QUẢN LÝ VIỆC ỨNG DỤNG TRÍ TUỆ NHÂN TẠO TẠO SINH TRONG BIÊN SOẠN CÂU HỎI TRẮC NGHIỆM ĐỌC HIỂU TẠI CƠ SỞ GIÁO DỤC ĐẠI HỌC: NGHIÊN CỨU ĐỊNH TÍNH

Mai Thu Phương
Trường Đại học Ngoại ngữ - Đại học Quốc gia Hà Nội

Tóm tắt: *Sự bùng nổ của Trí tuệ nhân tạo tạo sinh (GenAI) đã mở ra những hướng đi mới trong việc tự động hóa thiết kế các công cụ đánh giá ngôn ngữ. Nghiên cứu định tính này tìm hiểu việc quản lý ứng dụng GenAI trong xây dựng câu hỏi trắc nghiệm (MCQs) kỹ năng Đọc hiểu ở bậc đại học. Dựa trên kết quả phỏng vấn bán cấu trúc với 08 giảng viên tiếng Anh giàu kinh nghiệm trong việc biên soạn đề thi chuẩn hóa, dữ liệu được phân tích bằng phương pháp Phân tích chủ đề (Thematic Analysis). Kết quả cho thấy mặc dù GenAI hỗ trợ hiệu quả trong việc tinh chỉnh độ khó của văn bản và tạo các phương án nhiễu, nhưng việc thiếu hụt các khung giảng dạy AI cũng như cơ chế kiểm soát chất lượng tại các cơ sở giáo dục đang gây ra những rủi ro đối với tính giá trị của bài thi. Nghiên cứu đề xuất mô hình quản lý ba cấp độ dựa trên khung Trí tuệ nhân tạo lấy con người làm trung tâm (HCAI) nhằm đảm bảo tính liêm chính và chuẩn hóa trong đánh giá ngoại ngữ.*

Từ khóa: *GenAI, quản lý giáo dục, câu hỏi trắc nghiệm Đọc hiểu, đảm bảo chất lượng, con người kiểm soát (Human-in-the-loop).*

MANAGING THE APPLICATION OF ARTIFICIAL INTELLIGENCE IN THE CREATION OF READING COMPREHENSION MULTIPLE-CHOICE QUESTIONS IN HIGHER EDUCATION INSTITUTIONS: QUALITATIVE RESEARCH

Abstract: *The proliferation of Generative Artificial Intelligence (GenAI) has pioneered new avenues in automating the design of language assessment tools. This qualitative study explores the management of GenAI applications in developing multiple-choice questions (MCQs) for Reading skills in higher education. Based on semi-structured interviews with 08 experienced English lecturers in standardized test construction, data were analyzed using Thematic Analysis. The findings reveal that while GenAI effectively assists in fine-tuning text readability and generating distractors, the lack of AI pedagogical frameworks and institutional quality control mechanisms poses risks to test validity. The study proposes a three-tier management model grounded in the Human-Centered AI (HCAI) framework to ensure integrity and standardization in foreign language assessment.*

Keywords: *GenAI, educational management, Reading MCQs, quality assurance, Human-in-the-loop.*

Nhận bài: 18/01/2026 *Phản biện: 25/02/2026* *Duyệt đăng: 28/02/2026*

I. ĐẶT VẤN ĐỀ

Các mô hình ngôn ngữ lớn (LLMs) như GPT-4, Llama và Mistral đã trở thành những công cụ đắc lực trong việc phát triển học liệu và hỗ trợ đánh giá trong kỷ nguyên chuyển đổi số. Đặc thù của việc soạn thảo câu hỏi thủ công vốn tiêu tốn nhiều nguồn lực, thời gian và áp lực trí tuệ, nay có thể được giải quyết nhờ khả năng tự động hóa quy trình xây dựng câu hỏi trắc nghiệm (MCQs) của GenAI. Một nghiên cứu thực nghiệm của Biancini và cộng sự (2024) chỉ ra rằng, trong khi việc biên soạn MCQs truyền thống đòi hỏi nguồn lực khổng lồ, các mô hình LLM đã chứng minh khả năng tạo ra các câu hỏi có tính thách thức và giá trị sư phạm cao, giúp tinh giản quy trình này một cách đáng kể. Đối với các cơ sở giáo dục đại học, GenAI không chỉ đóng vai trò trợ lý giảm tải khối lượng công việc cho giảng viên mà còn tiềm tàng khả năng tạo ra các học liệu chất lượng cao. Đáng chú ý, kết quả so sánh cặp cho thấy các chuyên gia nội dung (SMEs) có xu hướng ưu tiên các câu hỏi do AI tạo ra hơn là do con người soạn thảo, đặc biệt trong các lĩnh vực chuyên biệt như Phân tích dữ liệu khoa học và Cơ sở thống kê (Kowal và cộng sự, 2025).

