VNU Journal of Foreign Studies

Journal homepage: https://jfs.ulis.vnu.edu.vn/

VERIFICATION NORMS AND GENERATIVE AI USE: A SURVEY OF ENGLISH LANGUAGE UNDERGRADUATES

Tran Thi Hoai Nam[1,*], Phan Van Hoa[2]

[1]*FPT University, FPT Polytechnic, Da Nang Campus,*
FPT Technology Urban Area, Da Nang, Ngu Hanh Son Ward, Da Nang, Vietnam
[2]*University of Foreign Language Studies, The University of Danang,*
No. 131 Luong Nhu Hoc Street, Cam Le Ward, Da Nang, Vietnam

Received 05 June 2026
Revised 23 September 2026; Accepted 23 October 2026

Abstract: This cross-sectional survey examined generative AI use and views on output verification among English Language undergraduates at the University of Foreign Language Studies, the University of Danang. The analyses used closed-ended items and responses to four open-ended questions. Of 210 responses, 159 passed both attention checks. In the retained sample, 62.3% reported using ChatGPT daily and 89.3% at least weekly. AI use was most commonly reported for writing, grammar and vocabulary, and speaking. Among 158 cases with complete data, 41 (25.9%) endorsed a requirement for teachers to verify AI output and also reported using AI output themselves without checking it. Without attention-check exclusions, the corresponding proportion was 56/209 (26.8%). The two items had different referents, so their joint endorsement did not establish an inconsistency between students' beliefs and behaviour. Epistemic asymmetry was therefore treated as a tentative interpretation. Scores for general readiness to use AI in teaching were higher than those for confidence in pedagogical prompt design. Self-reported confidence in identifying AI hallucinations and cultural bias was positively associated with confidence in prompt design. In addition, 93.7% reported no formal AI-related content in their programme. The findings described self-reported perceptions and practices at one institution, and the cross-sectional design did not permit causal inference.

Keywords: generative AI, English-language practice, verification norms, epistemic asymmetry, prompt design

* Corresponding author.
Email address: tran.hoainam1111@gmail.com
https://doi.org/10.63023/2525-2445/jfs.ulis.5894

CHUẨN MỰC KIỂM CHỨNG VÀ VIỆC SỬ DỤNG AI TẠO SINH: KHẢO SÁT SINH VIÊN NGÀNH NGÔN NGỮ ANH

Trần Thị Hoài Nam[1], Phan Văn Hòa[2]

[1]Trường Đại học FPT, Trường Cao đẳng FPT, Đà Nẵng,
Khu đô thị công nghệ FPT Đà Nẵng, Phường Ngũ Hành Sơn, Đà Nẵng, Việt Nam
[2]Trường Đại học Ngoại ngữ, Đại học Đà Nẵng, Số 131 Lương Nhữ Hộc, Phường Cẩm Lệ, Đà Nẵng, Việt Nam

Nhận bài ngày 05 tháng 6 năm 2026
Chỉnh sửa ngày 23 tháng 9 năm 2026; Chấp nhận đăng ngày 23 tháng 10 năm 2026

Tóm tắt: Khảo sát cắt ngang này tìm hiểu việc dùng AI tạo sinh và quan niệm về kiểm chứng đầu ra của AI ở sinh viên ngành Ngôn ngữ Anh, Trường Đại học Ngoại ngữ, Đại học Đà Nẵng. Dữ liệu gồm các mục đóng và bốn câu hỏi mở; 159/210 phản hồi đạt cả hai mục kiểm tra mức độ chú ý. Có 62,3% sinh viên dùng ChatGPT hằng ngày và 89,3% dùng ít nhất hằng tuần. AI được dùng nhiều nhất cho viết, ngữ pháp, từ vựng và nói. Trong 158 trường hợp đủ dữ liệu, 41 người (25,9%) vừa tán thành yêu cầu giáo viên kiểm chứng đầu ra, vừa tự nhận dùng đầu ra không kiểm tra lại. Khi không sàng lọc, tỷ lệ là 56/209 (26,8%). Do hai mục khác đối tượng quy chiếu, kết quả chưa đủ để kết luận có sự bất nhất giữa niềm tin và hành vi. Vì vậy, bất đối xứng nhận thức luận chỉ là gợi ý diễn giải. Điểm sẵn sàng dùng AI dạy học cao hơn điểm tự tin thiết kế câu lệnh sư phạm. Mức tự tin nhận diện ảo giác AI và thiên kiến văn hóa có tương quan thuận với mức tự tin thiết kế câu lệnh. Có 93,7% sinh viên cho biết chưa học nội dung AI chính thức. Kết quả dựa trên dữ liệu tự báo cáo tại một cơ sở và thiết kế cắt ngang không cho phép suy luận nhân quả.

Từ khóa: AI tạo sinh, thực hành tiếng Anh, chuẩn mực kiểm chứng, bất đối xứng nhận thức luận, thiết kế câu lệnh

1. Giới thiệu

Sinh viên đại học không chỉ thực hành tiếng Anh trong lớp học hay trong các hoạt động thuộc chương trình đào tạo chính thức. Sự phổ biến của AI tạo sinh mang lại cho người học thêm một nguồn hỗ trợ ngoài giờ lên lớp: họ có thể đặt câu hỏi, nhận phản hồi và nhờ chỉnh sửa ngôn ngữ (Xiao & Zhi, 2023). Góc nhìn sinh thái của van Lier (2004) chú trọng mối quan hệ giữa người học và môi trường trong quá trình học ngôn ngữ. Theo góc nhìn này, bài viết xem AI tạo sinh là một nguồn lực trong môi trường thực hành tiếng Anh của sinh viên. Ngoài tần suất và mục đích sử dụng, nghiên cứu còn tìm hiểu quan niệm của sinh viên về việc kiểm chứng đầu ra của AI, vì riêng tần suất sử dụng chưa thể cho biết người học kiểm tra thông tin ra sao.

Nghiên cứu này được thực hiện với đối tượng là sinh viên ngành Ngôn ngữ Anh tại Trường Đại học Ngoại ngữ, Đại học Đà Nẵng (UFLS-UD). Nghiên cứu tiến hành so sánh điểm sẵn sàng sử dụng AI trong dạy học với điểm tự tin thiết kế câu lệnh sư phạm, đồng thời xem xét mối liên hệ giữa mức tự tin nhận diện ảo giác AI và thiên kiến văn hóa với mức tự tin thiết kế câu lệnh. Đây là mức tự tin do sinh viên tự đánh giá, không phải năng lực được đo trực tiếp. Về kiểm chứng đầu ra, nghiên cứu xác định tỷ lệ sinh viên vừa tán thành với mục E2 trong bảng hỏi ("Giáo viên nên kiểm tra và điều chỉnh đầu ra của AI trước khi sử dụng, thay vì dùng trực tiếp"), vừa đồng ý với mục E9 ("Tôi thường tin tưởng và sử dụng trực tiếp đầu ra của AI mà không kiểm tra lại"). Hai phát biểu này có đối tượng quy chiếu khác nhau: E2 nói về yêu cầu đối với giáo viên, còn E9 nói về cách chính sinh viên sử dụng đầu ra của AI. Do đó, việc đồng

ý với cả hai chưa đủ để xác lập sự bất nhất giữa niềm tin và hành vi của người trả lời, nhất là khi nghiên cứu không có dữ liệu quan sát trực tiếp hành vi kiểm chứng. Khái niệm bất đối xứng nhận thức luận vì thế chỉ được bàn đến ở phần Thảo luận như một gợi ý diễn giải, không phải một cấu trúc tâm lý được đo lường.

2. Cơ sở lý thuyết và tổng quan nghiên cứu

2.1. Hệ sinh thái ngôn ngữ và vị trí của AI tạo sinh

Theo góc nhìn sinh thái, việc học ngôn ngữ diễn ra trong mối quan hệ giữa người học và môi trường tương tác của họ (van Lier, 2004). Trong bài viết này, khái niệm cơ hội hành động (*affordances*) được hiểu là những khả năng sử dụng ngôn ngữ mà người học nhận ra và tận dụng được trong môi trường ấy. Hiểu theo cách đó, AI tạo sinh có thể được xem là một nguồn lực trong môi trường thực hành tiếng Anh của người học.

