ỨNG DỤNG AI-CHEM HỖ TRỢ CÁ NHÂN HÓA PHẢN HỒI TRONG ÔN TẬP HÓA HỌC 11: MỘT NGHIÊN CỨU BÁN THỰC NGHIỆM

THE AI-CHEM APPLICATION SUPPORTING PERSONALIZED FEEDBACK IN GRADE 11 CHEMISTRY REVIEW: A QUASI-EXPERIMENTAL STUDY

Lê Thị Thúy Phượng[1],
Hồ Thị Bảo Châu[2],
Phạm Ngọc Sơn[3],
Đỗ Thị Ngọc Loan[4],
Đặng Thị Thuận An[4,+]

[1]Trường THPT chuyên Quốc Học Huế, thành phố Huế;
[2]Trường Đại học Văn Hiến;
[3]Trường Đại học Thủ đô Hà Nội;
[4]Trường Đại học Sư phạm - Đại học Huế
+Tác giả liên hệ ● Email: dtthuanan@hueuni.edu.vn

Article history
Received: 19/6/2026
Accepted: 16/7/2026
Published: 15/8/2026

Keywords
Large language models, AI-Chem, personalized learning, Chemistry 11, quasi-experimental study

ABSTRACT
The development of large language models offers the potential to personalize the learning process; however, quantitative evidence regarding the use of this technology as a learning scaffold in high school chemistry education in Vietnam remains limited. This study developed the AI-Chem system, which integrates custom "Gems" based on the Gemini platform using the Socratic method, to support Grade 11 chemistry review through step-by-step questions and hints. A quasi-experimental design with a non-equivalent control group was employed over a 12-week period involving two Grade 11 classes at Quoc Hoc Hue Gifted High School (Thuan Hoa Ward, Hue City); the experimental group used AI-Chem for self-study, while the control group followed conventional study methods. Preliminary results demonstrate AI-Chem's potential to support review and personalize feedback in Grade 11 chemistry instruction. However, due to the small sample size, the non-equivalence of the groups, and certain uncontrolled confounding factors, the results are insufficient to establish a causal relationship. Consequently, future research should expand the sample size, improve the control of confounding variables, and analyze interaction data to better evaluate the effectiveness of the Socratic approach.

1. Mở đầu

Theo Chương trình giáo dục phổ thông 2018, "năng lực tự chủ và tự học" là một trong những năng lực cốt lõi, đòi hỏi người học tự xác định mục tiêu, lập kế hoạch và điều chỉnh phương pháp học tập phù hợp (Bộ GD-ĐT, 2018). Cá nhân hóa học tập hướng đến việc đáp ứng sự khác biệt về đặc điểm và nhu cầu học tập của từng HS thay vì áp dụng một cách thức học tập đồng nhất cho tất cả người học (Wang và cộng sự, 2024). Sự phát triển của các mô hình ngôn ngữ lớn (Large Language Model - LLM) như ChatGPT, Gemini,... mở ra khả năng ứng dụng AI để hỗ trợ và cá nhân hóa quá trình học tập. Trong nghiên cứu này, AI-Chem được xây dựng như một hệ thống hỗ trợ ôn tập Hóa học 11, tích hợp các Gem - những trợ lí AI được tùy chỉnh trên nền tảng Gemini - được thiết kế theo phương pháp Socratic, trong đó AI không cung cấp trực tiếp đáp án mà sử dụng chuỗi câu hỏi gợi mở, truy vấn và phản hồi để hướng dẫn HS từng bước phân tích vấn đề, tự phát hiện sai sót và điều chỉnh cách hiểu. Hệ thống được thiết kế theo định hướng "giàn giáo học tập", dựa trên quan điểm về "vùng phát triển gần nhất" của Vygotsky (1978), trong đó HS được hỗ trợ thông qua các câu hỏi và gợi ý từng bước phù hợp với quá trình giải quyết nhiệm vụ học tập. Nhiều phân tích tổng hợp ghi nhận hiệu ứng tích cực ở mức trung bình đến lớn của LLM và các hệ thống hỗ trợ ôn tập đối với kết quả học tập (Wu và cộng sự, 2026; Wang và cộng sự, 2024; Ma và cộng sự, 2014). Tuy nhiên, các tổng quan này bao quát nhiều bối cảnh, môn học và thiết kế nghiên cứu khác nhau nên kết quả chủ yếu cho thấy tiềm năng chung của các công nghệ trong hỗ trợ học tập.

Ở cấp độ can thiệp cụ thể hơn, kĩ thuật tạo lệnh (prompt engineering) theo phương pháp Socratic - định hướng AI đặt câu hỏi dẫn dắt thay vì cung cấp ngay đáp án - được xem là một cách thức để LLM đóng vai trò hỗ trợ học tập thay vì chỉ cung cấp lời giải và được ghi nhận có liên hệ với việc cải thiện hiệu quả giải quyết vấn đề cũng như cá nhân hóa phản hồi ở nhiều bậc học (Cain, 2024; Qian, 2025; Zhou và cộng sự, 2025; Malik và cộng sự, 2025;