Tuy nhiên, thách thức trọng tâm vẫn nằm ở rủi ro về tính chính xác và sự thiếu hụt một khung pháp lý chính thức. Việc giảng viên sử dụng AI một cách tự phát có thể đối mặt với hiện tượng ảo giác (Hallucinations), dẫn đến các sai sót về kiến thức chuyên môn. Để giảm thiểu rủi ro này, các nghiên cứu mới đề xuất chuyển dịch từ việc phụ thuộc hoàn toàn vào kiến thức nội tại của mô hình sang phương pháp nạp kiến thức (Knowledge injection) thông qua kỹ thuật đặt câu lệnh (Prompting), cho phép giảng viên duy trì quyền kiểm soát tuyệt đối đối với văn bản nguồn của bài thi (Biancini và cộng sự, 2024). Thêm vào đó, dù AI có thể tạo câu hỏi nhanh chóng, việc thẩm định và tinh chỉnh bởi con người vẫn là yếu tố then chốt để đảm bảo các tiêu chuẩn đo lường tâm lý học (Psychometric standards) và tính công bằng, tránh tạo ra các câu hỏi mơ hồ hoặc có nhiều đáp án đúng ngoài ý muốn.

Tại Việt Nam, các cơ sở giáo dục đại học đang đứng trước một câu hỏi cấp thiết: làm thế nào để khai thác sức mạnh của GenAI mà vẫn duy trì được các tiêu chuẩn đánh giá nghiêm ngặt? Tổng quan các nghiên cứu cho thấy hiện vẫn tồn tại những khoảng trống tri thức đáng kể trong bối cảnh giáo dục đại học tại Việt Nam. Thứ nhất, các nghiên cứu hiện tại chủ yếu tập trung vào năng lực giải đề hoặc hiệu quả kỹ thuật của AI, mà thiếu đi góc nhìn về quản lý quy trình. Thứ hai, có sự đứt gãy giữa việc giảng viên sử dụng AI một cách tự phát và các quy định chính thức về Đảm bảo chất lượng của cơ sở đào tạo. Sự thiếu hụt trong quản lý chuẩn hóa này có nguy cơ làm tổn hại đến độ tin cậy của các ngân hàng câu hỏi thi. Xuất phát từ thực tiễn trên, bài viết này tập trung khảo sát nhận thức của giảng viên và đề xuất một mô hình quản lý thực tiễn dựa trên triết lý con người làm chủ (Humans in the lead). Cụ thể, nghiên cứu tập trung trả lời hai câu hỏi nghiên cứu sau:

- Câu hỏi nghiên cứu 1: Giảng viên tiếng Anh nhận thức và thực hiện vai trò con người kiểm soát (Human-in-the-loop) như thế nào khi sử dụng GenAI để thiết kế câu hỏi trắc nghiệm Đọc hiểu?

- Câu hỏi nghiên cứu 2: Các cơ sở giáo dục đại học cần thiết lập cơ chế quản lý ba cấp độ như thế nào để chuẩn hóa việc ứng dụng GenAI trong công tác khảo thí?

II. NỘI DUNG NGHIÊN CỨU

2.1. Tổng quan nghiên cứu

Những năm gần đây, sự giao thoa giữa AI và năng lực sư phạm của con người đã trở thành tâm điểm của các nghiên cứu thực nghiệm.