Nguồn lực này thể hiện rõ nhất ở khả năng duy trì tương tác bằng ngôn ngữ. Khi dùng ChatGPT, người học có thể đặt câu hỏi, nhận phản hồi rồi điều chỉnh câu lệnh hoặc cách diễn đạt của mình để tiếp tục tương tác. Các quan điểm dựa trên sử dụng (usage-based) cho rằng tần suất và đặc điểm của ngôn ngữ mà người học tiếp xúc có liên quan đến quá trình học ngôn ngữ (Ellis, 2002, 2008). Theo hướng này, tương tác với AI vì thế có thể tạo thêm cơ hội tiếp xúc và sử dụng tiếng Anh. Dù vậy, bản thân tần suất sử dụng không thể cho biết người học tham gia vào các tương tác ấy như thế nào hay học được gì từ đó.

Vấn đề này cũng gắn với cách lý thuyết văn hóa - xã hội nhìn nhận vai trò của công cụ trong hoạt động học tập. Theo Vygotsky (1978) cũng như Lantolf và Thorne (2006), công cụ giữ vai trò trung gian trong hoạt động của người học, nên việc có sẵn một công cụ chưa đủ để suy ra kết quả học tập; điều cần xem xét là người học sử dụng công cụ và tham gia vào hoạt động với công cụ ấy ra sao. Khảo sát này không đo trực tiếp mức độ tham gia nhận thức trong từng lần tương tác, cũng không đo sự thay đổi năng lực tiếng Anh theo thời gian.

2.2. Kiểm chứng đầu ra AI và cách diễn giải bất đối xứng nhận thức luận

Qua hình ảnh "con vẹt ngẫu nhiên" (*stochastic parrot*), Bender và cộng sự (2021) lưu ý rằng văn bản do mô hình ngôn ngữ tạo ra có thể trôi chảy và mạch lạc, nhưng những đặc điểm hình thức ấy không bảo đảm ý nghĩa hay độ chính xác của nội dung. Đối với người dùng, điều này cho thấy sự trôi chảy của đầu ra chưa phải là căn cứ đủ để coi thông tin là đáng tin cậy.

Các nghiên cứu về tính dễ xử lý thông tin (*processing fluency*) giúp lý giải hiện tượng này: cảm giác dễ xử lý một thông tin có thể ảnh hưởng đến phán đoán về tính đúng của thông tin đó (Alter & Oppenheimer, 2009; Reber & Unkelbach, 2010). Với đầu ra của AI, sự mạch lạc của văn bản có thể ảnh hưởng đến cách người đọc đánh giá độ tin cậy. Nghiên cứu này không đo trực tiếp tính dễ xử lý hay mức độ thuyết phục mà người học cảm nhận từ đầu ra của AI.

Từ những cơ sở trên, khái niệm bất đối xứng nhận thức luận (*epistemic asymmetry*) được dùng như một gợi ý để diễn giải cho việc hai loại phản hồi cùng xuất hiện trong dữ liệu khảo sát: sinh viên tán thành việc giáo viên cần kiểm chứng đầu ra AI, đồng thời cho biết bản thân dùng đầu ra mà không kiểm tra lại. Như đã nêu ở Mục 1, E2 và E9 khác nhau về đối tượng quy chiếu, và nghiên cứu cũng không đo trực tiếp năng lực đánh giá đầu ra hay hành vi kiểm chứng.

Vì vậy, khái niệm này không được dùng như một cấu trúc tâm lý hay cơ chế nhận thức đã được kiểm định. Trong bài viết này, khái niệm chỉ giúp định hướng cho việc thảo luận mẫu hình phản hồi và gợi mở hướng nghiên cứu tiếp theo. Mối quan hệ giữa phản hồi tự báo cáo và quá trình đánh giá đầu ra trên thực tế vẫn cần được kiểm tra bằng dữ liệu hành vi.

2.3. Thiết kế câu lệnh và sự chuẩn bị cho việc sử dụng AI trong dạy học

Hwang và cộng sự (2023) dùng khái niệm *prompt literacy* để chỉ khả năng xây dựng câu lệnh, diễn giải đầu ra và điều chỉnh câu lệnh qua các lần tương tác với AI tạo sinh. Theo khái niệm này, xây dựng câu lệnh là một quá trình gồm: diễn đạt yêu cầu, xem xét đầu ra và điều chỉnh cách tương tác với công cụ. Trên cơ sở đó, bài viết phân biệt ba khía cạnh để phục vụ thảo luận: kỹ thuật, sư phạm và đánh giá. Cách phân chia này do nhóm tác giả đề xuất cho nghiên cứu hiện tại, không phải là mô hình ba nhân tố đã được Hwang và cộng sự (2023) kiểm định.

Mỗi khía cạnh có một trọng tâm riêng. Khía cạnh kỹ thuật liên quan đến việc diễn đạt rõ yêu cầu đối với công cụ; khía cạnh sư phạm đặt yêu cầu ấy trong quan hệ với mục tiêu học tập và trình độ người học; khía cạnh đánh giá hướng đến độ chính xác, mức độ phù hợp văn hóa và giá trị sử dụng của đầu ra. Nghiên cứu không đo ba khía cạnh này bằng các thang đo riêng và cũng không kiểm định chúng như những cấu trúc tách biệt.

Nhu cầu chuẩn bị cho việc sử dụng AI trong dạy học cũng được đề cập trong nhiều nghiên cứu gần đây. Crompton và cộng sự (2024) phân tích cả tiềm năng sử dụng và những giới hạn của AI trong dạy tiếng Anh, còn Li và cộng sự (2025) ghi nhận số nghiên cứu thực nghiệm về AI tạo sinh trong dạy và học ngôn ngữ tăng lên trong giai đoạn 2023–2024. Tuy nhiên, nhiều nghiên cứu trong tổng quan của Li và cộng sự có cỡ mẫu nhỏ hoặc thời gian triển khai ngắn, trong khi nghiên cứu dọc còn ít.

Ở Việt Nam, việc chuẩn bị cho người học sử dụng AI còn gắn với vấn đề liêm chính học thuật và hướng dẫn sử dụng có trách nhiệm. Dựa trên các phản hồi mở của giáo viên EFL, Cong-Lem và cộng sự (2024) ghi nhận những lo ngại về liêm chính học thuật khi người học dùng AI, kèm theo đề xuất là cần có quy định rõ ràng hơn và hướng dẫn sử dụng AI có trách nhiệm. Những kết quả này chỉ được dùng làm bối cảnh để xem xét mức sẵn sàng và mức độ tự tin liên quan đến AI trong nghiên cứu hiện tại và không phản ánh trải nghiệm của nhóm sinh viên được khảo sát.

2.4. Nghiên cứu về việc đánh giá và sử dụng ChatGPT của người học EFL

Các nghiên cứu gần đây về người học tiếng Anh như một ngoại ngữ (EFL) đã tiếp cận việc sử dụng và đánh giá đầu ra của AI bằng những nguồn dữ liệu khác nhau. Alsadoon (2026) kết hợp bảng hỏi với ghi hình màn hình trong lúc sinh viên EFL dùng ChatGPT để viết luận, đồng thời đối chiếu văn bản do AI tạo ra với bài viết cuối cùng nhằm xác định mức độ sử dụng lại và chỉnh sửa. Thiết kế này cho phép so sánh dữ liệu tự báo cáo với hành vi trong một nhiệm vụ viết cụ thể, điều mà nghiên cứu hiện tại chưa thực hiện được.

Xiao và Zhi (2023) lại tiếp cận vấn đề bằng dữ liệu định tính về trải nghiệm của người học. Người tham gia mô tả cách họ dùng ChatGPT, điều chỉnh câu lệnh, kiểm tra thông tin và lựa chọn đầu ra. Dữ liệu loại này cho thấy người học kể lại trải nghiệm và cách xử lý đầu ra như thế nào, nhưng không cho phép đối chiếu trực tiếp lời kể với hành vi quan sát được trong nhiệm vụ.

Hai nghiên cứu trên đưa ra hai loại bằng chứng khác nhau về việc người học EFL sử dụng và đánh giá đầu ra của AI. Nghiên cứu hiện tại khảo sát sinh viên ngành Ngôn ngữ Anh tại UFLS-UD, tập trung vào tần suất và mục đích sử dụng AI, quan niệm về kiểm chứng đầu ra và mức tự tin đối với một số nhiệm vụ sử dụng AI cho mục đích sư phạm. Với E2 và E9, nghiên cứu xét đồng thời phản hồi của từng người ở hai mục để xác định số người đồng ý với cả hai. Tuy nhiên, dữ liệu khảo sát không cho phép đối chiếu các phản hồi này với hành vi kiểm chứng được quan sát trực tiếp.