Sharma và cộng sự, 2025). Trong lĩnh vực Hóa học, một số nghiên cứu bán thực nghiệm (TN) và khảo sát tương quan ghi nhận việc sử dụng ChatGPT có định hướng có thể góp phần cải thiện hứng thú, tư duy phản biện và kết quả học tập (Babalola và cộng sự, 2024; Alneyadi và Wardat, 2023; Ng và cộng sự, 2024). Tuy nhiên, tại Việt Nam, phần lớn bằng chứng nêu trên đến từ bậc đại học hoặc dạy học các môn Toán, Khoa học tự nhiên ở phổ thông; nghiên cứu định lượng về việc sử dụng LLM làm giàn giáo học tập trong dạy học môn Hóa học còn hạn chế. Các công bố gần đây chủ yếu tập trung vào nhận thức và khả năng sử dụng ChatGPT ở đại học (Maheshwari, 2024; Nguyen và và cộng sự, 2024). Trong bài báo này, chúng tôi thiết kế hệ thống AI-Chem theo nguyên tắc giàn giáo Socratic và đánh giá tác động của hệ thống đến kết quả học tập và trải nghiệm ôn tập của HS lớp 11. Bài báo tập trung trả lời hai câu hỏi: (1) Việc ứng dụng LLM như một giàn giáo học tập (thông qua prompt engineering theo phương pháp Socratic) có cải thiện kết quả học tập môn Hóa học 11 của nhóm TN so với nhóm đối chứng (ĐC) ở thời điểm kết thúc can thiệp sau khi kiểm soát sự khác biệt về điểm nền (điểm trung bình (ĐTB) ở Học kì 1) hay không? (2) HS đánh giá như thế nào về trải nghiệm sử dụng, tính cá nhân hóa và động lực học tập khi tương tác với hệ thống AI-Chem?

2. Phương pháp nghiên cứu

2.1. Cơ sở thiết kế can thiệp

Thiết kế hệ thống AI-Chem dựa trên lí thuyết giàn giáo học tập, kết hợp bốn khung lí thuyết bổ trợ: tải nhận thức của Sweller (1988) làm cơ sở cho việc chia nhỏ kiến thức thành các gợi ý tăng dần độ khó; phản hồi học tập của Hattie và Timperley (2007) làm cơ sở cho việc giải thích nguyên nhân sai; tự điều chỉnh học tập của Zimmerman (2002) và tự quyết định của Deci và Ryan (2000) làm cơ sở cho tương tác tức thời, cá nhân hóa nhằm đáp ứng nhu cầu về năng lực và quyền tự chủ. Bốn khung này được sử dụng để thiết kế can thiệp và diễn giải kết quả; nghiên cứu không đo trực tiếp tải nhận thức, tự điều chỉnh hay tự quyết định nên các cơ chế này chưa được kiểm định.

Để bổ sung cơ sở thực tiễn cho việc thiết kế can thiệp, nhóm nghiên cứu khảo sát nhu cầu của toàn bộ HS thuộc hai lớp tham gia nghiên cứu (N = 67) bằng phiếu hỏi trực tuyến ngay trước khi bắt đầu can thiệp. HS tham gia khảo sát tự nguyện và ẩn danh; phiếu hỏi gồm các câu hỏi lựa chọn và thang đo Likert 5 mức về thời lượng tự học, rào cản khi tự học, thói quen sử dụng công cụ AI và kì vọng đối với hệ thống hỗ trợ. Dữ liệu được xử lí bằng thống kê mô tả.

2.2. Thiết kế nghiên cứu

Nghiên cứu sử dụng thiết kế bán TN với nhóm ĐC không tương đương, được triển khai trên hai lớp nguyên vẹn. Kết quả học tập của hai nhóm được so sánh tại thời điểm kết thúc can thiệp, đồng thời ĐTB Học kì 1 được sử dụng làm điểm nền để kiểm soát sự khác biệt ban đầu giữa hai nhóm. Do HS không được phân ngẫu nhiên vào hai nhóm, thiết kế bán TN với các nhóm nguyên vẹn tiềm ẩn một số nguy cơ đối với giá trị nội tại của nghiên cứu (Shadish và cộng sự, 2002), như sự khác biệt về ban chuyên, định hướng lựa chọn tổ hợp xét tuyển đại học, chênh lệch điểm nền, hiệu ứng lớp và hiệu ứng mới lạ. Trong đó, sự khác biệt về điểm nền giữa hai nhóm được kiểm soát một phần bằng ANCOVA; các yếu tố nhiễu còn lại chưa được kiểm soát trực tiếp và được xem xét khi diễn giải kết quả nghiên cứu.

2.3. Đối tượng và bối cảnh nghiên cứu

TN được tiến hành trên HS khối 11 tại Trường THPT Chuyên Quốc Học Huế trong học kì II năm học 2025-2026, theo SGK Hóa học 11 (bộ Chân trời Sáng tạo), với 6 chương trọng tâm: Cân bằng hóa học; Nitrogen - sulfur; Đại cương hóa học hữu cơ; Hydrocarbon; Dẫn xuất halogen - alcohol - phenol; Hợp chất carbonyl - carboxylic acid. Nhóm TN là lớp 11/6 (ban Anh 2, Pháp, Nhật; N = 32), sử dụng AI-Chem khi tự học; nhóm ĐC là lớp 11/3 (ban Lí 1, Sử, Địa; N = 35), học theo cách truyền thống. Một HS thuộc nhóm TN vắng ở đánh giá thường xuyên (ĐGTX) lần 2 nên các phân tích tại thời điểm này sử dụng N = 31; đây cũng là số HS tham gia khảo sát trải nghiệm. Hai lớp do cùng một GV Hóa học phụ trách, học theo cùng phân phối chương trình và thời lượng chính khóa; hai bài ĐGTX sử dụng chung đề, đáp án - biểu điểm và quy trình chấm, qua đó góp phần kiểm soát ảnh hưởng của yếu tố GV và công cụ đánh giá. Khoảng 20% HS nhóm TN có định hướng lựa chọn tổ hợp xét tuyển đại học có môn Hóa học (khối A, B), so với khoảng 10% ở nhóm ĐC. Sự khác biệt này có thể ảnh hưởng đến động lực học tập độc lập với tác động của hệ thống AI-Chem và được xem xét trong phần hạn chế nghiên cứu.