Về hiệu quả và chất lượng, Moore và cộng sự (2023) nhận thấy AI giúp giảm thời gian soạn thảo từ 50-70%, mặc dù chất lượng của các phương án nhiễu phụ thuộc rất lớn vào kỹ thuật đặt câu lệnh (Prompt engineering). Một nghiên cứu mang tính bước ngoặt sử dụng phương pháp so sánh cặp đã chỉ ra rằng, các chuyên gia nội dung (SMEs) thường có xu hướng ưu tiên một cách có ý nghĩa thống kê các câu hỏi do LLM tạo ra so với các câu hỏi do con người soạn thảo, đặc biệt là trong các lĩnh vực kỹ thuật như Cơ sở Thống kê và Phân tích Dữ liệu Khoa học (Kowal và cộng sự, 2025). Điều này cho thấy GenAI đã đạt đến mức độ trưởng thành về đo lường tâm lý học (Psychometric maturity) tương đương hoặc thậm chí vượt qua phương pháp soạn thảo thủ công truyền thống.

Về nhận thức của giảng viên và các thách thức, bất chấp chất lượng đầu ra của AI, sự mâu thuẫn nhận thức (Cognitive dissonance) vẫn tồn tại trong cộng đồng giáo dục (Kasneci và cộng sự, 2023). Các nhà giáo dục đánh giá cao khả năng gợi mở ý tưởng của AI nhưng đồng thời lo ngại về việc mất kiểm soát sư phạm và nguy cơ xuất hiện nội dung ảo giác (Bender và cộng sự, 2021). Sự thiếu hụt niềm tin này thường bắt nguồn từ việc vắng bóng một khung quản lý cấu trúc nhằm đảm bảo tính giá trị khoa học (Moreno và cộng sự, 2014).

Về quy trình quản lý, mô hình tạo câu hỏi tự động (Automatic item generation) truyền thống đang dần được thay thế bởi quy trình con người kiểm soát (Human-in-the-loop). Biancini và cộng sự (2025) nhấn mạnh rằng hình thức tương tác hiệu quả nhất là nạp kiến thức (Knowledge injection), trong đó giảng viên cung cấp văn bản nguồn để đảm bảo đầu ra của AI có căn cứ và có thể kiểm soát được. Điều này nhất quán với quy trình 4 bước do Liu và cộng sự (2023) đề xuất: (1) Thiết kế ma trận bởi con người; (2) AI soạn thảo bản thảo; (3) Con người tinh chỉnh; và (4) Hội đồng độc lập phê duyệt. Mô hình cộng tác này xác định vị thế của giảng viên không phải là người đứng ngoài cuộc, mà là người kiểm soát chất lượng chiến lược, đảm bảo mỗi câu hỏi đều tuân thủ các đặc tính về độ chuẩn xác và sự điều chỉnh phù hợp.

2.2. Cơ sở lý luận

2.2.1. Các yêu cầu về tính giá trị trong xây dựng câu hỏi trắc nghiệm đọc hiểu

Trong đánh giá ngôn ngữ, thiết kế câu hỏi trắc nghiệm (MCQs) cho kỹ năng Đọc hiểu là một quy trình kỹ thuật nghiêm ngặt nhằm đo lường việc xử lý văn bản ở nhiều cấp độ nhận thức khác nhau: từ truy xuất thông tin hiển ngôn, hiểu ý chính đến suy luận logic phức tạp. Theo khung giá trị khoa học do Moreno và cộng sự (2014) đề xuất, chất lượng của một công cụ MCQ được chi phối bởi ba đặc tính cơ bản: (1) Tính tương thích (Adjustment - tính đại diện của nội dung), (2) Tính chính xác (Precision - sự rõ ràng, không mơ hồ) và (3) Tính phân biệt (Differentiation - sự loại trừ lẫn nhau của các phương án). Theo các tiêu chuẩn này, hiệu quả của một câu hỏi MCQ phụ thuộc chủ yếu vào tính hợp lý của các phương án nhiễu (Distractors). Phương án nhiễu không được là những lựa chọn ngẫu nhiên hay tầm thường; thay vào đó, chúng phải được thiết kế một cách hệ thống để phản ánh các lỗi sai phổ biến hoặc các ngộ nhận về nhận thức của người học (Haladyna và cộng sự, 2002; Moreno và cộng sự, 2014). Mặc dù các mô hình

ngôn ngữ lớn (LLMs) hiện nay có thể tự động hóa quy trình tạo câu hỏi này để giảm bớt gánh nặng về thời gian và trí tuệ, sự giám sát của con người vẫn là nhân tố quyết định trong việc đảm bảo tính giá trị khoa học.