3. Câu hỏi nghiên cứu

Nghiên cứu này tập trung trả lời 3 câu hỏi sau:

RQ1. *Sinh viên ngành Ngôn ngữ Anh sử dụng AI tạo sinh với tần suất như thế nào và cho những kỹ năng, hoạt động thực hành tiếng Anh nào?*

RQ2. *Có bao nhiêu sinh viên vừa tán thành việc giáo viên cần kiểm chứng đầu ra AI, vừa cho biết bản thân dùng đầu ra mà không kiểm tra lại? Các câu trả lời mở cung cấp thêm thông tin gì về quan niệm và trải nghiệm của người học đối với việc kiểm chứng đầu ra của AI?*

RQ3. *Điểm sẵn sàng dùng AI trong dạy học khác điểm tự tin thiết kế câu lệnh sư phạm như thế nào? Mức tự tin trong việc nhận diện ảo giác AI và thiên kiến văn hóa có liên hệ ra sao với mức tự tin trong thiết kế câu lệnh sư phạm?*

4. Phương pháp nghiên cứu

4.1. Thiết kế, bối cảnh và đạo đức nghiên cứu

Nghiên cứu sử dụng thiết kế khảo sát cắt ngang mang tính thăm dò, kết hợp các mục đóng với câu hỏi mở về trải nghiệm sử dụng AI. Dữ liệu được thu thập trong học kỳ II năm học 2025–2026. Dữ liệu được trích từ một khảo sát rộng hơn về trải nghiệm sử dụng AI trong học tập và định hướng nghề nghiệp của sinh viên tại UFLS-UD. Lãnh đạo Khoa Tiếng Anh đã chấp thuận việc thu thập dữ liệu và giảng viên phụ trách các lớp tham gia đã đồng ý cho triển khai khảo sát. Trước khi trả lời bảng hỏi, người học được cung cấp thông tin về mục đích nghiên cứu, điều kiện tham gia, thời gian trả lời dự kiến và cách bảo mật dữ liệu. Chỉ những người xác nhận đã đọc và hiểu thông tin nghiên cứu, tự nguyện đồng ý tham gia và đủ 18 tuổi mới được tiếp tục trả lời. Khảo sát được thực hiện ẩn danh và không thu thập thông tin định danh cá nhân. Tại thời điểm thu thập dữ liệu, nghiên cứu chưa được thẩm định bởi hội đồng đạo đức nghiên cứu độc lập ở cấp trường hoặc liên trường, và nhóm tác giả cũng không có văn bản xác nhận việc miễn thẩm định. Giới hạn này được trình bày ở Mục 8.

4.2. Người tham gia và quy trình sàng lọc

Nghiên cứu thu được 210 phản hồi từ sinh viên năm thứ 2 và năm thứ 3 ngành Ngôn ngữ Anh tại UFLS-UD. Bảng hỏi có hai mục kiểm tra mức độ chú ý (KT-1 và KT-2) được thiết kế theo hướng dẫn kiểm soát phản hồi thiếu chú ý của Ward và Meade (2023). Cả hai mục đều yêu cầu người trả lời chọn mức "Đồng ý", tương ứng với mức 4 trên thang 1–5. KT-1 được đặt sau C4; KT-2 được đặt sau G1–G6 và trước G7. Chỉ những trường hợp có KT-1 = 4 và KT-2 = 4 được giữ lại cho các phân tích chính. Sau bước sàng lọc, dữ liệu được kiểm tra theo từng biến. Mỗi phân tích sử dụng các trường hợp có đủ dữ liệu ở những biến liên quan, thay vì loại toàn bộ trường hợp chỉ vì thiếu dữ liệu ở một biến khác. Bảng 1 trình bày số phản hồi trước và sau sàng lọc, nội dung các mục khảo sát được trình bày ở Phụ lục.

Bảng 1

Số phản hồi trước và sau sàng lọc

Giai đoạn sàng lọc	n	Diễn giải
Phản hồi ban đầu	210	Toàn bộ phản hồi thu được
Đạt cả KT-1 = 4 và KT-2 = 4	159	Mẫu được giữ lại cho các phân tích chính
Không đạt ít nhất một trong hai mục kiểm tra mức độ chú ý	51	Bị loại khỏi mẫu phân tích chính; dùng trong so sánh sàng lọc ở Bảng 2

** Ghi chú: Các phân tích định lượng chính sử dụng từ 154 đến 159 trường hợp, tùy theo dữ liệu có sẵn ở các biến liên quan. Số phản hồi không để trống đối với các câu hỏi mở được báo cáo riêng ở Mục 5.5.*

Trong mẫu được giữ lại (*n* = 159), sinh viên năm thứ 3 chiếm 69,8% (*n* = 111) và sinh viên năm thứ 2 chiếm 30,2% (*n* = 48); nữ giới chiếm 84,9% (*n* = 135). Có 149 sinh viên (93,7%) cho biết chưa được học nội dung chính thức nào về AI trong chương trình đào tạo.

Người tham gia được chọn theo phương pháp lấy mẫu thuận tiện theo lớp học và việc tham gia khảo sát là tự nguyện. Do không có thông tin tương ứng về những sinh viên không tham gia, nghiên cứu không thể xác định mức độ đại diện của mẫu đối với toàn bộ sinh viên ngành Ngôn ngữ Anh tại cơ sở nghiên cứu hay các nhóm sinh viên ở các chương trình khác. Giới hạn này được tính đến khi bàn về khả năng khái quát hoá của kết quả.

Để mô tả khác biệt giữa hai nhóm sau sàng lọc, nhóm được giữ lại (*n* = 159) và nhóm bị loại (*n* = 51) được so sánh theo năm học, giới tính, việc đã học nội dung chính thức về AI hay chưa, tần suất sử dụng ChatGPT và điểm nhận thức về AI (B1). Kết quả được trình bày ở Bảng 2.

Bảng 2

So sánh mẫu được giữ lại (n = 159) và mẫu bị loại (n = 51)

Biến so sánh	Giữ lại (n = 159)	Bị loại (n = 51)	Kiểm định	Kết quả
Năm thứ 2/Năm thứ 3 (%)	30,2% / 69,8%	70,6% / 29,4%	χ^2 (1)	24,60, p < .001
Nữ giới (%)	84,9%	88,2%	χ^2 (1)	0,13, p = .718
Chưa học nội dung AI chính thức (%)	93,7%	96,1%	Fisher's exact	p = .735
Tần suất dùng ChatGPT (M, thang 0–4)	3,47	3,33	Mann–Whitney U	U = 4391,5, p = .306
Nhận thức về AI – B1 (M, thang 1–5)	3,91	3,54	t (Welch)	t (69,09) = 3,01, p = .004

** Ghi chú: Tỷ lệ phần trăm được tính riêng trong từng nhóm (n = 159 và n = 51). Kiểm định χ^2 cho các bảng 2 × 2 có áp dụng hiệu chỉnh liên tục Yates. Với biến đã học nội dung chính thức về AI hay chưa, kiểm định chính xác Fisher hai phía được sử dụng vì có ô có tần suất kỳ vọng thấp. Các so sánh này chỉ mô tả khác biệt giữa hai nhóm sau sàng lọc, chúng không kiểm định tính tương đương giữa hai nhóm hay mức độ đại diện của mẫu đối với quần thể sinh viên.*

Nhóm bị loại có tỷ lệ sinh viên năm thứ 2 cao hơn và điểm nhận thức về AI (B1) thấp hơn nhóm được giữ lại; cả hai khác biệt này đều đạt mức ý nghĩa thống kê. Ngược lại, sự khác biệt về giới tính, việc đã học nội dung chính thức về AI và tần suất sử dụng ChatGPT không có ý nghĩa thống kê. Những khác biệt này được tính đến khi diễn giải kết quả. Phân tích độ nhạy ở Mục 5.6 kiểm tra thêm mức độ ổn định của kết quả. Phân tích này lặp lại các tính toán trên toàn bộ dữ liệu ban đầu, không áp dụng tiêu chí loại trừ.