2.4. Công cụ đo lường

Kết quả học tập được đo bằng ĐTB Học kì 1 - mốc đo duy nhất hoàn toàn trước can thiệp do AI-Chem được kích hoạt ngay từ tuần 1 với điểm ĐGTX lần 1 (giữa can thiệp, sau khoảng 5 tuần sử dụng AI-Chem) và lần 2 (kết thúc can thiệp), theo thang điểm 10 hiện hành. Trải nghiệm và nhận thức của HS về hệ thống AI-Chem được đo bằng

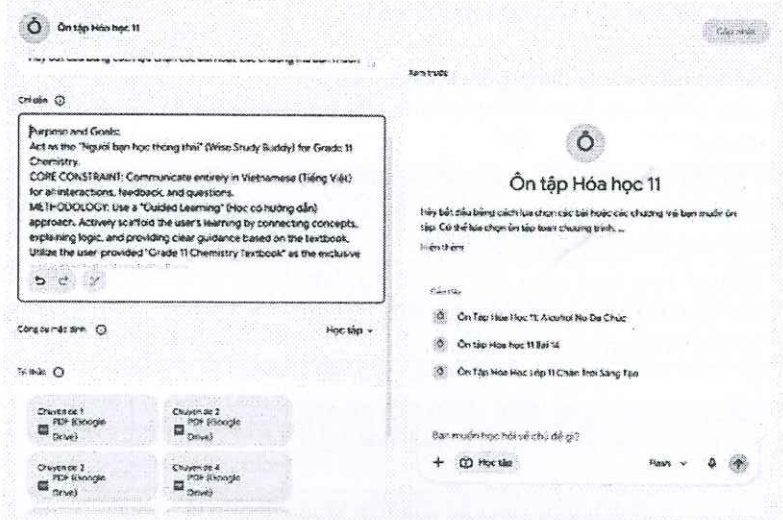
bảng hỏi Likert 5 mức, gồm 12 tiêu chí thuộc bốn phương diện: trải nghiệm kĩ thuật, tính cá nhân hóa, hiệu quả học tập và thái độ - động lực. Độ tin cậy nội tại của thang đo được kiểm định bằng hệ số Cronbach's Alpha.

2.5. Quy trình can thiệp sư phạm

Can thiệp sư phạm được triển khai qua ba bước: (1) Thiết kế tiến trình dạy học và thiết kế học liệu số: prompt engineering được sử dụng để thiết lập kịch bản sư phạm cho các "gia sư ảo", định hướng LLM thực hiện vai trò giàn giáo học tập thay vì chỉ cung cấp thông tin hoặc lời giải. Nhóm nghiên cứu sử dụng tính năng AI tùy chỉnh (Gem) của Google; mỗi Gem được cung cấp sách giáo khoa Hóa học 11 cùng các đề và chuyên đề ôn tập đã qua kiểm duyệt làm nguồn tài liệu tham chiếu nhằm hạn chế nguy cơ cung cấp thông tin sai lệch; (2) Tổ chức hoạt động TN: Môi trường tự học được xây dựng trên Google Sites với tên miền AI-Chem (chem.aii.edu.vn), thuộc hệ sinh thái Google Workspace for Education do Sở GD-ĐT cung cấp. Giao diện gồm sáu "trạm ôn tập", tương ứng với sáu chương trọng tâm, mỗi trạm tích hợp một Gem chuyên biệt tương ứng với một chương. Ngoài sáu Gem chuyên đề, hệ thống còn có một Gem tổng "Ôn tập Hóa học 11", được sử dụng trong giai đoạn ôn tập tích hợp cuối can thiệp. HS đăng nhập bằng email để ôn tập ngoài giờ lên lớp. Lịch sử hội thoại được lưu trên tài khoản do nhà trường quản lí; chỉ nhóm nghiên cứu và GV bộ môn được truy cập. Dữ liệu nghiên cứu được trích xuất và ẩn danh hóa ngay sau khi kết thúc can thiệp, sau đó được xóa sau khi công bố kết quả; (3) Kiểm tra, đánh giá và điều chỉnh: GV theo dõi quá trình sử dụng hệ thống, kiểm tra kết quả và thực hiện các điều chỉnh cần thiết trong quá trình can thiệp.

2.6. Thiết kế câu lệnh (prompt) cho AI-Chem

Kịch bản hệ thống của AI-Chem được thiết kế theo phương pháp Socratic và giới hạn trong phạm vi Chương trình Hóa học 11 (bộ Chân trời Sáng tạo), với 4 nguyên tắc cốt lõi: (1) Không cung cấp đáp án, công thức hoặc lời giải hoàn chỉnh ngay trong lần tương tác đầu tiên; (2) Chẩn đoán trước - yêu cầu HS trình bày cách hiểu hoặc hướng giải quyết hiện tại trước khi đưa ra gợi ý; (3) Khi HS trả lời sai, chỉ ra điểm chưa phù hợp trong lập luận thay vì chỉ thông báo "sai"; (4) Mỗi lượt tương tác chỉ đưa ra một gợi ý hoặc câu hỏi định hướng và chỉ tăng mức độ khó khi HS trả lời đúng câu hỏi trước đó. Ngoài bốn nguyên tắc trên, kịch bản còn giới hạn phạm vi phản hồi của AI-Chem trong nội dung Hóa học 11 theo bộ sách Chân trời Sáng tạo; yêu cầu hệ thống từ chối trả lời các câu hỏi ngoài phạm vi này và sử dụng danh pháp, thuật ngữ hóa học theo khuyến nghị của Liên minh Quốc tế về Hóa học thuần túy và Ứng dụng (International Union of Pure and Applied Chemistry - IUPAC). Nguyên văn kịch bản hệ thống (system prompt) được nạp cho mỗi Gem và kịch bản của bảy Gem (xem hình 1).