2.2.2. Từ AIG truyền thống đến sự trỗi dậy của GenAI

Trước đây, việc Tạo câu hỏi tự động (AIG) chủ yếu dựa vào các mô hình nhận thức và các thuật toán được lập trình cứng (Gierl & Lai, 2013). Tuy nhiên, AI tạo sinh (GenAI) như ChatGPT hay Gemini đã hoàn toàn thay đổi cục diện nhờ khả năng tạo lập ngôn ngữ tự nhiên và thiết kế câu hỏi dựa trên các ngữ cảnh ngôn ngữ phức tạp mà không cần các biểu mẫu cứng nhắc. Mặc dù vậy, thách thức chính vẫn là hiện tượng ảo giác (Hallucination), khi các mô hình tạo ra thông tin sai lệch về mặt học thuật nhưng lại có vẻ rất logic (Bender và cộng sự, 2021). Để khắc phục điều này, các phương pháp tiếp cận hiện đại đề xuất việc nạp kiến thức (Knowledge injection) vào các câu lệnh, đảm bảo giảng viên duy trì quyền kiểm soát đối với văn bản nguồn của bài thi (Biancini và cộng sự, 2024).

2.2.3. Khung lý thuyết Trí tuệ nhân tạo lấy con người làm trung tâm (HCAI)

Nghiên cứu này áp dụng khung lý thuyết Trí tuệ nhân tạo lấy con người làm trung tâm (HCAI) do Shneiderman (2020) đề xuất, một sự chuyển dịch mô hình được mô tả như là "Cuộc cách mạng Copernicus thứ hai" trong thiết kế hệ thống. Khác với các cách tiếp cận AI truyền thống nhằm tạo ra các tác nhân tự trị mô phỏng hành vi con người, HCAI định nghĩa lại AI như một tập hợp các công cụ mạnh mẽ nhằm tăng cường năng lực tự chủ của con người, khuyến khích sự sáng tạo và làm rõ trách nhiệm (Shneiderman, 2020). Trong bối cảnh quản lý đánh giá, khung lý thuyết này giới thiệu mô hình HCAI hai chiều, thách thức quan điểm truyền thống về sự đánh đổi giữa tự động hóa và kiểm soát. Mô hình chứng minh rằng chúng ta hoàn toàn có thể đạt được mức độ tự động hóa máy tính cao để tối ưu hiệu suất tạo câu hỏi, đồng thời duy trì quyền kiểm soát cao của con người để đảm bảo chất lượng sư phạm và các tiêu chuẩn đạo đức (Shneiderman, 2020). Bằng cách đặt con người vào vị trí trung tâm, khung lý thuyết này đảm bảo rằng AI xoay quanh giảng viên, người giữ thẩm quyền cuối cùng trong quy trình ra quyết định. Để vận hành lý thuyết này trong môi trường tổ chức, nghiên cứu tích hợp cấu trúc quản trị ba cấp độ của Shneiderman:

Cấp độ nhóm (Hệ thống tin cậy): Sử dụng các thực hành kỹ thuật đã được kiểm chứng, chẳng hạn như nạp kiến thức, để xây dựng quy trình tạo câu hỏi đáng tin cậy và có thể dự đoán được.

Cấp độ tổ chức (Văn hóa an toàn): Thiết lập văn hóa quản lý, trong đó sự giám sát của khoa/phòng ban đảm bảo tính an toàn và công bằng của các bài đánh giá do AI tạo ra.

Cấp độ ngành (Chứng nhận đáng tin cậy): Kết nối với các tiêu chuẩn kiểm định và chứng nhận độc lập để thúc đẩy sự tin cậy của tổ chức (Shneiderman, 2020).