4.3. Công cụ khảo sát

Bảng hỏi được xây dựng trong khuôn khổ một khảo sát rộng hơn về việc sử dụng AI trong học tập, giảng dạy và định hướng nghề nghiệp tại UFLS-UD. Khảo sát có hai phiên bản dành cho giảng viên và sinh viên, gồm một số nội dung chung và một số nội dung dành riêng cho từng nhóm. Phiên bản dành cho sinh viên được triển khai với sinh viên hai ngành Ngôn ngữ Anh và Sư phạm tiếng Anh; nghiên cứu này chỉ sử dụng dữ liệu của sinh viên ngành Ngôn ngữ Anh. Phạm vi áp dụng của từng phần trong bảng hỏi được trình bày ở Phụ lục.

Phiên bản dành cho sinh viên gồm 10 phần: thông tin nền (A); nhận thức và sẵn sàng sử dụng AI (B); tần suất và mục đích sử dụng (C); niềm tin sư phạm (D); quan niệm về AI (E); hỗ trợ thể chế (F); rào cản (G); tự tin sử dụng AI trong dạy học (H); sử dụng AI trong thiết kế bài dạy (I); phát triển nghề nghiệp và bản sắc nghề nghiệp (J). Các mục về tần suất, mức độ đồng ý và mức độ tự tin dùng thang trả lời năm mức, với nhãn và cách mã hóa được quy định riêng cho từng phần. Bảng hỏi có năm câu hỏi mở (C4, G7, I3, J3 và J4); bài viết chỉ phân tích câu trả lời cho C4, I3, J3 và J4, không đưa G7 vào phân tích.

Mục I1 yêu cầu người tham gia tự chọn mức mô tả đúng nhất cách họ đang dùng AI trong thiết kế hoặc thực hiện bài dạy theo bốn mức của khung SAMR: Thay thế, Tăng cường, Chỉnh sửa và Tái định nghĩa. Ngoài ra, bảng hỏi còn có phương án chưa dùng AI cho mục đích này. I1 là một mục tự phân loại và chỉ dùng để mô tả phản hồi. Đây không phải một thang đo SAMR gồm nhiều mục, cũng không phải phép đo trực tiếp chất lượng tích hợp AI.

Phần H gồm năm nhóm, mỗi nhóm có bốn mục đo mức độ tự tin: hiểu cơ chế tạo sinh (H1), nhận diện ảo giác AI và thiên kiến văn hóa (H2), thiết kế câu lệnh sư phạm (H3), phối hợp người – AI trong dạy học (H4), và quản trị đạo đức AI trong dạy học (H5). Năm nhóm này được xây dựng để bao quát các nội dung từ hiểu cơ chế và đánh giá đầu ra đến thiết kế tương tác và quản trị việc sử dụng AI cho mục đích sư phạm. Người tham gia trả lời theo mức độ tự tin chứ không theo mức độ đồng ý.

Các mục thuộc năm nhóm H1–H5 do nhóm tác giả tự xây dựng theo mục tiêu nghiên cứu, không lấy nguyên văn hoặc chuyển thể từ một thang đo đã được công bố. Nội dung được ba chuyên gia TESOL rà soát, sau đó được thử nghiệm với 25 sinh viên không thuộc mẫu nghiên cứu chính. Vì bảng hỏi được xây dựng và triển khai bằng tiếng Việt nên không cần quy trình dịch và dịch ngược. Việc rà soát chuyên gia và thử nghiệm giúp hoàn thiện nội dung của công cụ. Tuy nhiên, cấu trúc đo lường của năm nhóm H chưa được kiểm định đầy đủ trên một mẫu độc lập. Giới hạn này được bàn thêm ở Mục 8.

4.4. Phân tích dữ liệu

Phân tích định lượng gồm: thống kê mô tả, hệ số Cronbach’s alpha, các phép so sánh nhóm được giữ lại với nhóm bị loại và các phân tích theo từng câu hỏi nghiên cứu. Với RQ2, phân tích chính xem xét việc hai phản hồi E2 và E9 cùng xuất hiện ở từng người trả lời, trọng tâm là tỷ lệ trường hợp đồng thời đạt E2 ≥ 4 và E9 ≥ 4. Kiểm định *t* cặp đôi được dùng như một phân tích bổ trợ để mô tả chênh lệch điểm giữa hai mục. Vì E2 nói về yêu cầu đối với giáo viên, còn E9 nói về cách chính người trả lời dùng đầu ra của AI, chênh lệch điểm giữa hai mục không được hiểu là khoảng cách giữa niềm tin và hành vi của sinh viên. Với RQ3, điểm B2.1 và H3 được so sánh bằng kiểm định *t* cặp đôi; mối liên hệ giữa H2 và H3 được phân tích bằng hệ số tương quan Pearson. Mỗi phân tích dùng các trường hợp có đủ dữ liệu ở các biến liên quan: 158 trường hợp cho E2–E9 và 157 trường hợp cho các phân tích RQ3.

Điểm của mỗi nhóm là trung bình cộng các mục thành phần; chẳng hạn, H3 là trung bình của H3.1–H3.4. Điểm nhóm chỉ được tính khi trường hợp có đủ dữ liệu ở tất cả các mục trong nhóm. Một trường hợp thiếu dữ liệu ở nhóm này vẫn được đưa vào các phân tích khác nếu đủ dữ liệu ở những biến cần thiết. Cronbach’s alpha của từng nhóm được tính trên những trường hợp có đủ dữ liệu ở tất cả các mục thành phần của nhóm đó. Các phân tích thống kê được thực hiện bằng IBM SPSS Statistics 29.

Trước khi thực hiện kiểm định *t* cặp đôi, phân phối của điểm chênh lệch được kiểm tra về độ lệch và giá trị ngoại lai. Với E2–E9 (n = 158), độ lệch gần bằng 0 (skewness = −0,003) và không có giá trị ngoại lai theo tiêu chí 1,5 × IQR. Với B2.1–H3 (n = 157), độ lệch ở mức

nhẹ (skewness = 0,15), với 9 trường hợp (5,7%) được xác định là ngoại lai nhẹ theo cùng tiêu chí. Kết quả so sánh trung bình được đối chiếu thêm bằng kiểm định Wilcoxon signed-rank (E2–E9: $W = 474{,}5$, $p < .001$; B2.1–H3: $W = 1673{,}0$, $p < .001$). Cohen's d_z được tính bằng trung bình điểm chênh lệch chia cho độ lệch chuẩn của điểm chênh lệch.

Để đánh giá mức độ phụ thuộc của kết quả vào quy tắc sàng lọc, phân tích độ nhạy được thực hiện trên toàn bộ 210 phản hồi ban đầu, không áp dụng tiêu chí KT-1 và KT-2. Mỗi phép phân tích vẫn dùng các trường hợp có đủ dữ liệu ở những biến liên quan; 209 trường hợp cho E2–E9 và 207 trường hợp cho các phân tích RQ3. Các kết quả được đối chiếu gồm: tỷ lệ đồng thời đạt E2 ≥ 4 và E9 ≥ 4, chênh lệch B2.1–H3, tương quan H2–H3 và tần suất sử dụng ChatGPT. Phân tích này cho biết các ước lượng thay đổi đến đâu khi bỏ quy tắc sàng lọc, nhưng không hiệu chỉnh được sai lệch chọn mẫu.

Các mục D1–D6 được phân tích riêng vì mỗi mục phản ánh một khía cạnh khác nhau của niềm tin sư phạm liên quan đến AI; nghiên cứu không tính một điểm tổng hợp cho nhóm D. Điểm hỗ trợ thể chế được tính từ F1–F5. Mục F6 ("*Tôi chưa thực sự cần hỗ trợ thể chế nào để có thể tự sử dụng AI trong dạy học*") phản ánh nhận định của cá nhân về nhu cầu được hỗ trợ, khác với nội dung về điều kiện và sự hỗ trợ thể chế của cơ sở ở F1–F5, nên được báo cáo riêng và không đảo chiều điểm để gộp vào điểm tổng hợp. Các mục E cũng được phân tích riêng và giữ nguyên điểm gốc. E2 hỏi mức đồng ý với yêu cầu giáo viên cần kiểm tra và điều chỉnh đầu ra của AI, còn E9 hỏi cách chính người trả lời sử dụng đầu ra. E4 và E7 phản ánh quan niệm về độ tin cậy và điều người trả lời tự nhận thấy khi tiếp nhận thông tin do AI cung cấp, không phải tần suất hành vi quan sát được. Với RQ2, mức 4–5 ở E2 và E9 tương ứng với *Đồng ý*/*Hoàn toàn đồng ý*. Nghiên cứu báo cáo tỷ lệ đồng thời đạt hai ngưỡng trên trong toàn bộ các trường hợp đủ dữ liệu và tỷ lệ tương ứng trong nhóm có E2 ≥ 4.

