DOI: 10.31276/VJST.2026.3884 ***Khoa học tự nhiên*** | *Khoa học máy tính và thông tin*
Khoa học kỹ thuật và công nghệ | *Các lĩnh vực khác của khoa học kỹ thuật và công nghệ*

RFO-HITL: Kiến trúc tác tử AI hỗ trợ ra quyết định cho giảng viên trong giáo dục đại học

Phạm Thu Thuận*
Trường Đại học Thành Đông, 3 Vũ Công Đán, phường Tứ Minh, TP. Hải Phòng, Việt Nam
Ngày nhận bài 23/3/2026; ngày chuyển phản biện 24/3/2026; ngày nhận phản biện 24/4/2026; ngày chấp nhận đăng 6/5/2026

Tóm tắt:
Nhằm giải quyết thách thức về tính minh bạch và trách nhiệm giải trình trong bối cảnh tích hợp trí tuệ nhân tạo tạo sinh (Generative Artificial Intelligence - GenAI) vào giáo dục đại học, nghiên cứu này đề xuất kiến trúc RFO-HITL (Rule-First Orchestration with Human-in-the-Loop, điều phối ưu tiên quy tắc với cơ chế con người giám sát). Đây là một kiến trúc tác tử trí tuệ nhân tạo (AI) dung hòa giữa hiệu quả công nghệ và sự giám sát của con người, ưu tiên sử dụng bộ quy tắc tường minh để ra quyết định, giới hạn vai trò của mô hình ngôn ngữ lớn (LLM). Kiến trúc RFO-HITL được triển khai trên nền tảng không lập trình/ít lập trình với cơ chế con người giám sát (HITL) bắt buộc. Dựa trên phương pháp nghiên cứu khoa học thiết kế (DSR), kiến trúc được phát triển và đánh giá trên tập dữ liệu tổng hợp (N=100). Kết quả định lượng cho thấy, RFO-HITL đạt hiệu quả khả quan trong xác định rủi ro (F1-score 88,9%) và tạo nội dung hỗ trợ chất lượng. Tỷ lệ vi phạm ràng buộc an toàn ở mức 6%, chủ yếu liên quan đến giới hạn độ dài và sắc thái văn phong, qua đó khẳng định một cách thực chứng tầm quan trọng của cơ chế HITL. Nghiên cứu đóng góp vào việc xây dựng kiến trúc RFO-HITL, cung cấp tri thức thiết kế có giá trị thực tiễn cho việc triển khai GenAI có trách nhiệm, đảm bảo tính tường minh và duy trì phán đoán chuyên môn sư phạm của giảng viên.
Từ khóa: **con người giám sát, giáo dục đại học, nền tảng không lập trình/ít lập trình, quy trình tác tử trí tuệ nhân tạo, RFO-HITL.**
Chỉ số phân loại: **1.2, 2.11**

RFO-HITL: An AI agentic architecture for lecturer decision support in higher education

Thu Thuan Pham*
Thanh Dong University, 3 Vu Cong Dan Street, Tu Minh Ward, Hai Phong City, Vietnam
Received 23 March 2026; revised 24 April 2026; accepted 6 May 2026

Abstract:
In response to the challenges of transparency and accountability in integrating generative artificial intelligence (GenAI) into higher education, this study proposes the RFO-HITL (Rule-first orchestration with human-in-the-loop) architecture. This AI agent framework balances technological efficiency with human oversight

by prioritising the adoption of explicit rule sets for decision-making, and limiting the role of large language models (LLMs). In addition, this RFO-HITL architecture is deployed using a no-code/low-code (NCLC) platform. In combination with the mandatory human-in-the-loop (HITL) mechanism. Based on the design science research (DSR) methodology, the architecture was systematically developed and evaluated using a synthetic dataset (N=100). Quantitative results demonstrate that the RFO-HITL architecture is promising in identifying at-risk cases, achieving an F1-score of 88.9%, and generating high-quality student support content. A safety violation rate of 6% - mostly pertaining to length constraints and stylistic or tonal subtleties - was observed, thereby empirically underlining the significance of the HITL mechanism. This research primarily contributes to building the AI agent architecture, offering actionable design knowledge for the responsible implementation of GenAI in educational settings, thereby promoting transparency while preserving lectures pedagogical professional judgment.

Keywords: **artificial intelligence, AI agent workflow, higher education, human-in-the-loop, no-code/low-code.**
Classification numbers: **1.2, 2.11**

1. Đặt vấn đề

Giáo dục đại học đang đối mặt với thách thức nội tại: Làm thế nào để đảm bảo hỗ trợ kịp thời và cá nhân hóa cho sinh viên trong bối cảnh quy mô ngày càng mở rộng. Mặc dù các phương pháp phân tích học tập truyền thống đã đạt được tiến bộ trong việc dự báo rủi ro học thuật bởi M.J. Khan và cs (2023) [1], sự xuất hiện của trí tuệ nhân tạo tạo sinh (GenAI) và các mô hình ngôn ngữ lớn (LLM) đã mở ra tiềm năng to lớn trong việc tự động hóa các quy trình hỗ trợ, đồng thời giảm tải công việc hành chính cho giảng viên.

Tuy nhiên, việc triển khai GenAI trong giáo dục, một lĩnh vực có mức độ ảnh hưởng cao, đòi hỏi sự cẩn trọng tối đa. Các rủi ro của LLM, bao gồm khả năng tạo ra thông tin sai lệch hoặc bịa đặt (hallucinations) và tính chất "hộp đen" (black boxes) trong suy luận, đang đặt ra những thách thức nghiêm trọng về đạo đức và quản trị theo nghiên cứu của N.R. Abdurohman (2025) [2], A.K. Singh và cs (2024) [3]. Điều này làm nổi bật một khoảng trống nghiên cứu quan trọng: sự thiếu hụt các kiến trúc triển khai cụ thể được thiết kế để khai thác tiềm năng của GenAI trong khi vẫn đảm bảo tính minh bạch, trách nhiệm giải trình và đặt sự giám sát của con người làm trung tâm.

