

XÂY DỰNG BỘ TIÊU CHÍ VÀ ỨNG DỤNG MÔ HÌNH NGÔN NGỮ LỚN TRONG ĐÁNH GIÁ CHẤT LƯỢNG BÀI GIẢNG ĐIỆN TỬ PHỤC VỤ GIÁO DỤC MỞ TẠI TRƯỜNG ĐẠI HỌC MỞ HÀ NỘI

Nguyễn Thị Khánh Quyên¹, Nguyễn Thị Thúy¹, Chu Minh¹, Trần Thanh Tùng¹
Email: nguyenkhanhquyen@hou.edu.vn. ORCID: 0009-0006-1347-9761

Ngày tòa soạn nhận được bài báo: 06/07/2026

Ngày hoàn thành đánh giá: 05/10/2026

Ngày bài báo được duyệt đăng: 19/10/2026

DOI: 10.59266/houjs.2026.1488

Tóm tắt: Trường Đại học Mở Hà Nội hiện thẩm định chất lượng bài giảng điện tử chủ yếu bằng rà soát thủ công, cách làm tốn nhân lực và khó cho kết quả thống nhất vì chuẩn mới dừng ở mức mô tả, chưa thành bộ tiêu chí chấm cụ thể. Nghiên cứu có hai mục tiêu: xây bộ tiêu chí bám theo Quyết định 5005/QĐ-ĐHM, và thử cho mô hình ngôn ngữ lớn (LLM) chấm phần tiêu chí phù hợp. Điều 4 đến 10 được chuyển thành 37 tiêu chí trên bảy trục, gán hai nhãn: bản chất xác minh và dữ liệu đầu vào cần kiểm. Thực nghiệm đặt trên nhóm tiêu chí đọc được trên kịch bản, dùng các bài giảng tự soạn gieo lỗi sẵn để so sánh hai cách viết câu lệnh, đối chiếu đáp án chuẩn. Cách buộc mô hình nêu bằng chứng đạt độ nhạy trung bình 96,7%, cao hơn mức 83,3% của cách chấm trực tiếp và lưu được bằng chứng kiểm tra, đổi lại báo động giả tăng nhẹ. Do quy mô thử nghiệm nhỏ, kết quả này mới là chỉ dấu ban đầu. Nghiên cứu đề xuất bản đồ ủy thác hai chiều cho 37 tiêu chí, giữ quyết định cho con người ở xác minh sự thật và các tiêu chí âm thanh, hình ảnh, tương tác. Bộ tiêu chí và bản đồ ủy thác là tài liệu tham chiếu khi nghiệm thu học liệu điện tử của Nhà trường.

Từ khóa: bài giảng điện tử, bảo đảm chất lượng, bộ tiêu chí đánh giá, chuẩn học liệu điện tử, mô hình ngôn ngữ lớn

I. Đặt vấn đề

Sự phát triển mạnh mẽ của giáo dục số kéo theo gia tăng nhu cầu thẩm định bài giảng điện tử phục vụ đào tạo trực tuyến và kho tài nguyên giáo dục mở. Nhằm kiểm soát chất lượng, Trường Đại học Mở Hà Nội đã ban hành Quyết định 5005/

QĐ-ĐHM (2023) về tiêu chuẩn và tổ chức sản xuất học liệu điện tử, trong đó quy định chi tiết các tiêu chí từ cấu trúc, nội dung, hình thức, nhận diện thương hiệu đến kỹ thuật đóng gói, âm thanh và đánh giá. Tuy nhiên, các tiêu chuẩn này còn ở dạng mô tả, chưa thành bộ tiêu chí chấm cụ thể để nhiều thẩm định viên ra cùng

¹ Trường Đại học Mở Hà Nội, Hà Nội, Việt Nam

kết quả, quy trình rà soát vì vậy tốn công, khó thống nhất. Với một trường đại học mở, nơi phần lớn học liệu được số hóa và phát hành trực tuyến, chất lượng bài giảng điện tử gắn liền với chất lượng của cả hệ thống giáo dục mở. Với sự phát triển của mô hình ngôn ngữ lớn (LLM) có thể hỗ trợ trong công tác này, nhưng nghiên cứu quốc tế cho thấy LLM đánh giá giáo dục còn yếu trong giám sát con người, mức độ ổn định và minh bạch trong quy trình xử lý. Nghiên cứu hướng đến hai mục tiêu chính là xây bộ tiêu chí bám sát toàn bộ Quyết định 5005 và thử nghiệm cho LLM chấm theo chính bộ tiêu chí đó, sau đây từ kết quả đề xuất mức ủy thác phù hợp cho từng nhóm tiêu chí.