Hình 1. Minh họa kịch bản hệ thống cho mỗi Gem và hệ thống Gem (Nguồn: Tác giả)

2.7. Xử lí số liệu

Số liệu được xử lí bằng các phương pháp thống kê sau: (1) ANOVA một yếu tố được sử dụng để so sánh ĐTB giữa nhóm TN và nhóm ĐC tại từng thời điểm đánh giá; với hai nhóm, kết quả của ANOVA một yếu tố tương đương về mặt toán học với kiểm định t hai mẫu độc lập; (2) ANOVA đo lặp được sử dụng để phân tích sự biến thiên điểm số của nhóm TN qua ba thời điểm: điểm nền Học kì 1, ĐGTX1 và ĐGTX2, trên dữ liệu ghép của 31 HS có đầy đủ kết quả ở cả ba thời điểm. Giả định cầu tính được kiểm tra bằng kiểm định Mauchly; nếu giả định này bị vi phạm, hiệu chỉnh Greenhouse-Geisser được sử dụng.

Kích thước hiệu ứng (Cohen's d) và độ lệch chuẩn (ĐLC) của hai nhóm được tính trực tiếp từ dữ liệu điểm số của từng HS (xem bảng 2a). Do hai nhóm có khác biệt có ý nghĩa thống kê về điểm nền Học kì 1, nghiên cứu thực hiện thêm phân tích hiệp phương sai (ANCOVA) đối với điểm ĐGTX lần 2, với ĐTB Học kì 1 làm biến hiệp phương sai, sau khi kiểm tra các giả định về tuyến tính, đồng nhất hệ số hồi quy và đồng nhất phương sai.

2.8. Tiến trình triển khai

Can thiệp được triển khai trong 12 tuần của học kì II năm học 2025-2026 (03/02-26/4/2026). Với mỗi mô đun, quy trình lặp lại theo ba bước: GV kích hoạt Gem tương ứng ngay khi bắt đầu nội dung mới để HS nhóm TN tự ôn tập tại nhà; GV theo dõi báo cáo truy vấn để phát hiện lỗ hổng kiến thức phổ biến; trước mỗi bài ĐGTX, GV tinh chỉnh lại kịch bản Gem liên quan. Quá trình can thiệp gồm ba giai đoạn: Giai đoạn 1 (tuần 1-5, thời gian từ 03/02-08/3/2026) khởi động hệ thống, kích hoạt Gem 1-4 cho các chương Cân bằng hóa học, Nitrogen - sulfur, Đại cương hóa học hữu cơ và Hydrocarbon. Giai đoạn 2 (tuần 6-10, thời gian từ 10/3-02/4/2026) thực hiện ĐGTX1 đồng thời cho hai nhóm: nhóm TN đã sử dụng AI-Chem khoảng 5 tuần và kích hoạt Gem 5-6 cho hai chương "Dẫn xuất halogen - alcohol - phenol" và "Hợp chất carbonyl - carboxylic acid". Giai đoạn 3 (tuần 11-12, thời gian từ 06/4-26/4/2026) ôn tập tích hợp với Gem "Ôn tập Hóa học 11", thực hiện ĐGTX2 và khảo sát Likert.

Việc tinh chỉnh các Gem trong quá trình can thiệp tập trung vào ba nội dung chính: hạn chế cung cấp sẵn phương trình hóa học đã cân bằng; tăng cường hỗ trợ nhận diện và khắc phục nhầm lẫn về nhóm chức chính; mở rộng hệ thống giàn giáo theo nhiều bước đối với các bài toán tổng hợp hữu cơ.

3. Kết quả nghiên cứu

3.1. Trải nghiệm và đánh giá của học sinh đối với hệ thống AI-Chem

Khảo sát sau 12 tuần trải nghiệm AI-Chem trên nhóm TN bằng thang đo Likert 5 mức cho thấy, đánh giá tích cực ở cả bốn phương diện (xem bảng 1), với ĐTB các tiêu chí đều trên 4,3/5,0. Hệ số alpha toàn thang đo (12 tiêu chí) đạt 0,904 (ở mức "Rất tốt"); bốn tiểu thang thành phần đạt lần lượt 0,806 (Trải nghiệm kĩ thuật), 0,798 (Tính cá nhân hóa), 0,836 (Hiệu quả học tập) và 0,854 (Thái độ, động lực) - tất cả đều trong khoảng 0,798-0,854, đạt ngưỡng "tốt" trở lên theo quy ước phổ biến (Cronbach's Alpha ≥ 0,70). Kết quả này cho thấy thang đo có độ nhất quán nội tại tốt trong mẫu nghiên cứu.