Để giải quyết khoảng trống này, nghiên cứu đề xuất và đánh giá kiến trúc tham chiếu RFO-HITL. Kiến trúc này cung cấp một lộ trình thực tiễn để tích hợp GenAI vào

quy trình hỗ trợ sinh viên một cách an toàn và có trách nhiệm, cân bằng giữa hiệu quả công nghệ và trách nhiệm giải trình trong giáo dục. Cụ thể, kiến trúc ưu tiên tính tường minh và sự giám sát của con người, đồng thời giới hạn vai trò của LLM như một trợ lý nhận thức thay vì một tác nhân tự chủ.

Nghiên cứu giải quyết ba vấn đề trọng tâm. Thứ nhất, đề xuất một giải pháp thiết kế cho quy trình tác tử AI nhằm đảm bảo tính tường minh và cơ chế giám sát bắt buộc của con người. Thứ hai, kiến trúc đề xuất được đánh giá về hiệu quả kỹ thuật, chất lượng nội dung và sự tuân thủ các ràng buộc an toàn. Thứ ba, khả năng hỗ trợ truy vết quyết định và trách nhiệm giải trình của kiến trúc cũng được phân tích một cách toàn diện.

Phần còn lại của bài báo được bố cục như sau: phần 2 trình bày chi tiết kiến trúc RFO-HITL và cơ sở lựa chọn thiết kế; phần 3 báo cáo kết quả đánh giá kỹ thuật và thảo luận các phát hiện chính, đóng góp học thuật, ý nghĩa thực tiễn; cuối cùng, phần 4 đưa ra kết luận.

2. Kiến trúc đề xuất RFO-HITL

Mục tiêu của nghiên cứu là thiết kế một quy trình tác tử chủ động (AI agentic workflow) RFO-HITL nhằm tự động hóa việc hỗ trợ ra quyết định cho giảng viên, đảm bảo sự cân bằng giữa hiệu quả công nghệ và quyền tự chủ của nhà giáo dục được đề xuất bởi B. Shneiderman (2020) [4]. Kiến trúc được xây dựng trên tiền đề: trong các lĩnh vực có mức độ tác động cao như giáo dục, việc duy trì tính minh bạch và trách nhiệm giải trình là yêu cầu tiên quyết.

Nghiên cứu này áp dụng phương pháp luận nghiên cứu khoa học thiết kế (DSR) để xây dựng và đánh giá kiến trúc RFO-HITL. Quy trình DSR, theo khung đề xuất của K. Peffers và cs (2007) [5], được lựa chọn vì nó cung cấp một lộ trình hệ thống và chặt chẽ để giải quyết các vấn đề thực tiễn thông qua việc kiến tạo và đánh giá các giải pháp đổi mới (artifacts).

2.1. Công trình nghiên cứu liên quan

2.1.1. Bối cảnh công nghệ và vấn đề đạo đức AI trong giáo dục đại học

Trí tuệ nhân tạo, khởi đầu với các mô hình học máy dự báo (Predictive AI), đang định hình lại các quy trình cốt lõi trong giáo dục đại học. Lĩnh vực phân tích học tập và hệ thống cảnh báo sớm cho phép dự báo kết quả học tập và hỗ trợ can thiệp kịp thời trong nghiên cứu gần đây của M.J. Khan và cs (2023) [1], W. Moore và cs (2024) [6]. Tuy nhiên, các hệ thống này thường hoạt động như những "hộp đen". Sự thiếu hụt khả năng giải thích theo A. Adadi và cs (2018) [7] không chỉ hạn chế trách nhiệm giải trình mà còn cản trở giảng viên chuyển đổi cảnh báo thành các hành động can thiệp sư phạm cụ thể và khả thi. Hơn nữa, nguy cơ thiên kiến thuật toán (algorithmic bias) luôn hiện hữu, đòi hỏi các cơ chế quản trị chặt chẽ [2].

Sự phát triển nhanh chóng của GenAI và LLM đã tạo ra một sự dịch chuyển mô hình. Khác với mô hình AI dự báo, các công nghệ mới này có khả năng tạo ra phản hồi chi tiết, cá nhân hóa cao và đề xuất chiến lược can thiệp phức tạp bằng ngôn ngữ tự nhiên. Sự chuyển đổi này nâng cao tiềm năng tương tác và chuyển dịch vai trò của hệ thống hỗ trợ từ cung cấp thông tin sang tham gia vào quá trình giao tiếp sư phạm.

Tuy nhiên, sự dịch chuyển này làm gia tăng các thách thức về đạo đức và quản trị. Theo W. Moore và cs (2024) [2], các rủi ro mới phát sinh bao gồm khả năng tạo ra thông tin sai lệch nhưng có vẻ đáng tin cậy (hallucinations), ngôn ngữ không phù hợp về mặt sư phạm và nguy cơ gia tăng các định kiến xã hội. Thêm vào đó, sự phụ thuộc quá mức vào các hệ thống tự động có thể làm suy giảm tư duy phản biện và quyền tự chủ trong học thuật [3]. Do đó, yêu cầu cấp thiết là phát triển các mô hình triển khai cụ thể nhằm đảm bảo tính minh bạch và sự cân bằng giữa tối ưu hóa công nghệ và việc bảo vệ các yếu tố con người cốt lõi trong giáo dục [2, 8].

2.1.2. Vai trò của nền tảng no-code/low-code và cơ chế con người giám sát

Sự phát triển của các nền tảng không lập trình hoặc ít lập trình (no-code/low-code - NCLC) đang chuyển đổi công nghệ giáo dục (EdTech) bằng cách trao quyền cho nhà giáo dục tự thiết kế và quản lý các ứng dụng phức tạp qua giao diện trực quan, không cần kỹ năng lập trình chuyên sâu [9, 10]. Gần đây, sự kết hợp giữa NCLC và LLM đã thúc đẩy sự phát triển của các quy trình tác tử chủ động. Tuy nhiên, các khung agentic hiện nay thường được thiết kế để tối đa hóa khả năng tự động hóa và sự tự chủ của AI, dẫn đến những hạn chế đáng kể về tính minh bạch và khả năng kiểm soát [11, 12]. Mức độ tự chủ quá cao đặt ra rủi ro lớn trong bối cảnh giáo dục, nơi các quyết định tự động đòi hỏi khả năng kiểm soát chặt chẽ và sự giám sát bắt buộc của con người để đảm bảo trách nhiệm giải trình.

