasneji, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneji, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kim, S., Shin, J., Cho, Y., Jang, J., Longpre, S., Lee, H., Yun, S., Shin, S., Kim, S., Thorne, J., & Seo, M. (2024). PROMETHEUS: Inducing fine-grained evaluation capability in language models. *International Conference on Learning Representations (ICLR)*.
- Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). G-EVAL: NLG evaluation using GPT-4 with better human alignment. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Mansour, W. A., Albatarni, S., Eltanbouly, S., & Elsayed, T. (2024). Can large language models automatically score the proficiency of written essays? *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2777-2786.
- Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2), 100050. <https://doi.org/10.1016/j.rmal.2023.100050>
- Mughal, N., Imran, A. S., Daudpota, S. M., Kastrati, Z., & Noor, W. (2026). Exploring potential of large language models for automated essay scoring in education. *Discover Artificial Intelligence*, 6, Article 166. <https://doi.org/10.1007/s44163-026-01002-y>
- Nguyen, T.-H., Le, A.-C., & Nguyen, V.-C. (2024). ViLLM-Eval: A comprehensive evaluation suite for Vietnamese large language models [Preprint]. *arXiv*.
- Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6, 100234. <https://doi.org/10.1016/j.caeai.2024.100234>
- Shibata, T., & Miyamura, Y. (2025). LCES: Zero-shot automated essay scoring via pairwise comparisons using large language models. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 29988-30001. <https://doi.org/10.18653/v1/2025.emnlp-main.1523>
- Su, J., Yan, Y., Fu, F., Han, Z., Ye, J., Liu, X., Huo, J., Zhou, H., & Hu, X. (2025). EssayJudge: A multi-granular benchmark for assessing automated essay scoring capabilities of multimodal large language models. *Findings of the Association for Computational Linguistics: ACL 2025*, 6363-6389. <https://doi.org/10.18653/v1/2025.findings-acl.329>

- Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, 10, Article 15. <https://doi.org/10.1186/s40561-023-00237-x>
- Tran, M.-N., Nguyen, P.-V., Nguyen, L., & Dinh, D. (2024). ViGLUE: A Vietnamese general language understanding benchmark and analysis of Vietnamese language models. *Findings of the Association for Computational Linguistics: NAACL 2024*, 4174-4189.
- UNESCO. (2023). *Guidance for generative AI in education and research*. <https://doi.org/10.54675/EWZM9535>
- Yoshida, L. (2024). The impact of example selection in few-shot prompting on automated essay scoring using GPT models. In *Artificial Intelligence in Education (AIED 2024)* (Lecture Notes in Computer Science, Vol. 14830, pp. 61-73). Springer. https://doi.org/10.1007/978-3-031-64315-6_5
- Yoshida, L. (2025). Do we need a detailed rubric for automated essay scoring using large language models? In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds.), *Artificial intelligence in education* (Lecture Notes in Computer Science, Vol. 15882, pp. 60-67). Springer. https://doi.org/10.1007/978-3-031-98465-5_8

LARGE LANGUAGE MODELS IN AUTOMATED ESSAY SCORING: A THEORETICAL REVIEW AND PROPOSED FRAMEWORK FOR VIETNAMESE HIGHER EDUCATION

Mai Thi Thuy Ha¹, Nguyen Thuy Linh², Trinh Thi Xuan², Le Thi Thanh Thuy²

Abstract: *This paper reviews large-language-model-based Automated Essay Scoring (AES) and proposes a framework for Vietnamese higher education. A structured narrative review analyzes 19 core sources, including six studies published in 2025-2026. Evidence shows that LLM-human agreement varies with datasets, models, rubrics, and prompts; reliability alone does not establish validity. The proposed LLM-Rubric-LMS-Human-in-the-Loop framework is an unvalidated socio-technical process for Vietnamese essays: the LLM proposes scores and feedback, while the lecturer retains final authority. It includes an illustrative rubric and prompt, risk-mitigation mechanisms, and an empirical validation plan.*

Keywords: *automated essay scoring, higher education, large language models, rubric-based assessment, Vietnamese*

¹ University of Phuong Dong, Hanoi, Vietnam

² Hanoi Open University, Hanoi, Vietnam