II. Cơ sở lý thuyết

2.1. Các nghiên cứu liên quan.

Yan và cộng sự (2024) tổng quan những thách thức thực tiễn và đạo đức khi dùng LLM trong giáo dục. Nhóm này nhận thấy phần lớn nghiên cứu khó tái lập và thiếu minh bạch. Trong khi đó, nghiên cứu của Grévisse (2024) thử cho LLM chấm câu trả lời ngắn trong đào tạo y khoa; kết quả dùng được nhưng vẫn cần con người giám sát. Công trình nghiên cứu của Pack và cộng sự (2024) ghi nhận kết quả của LLM thay đổi giữa các lần chấm dù dùng cùng một bộ tiêu chí. Nghiên cứu của López-Acosta và cộng sự (2026) nhấn mạnh còn thiếu khung bảo đảm chất lượng gắn cả sự phạm, công nghệ và quy chuẩn thể chế.

Gần đây tài này nhất là nghiên cứu của Kim (2025) về ảnh hưởng của cách viết lệnh tới độ ổn định của LLM, cho rằng chỉ nên dùng LLM ở vai trò hỗ trợ. García-Varela và cộng sự (2025) nhận

thấy điểm số và tiêu chí chấm thay đổi giữa các phiên khi thiếu hướng dẫn, và chỉ nhất quán khi bảng tiêu chí chấm (rubric) có ví dụ, định dạng đầu ra được chuẩn hóa và nhiệt độ đặt thấp; điều này cho thấy bộ tiêu chí giao cho LLM chấm phải viết thành các mục kiểm tra rõ ràng. Hawk và Soujon (2026) thấy mô hình gặp khó khi phải suy luận sự phạm ở mức sâu, nên vai trò con người vẫn cần thiết. Các công trình này có một điểm chung: chúng đo mức khớp nhau giữa máy và người trên những bảng tiêu chí đánh giá có sẵn, nhưng ít mô tả cách xây dựng bộ tiêu chí trước khi giao cho LLM. Đây chính là khoảng trống mà nghiên cứu này hướng tới.

Ngoài ra, tại Trường Đại học Mở Hà Nội, Nguyễn (2026) khảo sát sinh viên học trực tuyến và cho thấy chất lượng học liệu là yếu tố ảnh hưởng mạnh nhất đến trải nghiệm học tập số. Kết quả này cho thấy Trường cần một bộ tiêu chí rõ ràng để kiểm soát chất lượng bài giảng điện tử. Về hướng dùng LLM làm người chấm, Zheng và cộng sự (2023) chỉ ra mô hình khi chấm bị thiên lệch theo vị trí và theo độ dài của câu trả lời. So với các công trình này, nghiên cứu khác ở ba điểm: xây dựng tiêu chí từ một quy định cụ thể trước khi giao cho máy; buộc mô hình trích nguyên văn bằng chứng; đo khả năng phát hiện lỗi trên ngữ liệu có vị trí lỗi biết trước.

2.2. Khung lý thuyết đề xuất.

Nghiên cứu dựa bốn nền tảng: Thứ nhất, lý thuyết giá trị dựa trên lập luận (Kane, 2013) yêu cầu mỗi tiêu chí có chuỗi suy luận rõ từ minh chứng đến kết luận; thứ hai, nguyên lý tương thích kiến tạo của Biggs (1996) biện minh kiểm tra nhất quán giữa mục tiêu, chuẩn đầu ra, câu

hỏi luyện tập; Thứ ba, lý thuyết về mức độ tự động hóa và ủy thác giữa người với máy cho một thang từ tự động hoàn toàn, qua máy đề xuất còn người quyết định, đến chỉ con người làm. Thang này biến câu hỏi “LLM có thay được người không” thành câu hỏi thiết thực hơn: mỗi nhóm tiêu chí nên ủy thác cho máy tới mức nào. Thứ tư, kỹ thuật gieo lỗi có chủ đích, vốn xuất phát từ kiểm thử phần mềm, cho phép tạo một tập đáp án chuẩn với vị trí lỗi biết trước, nhờ đó đo được độ nhạy phát hiện lỗi mà không cần huy động hội đồng chuyên gia quy mô lớn.