Sự tham gia của con người (human-in-the-loop - HITL) được xem là nguyên tắc quan trọng khi tích hợp AI vào giáo dục đại học nhằm bảo đảm minh bạch, trách nhiệm giải trình và công bằng theo nghiên cứu của B. Memarian và cs (2024) [13]. Sự giám sát của con người giúp các quyết định do AI hỗ trợ phù hợp với giá trị và tiêu chuẩn đạo đức nghề nghiệp, đồng thời làm rõ cơ chế vận hành AI, góp phần xây dựng lòng tin và giảm thiên kiến thuật toán [14, 15].

2.1.3. Khung lý thuyết và khoảng trống nghiên cứu

Tổng quan tài liệu cho thấy sự hội tụ của nhiều xu hướng công nghệ trong việc tích hợp AI vào giáo dục đại học. Để hiểu sâu hơn về động lực và thách thức khi triển khai các công nghệ này, nghiên cứu sử dụng hai lý thuyết nền tảng. Thứ nhất, lý thuyết hoạt động (activity theory - AT) bởi Y. Engeström (2014) [16] cung cấp góc nhìn để phân tích hệ thống kỹ thuật - xã hội, xem xét cách các công cụ AI đóng vai trò trung gian

trong hoạt động hỗ trợ sinh viên của giảng viên. AT nhấn mạnh rằng việc tích hợp công nghệ mới phải duy trì quyền tự chủ và phán đoán chuyên môn của con người. Thứ hai, lý thuyết tải nhận thức (cognitive load theory - CLT) bởi J. Sweller (1988) [18] giải thích rằng hiệu quả ra quyết định phụ thuộc vào việc quản lý năng lực xử lý thông tin hạn chế của con người. CLT gợi ý rằng các hệ thống hỗ trợ nên giảm "tải nhận thức ngoại lai" (ví dụ như việc tổng hợp dữ liệu hay soạn thảo văn bản) giúp giảng viên có thể tập trung ưu tiên nguồn lực tư duy cho các quyết định và nhiệm vụ sư phạm phức tạp hơn.

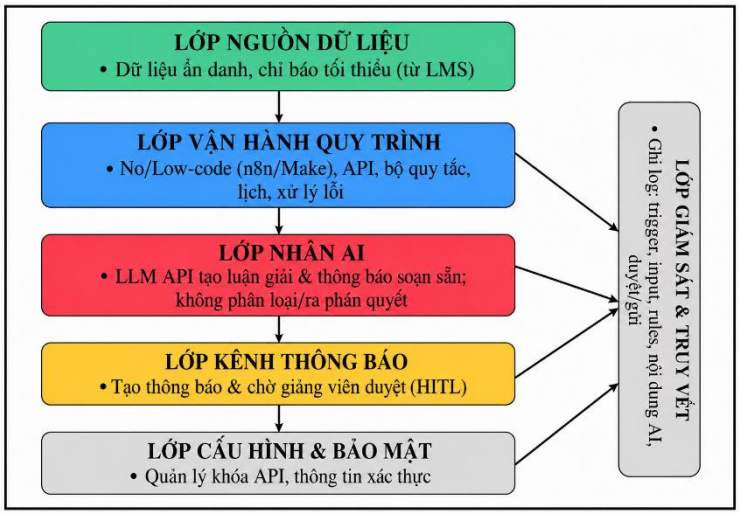
Khi tổng hợp các góc nhìn công nghệ, quản trị và lý thuyết nền (AT, CLT), khoảng trống nghiên cứu trở nên rõ ràng: sự thiếu hụt các kiến trúc tích hợp cần được định hướng bởi cơ sở lý luận vững chắc cho việc tích hợp LLM vào hoạt động hỗ trợ sinh viên. Các nghiên cứu hiện tại chưa giải quyết đầy đủ các rủi ro về đạo đức và vận hành do thiếu sự phối hợp của ba yếu tố then chốt: (1) các nguyên tắc đảm bảo tính tường minh, minh bạch và khả năng kiểm chứng (giải quyết vấn đề "hộp đen"); (2) cơ chế HITL bắt buộc để duy trì trách nhiệm giải trình và đảm bảo quyền tự chủ của giảng viên (phù hợp với AT); (3) chiến lược khai thác NCLC để giảm gánh nặng trong việc xử lý thông tin (phù hợp với CLT) và tăng khả năng tiếp cận cho giảng viên.

Nhằm khắc phục các hạn chế này, bài nghiên cứu đề xuất mô hình RFO-HITL như một giải pháp tích hợp các yếu tố trên. Kiến trúc lồng ghép bộ quy tắc tường minh nhằm đảm bảo tính minh bạch với khả năng xử lý ngôn ngữ linh hoạt của LLM để giảm tải các công việc lặp lại. Bằng cách tích hợp các điểm kiểm soát bắt buộc để con người phê duyệt và sử dụng nền tảng NCLC, RFO-HITL có tiềm năng thiết lập các quy trình hỗ trợ ra quyết định có trách nhiệm, đáng tin cậy và phù hợp với thực tiễn giáo dục đại học [18, 19] .

2.2. Kiến trúc tổng quan và cơ sở lựa chọn thiết kế

2.2.1. Kiến trúc tổng quan

Hiện thực hóa các nguyên tắc đã nêu, một kiến trúc nhiều lớp được đề xuất phân tách rõ các vai trò chức năng và nâng cao tính mô-đun. Quy trình tác tử được tạo thành từ sáu lớp logic phối hợp bao gồm: lớp nguồn dữ liệu, lớp vận hành quy trình, lớp nhân AI, lớp kênh thông báo, lớp giám sát và truy vết, lớp cấu hình và bảo mật. Hình 1 minh họa cấu trúc tổng thể.



Hình 1. Kiến trúc hệ thống 6 lớp của quy trình tác tử RFO-HITL.