Bảng 1. Đánh giá của HS về hệ thống AI-Chem theo bốn phương diện (3 tiêu chí/phương diện; N = 31)

Phương diện	Các tiêu chí khảo sát thành phần	ĐTB (khoảng)	ĐLC (khoảng)	Tỉ lệ % đồng ý và rất đồng ý (khoảng)
1. Trải nghiệm kĩ thuật (α = 0,806)	Dễ truy cập và sử dụng trên website; tốc độ phản hồi nhanh; giao diện thân thiện, câu hỏi rõ ràng	4,34-4,74	0,51-0,75	83,8-96,7
2. Cá nhân hóa (α = 0,798)	Câu hỏi phù hợp với năng lực hiện tại; khi trả lời sai, AI giải thích và gợi ý dễ hiểu; độ khó tăng dần, không nhàm chán	4,52-4,87	0,34-0,68	87,1-100
3. Hiệu quả học tập (α = 0,836)	Hiểu rõ hơn khái niệm khó (cân bằng, hữu cơ); phát hiện lỗ hổng kiến thức chưa biết; tự tin hơn khi làm bài kiểm tra.	4,48-4,71	0,53-0,68	87,1-96,7
4. Thái độ, động lực (α = 0,854)	Học với AI Gems thú vị hơn làm bài trên giấy/PDF; chủ động dành nhiều thời gian tự học hơn; sẽ tiếp tục sử dụng và giới thiệu cho bạn bè	4,55-4,83	0,45-0,62	90,3-96,7

Các kết quả này phù hợp với định hướng thiết kế dựa trên khung lí thuyết ở mục 2.1: nhóm "hiệu quả học tập" (ĐTB từ 4,48-4,71) phù hợp với lí thuyết phản hồi của Hattie và Timperley (2007); nhóm "Thái độ, động lực" (ĐTB từ 4,55-4,83) phản ánh lí thuyết tự quyết của Deci và Ryan (2000); việc 90,3% HS tự báo cáo dành nhiều thời gian tự học hơn là phù hợp với định hướng thiết kế theo hướng tự điều chỉnh học tập của Zimmerman (2002), song đây là dữ liệu tự báo cáo nên chưa phải bằng chứng trực tiếp về năng lực tự điều chỉnh. Cần lưu ý toàn bộ số liệu ở bảng 1 là tự báo cáo, nên phản ánh cảm nhận của HS chứ không xác nhận mức độ tuân thủ kịch bản Socratic.

3.2. Ghi nhận của giáo viên trong quá trình triển khai AI-Chem

Quá trình vận hành AI-Chem cho thấy một số vấn đề liên quan đến hoạt động học tập của HS, được GV ghi nhận qua nhật kí dạy học. Các ghi nhận này mang tính thực hành, chưa phải dữ liệu định tính được thu thập và phân tích

theo một quy trình có cấu trúc. Những lỗi sai được ghi nhận tập trung chủ yếu ở hai nội dung: Hydrocarbon (đánh số mạch chính, xác định đồng phân, phân biệt phản ứng thế và phản ứng cộng) và Cân bằng hóa học (vận dụng nguyên lí Le Chatelier để dự đoán chiều chuyển dịch cân bằng). Các lỗi này là căn cứ thực tiễn để GV điều chỉnh và tinh chỉnh các Gem trong quá trình can thiệp. Bên cạnh đó, việc một số HS có xu hướng đưa nguyên đề bài và yêu cầu AI cung cấp đáp án cho thấy sự cần thiết của nguyên tắc "chẩn đoán trước", theo đó Gem yêu cầu HS trình bày cách hiểu hoặc hướng giải quyết trước khi đưa ra gợi ý.

3.3. Hiệu quả đối với kết quả học tập

Ở thời điểm nền (Học kì 1), nhóm ĐC có ĐTB cao hơn nhóm TN (8,34 so với 7,66; F = 8,310; p = 0,0053). Đây là số liệu duy nhất được thu thập hoàn toàn trước can thiệp và được sử dụng trong phân tích ANCOVA. Tại ĐGTX1, ĐTB của hai nhóm gần như tương đương (7,46 và 7,45; F = 0,001; p = 0,976). Tuy nhiên, theo tiến trình triển khai ở mục 2.8, AI-Chem đã được đưa vào sử dụng từ tuần đầu tiên, trong khi ĐGTX1 được thực hiện trong khoảng tuần 6-10. Vì vậy, ĐGTX lần 1 không được xem là phép đo trước can thiệp và sự thu hẹp chênh lệch điểm giữa hai nhóm chỉ được xem xét ở mức độ mô tả. Hiện tượng này có thể liên quan đến nhiều yếu tố, như việc nhóm TN đã sử dụng AI-Chem trước ĐGTX1, phạm vi kiến thức của ĐGTX1 hẹp hơn so với ĐTB Học kì 1; tuy nhiên, nghiên cứu hiện tại chưa đủ cơ sở để xác định mức độ đóng góp của từng yếu tố (xem bảng 2a, 2b).

Bảng 2a. So sánh liên nhóm (ANOVA một yếu tố, hai nhóm độc lập) tại từng thời điểm

Thời điểm	N (ĐC /TN)	ĐTB (ĐLC) ĐC	ĐTB (ĐLC) TN	F (df)	p	Cohen's d	Kết luận
ĐGTX lần 1 (giữa can thiệp)	35/32	7,46 (1,44)	7,45 (1,33)	0,001 (1,65)	0,976	−0,01	Không khác biệt
ĐGTX lần 2 (sau can thiệp)	35 / 31	8,13 (1,54)	8,82 (1,05)	4,467 (1,64)	0,038	0,52	Khác biệt có ý nghĩa (mức trung bình)

Bảng 2b. Biến thiên điểm số nội nhóm TN qua ba giai đoạn (ANOVA đo lặp trên dữ liệu ghép theo từng HS, N = 31)

Giai đoạn (lớp TN)	N	ĐTB	ĐLC	F (df); p - đo lặp 3 giai đoạn
Nền tảng (ĐTB Học kì 1), trước can thiệp	31	7,71	0,94	F(2, 60) = 17,358; p < 0,001; η^2p = 0,367
Giữa can thiệp (ĐGTX1)	31	7,53	1,25	
Sau can thiệp (ĐGTX2)	31	8,82	1,05	

Phân tích ĐTB Học kì 1 làm biến hiệp phương sai cho thấy, khác biệt giữa hai nhóm ở ĐGTX2 vẫn có ý nghĩa thống kê và rõ hơn so với so sánh thô: F(1, 63) = 9,359; p = 0,003; η^2p = 0,129; ĐTB hiệu chỉnh của nhóm TN là 8,99 so với 7,98 của nhóm ĐC. Giả định đồng nhất hệ số hồi quy giữa hai nhóm được kiểm tra và không bị vi phạm (F(1, 62) = 3,493; p = 0,066).

Điểm của nhóm TN biến thiên có ý nghĩa thống kê qua ba giai đoạn: 7,71 (điểm nền Học kì 1), 7,53 (ĐGTX1, giữa can thiệp) và 8,82 (ĐGTX2), với F(2, 60) = 17,358; p < 0,001; η^2p = 0,367 (xem bảng 2b); kiểm định Mauchly W = 0,985; p = 0,808 nên giả định cầu tính được đáp ứng. So sánh hậu kiểm có hiệu chỉnh Bonferroni cho thấy, mức tăng tập trung ở giai đoạn cuối: điểm gần như không đổi từ Học kì 1 đến ĐGTX1 (p = 1,000), nhưng tăng 1,29 điểm từ ĐGTX1 đến ĐGTX2 (t(30) = 5,154; p < 0,001). Tuy nhiên, nhóm ĐC cũng tiến bộ trong cùng khoảng thời gian (tăng 0,67 điểm; t(34) = 3,079; p = 0,004), nên mức tăng nội nhóm không được quy hoàn toàn cho AI-Chem; căn cứ chính vẫn là so sánh liên nhóm ở ĐGTX2, nơi khác biệt có ý nghĩa thống kê với kích thước hiệu ứng trung bình (d = 0,52; xem bảng 2a) và vẫn giữ nguyên sau khi kiểm soát điểm nền bằng ANCOVA. ĐLC của nhóm TN giảm từ 1,25 xuống 1,05, nhưng đây là quan sát mô tả chưa được kiểm định.

Xét theo mức điểm, tại ĐGTX1 (giữa can thiệp), có 3/32 HS (9,4%) nhóm TN đạt dưới 5 điểm; ở ĐGTX2 không còn HS nào ở mức này (0/31), trong khi nhóm ĐC vẫn còn 2/35 HS (5,7%). Tỉ lệ HS đạt từ 8 điểm trở lên ở ĐGTX2 cũng cao hơn ở nhóm TN (26/31; chiếm 83,9%) so với nhóm ĐC (25/35; chiếm 71,4%). Do số trường hợp rất nhỏ và các thay đổi về phân loại chưa được kiểm định thống kê, đây chỉ là xu hướng mô tả, chưa cho phép quy kết nhân quả cho hệ thống AI-Chem. Quá trình TN cũng ghi nhận rủi ro về "ảo giác AI" khi mô hình đôi lúc nhầm lẫn ở danh pháp phức tạp cùng độ trễ đường truyền. Ảo giác AI được xem là hệ quả cấu trúc của cách LLM sinh văn bản, dễ

xảy ra với thuật ngữ chuyên ngành hẹp như danh pháp hóa học (Danyaro và cộng sự, 2025); vì vậy AI là công cụ bổ trợ, đòi hỏi kĩ năng kiểm chứng chéo từ cả người dạy và người học.

3.4. Một số đánh giá chung

Nhìn chung, kết quả nghiên cứu của bài báo cho thấy, nhóm sử dụng AI-Chem có sự cải thiện về kết quả học tập và HS có phản hồi tích cực đối với quá trình tương tác với hệ thống. Chiều hướng cải thiện kết quả học tập phù hợp với một số bằng chứng quốc tế gần đây về việc ứng dụng AI trong giáo dục. Nghiên cứu bán TN của Alneyadi và Wardat (2023) trên HS lớp 11 môn Vật lí tại UAE ghi nhận tác động tích cực của công cụ AI đối với kết quả học tập; phân tích tổng hợp 35 nghiên cứu TN của Wu và cộng sự (2026), với tổng cỡ mẫu 4 193 người học, cũng cho thấy hiệu quả tích cực ($g = 0,670$), trong đó hiệu ứng lớn hơn ở bậc trung học. Nghiên cứu của Babalola và cộng sự (2024) về sinh viên học Hóa học hữu cơ sử dụng thiết kế khảo sát tương quan nên kết quả không cho phép suy luận quan hệ nhân quả. Do các nghiên cứu trên khác nhau về môn học, cấp học và loại công cụ AI, sự tương đồng này chỉ cho thấy kết quả của nghiên cứu hiện tại phù hợp với xu hướng chung được ghi nhận trong các nghiên cứu trước, chưa đủ cơ sở để khái quát về hiệu quả của giàn giáo Socratic trong dạy học Hóa học 11.