Bằng cách chuyển trọng tâm từ việc mô phỏng con người sang việc trao quyền cho con người, khung HCAI cung cấp nền tảng vững chắc cho quy trình làm việc cộng tác Người-AI. Điều này đảm bảo rằng AI tạo sinh đóng vai trò là một công cụ hiệu suất cao dưới sự giám sát chiến lược của giảng viên, đáp ứng các kỳ vọng cốt lõi về sự tham gia xã hội và tính liêm chính trong học thuật.

2.3. Phương pháp nghiên cứu

2.3.1. Thiết kế nghiên cứu

Nghiên cứu vận dụng phương pháp định tính khám phá (Exploratory qualitative). Cách tiếp cận này cho phép tác giả đi sâu tìm hiểu nhận thức và thực hành của giảng viên trong một lĩnh vực mới nổi là ứng dụng GenAI vào khảo thí, nơi các biến số chưa được xác định rõ ràng (Creswell, 2014).

2.3.2. Khách thể nghiên cứu

Thông qua phương pháp chọn mẫu có chủ đích (Purposive sampling), 08 giảng viên tiếng Anh (GV1-GV8) đang công tác tại các cơ sở giáo dục đại học ở Hà Nội đã được mời tham gia. Tiêu chí lựa chọn bao gồm: (1) Có trên 03 năm kinh nghiệm biên soạn MCQ Đọc hiểu; (2) Trực tiếp tham gia xây dựng ngân hàng câu hỏi chuẩn hóa; (3) Đã từng sử dụng ít nhất một công cụ GenAI (ChatGPT, Gemini, Claude) trong công việc. Quy mô mẫu này đảm bảo đạt đến điểm bão hòa dữ liệu (Data saturation) khi các chủ đề nghiên cứu bắt đầu lặp lại và không xuất hiện mã mới (Guest et al., 2006).

2.3.3. Thu thập và phân tích dữ liệu

- Thu thập: Sử dụng phỏng vấn bán cấu trúc (Semi-structured interview) kéo dài từ 45-60 phút. Nội dung tập trung vào: quy trình tích hợp AI, kỹ thuật kiểm soát lỗi gây ảo giác, và các rào cản quản lý thực tế.

- Phân tích: Dữ liệu được xử lý bằng phương pháp Phân tích chủ đề (Thematic analysis) theo quy trình 06 bước của Braun và Clarke (2006).

Quá trình mã hóa được thực hiện theo lối quy nạp, giúp các chủ đề nảy sinh khách quan từ trải nghiệm của giảng viên.

2.4. Kết quả và thảo luận

2.4.1. Chủ đề 1: Sự cộng tác giữa năng lực sư phạm và hiệu suất của GenAI

Chủ đề này tập trung vào cách thức giảng viên sử dụng AI như một công cụ hỗ trợ thay vì thay thế. Dữ liệu cho thấy một quy trình lai đang hình thành một cách tự phát.

Tối ưu hóa ngữ liệu và độ khó văn bản: Việc tìm kiếm một văn bản đọc hiểu vừa đảm bảo tính xác thực, vừa khớp với khung năng lực ngoại ngữ (CEFR) là một thách thức lớn. GV2 chia sẻ: *"Trước đây, để biên tập một bài báo từ BBC về mức B1, tôi phải mất hàng giờ để thay thế các cấu trúc ngữ pháp phức tạp và từ vựng C1. Với ChatGPT, tôi chỉ cần dùng lệnh "Simplify this text to B1 level using Oxford 3000 keywords". Kết quả đạt khoảng 80% yêu cầu, tôi chỉ cần chỉnh sửa lại các liên từ để đảm bảo tính mạch lạc"*. Điều này cho thấy GenAI đã giải quyết được nút thắt về mặt thời gian trong khâu chuẩn bị tài liệu, một phần quan trọng của quản lý nguồn lực trong khảo thí.

Thiết kế phương án nhiễu dựa trên lỗi tư duy: Một câu hỏi MCQ chất lượng phụ thuộc vào các phương án nhiễu. GV4 nhận định: "*AI rất giỏi trong việc tìm các từ đồng nghĩa hoặc các chi tiết gây nhiễu trong bài. Khi tôi yêu cầu AI "Tạo 3 phương án sai dựa trên việc hiểu nhầm nghĩa của từ X trong đoạn 2",* nó đưa ra các lựa chọn rất sát với lỗi thực tế của sinh viên". Việc này giúp tăng cường độ phân hóa của câu hỏi mà không đòi hỏi giảng viên phải suy nghĩ quá nhiều kịch bản gây nhiễu thủ công.