2.2.2. Cơ sở lựa chọn thiết kế

Kiến trúc RFO-HITL kết hợp có chủ đích các lựa chọn thiết kế nhằm ưu tiên quản trị và an toàn hơn là tối đa hóa tự động hóa. Các lựa chọn này được lý giải dựa trên các nguyên tắc thiết kế và khung lý thuyết nền tảng:

Điều phối ưu tiên quy tắc: Khác với các khung tác tử hiện đại (ReAct, AutoGen) tập trung vào khả năng tự lập kế hoạch của LLM [11, 12], RFO-HITL sử dụng các quy tắc tường minh làm trung tâm của việc ra quyết định. Lựa chọn này trực tiếp hiện thực hóa tính tường minh. Nó đảm bảo rằng logic cốt lõi của hệ thống hoàn toàn minh bạch và có thể kiểm chứng, giải quyết các lo ngại về "hộp đen" thường thấy trong các hệ thống cảnh báo sớm truyền thống.

Tách biệt vai trò và giới hạn LLM: Vai trò của LLM bị ràng buộc nghiêm ngặt, tuân thủ quyền tự chủ có giới hạn. Việc tách biệt vai trò tạo nội dung và vai trò ra quyết định giúp giảm thiểu đáng kể rủi ro do thông tin bịa đặt hoặc thiên kiến thuật toán. Theo lăng kính của CLT, việc tự động hóa soạn thảo văn bản giúp giảm "tải nhận thức ngoại lai" cho giảng viên.

Tích hợp NCLC và HITL bắt buộc: Toàn bộ quy trình được điều phối trên nền tảng NCLC, hiện thực hóa khả năng tiếp cận, giúp giảng viên không chuyên lập trình có thể kiểm soát logic hệ thống. Đồng thời, cơ chế HITL là bắt buộc trước mọi tác vụ. Theo lăng kính của AT, thiết kế này đảm bảo giảng viên vẫn là "chủ thể" kiểm soát "công cụ trung gian" (AI), duy trì quyền tự chủ chuyên môn.

RFO-HITL là một kiến trúc thiết kế nhằm tăng cường và không thay thế phán đoán chuyên môn của giảng viên. Bảng 1 so sánh kiến trúc này với các cách tiếp cận liên quan.

Bảng 1. So sánh RFO-HITL với các cách tiếp cận liên quan.

Tiêu chí	Hệ thống cảnh báo sớm truyền thống	Khung agentic tổng quát	RFO-HITL (đề xuất)
Cơ chế ra quyết định	Mô hình học máy (hộp đen)	LLM tự lập kế hoạch và thực thi	Quy tắc tường minh (rule engine)
Vai trò của LLM	Thường không áp dụng	Trung tâm (ra quyết định, thực thi)	Giới hạn (chỉ tạo lý giải và bản nháp)
HITL (con người kiểm soát)	Tùy chọn hoặc thụ động (xem dashboard)	Tùy chọn hoặc hạn chế	Bắt buộc trước mọi hành động
Nền tảng triển khai	Yêu cầu lập trình chuyên sâu	Yêu cầu kỹ năng lập trình	NCLC (No-code/Low-code)
Minh bạch & kiểm soát	Thấp đến trung bình	Thấp (khó truy vết chuỗi suy luận LLM)	Cao (quy tắc rõ ràng, audit log chuẩn hóa)

2.3. Các thành phần chính và luồng hoạt động

Phần này trình bày chi tiết các thành phần kỹ thuật cốt lõi cấu thành kiến trúc RFO-HITL và luồng hoạt động của chúng.

Nguồn dữ liệu và cách tiếp cận bảo vệ quyền riêng tư: Nền tảng của kiến trúc là dữ liệu đầu vào. Tuân thủ nguyên tắc tối thiểu hóa dữ liệu và đạo đức ngay từ khâu thiết kế [20, 21], hệ thống chỉ sử dụng một tập hợp tối thiểu các chỉ báo hành vi từ LMS. Toàn bộ dữ liệu được phi định danh hóa (pseudonymised) ngay tại nguồn bằng mã định danh đã được mã hóa băm (hashed ID), đảm bảo quy trình AI không xử lý thông tin nhận dạng cá nhân.

Nền tảng vận hành và bộ quy tắc: Lớp vận hành quy trình đóng vai trò là trung tâm điều phối. Việc sử dụng một bộ quy tắc tường minh để phân loại và ra quyết định đảm bảo tính tường minh (interpretability) và khả năng truy vết (auditability) (ví dụ: *NẾU điểm_trung_bình < 50 VÀ tỷ_lệ_nộp_bài < 0.6 THÌ phân_loại = 'Rủi_ro_cao'*).

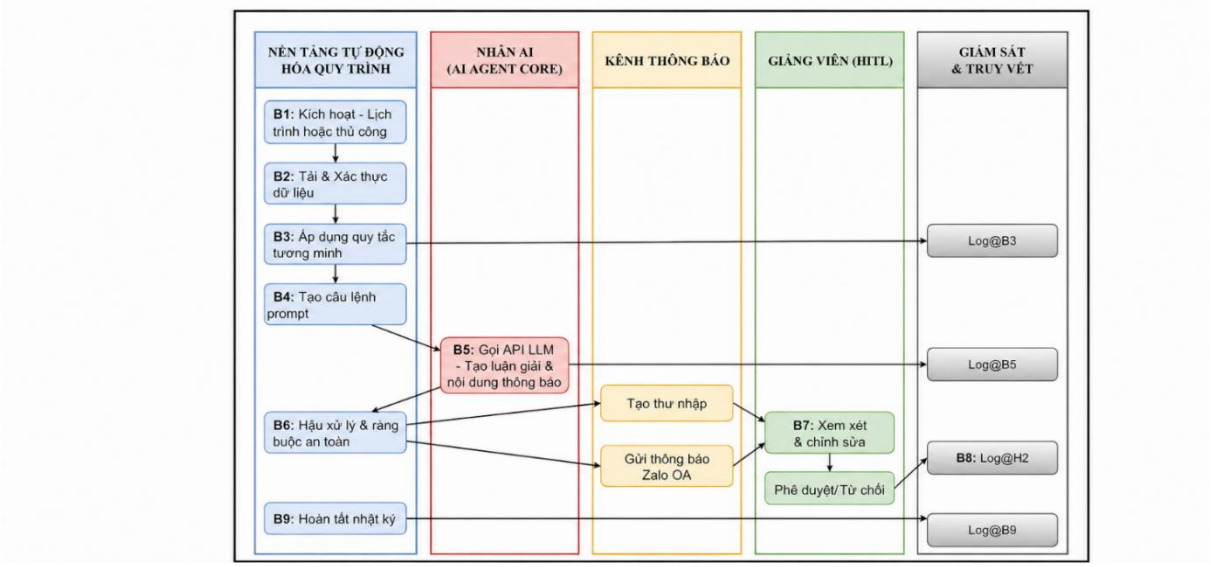
Lớp nhân AI và kỹ thuật thiết kế câu lệnh prompt: Vai trò LLM được giới hạn nghiêm ngặt theo nguyên tắc "tự chủ có giới hạn". LLM hoạt động như một trợ lý nhận thức, chịu trách nhiệm tổng hợp thông tin và soạn thảo văn bản. Để đảm bảo đầu ra nhất quán, kỹ thuật thiết kế câu lệnh (prompt engineering) có cấu trúc chặt chẽ được áp dụng, bao gồm các thành phần như chỉ thị hệ thống, ngữ cảnh, nhiệm vụ và ràng buộc đầu ra. Bảng 2 minh họa một ví dụ về cấu trúc câu lệnh prompt được sử dụng trong nghiên cứu.

Bảng 2. Cấu trúc prompt chi tiết và lược đồ đầu ra JSON.

Phần Prompt	Nội dung ví dụ (minh họa)
System	Bạn là một trợ lý AI cho giảng viên, luôn giao tiếp với giọng văn chuyên nghiệp, tích cực và hỗ trợ.
Context	Dữ liệu sinh viên: *{ "avg_score":4.5, "assignment_submission_rate": 0.5, "risk_level": "high" }*
Task	Hãy tạo một bản luận giải ngắn gọn giải thích các chỉ số trên và soạn nội dung thông báo gửi đến sinh viên để đề nghị hỗ trợ.
Constraints	Đầu ra phải là một đối tượng định dạng JSON. Không sử dụng các từ tiêu cực như "thất bại", "yếu kém". Email không được dài quá 150 từ. *Schema: { "rationale": string, "email_draft": string }*

Kênh thông báo và giao diện HITL: Thành phần này là cơ chế thực thi cho nguyên tắc HITL. Hệ thống được thiết kế để tích hợp liền mạch vào quy trình làm việc hiện có của giảng viên (tạo thư nháp trực tiếp trong email và gửi thông báo tức thời), giúp giảm tải công việc thủ công đồng thời duy trì quyền tự chủ chuyên môn.

Luồng hoạt động của quy trình RFO-HITL diễn ra qua một chuỗi các giai đoạn tuần tự. Sơ đồ tuần tự trong hình 2 minh họa chi tiết các tương tác giữa các thành phần hệ thống và người dùng cuối (giảng viên). Quy trình hoạt động bao gồm các giai đoạn chính sau: *Kích hoạt và thu thập dữ liệu (B1-B2); Phân tích dựa trên quy tắc (B3); Tạo nội dung hỗ trợ bởi AI (B4-B6); Kiểm soát và thực thi bởi giảng viên (B7-B8); Giám sát và truy vết (B9)*.



Hình 2. Sơ đồ tuần tự tương tác giữa các thành phần trong hệ thống.

Luồng hoạt động này thể hiện rõ đặc trưng của kiến trúc RFO-HITL: quyết định dựa trên quy tắc (giai đoạn 2), nội dung được tăng cường bởi AI (giai đoạn 3) và hành động cuối cùng luôn được kiểm soát bởi con người (giai đoạn 4).

2.4. Xây dựng tập dữ liệu tổng hợp

Nhằm đảm bảo quyền riêng tư của sinh viên và cho phép kiểm soát chặt chẽ các điều kiện thử nghiệm trong giai đoạn đánh giá kỹ thuật, nghiên cứu sử dụng một tập dữ liệu tổng hợp. Tập dữ liệu gồm 100 hồ sơ sinh viên được tạo ra bằng ngôn ngữ lập trình Python (sử dụng thư viện Faker và NumPy) để mô phỏng các chỉ báo hành vi dựa trên các phân phối thống kê thường thấy trong phân tích học tập.

Tập dữ liệu có cấu trúc chủ đích để bao quát các kịch bản đa dạng: 30% trường hợp rủi ro cao, 40% rủi ro trung bình, 20% rủi ro thấp và 10% là các trường hợp ngoại lệ (ví dụ: dữ liệu thiếu hoặc giá trị bất thường) để kiểm tra hiệu suất của hệ thống.

3. Kết quả và bàn luận

Phần này trình bày kết quả chứng minh tính khả thi của kiến trúc thông qua một kịch bản sử dụng cụ thể, tiếp theo là các phát hiện từ quá trình đánh giá kỹ thuật dựa trên các tiêu chí đã xác định.

3.1. Mô tả kịch bản minh họa

Quy trình hoạt động đầu-cuối (end-to-end) của kiến trúc RFO-HITL được minh họa thông qua một kịch bản xử lý hồ sơ từ tập dữ liệu tổng hợp (*Sinh viên SV042*), mô phỏng tình huống có rủi ro học thuật cao. Kịch bản này chứng minh tính khả thi của thiết kế.

- Giai đoạn 1 và 2: Thu thập dữ liệu và áp dụng quy tắc.

Quy trình được kích hoạt theo lịch trình. Dữ liệu ẩn danh của sinh viên SV042 được tải vào hệ thống (hình 3). Các chỉ số cho thấy dấu hiệu cần quan tâm: điểm trung bình (*avg_score*): 4,5; tỷ lệ nộp bài (*assignment_submission_rate*): 0,5 (50%) và số lần đăng nhập (*login_count*): 3 (thấp).

Bộ quy tắc tường minh trên nền tảng NCLC áp dụng logic nghiệp vụ:

NẾU (avg_score <50) VÀ (assignment_submission_rate <0.6) THÌ risk_level = 'High'

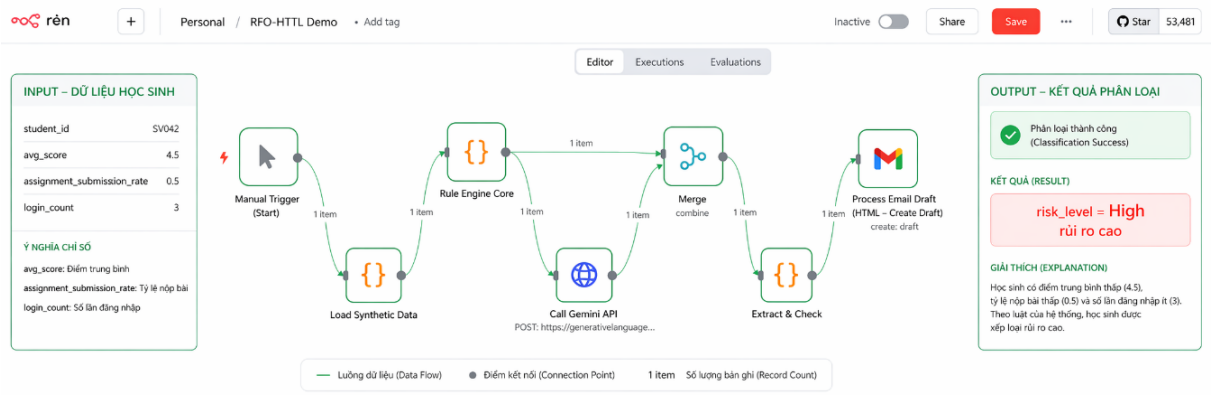
Do thỏa mãn điều kiện, hệ thống phân loại SV042 vào nhóm "rủi ro cao". Quyết định này và dữ liệu đầu vào được ghi nhận vào nhật ký kiểm chứng.

- Giai đoạn 3: Tạo nội dung hỗ trợ bởi AI.

Dữ liệu có cấu trúc và mức độ rủi ro được đưa vào câu lệnh prompt được kiểm soát chặt chẽ và gửi đến lớp nhân AI (Gemini). LLM trả về một đối tượng JSON gồm:

(1) Luận giải cho giảng viên: "Sinh viên Nguyễn Phương N hiện có điểm trung bình… Điều này cho thấy em cần hỗ trợ để cải thiện kết quả học tập.".

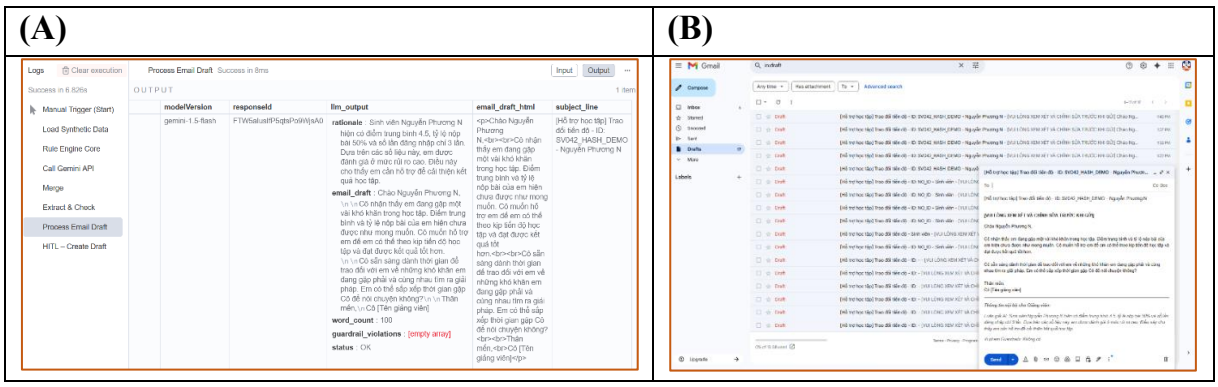
(2) Email nháp gửi sinh viên: "Chào Nguyễn Phương N,\n\nCô nhận thấy em đang gặp một vài khó khăn trong học tập… Em có thể sắp xếp thời gian gặp Cô để nói chuyện không?\n\nThân mến,\n\nCô [Tên giảng viên]".



Hình 3. Minh họa dữ liệu đầu vào và kết quả phân loại của bộ quy tắc tường minh, hồ sơ SV042 được xác định là "rủi ro cao" (risk_level) dựa trên các chỉ số hiệu suất học tập.

- Giai đoạn 4: Kiểm soát bởi giảng viên (HITL).

Hệ thống không tự động gửi email. Thay vào đó, nó tạo một thư nháp trong Gmail của giảng viên (hình 4). Giảng viên xem xét luận giải, thực hiện chỉnh sửa để cá nhân hóa thông điệp trước khi quyết định gửi đi.



Hình 4. Giao diện HITL hiển thị một thư nháp trong Gmail. (A) Thông tin nội bộ dành riêng cho giảng viên tham khảo; **(B)** Email được soạn sẵn để gửi cho sinh viên.

3.2. Kết quả đánh giá

Kết quả đánh giá kỹ thuật (bảng 3) trên tập dữ liệu tổng hợp (N=100) cho thấy, kiến trúc RFO-HITL thể hiện hiệu quả hứa hẹn trong việc hỗ trợ ra quyết định, tuy nhiên vẫn tồn tại những hạn chế cần cải thiện về độ chính xác, tuân thủ ràng buộc và hiệu năng.

Về hiệu quả và chất lượng nội dung, hệ thống tạo ra các đề xuất hỗ trợ được đánh giá cao, với điểm rubric trung bình đạt 4,3/5 và độ tin cậy đảm bảo (Cohen's Kappa

0,82). Điều này khẳng định vai trò hiệu quả của LLM như một trợ lý nhận thức trong việc soạn thảo nội dung phù hợp với ngữ cảnh sư phạm. Tuy nhiên, tính chính xác của bộ quy tắc cốt lõi trong việc phân loại rủi ro (F1-score 88,9%) chưa đạt mục tiêu (>90%), gợi ý cần tinh chỉnh logic nghiệp vụ để cải thiện khả năng bao quát các trường hợp phức tạp hoặc trường hợp biên.

Về tuân thủ an toàn, tỷ lệ vi phạm tổng thể (6,0%) vượt ngưỡng cho phép (<5%). Mặc dù hệ thống ngăn chặn thành công việc sử dụng từ ngữ cấm (0%), các vi phạm chủ yếu liên quan đến giới hạn độ dài (4,0%) và các sai lệch nhỏ về sắc thái giọng điệu (2,0%). Đây là một phát hiện quan trọng, chỉ ra thách thức trong việc kiểm soát hoàn toàn các sắc thái ngôn ngữ phức tạp của LLM, đồng thời củng cố mạnh mẽ sự cần thiết của cơ chế HITL để đảm bảo chất lượng và sự phù hợp sư phạm.

Về hiệu suất kỹ thuật và quản trị, kiến trúc chứng tỏ tính vững chắc khi xử lý các trường hợp biên (85%) và đảm bảo khả năng truy vết tuyệt đối (100%), cung cấp nền tảng vững chắc cho trách nhiệm giải trình và quản trị. Hạn chế đáng kể nhất là hiệu năng, với độ trễ P95 (3,3 giây) cao hơn nhiều so với mục tiêu (<2 giây), cho thấy việc gọi API của LLM là điểm nghẽn chính cần được tối ưu hóa.

Bảng 3. Tóm tắt kết quả đánh giá kỹ thuật (N=100).

Tiêu chí đánh giá	**Chỉ số đo lường**	**Mục tiêu**	**Kết quả thực tế**	**Trạng thái**
Tính chính xác của bộ quy tắc	F1-score (nhóm rủi ro cao)	>90%	88,9%	Chưa đạt
Chất lượng nội dung AI	B.0. IRR (Cohen's Kappa)	>0,8	0,82	Đạt
	B.1. Điểm Rubric trung bình (1-5)	>4,0	4,3/5	Đạt
Tuân thủ ràng buộc an toàn	Tỷ lệ vi phạm tổng thể	<5%	6,0%	Chưa đạt
	Chi tiết:			
	C.1. Sử dụng từ ngữ cấm	0%	0%	Đạt
	C.2. Giới hạn độ dài	<3%*	4,0%	Chưa đạt
	C.3. Vi phạm giọng điệu tinh vi	<3%*	2,0%	Đạt
Tính vững chắc	Tỷ lệ xử lý thành công trường hợp biên	>80%	85%	Đạt
Hiệu năng	Độ trễ P95	<2 giây	3,3 giây	Chưa đạt
Khả năng truy vết	Tỷ lệ quyết định có thể truy vết	100%	100%	Đạt

3.3. Bàn luận

Mục tiêu tổng quát của nghiên cứu là thiết kế và đánh giá một mẫu quy trình agentic (RFO-HITL) ưu tiên tính minh bạch, khả năng kiểm soát và phù hợp với bối cảnh giáo dục đại học. Kết quả đánh giá cung cấp bằng chứng thực nghiệm để trả lời các câu hỏi nghiên cứu đã đặt ra. Kiến trúc đã chứng tỏ hiệu quả cao trong các chức năng cốt lõi. Mặc dù độ chính xác của bộ quy tắc (F1-score 88,9%) thấp hơn so với mục tiêu, chất lượng nội dung do LLM tạo ra lại vượt trội, được đánh giá rất cao (điểm rubric 4,3/5) với độ tin cậy ở mức tốt (Cohen's Kappa =0,82). Về mặt quản trị, kiến trúc thể hiện thế mạnh vượt trội với khả năng truy vết tuyệt đối (100%) và tính mạnh mẽ trong vận hành (xử lý 85% trường hợp biên), khẳng định nền tảng kỹ thuật vững chắc cho việc đảm bảo trách nhiệm giải trình.

Tuy nhiên, phát hiện quan trọng nhất nằm ở các chỉ số chưa đạt mục tiêu về an toàn và hiệu năng. Tỷ lệ vi phạm ràng buộc an toàn là 6,0%, chủ yếu do vượt giới hạn độ dài (4,0%), đã cho thấy một góc nhìn có giá trị: ngay cả với các kỹ thuật prompt chặt chẽ, việc kiểm soát hoàn toàn đầu ra của LLM vẫn là một thách thức. Phát hiện này không chỉ là một giới hạn kỹ thuật mà còn là một sự xác thực thực nghiệm cho nguyên tắc thiết kế cốt lõi của RFO-HITL: sự cần thiết không thể thiếu của cơ chế HITL. Chính những ngoại lệ này chứng minh rằng việc loại bỏ con người khỏi quy trình là một rủi ro. Về hiệu năng, độ trễ P95 (3,3 giây) cho thấy việc gọi API của LLM là một điểm nghẽn, phù hợp cho các quy trình chạy nền nhưng cần tối ưu hóa cho các ứng dụng quy mô lớn hơn.

Khi đánh giá qua lăng kính lý thuyết, các kết quả này hoàn toàn phù hợp. Việc tự động hóa các tác vụ soạn thảo đã giúp giảm "tải nhận thức ngoại lai" cho giảng viên (phù hợp với CLT), trong khi cơ chế HITL bắt buộc đảm bảo giảng viên luôn là "chủ thể" kiểm soát công cụ (phù hợp với AT). Tóm lại, quá trình đánh giá đã chứng minh RFO-HITL là một kiến trúc cân bằng, thành công trong việc đánh đổi một phần tự động hóa để đạt được sự an toàn, minh bạch và quyền kiểm soát của con người.

4. Kết luận

Nghiên cứu đáp ứng thách thức cấp bách trong việc tích hợp trí tuệ nhân tạo tạo sinh vào các quy trình hỗ trợ ra quyết định trong giáo dục đại học, một lĩnh vực đòi hỏi nghiêm ngặt về tính minh bạch, trách nhiệm giải trình và sự giám sát của con người. Áp dụng phương pháp luận nghiên cứu khoa học thiết kế, nghiên cứu đã phát triển và đánh giá ban đầu kiến trúc RFO-HITL. Đóng góp chính là một kiến trúc thiết kế được xây dựng dựa trên nền tảng lý thuyết hoạt động và lý thuyết tải nhận thức. Kiến trúc RFO-HITL đặc trưng bởi sự kết hợp của ba trụ cột: (1) điều phối ưu tiên quy tắc tường minh để đảm bảo tính minh bạch; (2) giới hạn vai trò của các mô hình ngôn ngữ lớn trong

việc hỗ trợ nhận thức nhằm giảm thiểu rủi ro; (3) tích hợp nền tảng NCLC với cơ chế HITL bắt buộc để đảm bảo khả năng tiếp cận và trách nhiệm giải trình cuối cùng thuộc về con người.