Dữ liệu từ các câu hỏi mở được xử lý bằng phương pháp phân tích nội dung quy nạp trên mẫu được giữ lại ($n = 159$). Mỗi câu trả lời cho một câu hỏi mở được xem là một đơn vị mã hóa và có thể được gán nhiều mã nếu đề cập đến nhiều nội dung. Số câu trả lời không để trống được xác định sau khi loại bỏ khoảng trắng thừa, các câu ngắn hoặc câu phủ định vẫn được tính là có trả lời. Mỗi câu trả lời được đọc trong ngữ cảnh của từng câu hỏi tương ứng. Những câu cung cấp thông tin về mục đích sử dụng, quan niệm, trải nghiệm hoặc đề xuất được dùng để tổng hợp nội dung. Những câu chỉ cho biết người trả lời không có ý kiến vẫn được tính là có trả lời nhưng không được dùng làm căn cứ cho một nội dung phân tích nào. Những câu trả lời có ý nghĩa nhưng chưa thuộc danh mục mã hiện có được giữ riêng kèm lý do không gán mã.

Hai thành viên của nhóm tác giả đọc và mã hóa độc lập các câu trả lời mở. Sau vòng mã hóa ban đầu, các nhãn mã có nội dung tương đương được đối chiếu để hình thành danh mục mã chung. Những điểm bất đồng được thảo luận dựa trên câu trả lời gốc trước khi quyết định mã chung; sau đó, toàn bộ dữ liệu được rà soát lại theo danh mục đã thống nhất. Các đoạn trích được đối chiếu với câu trả lời gốc và ghi kèm mã hồ sơ, mã câu hỏi. Phần lược bỏ được đánh dấu bằng […]. Nội dung và cách viết của người trả lời được giữ nguyên; nếu có hiệu chỉnh tối thiểu, thay đổi được ghi chú ngay tại đoạn trích. Nội dung tổng hợp từ nhiều câu trả lời được trình bày bằng lời của nhóm tác giả và không đặt trong dấu ngoặc kép.

5. Kết quả nghiên cứu

5.1. RQ1. Tần suất và mục đích sử dụng AI

Trong mẫu được giữ lại ($n = 159$), 99 sinh viên (62,3%) cho biết dùng ChatGPT hằng ngày và 142 sinh viên (89,3%) dùng ít nhất hằng tuần. Bảng 3 trình bày tần suất sử dụng sáu

nhóm công cụ AI trong bảng hỏi.

Bảng 3

Tần suất sử dụng các nhóm công cụ AI (n = 159)

Nhóm công cụ	Chưa bao giờ	<1 lần/tháng	1–3 lần/tháng	1–3 lần/tuần	Hằng ngày	*M*	*SD*
ChatGPT (OpenAI)	2	4	11	43	99	3,47	0,83
Google Gemini	20	22	35	44	38	2,36	1,32
Grammarly/QuillBot	68	32	35	17	7	1,14	1,21
MS Copilot	105	28	17	6	3	0,58	0,96
Claude (Anthropic)	132	8	12	4	3	0,35	0,87
Công cụ AI tiếng Việt	135	9	10	5	0	0,28	0,72

** Ghi chú: Các cột tần suất trình bày số người. M và SD được tính trên mã thứ bậc 0–4, với 0 = chưa bao giờ và 4 = hằng ngày, không phải số lần sử dụng thực tế. Mỗi nhóm công cụ có n = 159. Grammarly/QuillBot là một phương án gộp trong bảng hỏi gốc.*

AI được dùng nhiều nhất cho kỹ năng viết (89,9%), tiếp theo là ngữ pháp và từ vựng (88,1%), kỹ năng nói (72,3%), kỹ năng đọc (57,2%), thiết kế kiểm tra và đánh giá (47,2%) và kỹ năng nghe (39,6%). Câu hỏi cho phép chọn nhiều phương án, mỗi tỷ lệ được tính trên 159 người tham gia.

Ở I1, 7/159 sinh viên cho biết chưa dùng AI để thiết kế hoặc thực hiện bài dạy. 29 người tự xếp mình ở mức Thay thế (*Substitution*), 44 người ở mức Tăng cường (*Augmentation*), 54 người ở mức Chỉnh sửa (*Modification*) và 25 người ở mức Tái định nghĩa (*Redefinition*). Hai mức Chỉnh sửa và Tái định nghĩa chiếm 79/159 trường hợp (49,7%). Đây là kết quả tự phân loại của người tham gia theo I1.

5.2. Thống kê mô tả các nhóm mục và mục đơn

Bảng 4 trình bày giá trị trung bình và độ lệch chuẩn của các nhóm mục và mục đơn, hệ số Cronbach’s alpha chỉ được báo cáo cho các nhóm đa mục. Cỡ mẫu dao động từ 154 đến 159, trong đó, các nhóm H có 157 trường hợp.

Bảng 4

Thống kê mô tả các nhóm mục và mục đơn (thang 1–5)

Nhóm mục/Mục đơn	*n*	*M*	*SD*	α
Nhận thức về AI (B1)	159	3,91	0,61	.734
Sẵn sàng dùng AI trong dạy học – mục đơn (B2.1)	159	3,84	0,91	—
Hỗ trợ thể chế được nhận thức (F1–F5)	159	3,43	0,68	.827
Chưa cần hỗ trợ thể chế thêm – mục đơn (F6)	159	3,11	0,96	—
Rào cản khi tích hợp AI (G)	159	3,84	0,59	.808
H1: Hiểu cơ chế tạo sinh của AI (tự tin)	157	3,19	0,73	.807
H2: Nhận diện ảo giác AI và thiên kiến văn hóa (tự tin)	157	3,38	0,63	.779
H3: Thiết kế câu lệnh sư phạm (tự tin)	157	3,27	0,81	.877
H4: Phối hợp người – AI trong dạy học (tự tin)	157	3,29	0,68	.786
H5: Quản trị đạo đức AI trong dạy học (tự tin)	157	3,36	0,76	.861
Nhu cầu phát triển chương trình đào tạo giáo viên (J1)	154	3,96	0,66	.912

** Ghi chú: Dấu — cho biết Cronbach’s alpha không áp dụng cho hai mục đơn B2.1 và F6. F6 thể hiện mức đồng ý với việc chưa cần hỗ trợ thêm, nên điểm càng cao không có nghĩa là mức hỗ trợ cảm nhận càng thấp. Các nhóm H1–H5 đo mức tự tin tự báo cáo, không phải mức đồng ý hay năng lực được đo*

trực tiếp. J1 đề cập đến nhu cầu phát triển chương trình đào tạo giáo viên tiếng Anh, trong khi mẫu nghiên cứu gồm sinh viên ngành Ngôn ngữ Anh. Điểm J1 được tính trên 154 trường hợp có đủ dữ liệu ở cả năm mục J1.1–J1.5 (xem Mục 4.4). Nhóm D không có điểm tổng hợp nên không được trình bày trong bảng; nội dung D1–D6 có trong Phụ lục. B2.1 được mô tả trên 159 trường hợp (M = 3,84); còn phép so sánh với H3 ở Mục 5.4 sử dụng 157 trường hợp có đủ dữ liệu ở cả hai biến (M của B2.1 = 3,83).

Điểm tự tin trung bình ở năm nhóm H dao động từ 3,19 đến 3,38, H1 có điểm thấp nhất và H2 có điểm cao nhất.

5.3. RQ2. Phản hồi về chuẩn mực kiểm chứng và việc sử dụng đầu ra AI

Trong 158 trường hợp có đủ dữ liệu E2–E9, 133 sinh viên (84,2%) đạt E2 ≥ 4, 51 sinh viên (32,3%) đạt E9 ≥ 4 và 41 sinh viên (25,9%) đồng thời đạt cả hai ngưỡng. Trong nhóm 133 sinh viên đạt E2 ≥ 4, 41 người (30,8%) cũng đạt E9 ≥ 4.