III. Phương pháp, dữ liệu nghiên cứu

3.1. Quy trình xây dựng bộ tiêu chí đánh giá.

Bộ tiêu chí bám sát theo Quyết định 5005/QĐ-ĐHM, trong đó lấy 07 nhóm tiêu chuẩn làm 07 trục tiêu chí: cấu trúc bài giảng (Điều 4, 5), nội dung học thuật (Điều 5, 6), câu hỏi và đánh giá (Điều 10), hình thức và nhận diện thương hiệu (Điều 7), kỹ thuật đóng gói và tương tác (Điều 9), âm thanh và đồng bộ (Điều 5, 9), bài giảng văn bản hỗ trợ (Điều 8). Quy trình bốn bước: tổng hợp yêu cầu bắt buộc theo bảy trục kèm số hiệu điều khoản; viết lại mỗi yêu cầu thành phát biểu đối chiếu trực tiếp, kết quả nhị phân đạt/không đạt; phân loại theo hai chiều, bản chất xác minh (kiểm đếm cấu trúc, đối chiếu ngữ nghĩa, xác minh sự thật) và dữ liệu đầu vào (kịch bản, âm thanh, hình ảnh, tương tác, gói SCORM (Sharable Content Object Reference Model - tiêu chuẩn đóng gói nội dung bài giảng điện tử)); rà soát chéo bỏ tiêu chí trùng/mơ hồ, gộp yêu cầu nhỏ. Kết quả là bộ 37 tiêu chí ở Bảng 1. Bộ tiêu chí còn được kiểm tra khả năng vận hành ngay bằng chính thực nghiệm với LLM.

Nếu một tiêu chí liên tục gây tranh cãi giữa các lần chấm của cùng một cách viết câu lệnh (prompt), hoặc buộc mô hình từ chối kết luận vì thiếu căn cứ, thì đó là dấu hiệu tiêu chí ấy thuộc nhóm xác minh sự thật và cần giữ quyền quyết định cho con người, bất kể phân loại ban đầu.

3.2. Ngữ liệu và cách chọn mẫu.

Nhóm tự soạn sáu kịch bản mẫu (L1-L6) thuộc sáu học phần đại cương (Nguyên lý kế toán, Marketing căn bản, Pháp luật đại cương, Tin học đại cương, Tâm lý học đại cương và Kinh tế vi mô), dùng dữ liệu tự soạn thay cho học liệu nội bộ thật để tránh rủi ro thể chế và khó khăn về quyền truy cập dữ liệu. Thực nghiệm chỉ đặt trên mười tiêu chí đọc được trên kịch bản, đại diện cả ba bản chất xác minh, gồm năm tiêu chí kiểm đếm cấu trúc, bốn tiêu chí đối chiếu ngữ nghĩa và một tiêu chí xác minh sự thật (B4); hai tiêu chí xác minh sự thật còn lại (B3, C2) không đưa vào thực nghiệm. Mỗi kịch bản gieo lỗi có chủ đích, 0-7 lỗi/bài, tổng 20 lỗi; L5 không gieo lỗi, dùng để đo mức báo động giả cơ sở. Ma trận 6 bài x 10 tiêu chí = 60 ô; ô L4 với tiêu chí trích dẫn nguồn không áp dụng vì bài này không có mục tài liệu tham khảo, còn 59 ô hợp lệ (20 lỗi, 39 đạt).

3.3. Thiết kế thực nghiệm ứng dụng LLM.

Để kiểm tra khả năng ứng dụng của bộ tiêu chí, nghiên cứu thực hiện một thực nghiệm so sánh trên cùng một đối tượng: cùng một mô hình (Claude Haiku 4.5, truy cập tháng 8 năm 2026) chấm toàn bộ 59 ô theo hai cách viết câu lệnh (prompt). Cách A là chấm trực tiếp: câu lệnh nêu tiêu chí và kịch bản, mô hình đưa ngay kết luận đạt

hoặc không đạt cho từng cặp bài và tiêu chí, không giải thích, mô phỏng kiểu dùng LLM phổ biến nhất. Cách B là neo bằng chứng: câu lệnh buộc mô hình trích nguyên văn phần làm bằng chứng, ra kết luận dựa trên chính đoạn trích đó, và trả lời “không đủ căn cứ” thay vì đoán khi thiếu bằng chứng rõ; lượt trả lời “không đủ căn cứ” trên ô có lỗi được tính là không phát hiện. Mỗi cách chạy độc lập ba lần, tổng cộng 354 lượt chấm. Các chỉ số đo gồm: độ nhạy phát hiện lỗi, tỉ lệ báo động giả, độ lệch chuẩn qua ba lần chạy và tỉ lệ từ chối ở cách B. Toàn bộ kết quả chấm được đối chiếu thủ công với đáp án chuẩn; ba lần chạy của mỗi cách được gộp thành kết luận theo đa số rồi so sánh trên 20 ô có lỗi bằng kiểm định McNemar chính xác hai phía. Ngữ liệu, đáp án chuẩn, nguyên văn câu lệnh và kết quả từng lượt chạy được nhóm tác giả lưu trữ và sẵn sàng cung cấp khi có yêu cầu.

3.4. Hạn chế phương pháp.

Nghiên cứu có 05 hạn chế chính: Thứ nhất, quy mô sáu bài nhân ba lần chạy còn nhỏ, chưa thể khái quát cho toàn bộ kho học liệu của Trường. Thứ hai, lỗi được gieo nhân tạo nên độ nhạy đo được là

mức cận trên so với khả năng phát hiện lỗi phát sinh tự nhiên. Thứ ba, đáp án chuẩn do chính nhóm nghiên cứu lập, chưa qua hội đồng chuyên gia độc lập. Thứ tư, do cỡ mẫu nhỏ, kiểm định McNemar không đạt ý nghĩa thống kê, nên kết quả chỉ mang tính chỉ dẫn. Thứ năm, hiệu năng của LLM có thể thay đổi khi mô hình được cập nhật, vì vậy nhóm ghi lại tên, phiên bản và thời điểm truy cập mô hình để tiện truy vết. Thứ sáu, thực nghiệm chỉ phủ mười trong 37 tiêu chí; 27 tiêu chí còn lại, gồm toàn bộ tiêu chí về âm thanh, hình ảnh, tương tác và đóng gói, chưa được thử.

IV. Kết quả nghiên cứu và thảo luận

4.1. Bộ tiêu chí đánh giá hoàn chỉnh.

Áp dụng quy trình ở mục 3.1, bộ tiêu chí được thể hiện ở Bảng 1 dưới đây, bao gồm 23 tiêu chí kiểm đếm cấu trúc², 11 tiêu chí đối chiếu ngữ nghĩa³, 3 tiêu chí xác minh sự thật⁴. Theo đầu vào bao gồm ba nhóm: đọc được trên kịch bản; gói và định dạng (SCORM, video/âm thanh); âm thanh, hình ảnh, tương tác. Mô hình văn bản chỉ chạm nhóm đầu; hai nhóm còn lại cần công cụ đọc gói, nhận dạng giọng nói, thị giác máy hoặc kiểm thử khi chạy.

Bảng 1. Bộ tiêu chí đánh giá bài giảng điện tử ảnh xạ từ chuẩn trong Quy chế tuyển sinh của Trường Đại học Mở Hà Nội (Hou, 2022)

Mã	Nội dung tiêu chí (đạt/không đạt)	Điều khoản	Bản chất	Đầu vào	Ủy thác
Trục A: Cấu trúc bài giảng					
A1	Video giới thiệu học phần nêu giảng viên, mục đích, bố cục môn học và tài liệu học tập	4.1	CT	T	ĐX
A2	Mỗi bài có phần video giới thiệu gồm tên bài, mục tiêu bài học, nội dung bài học	4.2a	CT	T	TĐ
A3	Phần Nội dung đủ: lý thuyết, minh họa thực tế, câu hỏi vận dụng, tổng hợp	4.2b	NN	T	ĐX

² Kiểm đếm cấu trúc là kiểm tra sự hiện diện hoặc số lượng thành phần

³ Đối chiếu ngữ nghĩa là so khớp ý nghĩa giữa hai phần văn bản

⁴ Xác minh sự thật là đối chiếu với tri thức ngoài văn bản

A4	Phân đánh giá có tối thiểu 5 câu hỏi tương tác	4.2c	CT	T	TD
A5	Có phần Tổng kết gồm tóm lược nội dung và tài liệu tham khảo	4.2d	CT	T	TD
A6	Bài giảng đủ cấu trúc bốn phần theo chuẩn	5.1	CT	T	TD
A7	Thời lượng ≥ 90 phút/tin chỉ, chia ≥ 2 bài, mỗi bài ≤ 45 phút	5.2	CT	G	DX
Trục B: Nội dung học thuật					
B1	Nội dung phù hợp, thống nhất đề cương, giáo trình, đáp ứng chuẩn đầu ra	6.1	NN	T	DX
B2	Có tài liệu tham khảo, ví dụ minh họa, câu hỏi ôn tập, thảo luận, bài tập	6.2	CT	T	TD
B3	Kiến thức chính xác về khoa học, không sai sót	6.2; 10.1b	ST	T	CN
B4	Số liệu, dẫn chứng có căn cứ, không bịa đặt	6.2	ST	T	CN
B5	Nội dung không tự xây dựng có bản quyền, liên kết nguồn gốc	5.3	NN	T	DX
B6	Không vi phạm thuần phong mỹ tục, văn hóa, pháp luật	5.4	NN	T	DX
B7	Ngôn ngữ phù hợp ngôn ngữ giảng dạy của học phần	5.6	CT	T	TD
Trục C: Câu hỏi và đánh giá					
C1	Câu chữ rõ ràng, mạch lạc, đơn nghĩa, đúng chính tả	10.1a	NN	T	DX
C2	Câu hỏi bám giáo trình, không sai kiến thức, bao quát nội dung	10.1b	ST	T	CN
C3	Câu hỏi đánh giá lý thuyết, thực hành, vận dụng, mức đạt chuẩn đầu ra	10.1c	NN	T	DX
C4	Trắc nghiệm 4 phương án, chỉ 1 đáp án đúng	10.3	CT	T	TD
C5	Câu hỏi tự luận có 2-5 câu	10.2	CT	T	TD
C6	Ngân hàng câu hỏi đủ tối thiểu (tự luận 15, trắc nghiệm 50)	10.1d	CT	T	TD
Trục D: Hình thức và nhận diện thương hiệu					
D1	Có logo nhận diện thương hiệu của nhà trường, đoạn phim mở đầu, kết thúc theo mẫu	7.1	CT	V	CN
D2	Theo khuôn mẫu rõ ràng, có mục lục theo dõi	7.2	CT	V, I	CN
D3	Dùng hình ảnh, biểu tượng, biểu đồ, đồ họa thông tin (infographic), video hỗ trợ	7.3	NN	V	CN
D4	Hiển thị tốt trên máy tính, máy tính bảng, điện thoại	7.4	CT	V, I	CN
Trục E: Kỹ thuật đóng gói và tương tác đa phương tiện (richmedia)					
E1	Có kịch bản chi tiết được thẩm định, duyệt, ký xác nhận	9.1	CT	T, G	DX
E2	Đóng gói đúng chuẩn SCORM 1.2, phần mềm có bản quyền	9.2	CT	G	DX
E3	Phân giải $\geq 720p$, khung 16:9, toàn màn hình, menu góc trên phải	9.2; 9.3	CT	V, G	CN
E4	Đủ thanh điều hướng: tiến/lùi, tiến trình, âm lượng, mục lục	9.2	CT	I	CN
E5	Video chuẩn MPEG-4, H.264, tốc độ bit 3-15 Mbps	9.3	CT	G	TD
Trục F: Âm thanh và đồng bộ					
F1	Âm thanh MP3, stereo 2 kênh, nén 64-128 kbps	9.4	CT	G	TD
F2	Không xì nhiễu, tạp âm; lời giảng rõ, không lặp/vấp/sai	9.4; 5.5	CT	A	CN
F3	Âm thanh đồng bộ với tương tác, hoạt cảnh	9.5	NN	A, V, I	CN
F4	Âm thanh, hình ảnh, tương tác đồng bộ nội dung	9.6	NN	A, V, I	CN
F5	Giảng viên trang phục lịch sự, nói rõ ràng, mạch lạc	5.5	NN	A, V	CN
Trục G: Bài giảng dạng văn bản hỗ trợ					
G1	Nội dung văn bản phù hợp bài chính, không khác biệt/dư thừa	8.1	NN	T	DX
G2	20-30 trang A4, Times New Roman 13, giãn dòng 1.5, lề Normal, đánh số trang, lưu doc/docx và PDF	8.2	CT	T, G	TD
G3	Có tổng kết, ví dụ minh họa, bài tập tình huống có gợi ý giải	8.4	CT	T	TD

Chú giải: CT: kiểm đếm cấu trúc; NN: đối chiếu ngữ nghĩa; ST: xác minh sự thật; T: kịch bản; A: âm thanh; V: hình ảnh; I: tương tác; G: gói SCORM; TD: tự động; DX: máy đề xuất/người quyết định; CN: chỉ con người.