Thảo luận: Kết quả này ủng hộ quan điểm của Gierl & Lai (2013) về việc tự động hóa tạo mục hỏi giúp tăng hiệu suất và Kowal et al. (2025) về độ chính xác của các phần mềm AI, nhưng đồng thời nhấn mạnh vai trò điều phối của giảng viên. Tuy nhiên, sự phụ thuộc vào AI trong khâu này cũng đặt ra câu hỏi về tính nguyên bản của đề thi, một khía cạnh mà các nhà quản lý giáo dục cần lưu tâm khi xây dựng quy định về liêm chính học thuật.

2.4.2. Chủ đề 2: Rào cản về ảo giác AI và sự suy giảm độ giá trị nội dung

Bên cạnh những ưu điểm, chủ đề thứ hai làm nổi bật những rủi ro kỹ thuật mà AI gây ra, đòi hỏi cơ chế kiểm soát nghiêm ngặt.

Hiện tượng "hallucination" và lỗi logic bắc cầu: Lỗi nghiêm trọng nhất được ghi nhận là việc AI tự tạo ra thông tin không có trong văn bản để làm đáp án hoặc phương án nhiễu. GV5 thuật lại một trải nghiệm: *"Trong một bài đọc về biến đổi khí hậu, AI đã tự đưa ra một con số thống kê rất thuyết phục về nhiệt độ trái đất, nhưng khi tôi rà soát lại văn bản gốc thì hoàn toàn không có số liệu đó. Nếu một giảng viên thiếu kinh nghiệm hoặc đang chịu áp lực thời gian (như làm việc từ 8 giờ sáng đến 10 giờ đêm) mà bỏ qua bước đối soát này, độ tin cậy của bài thi sẽ bằng không".*

Tính đơn điệu và hạn chế ở tư duy bậc cao (HOTS): Dữ liệu chỉ ra rằng AI có xu hướng tạo ra các câu hỏi ở mức độ nhận biết và thông hiểu. GV7 nhận xét*: "AI rất khó thiết kế được những câu hỏi yêu cầu suy luận ẩn ý hoặc nhận diện thái độ của tác giả. Các câu hỏi của nó thường mang tính chất nhặt nhạnh thông tin bề mặt"*. Điều này tạo ra một rủi ro quản lý: nếu ngân hàng câu hỏi quá phụ thuộc vào AI, chuẩn đầu ra về tư duy phản biện của người học sẽ không được đánh giá chính xác.

Thảo luận: Sự tồn tại của ảo giác AI minh chứng cho cảnh báo của Bender et al. (2021) về các vẹt ngôn ngữ (stochastic parrots). Trong quản lý khảo thí, điều này đặt ra yêu cầu cấp thiết về việc thiết lập các chỉ số kiểm soát chất lượng riêng biệt cho các mục hỏi có sự hỗ trợ của AI, thay vì dùng chung quy trình kiểm định đề thi truyền thống.

2.4.3. Chủ đề 3: Sự đứt gãy giữa thực hành tự phát và cơ chế quản lý thiết chế

Chủ đề cuối cùng tập trung vào góc độ quản lý giáo dục, chỉ ra sự thiếu hụt hành lang pháp lý tại các cơ sở đào tạo.

Sự thiếu minh bạch và áp lực dùng chui: Do chưa có quy định chính thức, nhiều giảng viên sử dụng AI nhưng không dám công khai trong tờ trình ra đề. GV6 chia sẻ: *"Tôi cảm thấy mình đang làm điều gì đó sai trái dù thực tế nó giúp tôi làm việc hiệu quả hơn. Nhà trường chưa nói rõ chúng tôi được dùng bao nhiêu phần trăm AI trong một đề thi"*. Sự thiếu minh bạch này gây khó khăn cho phòng Đảm bảo chất lượng trong việc truy xuất nguồn gốc và thẩm định tính công bằng của đề thi.