Bảng 5

Các mục E liên quan đến quan niệm và phản ứng đối với đầu ra AI (n = 158)

Mục	Nội dung mục	M	SD
E2	Giáo viên nên kiểm tra và điều chỉnh đầu ra của AI trước khi sử dụng, thay vì dùng trực tiếp.	4,09	0,70
E4	Khi AI trả lời rất chi tiết và tự tin, điều đó đủ để tôi tin vào độ chính xác của thông tin.	3,16	1,07
E7	Tôi nhận thấy mình đôi khi dễ chấp nhận thông tin từ AI khi thông tin được trình bày mạch lạc.	3,65	0,88
E9	Tôi thường tin tưởng và sử dụng trực tiếp đầu ra của AI mà không kiểm tra lại.	2,82	1,14

** Ghi chú: Nội dung các mục được giữ nguyên theo bảng hỏi gốc. Các mục sử dụng thang 1–5, từ 1 = Hoàn toàn không đồng ý đến 5 = Hoàn toàn đồng ý. E2 nói về yêu cầu đối với giáo viên, trong khi E9 nói về cách chính người trả lời sử dụng đầu ra của AI. E4 và E7 phản ánh quan niệm về việc tiếp nhận thông tin do AI cung cấp, không phải tần suất hành vi quan sát được (Phụ lục).*

E2 có *M* = 4,09 (*SD* = 0,70), E9 có *M* = 2,82 (*SD* = 1,14). E2 phản ánh mức tán thành việc yêu cầu giáo viên kiểm tra và điều chỉnh đầu ra của AI, còn E9 phản ánh mức độ người trả lời tự nhận dùng đầu ra của AI mà không kiểm tra lại. Vì khác nhau về đối tượng quy chiếu, hai mục này không đo hai mặt tương ứng của cùng một hành vi.

Phép so sánh bổ trợ cho thấy điểm E2 cao hơn E9, $t(157) = 11{,}33$, $p < .001$, $d_z = 0{,}90$. Kiểm định Wilcoxon cho kết quả tương tự, $W = 474{,}5$, $p < .001$. Phép so sánh này chỉ mô tả chênh lệch điểm giữa hai mục, không cho thấy sự bất nhất giữa niềm tin và hành vi của sinh viên.

5.4. RQ3. Mức sẵn sàng và tự tin thiết kế câu lệnh sư phạm

Trong 157 trường hợp có đủ dữ liệu ở cả B2.1 và H3, B2.1 có *M* = 3,83 (*SD* = 0,91), còn H3 có *M* = 3,27 (*SD* = 0,81). Kiểm định *t* cặp đôi cho thấy điểm B2.1 cao hơn điểm H3, $t(156) = 6{,}82$, $p < .001$, $d_z = 0{,}54$. B2.1 là một mục đơn đo mức đồng ý với phát biểu về sự sẵn sàng, còn H3 gồm bốn mục đo mức tự tin trong thiết kế câu lệnh sư phạm. Vì hai phép đo khác nhau về nội dung và cách trả lời, chênh lệch điểm không được hiểu là chênh lệch về năng lực thực hiện. Điểm H2 và H3 có tương quan thuận, $r = .560$, $p < .001$, $n = 157$.

5.5. Nội dung các câu trả lời mở

Trong mẫu được giữ lại (*n* = 159), số câu trả lời không để trống là 130 ở C4, 38 ở I3, 33 ở J3 và 18 ở J4. Đây là số câu trả lời theo từng mục, một sinh viên có thể trả lời nhiều mục. Như đã nêu ở Mục 4.4, số câu trả lời không để trống chỉ phản ánh mức độ tham gia trả lời và

không trùng với số câu trả lời được dùng làm căn cứ cho từng nội dung phân tích.

Các câu trả lời ở C4 chủ yếu đề cập đến việc sửa lỗi, tìm ý tưởng và luyện kỹ năng. Ở J4, một sinh viên bày tỏ lo ngại rằng thông tin sai vẫn có thể được trình bày một cách thuyết phục:

> *"Trải nghiệm về việc sử dụng AI khá tệ mặc dù nó hỗ trợ không nhỏ nhưng cũng đưa ra nhiều thông tin sai lệch một cách chỉn chu, kín kẽ. […] Kỳ vọng: chỉ nên coi AI như một công cụ hỗ trợ, tham khảo, không nên kì vọng quá nhiều vào AI"* (UFLS_EN_P1_145, J4).

Sinh viên mang mã UFLS_EN_P1_145 vừa nêu lo ngại về thông tin sai trong J4, vừa chọn mức 5 ở cả E2 và E9. Câu trả lời này giúp hiểu thêm bối cảnh của hai lựa chọn ở mục đóng, nhưng một trường hợp đơn lẻ không thể giải thích mẫu hình chung của toàn bộ nhóm.

Ở J3, sinh viên đề xuất chương trình đào tạo giáo viên tiếng Anh cần tăng cường hướng dẫn kiểm chứng đầu ra và sử dụng AI có trách nhiệm. Hai trích dẫn dưới đây minh họa nội dung này, cả hai sinh viên đều không thuộc nhóm đồng thời đạt E2 ≥ 4 và E9 ≥ 4:

> *"Giáo viên nên được trau dồi kiến thức về AI. AI nên được xác thực, sử dụng các nguồn thông tin uy tín"* (UFLS_EN_P1_004, J3).

> *"Chương trình đào tạo giáo viên tiếng Anh tại Việt Nam cần kiểm soát, hướng dẫn chặt chẽ các cách thức sử dụng và kiểm tra đúng sai, hiểu được cách sử dụng, nhận dạng đúng sai khi AI đưa ra câu trả lời, nên sử dụng AI để hỗ trợ, tránh lạm dụng"* (UFLS_EN_P1_090, J3).

Các câu trả lời ở I3 nói về việc đã dùng hoặc mong muốn sử dụng AI để thiết kế bài tập, trò chơi và tài liệu phù hợp với trình độ người học. Đây là những cách sử dụng do sinh viên tự thuật lại, không phải kết quả đánh giá trực tiếp chất lượng của học liệu được tạo ra.

Độ dài và mức độ chi tiết của các câu trả lời mở không đồng đều, riêng J4 chỉ có 18 câu trả lời không để trống. Các trích dẫn chỉ minh họa quan niệm và trải nghiệm do người học tự thuật lại, không xác định được cơ chế khiến E2 và E9 cùng xuất hiện.

5.6. Phân tích độ nhạy

Khi không áp dụng tiêu chí sàng lọc theo hai mục kiểm tra mức độ chú ý, có 56/209 trường hợp đủ dữ liệu (26,8%) đồng thời đạt E2 ≥ 4 và E9 ≥ 4, so với 41/158 (25,9%) ở mẫu chính, chênh lệch tính từ hai tỷ lệ chưa làm tròn là 0,8 điểm phần trăm. Với B2.1–H3, Cohen's d_z là 0,49 trên 207 trường hợp, so với 0,54 ở mẫu chính. Hệ số tương quan H2–H3 lần lượt là r = .575 và r = .560. Tỷ lệ dùng ChatGPT hằng ngày là 60,5% ở mẫu ban đầu và 62,3% ở mẫu chính; tỷ lệ dùng ít nhất hằng tuần lần lượt là 88,1% và 89,3%. Các ước lượng thay đổi ít giữa hai cách xử lý mẫu. Dù vậy, phân tích độ nhạy không khắc phục được sai lệch chọn mẫu và cũng không xác nhận chất lượng của các phản hồi không đạt kiểm tra mức độ chú ý.

6. Thảo luận

6.1. Việc sử dụng AI trong thực hành tiếng Anh

Sinh viên cho biết dùng AI cho nhiều kỹ năng và hoạt động thực hành tiếng Anh. Từ góc nhìn sinh thái, AI có thể được xem là một nguồn lực trong môi trường thực hành ngôn ngữ của người học (van Lier, 2004). Tuy vậy, dữ liệu khảo sát chỉ phản ánh tần suất và phạm vi sử dụng, chưa cho biết chất lượng tương tác hay sự thay đổi năng lực tiếng Anh theo thời gian.

Kết quả ở I1 cho biết thêm cách sinh viên tự xếp loại việc dùng AI trong thiết kế hoặc thực hiện bài dạy: gần một nửa số người tham gia chọn hai mức Chỉnh sửa hoặc Tái định nghĩa. Tuy nhiên, tỷ lệ này cần được diễn giải thận trọng vì người trả lời có thể hiểu các mức của khung SAMR theo những cách khác nhau. Do I1 là một mục tự phân loại đơn lẻ và nghiên cứu không quan sát bài dạy hay đánh giá sản phẩm của sinh viên, kết quả này không cho biết mức

độ thay đổi nhiệm vụ hay chất lượng tích hợp AI trong thực tế. Để làm rõ mối liên hệ giữa việc sử dụng AI và sự phát triển năng lực ngôn ngữ, cần có nghiên cứu dọc hoặc thiết kế đo lường kết quả học tập ở nhiều thời điểm khác nhau.