4.2. Kết quả ứng dụng LLM trong đánh giá theo bộ tiêu chí.

Kết quả sáu lượt chạy được trình bày ở Bảng 2. Cách B đạt độ nhảy trung bình 96,7% (SD 5,8), cao hơn cách A 13,3 điểm phần trăm (cách A đạt 83,3%, SD 12,6), và độ lệch chuẩn chỉ bằng gần một nửa. Như vậy cách B phát hiện lỗi tốt hơn và dao động ít hơn trong ba lần chạy, dù ba lần chạy chưa đủ để kết luận về độ ổn

định. Bù lại, tỉ lệ báo động giả trung bình của B tăng nhẹ, từ 11,1% lên 14,5%. Một điểm cần nêu là báo động giả của A dao động rất rộng, từ 0,0% đến 17,9% (SD 9,7). Ở lần chạy thứ ba, A báo lỗi dè dặt hơn hẳn, không có báo nhầm nào nhưng bỏ sót 6 trong 20 lỗi thật. Đây là biểu hiện ngưỡng quyết định trôi mà không kiểm soát được, phù hợp với cảnh báo của Pack và cộng sự (2024) và García-Varela và cộng sự (2025).

Bảng 2. Độ nhảy phát hiện lỗi và tỉ lệ báo động giả theo hai cách viết câu lệnh

Cách viết câu lệnh	Lần chạy	Lỗi phát hiện đúng/20	Độ nhảy	Báo động giả/39	Tỉ lệ báo động giả
A (trực tiếp)	1	19/20	95,0%	6/39	15,4%
A (trực tiếp)	2	17/20	85,0%	7/39	17,9%
A (trực tiếp)	3	14/20	70,0%	0/39	0,0%
A trung bình (SD)	-	-	83,3% (SD 12,6)	-	11,1% (SD 9,7)
B (neo bằng chứng)	1	18/20*	90,0%	4/39	10,3%
B (neo bằng chứng)	2	20/20	100,0%	6/39	15,4%
B (neo bằng chứng)	3	20/20	100,0%	7/39	17,9%
B trung bình (SD)	-	-	96,7% (SD 5,8)	-	14,5% (SD 3,9)

Ghi chú: (*) lần chạy 1 của B, mô hình trả lời “không đủ căn cứ” ở 2/20 ô, đều thuộc tiêu chí B4 Hai cách viết câu lệnh mỗi cách chạy ba lần

Khi gộp ba lần chạy thành kết luận theo đa số, A phát hiện 18 trong 20 lỗi (90%), còn B phát hiện đủ 20 trong 20 (100%). Trên 20 ô có lỗi có 2 cặp bất đồng, đều là lỗi B phát hiện còn A bỏ sót; kiểm định McNemar chính xác hai phía cho $p = 0,50$, không đạt ý nghĩa thống kê nên chênh lệch này chỉ mang tính chỉ dấu. Về tính minh bạch, mọi kết luận không từ chối của B đều kèm trích dẫn nguyên văn làm bằng chứng, và khi kiểm tra ngẫu nhiên, các trích dẫn này khớp đúng nguyên văn ngữ liệu; ngược lại, A không để lại dấu vết nào để kiểm chứng độc lập. Cơ chế từ chối của B chỉ kích hoạt ở 2 trong 354 lượt chấm, cả hai đều ở lần chạy thứ nhất và rơi vào tiêu chí B4, một trong ba tiêu chí thuộc nhóm xác minh sự thật ở

Bảng 1 và là tiêu chí duy nhất của nhóm này được thử; ở lần chạy thứ hai và thứ ba, trên chính hai ô này mô hình kết luận không đạt mà không từ chối. Quan sát này phù hợp với cách phân nhóm ở mục 3.1, nhưng hai lượt từ chối không lặp lại chưa đủ để khẳng định cách phân nhóm khớp với hành vi của mô hình; B3 và C2 chưa được thử.

4.3. Khung phân tầng ủy thác đề xuất.

Cột “Ủy thác” ở Bảng 1 là đề xuất suy ra từ bản chất xác minh và dữ liệu đầu vào của mỗi tiêu chí, theo thang tự động hóa ở mục 2.2. Thực nghiệm chỉ đối chứng đề xuất này trên mười tiêu chí đọc được từ kịch bản; mức xếp của 27 tiêu chí còn lại là suy luận lý thuyết. Ba mức

gồm: máy sàng lọc tự động (TĐ), máy đề xuất và người quyết định (ĐX), chỉ con người (CN).

Mức TĐ gồm các tiêu chí kiểm đếm cấu trúc trên kịch bản và các tiêu chí định dạng tệp (E5, F1, G2) kiểm bằng công cụ đọc siêu dữ liệu; mọi kết luận không đạt vẫn chuyển cho thẩm định viên xác nhận. A6 được giữ riêng vì kiểm bài đủ bốn phần theo Điều 5.1, còn A2 đến A5 kiểm thành phần trong từng phần. A1, A7 xếp ĐX vì cần cân nhắc nội dung và thời lượng; E1, E2 xếp ĐX vì máy chỉ kiểm được sự có mặt của kịch bản, trang ký và chuẩn SCORM 1.2, còn chữ ký, thẩm quyền duyệt và bản quyền phần mềm phải do người xác nhận.

Mức ĐX còn gồm các tiêu chí đối chiếu ngữ nghĩa; với cách neo bằng chứng, thẩm định viên chỉ cần đọc đoạn trích để xác nhận hoặc bác. Mức CN gồm ba tiêu chí xác minh sự thật (B3, B4, C2), vì mô hình không có nguồn đối chiếu ngoài văn bản, và các tiêu chí về hình ảnh, âm thanh, tương tác (Trục D, E, F) nằm ngoài dữ liệu mô hình văn bản đọc được.

Khi vận hành, nếu máy kết luận không đạt, chuyên gia thẩm định đối chiếu đoạn trích rồi xác nhận hoặc bác bỏ, mỗi lần bỏ được ghi nhận là một báo động giả. Nếu máy trả lời “không đủ căn cứ”, ô đó chuyển cho người có chuyên môn, không chạy lại. Nếu máy kết luận đạt ở mức TĐ, thẩm định viên kiểm tra xác suất một phần số ô. Kết luận nghiệm thu luôn do chuyên gia thẩm định ký. Chi phí tăng theo số bài, số tiêu chí và công rà lại báo động giả (10-18%); nghiên cứu chưa đo chi phí thực tế. Về bảo mật, Trường cần quy định loại dữ liệu được gửi tới dịch vụ mô hình bên ngoài, loại bỏ thông tin cá nhân hoặc dùng mô hình triển khai nội bộ, và lưu nhật ký để truy vết.

V. Kết luận

Nghiên cứu đã đưa ra bộ 37 tiêu chí trên bảy trục, phân loại theo bản chất xác minh và dữ liệu đầu vào. Khi tiêu chí viết rõ và giao LLM chấm bằng cách neo bằng chứng, độ nhạy phát hiện lỗi cao hơn cách trực tiếp và lưu được bằng chứng kiểm tra. Cơ chế từ chối chỉ xuất hiện ở tiêu chí xác minh sự thật, phù hợp với dự đoán lý thuyết, nhưng số lượt còn quá ít để khẳng định. Nhìn từ góc quản trị, khung này phân vai rõ trong khâu nghiệm thu học liệu, phục vụ mở rộng học liệu số cho hệ thống giáo dục mở. Từ đó, nghiên cứu kiến nghị Trường đưa 37 tiêu chí vào quy trình nghiệm thu làm khung chấm chung, thí điểm LLM neo bằng chứng cho tiêu chí đọc được trên kịch bản kèm phiếu bằng chứng, dùng công cụ chuyên dụng cho phần gói, âm thanh, hình ảnh và tương tác, giữ quyết định cho con người ở tiêu chí xác minh sự thật và ghi nhận mô hình để truy vết. Hướng nghiên cứu tiếp theo là thử trên bài giảng thật với đáp án do chuyên gia độc lập lập, mở rộng sang các tiêu chí chưa thử và đo chi phí vận hành.

Tài liệu tham khảo

- Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, 32(3), 347-364. <https://doi.org/10.1007/BF00138871>
- García-Varela, F., Nussbaum, M., Mendoza, M., Martínez-Troncoso, C., & Bekerman, Z. (2025). ChatGPT as a stable and fair tool for automated essay scoring. *Education Sciences*, 15(8), Article 946. <https://doi.org/10.3390/educsci15080946>
- Grévisse, C. (2024). LLM-based automatic short answer grading in undergraduate medical education. *BMC Medical Education*, 24, Article 1060. <https://doi.org/10.1186/s12909-024-06026-5>

- Hauk, D., & Soujon, N. (2026). How reliable are large language models in analyzing the quality of written lesson plans? A mixed-methods study from a teacher internship program. *Computers and Education: Artificial Intelligence*, 10, Article 100538. <https://doi.org/10.1016/j.caeai.2025.100538>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73. <https://doi.org/10.1111/jedm.12000>
- Kim, Y. (2025). Automated essay scoring with GPT-4 for a local placement test: Investigating prompting strategies, intra-rater reliability, and alignment with human scores. *TESOL Quarterly*, 59(S1). <https://doi.org/10.1002/tesq.3405>
- López-Acosta, D. Y., Pacheco-Mendoza, S. R., Alvarado-Jaramillo, D. J., & Zurita-Flores, P. R. (2026). Quality assurance in generative AI-mediated education: A bibliometric and scoping review. *Frontiers in Artificial Intelligence*, 9, Article 1844517. <https://doi.org/10.3389/frai.2026.1844517>
- Nguyễn Đ. T. (2026). Nâng cao trải nghiệm học tập số của sinh viên thông qua nền tảng học liệu điện tử tại Trường Đại học Mở Hà Nội. *Tạp chí Khoa học Trường Đại học Mở Hà Nội*, 17A, 109-117. <https://doi.org/10.59266/houjs.2026.1354>
- 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, Article 100234. <https://doi.org/10.1016/j.caeai.2024.100234>
- Trường Đại học Mở Hà Nội. (2023). *Quyết định số 5005/QĐ-ĐHM ngày 29 tháng 12 năm 2023 ban hành Quy định về tiêu chuẩn học liệu điện tử và tổ chức sản xuất học liệu điện tử của Trường Đại học Mở Hà Nội*. Tài liệu lưu hành nội bộ.
- Yan, L., Sha, L., Zhao, L., Li, Y., Martínez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1), 90-112. <https://doi.org/10.1111/bjet.13370>
- Zheng, L. M., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36, 46595-46623. <https://doi.org/10.48550/arXiv.2306.05685>

BUILDING AN EVALUATION CRITERIA SET AND APPLYING LARGE LANGUAGE MODELS IN THE QUALITY ASSESSMENT OF E-LECTURES FOR OPEN EDUCATION AT HANOI OPEN UNIVERSITY

Nguyen Thi Khanh Quyen¹, Nguyen Thi Thuy⁵, Chu Minh⁵, Tran Thanh Tung⁵

Abstract: *Hanoi Open University currently assures the quality of its e-lectures mainly through manual review, an approach that is labour-intensive and hard to keep consistent because the standard remains descriptive rather than an operational scoring rubric. This study has two aims: to build a criteria set mapped to Decision 5005/QD-DHM, and to test a large language model (LLM) on the criteria it can handle. Articles 4 to 10 were turned into 37 criteria across seven axes, each tagged with two labels: verification nature and the input data to be checked. The experiment covered the script-readable criteria, using self-authored lectures with seeded errors to compare two prompting methods against ground truth. Evidence-anchored prompting reached a mean sensitivity of 96.7%, above the 83.3% of direct scoring, and left verifiable evidence behind, at the cost of a slightly higher false-alarm rate. Given the small scale of the experiment, these results are preliminary indications only. The study proposes a two-dimensional delegation map for the 37 criteria that keeps human judgement for fact verification and for the audio, visual, and interaction criteria. The criteria set and the delegation map can serve as a reference for the acceptance review of e-learning materials at the University.*

Keywords: *e-learning material standard, e-lecture, evaluation criteria set, large language model, quality assurance*

¹ Hanoi Open University, Hanoi, Vietnam