Bên cạnh kết quả học tập, phản hồi của HS cho thấy mức độ chấp nhận tích cực đối với AI-Chem, trong đó 100% HS đồng ý rằng khi trả lời sai, AI đưa ra giải thích và gợi ý dễ hiểu. Tuy nhiên, đây là dữ liệu tự báo cáo, phản ánh cảm nhận của HS chứ chưa phải bằng chứng khách quan về mức độ các Gem tuân thủ kịch bản Socratic. Đồng thời, kết quả về điểm số cần được diễn giải thận trọng do cỡ mẫu còn nhỏ ($N = 66$), mức ý nghĩa thống kê quan sát được khá gần ngưỡng ($p = 0,038$), hai nhóm không tương đương về điểm nền Học kì 1 và được hình thành từ hai lớp thuộc các ban học khác nhau. Mặc dù ANCOVA đã được sử dụng để kiểm soát sự khác biệt về điểm nền, phương pháp này không loại trừ được ảnh hưởng của các yếu tố nhiễu khác gắn với đặc điểm lớp học và người học. Vì vậy, các kết quả hiện tại cung cấp bằng chứng ban đầu về tiềm năng của AI-Chem trong hỗ trợ ôn tập Hóa học 11, nhưng cần được kiểm chứng trên mẫu lớn hơn và với thiết kế nghiên cứu có khả năng kiểm soát tốt hơn các yếu tố nhiễu.

4. Kết luận và bình luận

Bài báo đã xây dựng và triển khai AI-Chem như một hệ thống hỗ trợ ôn tập Hóa học 11, trong đó các Gem được cấu hình trên nền tảng Gemini theo phương pháp Socratic nhằm cung cấp phản hồi và gợi ý từng bước cho HS. Kết quả nghiên cứu cho thấy, nhóm TN có sự cải thiện về kết quả học tập trong quá trình can thiệp; sau khi kiểm soát điểm nền Học kì 1 bằng ANCOVA, điểm ĐGTX2 của nhóm TN cao hơn nhóm ĐC với sự khác biệt có ý nghĩa thống kê. Phản hồi của HS cũng cho thấy mức độ chấp nhận tích cực đối với AI-Chem, đặc biệt ở khả năng cung cấp giải thích và gợi ý khi HS gặp khó khăn. Những kết quả này cung cấp bằng chứng bước đầu về tiềm năng của AI-Chem trong hỗ trợ ôn tập và cá nhân hóa phản hồi trong dạy học Hóa học 11. Tuy nhiên, do nghiên cứu không có nhóm ĐC sử dụng LLM với câu lệnh thông thường, chưa thể xác định riêng tác động của kịch bản Socratic đối với kết quả thu được.

Tuy nhiên, bài báo còn một số hạn chế. Cỡ mẫu còn nhỏ và được thực hiện tại một trường THPT chuyên nên khả năng khái quát hóa chưa cao; hai nhóm được hình thành từ các lớp nguyên vẹn, không tương đương về điểm nền và có sự khác biệt về tỉ lệ HS định hướng lựa chọn tổ hợp xét tuyển đại học có môn Hóa học. ANCOVA giúp kiểm soát sự khác biệt về điểm nền nhưng không loại trừ được hiệu ứng lớp và các yếu tố nhiễu không quan sát được. Bên cạnh đó, ĐGTX1 được thực hiện sau khi AI-Chem đã được đưa vào sử dụng nên nghiên cứu không phải là thiết kế tiền kiểm - hậu kiểm đầy đủ; dữ liệu về trải nghiệm của HS chủ yếu là tự báo cáo; mức độ tuân thủ kịch bản Socratic của các Gem chưa được kiểm chứng bằng phân tích có hệ thống dữ liệu hội thoại; hiệu ứng mới lạ và sự khác biệt về thời lượng ôn tập giữa hai nhóm cũng chưa được kiểm soát. Vì vậy, kết quả hiện tại chưa đủ cơ sở để khẳng định quan hệ nhân quả giữa việc sử dụng AI-Chem và sự cải thiện kết quả học tập. Các nghiên cứu tiếp theo cần mở rộng quy mô mẫu trên nhiều lớp và nhiều trường THPT, bao gồm cả trường không chuyên; tăng cường kiểm soát các yếu tố nhiễu thông qua thiết kế TN chặt chẽ hơn và phân nhóm ngẫu nhiên ở từng cấp học hoặc lớp học khi điều kiện cho phép. Đồng thời, cần thu thập dữ liệu theo thời gian và dữ liệu tương tác để phân tích quá trình học tập của HS; phân tích lịch sử hội thoại theo rubric nhằm đánh giá mức độ các Gem tuân thủ kịch bản Socratic và ghi nhận thời lượng tự học thực tế; so sánh trực tiếp các chiến lược thiết kế câu lệnh khác nhau để xác định rõ hơn đóng góp của kịch bản Socratic đối với hiệu quả của AI-Chem.

Tuyên bố về vai trò của các tác giả: *Lê Thị Thúy Phượng: Lên ý tưởng nghiên cứu, thiết kế và xây dựng nền tảng AI-Chem trên Google Sites. Hồ Thị Bảo Châu: Lên ý tưởng nghiên cứu, triển khai thực nghiệm sư phạm, xử lí số liệu thống kê. Phạm Ngọc Sơn: Khảo sát thực trạng trước can thiệp, xử lí số liệu thống kê. Đỗ Thị Ngọc Loan: Tổng*

quan tài liệu và xây dựng khung lí thuyết. Đặng Thị Thuận An: Triển khai thực nghiệm, thiết kế công cụ đo lường, giám sát, điều phối chung và viết chính bản thảo. Tất cả tác giả cùng phân tích, diễn giải kết quả, chỉnh sửa và phê duyệt bản thảo cuối cùng.

Tuyên bố về GenAI và Quyền tác giả: *Các mô hình ngôn ngữ lớn được sử dụng với hai vai trò tách biệt: (1) Đối tượng nghiên cứu: Gemini được cấu hình thành "gia sư ảo" (Gems) để học sinh tương tác dưới sự kiểm soát của kịch bản sư phạm; (2) Công cụ hỗ trợ nghiên cứu: Sử dụng Gemini để hỗ trợ tổng hợp tài liệu tham khảo, lập dàn ý và viết mã nguồn phân tích dữ liệu. Các tác giả đã xác minh độc lập toàn bộ các thông tin do AI tạo ra và chịu trách nhiệm hoàn toàn về tính trung thực, đạo đức khoa học và kết luận của công trình.*

Tuyên bố về xung đột lợi ích: *Các tác giả tuyên bố không có xung đột lợi ích.*

Thông tin tài trợ: *Nghiên cứu này được tài trợ bởi Trường Đại học Sư phạm - Đại học Huế trong nhiệm vụ mã số NCTB-T.26-XH.503.01.*

Lời cảm ơn: *Nhóm tác giả cảm ơn Ban Giám hiệu, Tổ chuyên môn Hóa học và học sinh Trường THPT Chuyên Quốc Học Huế, phường Thuận Hóa, thành phố Huế đã hỗ trợ, đồng hành thực hiện các nội dung thực nghiệm của nghiên cứu này; cảm ơn sự tài trợ của Trường Đại học Sư phạm - Đại học Huế trong nhiệm vụ mã số NCTB-T.26-XH.503.01.*

Tài liệu tham khảo

Alneyadi, S., & Wardat, Y. (2023). ChatGPT: Revolutionizing student achievement in the electronic magnetism unit for eleventh-grade students in Emirates schools. *Contemporary Educational Technology*, *15*(4), ep448. https://doi.org/10.30935/cedtech/13417

Babalola, V. T., Ahmad, S. S., & Tafida, H. S. (2024). ChatGPT in organic chemistry classrooms: Analyzing the impacts of social environment on students' interest, critical thinking and academic achievement. International *Journal of Education and Teaching Zone*, *3*(1), 60-72. https://doi.org/10.57092/ijetz.v3i1.155

Bộ GD-ĐT (2018). *Chương trình giáo dục phổ thông môn Hóa học* (ban hành kèm theo Thông tư số 32/2018/TT-BGDĐT ngày 26/12/2018 của Bộ trưởng Bộ GD-ĐT).

Cain, W. (2024). Prompting change: Exploring prompt engineering in large language model AI and its potential to transform education. *TechTrends*, *68*(1), 47-57. https://doi.org/10.1007/s11528-023-00896-0

Danyaro, K. U., Abdullahi, S., Abdallah, A. S., & Chiroma, H. (2025). Hallucinations in large language models for education: Challenges and mitigation. *International Journal of Teaching, Learning and Education*, *4*(6), 13-19. https://doi.org/10.22161/ijtle.4.6.2

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*(4), 227-268. https://doi.org/10.1207/S15327965PLI1104_01

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112. https://doi.org/10.3102/003465430298487

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, *106*(4), 901-918. https://doi.org/10.1037/a0037123

Maheshwari, G. (2024). Factors influencing students' intention to adopt and use ChatGPT in higher education: A study in the Vietnamese context. *Education and Information Technologies*, *29*, 12167-12195. https://doi.org/10.1007/s10639-023-12333-z

Malik, R., Abdi, D., Wang, R., & Demszky, D. (2025). Scaffolding middle school mathematics curricula with large language models. *British Journal of Educational Technology*, *56*(3), 999-1027. https://doi.org/10.1111/bjet.13571

Ng, D. T. K., Tan, C. W., & Leung, J. K. L. (2024). Empowering student self-regulated learning and science education through ChatGPT: A pioneering pilot study. *British Journal of Educational Technology*, *55*(4), 1328-1353. https://doi.org/10.1111/bjet.13454

Nguyen, T. N. T., Lai, N. V., & Nguyen, Q. T. (2024). Artificial intelligence (AI) in education: A case study on ChatGPT's influence on student learning behaviors. *Educational Process: International Journal, 13*(2), 105-121. https://doi.org/10.22521/edupij.2024.132.7

Qian, Y. (2025). Prompt engineering in education: A systematic review of approaches and educational applications. *Journal of Educational Computing Research*, *63*(7-8), 1782-1818. https://doi.org/10.1177/07356331251365189

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton Mifflin.

Sharma, S., Mittal, P., Kumar, M., & Bhardwaj, V. (2025). The role of large language models in personalized learning: A systematic review of educational impact. *Discover Sustainability*, *6*(1), Article 243. https://doi.org/10.1007/s43621-025-01094-z

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257-285. https://doi.org/10.1207/s15516709cog1202_4

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wang, X., Huang, R., Sommer, M., Pei, B., Shidfar, P., Rehman, M. S., Ritzhaupt, A. D., & Martin, F. (2024). The efficacy of artificial intelligence-enabled adaptive learning systems from 2010 to 2022 on learner outcomes: A meta-analysis. *Journal of Educational Computing Research*, *62*(6), 1348-1383.

Wu, X., Zhu, P., Zhang, J., Yin, M., & Wang, Y. (2026). ChatGPT's impact on student learning outcomes: A meta-analysis of 35 experimental studies. *Humanities and Social Sciences Communications*, *13*, Article 684. https://doi.org/10.1057/s41599-026-07019-z

Zhou, R., He, X., Fan, Q., Li, Y., Li, Y., Xiao, X., & Fang, J. (2025). Exploring ChatGPT-facilitated scaffolding in undergraduates' mathematical problem solving. *Journal of Computer Assisted Learning*, *41*(4), e70077. https://doi.org/10.1111/jcal.70077

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, *41*(2), 91-100. https://doi.org/10.1207/s15430421tip4102_2