6.2. Diễn giải phản hồi về kiểm chứng đầu ra AI

Trong các trường hợp có đủ dữ liệu E2–E9, 25,9% đồng thời đạt E2 ≥ 4 và E9 ≥ 4. Kết quả này cần được đặt trong giới hạn nội dung của hai mục. E2 nói về yêu cầu đối với giáo viên, còn E9 nói về cách chính sinh viên dùng đầu ra của AI. Vì vậy, hai phản hồi này không nhất thiết mâu thuẫn: người trả lời có thể đặt ra cho giáo viên một yêu cầu khác với yêu cầu đặt ra cho chính mình. Khảo sát cũng không cho biết người trả lời hiểu phạm vi áp dụng của yêu cầu ở E2 như thế nào.

E4 và E7 cho biết thêm về cách sinh viên nhìn nhận đầu ra của AI. E4 liên quan đến độ chi tiết và vẻ tự tin của câu trả lời, còn E7 phản ánh việc người học tự nhận thấy mình đôi khi dễ chấp nhận thông tin được trình bày mạch lạc. Các kết quả này có thể được xem xét cùng với những nghiên cứu về tính dễ xử lý thông tin, theo đó, cách trình bày có thể ảnh hưởng đến phán đoán về tính đúng của thông tin (Alter & Oppenheimer, 2009; Reber & Unkelbach, 2010). Dù vậy, nghiên cứu không đo trực tiếp tính dễ xử lý, mức độ thuyết phục của đầu ra hay hành vi kiểm chứng, và cũng chưa phân tích xem nhóm đồng thời đạt E2 ≥ 4 và E9 ≥ 4 có khác các nhóm còn lại về E4 hoặc E7 hay không.

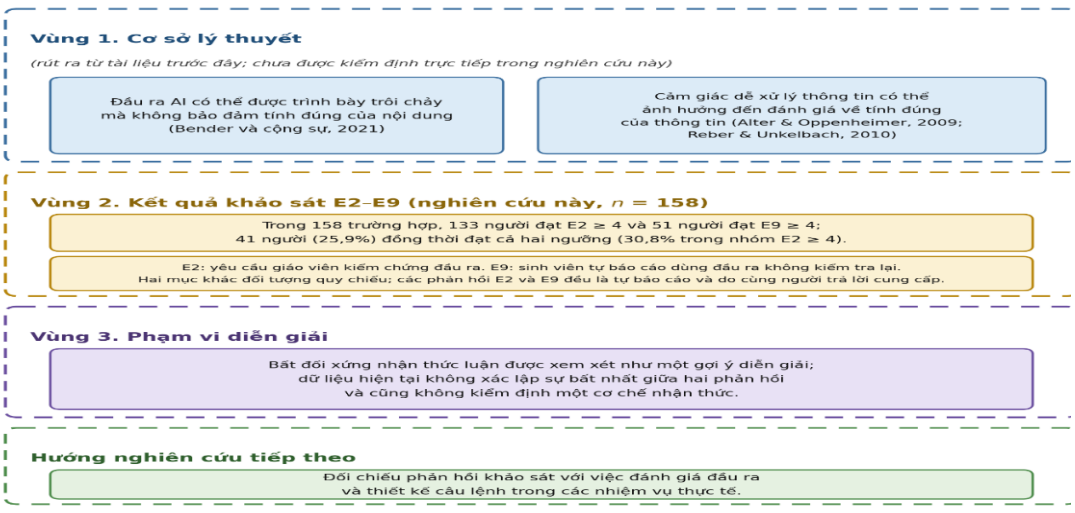
Trong phạm vi đó, bất đối xứng nhận thức luận là một cách diễn giải việc E2 và E9 cùng xuất hiện, chứ không phải một cơ chế đã được dữ liệu xác nhận. Có ít nhất hai khả năng diễn giải: (1) liên quan đến tính dễ xử lý của đầu ra AI; (2) liên quan đến sự khác biệt về đối tượng quy chiếu giữa E2 và E9. Dữ liệu hiện có chưa cho phép phân biệt hai khả năng này. Điều đó cần được kiểm tra thêm bằng các nghiên cứu quan sát trực tiếp quá trình người học đánh giá và kiểm chứng đầu ra của AI.

Câu trả lời J4 của sinh viên mang mã UFLS_EN_P1_145, người chọn mức 5 ở cả E2 và E9 (Mục 5.5), cung cấp thêm bối cảnh cho việc hai phản hồi cùng xuất hiện, nhưng một trường hợp đơn lẻ không đủ để giải thích nguyên nhân của cả 41 trường hợp được ghi nhận.

Hình 1 dưới đây tách bạch cơ sở lý thuyết, kết quả khảo sát và những cách giải thích chưa được kiểm định.

Hình 1

Cơ sở lý thuyết và phạm vi diễn giải kết quả E2–E9



** Ghi chú: Sơ đồ gồm ba vùng nhằm tách bạch cơ sở lý thuyết, kết quả quan sát được trong dữ liệu và phần diễn giải chưa được kiểm định, kèm theo một ô nêu hướng nghiên cứu tiếp theo. Trình tự từ Vùng 1 đến Vùng 3 chỉ thể hiện mạch lập luận, không biểu thị quan hệ nhân quả hay một mô hình đường dẫn đã được kiểm định.*

(Nguồn: Nhóm tác giả tổng hợp và trình bày trên cơ sở Bender và cộng sự (2021), Alter và Oppenheimer (2009), Reber và Unkelbach (2010) cùng dữ liệu khảo sát của nghiên cứu.)

6.3. Mức sẵn sàng và tự tin đối với việc sử dụng AI trong dạy học

Điểm tự tin ở năm nhóm H dao động từ 3,19 đến 3,38 trên thang 1–5, cao hơn một chút so với điểm giữa của thang đo. Đây là mức tự tin do sinh viên tự đánh giá đối với các nhiệm vụ liên quan đến AI trong dạy học, không phải năng lực thực tế được đánh giá trực tiếp. Dù ChatGPT được sử dụng thường xuyên trong mẫu, nghiên cứu chưa phân tích mối liên hệ giữa tần suất sử dụng và điểm ở các nhóm H, nên chưa thể kết luận liệu tần suất sử dụng AI có liên quan đến mức tự tin đối với các nhiệm vụ này hay không.

Crompton và cộng sự (2024) nhấn mạnh yêu cầu về hiểu biết và sự chuẩn bị của người dạy khi sử dụng AI trong dạy tiếng Anh. Trong nghiên cứu này, điểm B2.1 cao hơn điểm H3, nhưng hai phép đo khác nhau về nội dung và cách trả lời: B2.1 là mức đồng ý với một phát biểu về sự sẵn sàng, còn H3 gồm bốn mục đo mức tự tin trong thiết kế câu lệnh sư phạm. Kết quả này cho thấy nên phân biệt mức sẵn sàng chung với mức tự tin trong một nhiệm vụ cụ thể, nhưng không cho phép xem chênh lệch điểm là biểu hiện của sự thiếu hụt năng lực.

H2 và H3 có tương quan thuận (r = .560, p < .001). Trong mẫu nghiên cứu, sinh viên có mức tự tin cao hơn trong việc nhận diện ảo giác AI và thiên kiến văn hóa cũng có xu hướng tự tin hơn trong thiết kế câu lệnh sư phạm. Vì cả hai biến đều dựa trên tự đánh giá và được đo tại cùng một thời điểm, kết quả không cho phép kết luận biến nào tác động đến biến nào, cũng không cho phép suy ra năng lực thực hiện của sinh viên.

7. Một số định hướng cho chương trình đào tạo

Dựa trên kết quả khảo sát, chương trình đào tạo ngành Ngôn ngữ Anh tại UFLS-UD có thể xem xét bổ sung một số nội dung liên quan đến việc sử dụng AI. Những gợi ý dưới đây chỉ mang tính định hướng và cần được điều chỉnh theo mục tiêu, đặc điểm và nguồn lực của từng chương trình cụ thể.