Kết quả đánh giá kỹ thuật ban đầu xác nhận tính khả thi và hiệu quả cao của kiến trúc đề xuất. Đáng chú ý, các phát hiện về việc vi phạm các ràng buộc an toàn (sắc thái văn phong) trong đầu ra của LLM đã củng cố tầm quan trọng thiết yếu của sự giám sát bắt buộc bởi con người.

Kiến trúc RFO-HITL cung cấp một lộ trình thực tiễn cho việc triển khai AI có trách nhiệm trong giáo dục. Nó đề xuất một kiến trúc cân bằng, cho phép các cơ sở giáo dục khai thác sức mạnh của tự động hóa để tăng cường năng lực cho giảng viên, đồng thời đảm bảo các quyết định sư phạm quan trọng luôn nằm trong sự kiểm soát và phán quyết chuyên môn của con người.

TÀI LIỆU THAM KHẢO

[1] M.J. Khan, A. Mahade (2023), "Investigation to improve decision support systems with the help of artificial intelligence in higher education", DOI: 10.2139/ssrn.4530480.

[2] N.R. Abdurohman (2025), "Artificial intellegent in higher education: Opportunities and challenges", *Eurasian Science Review*, **2(Special Issue)**, pp.1683-1695, DOI: 10.63034/esr-334.

[3] A.K. Singh, B.K. Singh, S.P. Pandey (2024), "Artificial intelligence in education: Effects on decision-making, human effort, and risk management", *Noble Science Press*, **1**, pp.5-11, DOI: 10.52458/9788197112492.nsp.2024.eb.ch-02.

[4] B. Shneiderman (2020), "Human-centered artificial intelligence: Reliable, safe & trustworthy", *International Journal of Human-Computer Interaction*, **36(6)**, pp.495-504, DOI: 10.1080/10447318.2020.1741118.

[5] K. Peffers, T. Tuunanen, M.A. Rothenberger, et al. (2007), "A design science research methodology for information systems research", *Journal of Management Information Systems*, **24(3)**, pp.45-77, DOI: 10.2753/MIS0742-1222240302.

[6] W. Moore, L.S. Tsay (2024), "From data to decisions: Leveraging AI for proactive education strategies", *International Conference on AI Research*, **5(1)**, pp.281-288, DOI: 10.34190/icair.4.1.3082.

[7] A. Adadi, M. Berrada (2018), "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)", *IEEE Access*, **6**, pp.52138-52160, DOI: 10.1109/ACCESS.2018.2870052.

[8] F. Zhou, H. Lv, K. Zhou, et al. (2024), "Innovative application and societal impact of AI in student behavior early warning systems within smart campuses", *Philosophy and Social Science*, **1(5)**, pp.33-40, DOI: 10.62381/P243506.

[9] A. Kokala (2022), "Low-code platforms in BPM: How workflow and rules engines enable citizen development", *International Journal of Science and Research Archive*, **6(1)**, pp.343-346, DOI: 10.30574/ijsra.2022.6.1.0145.

[10] N. Upadhyaya (2025), "Low-code/no-code platforms and their impact on traditional software development: A literature review", DOI: 10.13140/RG.2.2.31585.72807.

[11] Q. Wu, G. Bansal, J. Zhang, et al. (2023), "Autogen: Enabling next-gen LLM applications via multi-agent conversation", *ArXiv*, DOI: 10.48550/arXiv.2308.08155.

[12] S. Yao, J. Zhao, D. Yu, et al. (2023), "React: Synergising reasoning and acting in language models", *ArXiv*, DOI: 10.48550/arXiv.2210.03629.

[13] B. Memarian, T. Doleck (2024), "Human-in-the-loop in artificial intelligence in education: A review and entity-relationship (ER) analysis", *Computers in Human Behavior: Artificial Humans*, **2(1)**, DOI: 10.1016/j.chbah.2024.100053.

[14] M. Ninaus, M. Sailer (2022), "Closing the loop - The human role in artificial intelligence for education", *Frontiers in Psychology*, **13**, DOI: 10.3389/fpsyg.2022.956798.

[15] O. Akinrinola, C.C. Okoye, O.C. Ofodile, et al. (2024), "Navigating and reviewing ethical dilemmas in AI development: Strategies for transparency, fairness, and accountability", *GSC Advanced Research and Reviews*, **18(3)**, pp.50-58, DOI: 10.30574/gscarr.2024.18.3.0088.

[16] Y. Engeström (2014), *Learning by Expanding: An Activity-theoretical Approach to Developmental Research,* Cambridge University Press, **2**, DOI: 10.1017/CBO9781139814744.

[17] J. Sweller (1988), "Cognitive load during problem solving: Effects on learning", *Cognitive Science*, **12(2)**, pp.257-285, DOI: 10.1207/s15516709cog1202_4.

[18] O. Al-Omari, A. Alyousef, S. Fati, et al. (2025), "Governance and ethical frameworks for ai integration in higher education: Enhancing personalized learning and legal compliance", *Journal of Ecohumanism*, **4(2)**, pp.80-86, DOI: 10.62754/joe.v4i2.5781.

[19] D. Kolovos, J.D. Lara, M. Tisi, et al. (2023), "4th international workshop on modeling in low-code development platforms (LowCode 2023)", *2023 ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C)*, pp.851-853, DOI: 10.1109/MODELS-C59198.2023.00134.

[20] K. Shilton (2013), "Values levers: Building ethics into design", *Science, Technology, & Human Values*, **38(3)**, pp.374-397**,** DOI: 10.1177/0162243912436985.

[21] K. Shilton (2018), "Values and ethics in human-computer interaction", *Foundations and Trends® in Human-Computer Interaction*, **12(2)**, pp.107-171, DOI: 10.1561/1100000073.