Nhu cầu về khung năng lực AI cho giảng viên: Hầu hết giảng viên tham gia phỏng vấn đều tự học cách dùng AI. GV1 đề xuất: *"Nhà trường cần có những buổi workshop về "Prompt Engineering" dành riêng cho khảo thí. Hiện tại mỗi người dùng một kiểu, không có sự chuẩn hóa trong cách ra*

lệnh cho AI, dẫn đến chất lượng ngân hàng câu hỏi không đồng đều".

Trách nhiệm giải trình: Một vấn đề quản lý hóc búa được đặt ra là khi có sự cố sai sót trong đề thi. GV8 đặt câu hỏi: *"Nếu một câu hỏi sai do AI tạo ra lọt qua được các vòng thẩm định, ai sẽ là người chịu trách nhiệm chính? Giảng viên, Hội đồng phản biện, hay chúng ta đổ lỗi cho công nghệ?"* Đây là khoảng trống lớn trong quy trình quản lý rủi ro hiện nay tại các trường đại học.

2.4.4. Thảo luận chung: Hướng tới mô hình Human-in-the-loop (HCAI)

Tổng hợp các kết quả trên, có thể thấy việc ứng dụng GenAI trong biên soạn MCQ Đọc hiểu không phải là một vấn đề kỹ thuật đơn thuần, mà là một bài toán quản lý sự thay đổi. Dựa trên khung lý thuyết của Shneiderman (2020), chúng tôi cho rằng để giải quyết các rào cản ở Chủ đề 2 và 3, các cơ sở giáo dục cần chuyển dịch sang mô hình Trí tuệ nhân tạo lấy con người làm trung tâm (HCAI). Trong mô hình này, mức độ tự động hóa của AI có thể rất cao, nhưng mức độ kiểm soát của con người cũng phải rất cao . Việc giảng viên làm việc từ 8 giờ sáng đến 10 giờ đêm như thực trạng hiện nay là một biến số nguy hiểm trong mô hình này. Khi con người kiệt sức, vai trò người gác cổng sẽ trở nên lỏng lẻo, tạo điều kiện cho các lỗi "ảo giác AI" lọt vào ngân hàng đề thi. Do đó, chính sách quản lý không chỉ dừng lại ở việc ban hành quy định, mà còn phải bao gồm việc tái cấu trúc khối lượng công việc để giảng viên có đủ không gian tinh thần thực hiện vai trò thẩm định chất lượng.

III. KẾT LUẬN

Nghiên cứu khẳng định việc tích hợp GenAI vào biên soạn câu hỏi MCQ Đọc hiểu là một xu hướng tất yếu, giúp tối ưu hóa hiệu suất làm việc cho giảng viên trong bối cảnh áp lực công việc ngày càng tăng cao. Tuy nhiên, ranh giới giữa việc sử dụng AI như một trợ lý sáng tạo và sự phụ thuộc dẫn đến sai lệch học thuật vẫn còn mong manh. Kết quả phỏng vấn cho thấy, mặc dù GenAI xuất sắc trong việc biến đổi ngữ liệu và tạo phương án nhiễu, nó vẫn bộc lộ những lỗ hổng về logic suy luận và tính xác thực. Khoảng trống lớn nhất hiện nay không nằm ở công nghệ, mà ở sự thiếu hụt các khung quản lý thiết chế và hướng dẫn đạo đức nghề nghiệp. Mô hình "Human-in-the-loop" là lựa chọn tối ưu, trong đó trí tuệ nhân tạo đóng vai trò thực thi, còn trí tuệ con người giữ quyền quyết định và thẩm định chất lượng cuối cùng.

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.

Biancini, A., et al. (2025). Knowledge injection and human-in-the-loop: Managing Generative AI in educational assessment. *Journal of Educational Computing Research. (In press).*

Kasneci, E., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274.

Kowal, M., et al. (2025). Expert preferences in Automated Item Generation: A paired comparison study of LLM vs. Human authors. *Assessment in Education: Principles, Policy & Practice.*

Liu, Z. (2025). Human-AI Co-Creation: A Framework for Collaborative Design in Intelligent Systems. *In Proceedings of the 16th International Conference on Applied Human Factors and Ergonomics (AHFE 2025). AHFE Open Access.*