- Trước hết, chương trình có thể chú trọng phát triển năng lực thiết kế câu lệnh gắn với mục tiêu học tập và trình độ của từng nhóm người học, đồng thời rèn luyện kỹ năng đánh giá đầu ra của AI theo các tiêu chí ngôn ngữ, văn hóa và sư phạm. Những nội dung này đặc biệt phù hợp với sinh viên có định hướng trở thành giáo viên tiếng Anh. Kết quả ở mục J1 cho thấy sinh viên tán thành việc đưa nội dung về AI vào chương trình đào tạo giáo viên tiếng Anh nói chung; tuy nhiên, mục này không hỏi trực tiếp về chương trình Ngôn ngữ Anh mà họ đang theo học. Vì vậy, trước khi điều chỉnh chương trình, có thể khảo sát thêm nhu cầu cụ thể của sinh viên ngành này.

- Bên cạnh hướng dẫn thiết kế câu lệnh, chương trình có thể xem xét các hoạt động phát triển kỹ năng đánh giá đầu ra của AI. Chẳng hạn, sinh viên có thể đối chiếu thông tin do AI cung cấp với các nguồn tư liệu đáng tin cậy, hoặc giải thích dựa trên tiêu chí rõ ràng vì sao chấp nhận, điều chỉnh hay loại bỏ đầu ra. Những hoạt động này vừa giúp sinh viên thực hành đánh giá, vừa tạo ra sản phẩm cụ thể để giảng viên nhận xét, qua đó bổ sung cho dữ liệu tự đánh giá mức tự tin.

- Một vấn đề khác là mức độ hiện diện của các nội dung đào tạo về AI trong chương

trình hiện hành. Có 93,7% sinh viên cho biết chưa học nội dung chính thức nào về AI. Thông tin này, cùng với các phản hồi về hỗ trợ thể chế (F1–F5), đặt ra câu hỏi về mức độ phù hợp của các nguồn lực và hoạt động hỗ trợ hiện có. Nhà trường có thể đối chiếu dữ liệu khảo sát với mục tiêu và nội dung của từng học phần, cũng như với các chính sách và biện pháp hỗ trợ đang triển khai, để xác định khoảng cách giữa nhu cầu của người học và những gì chương trình hiện cung cấp, từ đó đề xuất điều chỉnh phù hợp.

8. Giới hạn nghiên cứu và kết luận

Nghiên cứu mô tả việc sử dụng AI tạo sinh và các phản hồi về kiểm chứng đầu ra của sinh viên ngành Ngôn ngữ Anh tại một cơ sở đào tạo, dựa trên khảo sát cắt ngang. Khi diễn giải kết quả, cần lưu ý một số giới hạn sau:

- Thứ nhất, thiết kế cắt ngang với dữ liệu thu được tại một thời điểm không cho phép kết luận về sự thay đổi năng lực ngôn ngữ hay quan hệ nhân quả giữa việc sử dụng AI và sự tiến bộ của người học.

- Thứ hai, dữ liệu hoàn toàn dựa trên tự báo cáo nên có thể chịu ảnh hưởng của sai lệch hồi tưởng và xu hướng trả lời theo chuẩn mực xã hội. Nghiên cứu không quan sát trực tiếp hành vi kiểm chứng hay việc thực hiện nhiệm vụ.

- Thứ ba, do sử dụng mẫu thuận tiện từ một cơ sở và chỉ gồm sinh viên ngành Ngôn ngữ Anh, nghiên cứu không có cơ sở để xác định mức độ đại diện của mẫu đối với các cơ sở hoặc chương trình đào tạo khác.

- Thứ tư, các nhóm H do nhóm tác giả tự xây dựng. Dù đã được chuyên gia rà soát và thử nghiệm, cấu trúc đo lường của các nhóm này vẫn cần được kiểm định trên mẫu độc lập. Hệ số Cronbach’s alpha không thay thế cho việc đánh giá này.

- Thứ năm, nhóm bị loại khác nhóm được giữ lại về năm học và điểm nhận thức về AI (B1), nên cần thận trọng khi suy rộng kết quả.

- Cuối cùng, nghiên cứu chưa được hội đồng đạo đức nghiên cứu độc lập thẩm định và nhóm tác giả không có văn bản xác nhận việc miễn thẩm định (Mục 4.1).

Tóm lại, sinh viên ngành Ngôn ngữ Anh trong mẫu khảo sát dùng AI tạo sinh khá thường xuyên, đặc biệt là ChatGPT, cho nhiều kỹ năng và hoạt động thực hành tiếng Anh, trong khi phần lớn chưa được học nội dung chính thức về AI. Khoảng một phần tư số sinh viên có đủ dữ liệu vừa tán thành yêu cầu giáo viên kiểm chứng đầu ra AI, vừa tự nhận sử dụng đầu ra mà không kiểm tra lại. Do hai mục khác nhau về đối tượng quy chiếu, kết quả này chưa đủ để xác lập sự bất nhất giữa niềm tin và hành vi. Mức tự tin thiết kế câu lệnh sư phạm có tương quan thuận với mức tự tin nhận diện ảo giác AI và thiên kiến văn hóa, song cả hai biến đều dựa trên tự đánh giá. Trong giới hạn về mẫu và thiết kế, kết quả gợi ý rằng nội dung về thiết kế câu lệnh và đánh giá đầu ra AI có thể được cân nhắc đưa vào chương trình đào tạo. Các nghiên cứu tiếp theo cần kết hợp dữ liệu tự báo cáo với dữ liệu quan sát quá trình người học đánh giá đầu ra và thiết kế câu lệnh trong nhiệm vụ thực tế, đồng thời theo dõi người học qua nhiều thời điểm.

Tài liệu tham khảo

Alsadoon, R. (2026). Perceptions vs. practices: Academic integrity and actual ChatGPT use among EFL students. *Frontiers in Psychology, 17*, Article 1796737. https://doi.org/10.3389/fpsyg.2026.1796737

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13*(3), 219–235. https://doi.org/10.1177/1088868309341564

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability,*

and Transparency (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Cong-Lem, N., Tran, T. N., & Nguyen, T. T. (2024). Academic integrity in the age of generative AI: Perceptions and responses of Vietnamese EFL teachers. *Teaching English with Technology, 24*(1), 28–47. https://doi.org/10.56297/FSYB3031/MXNB7567

Crompton, H., Edmett, A., Ichaporia, N., & Burke, D. (2024). AI and English language teaching: Affordances and challenges. *British Journal of Educational Technology, 55*(6), 2503–2529. https://doi.org/10.1111/bjet.13460

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24*(2), 143–188. https://doi.org/10.1017/S0272263102002024

Ellis, N. C. (2008). The dynamics of second language emergence: Cycles of language use, language change, and language acquisition. *The Modern Language Journal, 92*(2), 232–249. https://doi.org/10.1111/j.1540-4781.2008.00716.x

Hwang, Y., Lee, J. H., & Shin, D. (2023). *What is prompt literacy? An exploratory study of language learners' development of new literacy skill using generative AI* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2311.05373

Lantolf, J. P., & Thorne, S. L. (2006). *Sociocultural theory and the genesis of second language development*. Oxford University Press. https://books.google.com.vn/books?id=gGhpAAAAMAAJ&hl=vi&source=gbs_citations_module_r&cad=4

Li, B., Tan, Y. L., Wang, C., & Lowell, V. (2025). Two years of innovation: A systematic review of empirical generative AI research in language learning and teaching. *Computers and Education: Artificial Intelligence, 9*, Article 100445. https://doi.org/10.1016/j.caeai.2025.100445

Reber, R., & Unkelbach, C. (2010). The epistemic status of processing fluency as source for judgments of truth. *Review of Philosophy and Psychology, 1*(4), 563–581. https://doi.org/10.1007/s13164-010-0039-7

van Lier, L. (2004). *The ecology and semiotics of language learning: A sociocultural perspective*. Kluwer Academic Publishers. https://doi.org/10.1007/1-4020-7912-5

Vygotsky, L. S. (1978). *Mind in society: Development of higher psychological processes* (M. Cole, V. Jolm-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press. https://doi.org/10.2307/j.ctvjf9vz4

Ward, M. K., & Meade, A. W. (2023). Dealing with careless responding in survey data: Prevention, identification, and recommended best practices. *Annual Review of Psychology, 74*, 577–596. https://doi.org/10.1146/annurev-psych-040422-045007

Xiao, Y., & Zhi, Y. (2023). An exploratory study of EFL learners' use of ChatGPT for language learning tasks: Experience and perceptions. *Languages, 8*(3), Article 212. https://doi.org/10.3390/languages8030212

PHỤ LỤC

Các mục trích từ bảng hỏi khảo sát xin truy cập mã QR dưới đây: