

TẠO SINH CÂU HỎI TRẮC NGHIỆM THEO CHUẨN ĐẦU RA HỌC PHẦN DỰA TRÊN MÔ HÌNH NGÔN NGỮ LỚN VÀ THANG NHẬN THỨC BLOOM

Ngô Văn Sơn^{1*}, Bùi Anh Tuấn¹, Nguyễn Quang Anh¹

*Tác giả liên hệ, Email: nvson@hou.edu.vn. ORCID: 0009-0004-8170-2588

Ngày tòa soạn nhận được bài báo: 15/06/2026

Ngày phản biện đánh giá: 22/06/2026

Ngày bài báo được duyệt đăng: 14/09/2026

DOI: 10.59266/houjs.2026.1437

Tóm tắt: Nghiên cứu đề xuất và thử nghiệm quy trình tạo sinh tự động câu hỏi trắc nghiệm khách quan (TNKQ) tiếng Việt dựa trên mô hình ngôn ngữ lớn (LLM), trong đó mỗi câu hỏi được gắn với một chuẩn đầu ra học phần (Course Learning Outcomes - CLO) và một mức nhận thức theo thang Bloom cải tiến, đồng thời được dựa vào nội dung giáo trình bằng kỹ thuật truy xuất BM25 nhằm hạn chế hiện tượng ảo giác nội dung. Khác với các nghiên cứu trước dựa vào giảng viên đánh giá thủ công hoặc dùng một LLM tự đánh giá bản thân, nghiên cứu này đề xuất cơ chế chấm chéo giữa hai LLM thuộc hai họ độc lập (LLaMA-3.3-70B và Qwen3-32B) làm phương pháp đánh giá tự động. Thử nghiệm trên học phần Mạng và truyền thông với 40 câu hỏi sinh được cho thấy hệ thống đạt mức đồng thuận đáp án tốt đến rất tốt (Cohen's $\kappa = 0,78-0,82$) và độ bám sát giáo trình rất cao ($0,97-0,99$), nhưng bộc lộ thách thức lớn khi phân loại mức nhận thức Bloom ($\kappa = 0,03-0,26$). Kết quả khẳng định xu hướng khả thi của AI trong hỗ trợ đánh giá đại học, đồng thời chỉ rõ ranh giới giữa năng lực tự động hóa và vai trò không thể thay thế của giảng viên.

Từ khóa: tạo sinh câu hỏi tự động, mô hình ngôn ngữ lớn, chuẩn đầu ra học phần, chấm chéo mô hình ngôn ngữ lớn, đánh giá giáo dục đại học

I. Đặt vấn đề

Trong giáo dục đại học, câu hỏi trắc nghiệm khách quan (TNKQ) là hình thức đánh giá phổ biến nhờ khả năng kiểm tra diện rộng, chấm điểm khách quan và triển khai trên quy mô lớn (CH và Saha, 2020; Haladyna và cộng sự, 2002). Tuy nhiên, soạn thủ công câu hỏi TNKQ chất lượng cao tốn nhiều thời gian và đòi hỏi chuyên

môn sâu cả về nội dung lẫn kỹ thuật ra đề (Kurdi và cộng sự, 2020). Áp lực này càng lớn khi các trường đại học Việt Nam phải duy trì ngân hàng câu hỏi đủ lớn để phục vụ đánh giá theo chuẩn đầu ra chương trình đào tạo (Bộ Giáo dục và Đào tạo, 2021).

Mô hình ngôn ngữ lớn (Large Language Models - LLM) đã mở ra xu hướng mới trong sinh tự động câu hỏi

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

(Automatic Question Generation - AQG): sinh câu hỏi tự nhiên, chất lượng cao và điều khiển được theo mục tiêu học tập. Doughty và cộng sự (2024) cho thấy câu hỏi do GPT-4 sinh đạt chất lượng tương đương câu do giảng viên soạn; Elkins và cộng sự (2024) chứng minh giáo viên ưu tiên dùng câu sinh bởi LLM khi xây dựng bài kiểm tra. Tuy nhiên, vẫn tồn tại bốn thách thức: (i) Thứ nhất, hầu hết nghiên cứu thực hiện trên tiếng Anh; (ii) Thứ hai, các nghiên cứu thường gắn câu hỏi với mục tiêu bài học chung, chưa ở cấp chuẩn đầu ra học phần (CLO) theo yêu cầu kiểm định tại Việt Nam (Bộ Giáo dục và Đào tạo, 2021); (iii) Thứ ba, hầu hết dựa vào giảng viên chấm tay - tốn kém, khó nhân rộng; (iv) Thứ tư, chưa có công trình nào sử dụng chấm chéo giữa hai họ LLM hoàn toàn độc lập như một phương pháp đánh giá thay thế có thể tái lập.

Xuất phát từ đó, nghiên cứu này nhằm: (1) thiết kế một quy trình tạo sinh câu hỏi TNKQ tiếng Việt gắn chặt CLO và mức Bloom, dựa vào giáo trình; (2) đề xuất và thử nghiệm khung đánh giá tự động dựa trên chấm chéo hai LLM độc lập với bộ chỉ số toán học cụ thể; (3) phân tích các xu hướng khả thi và thách thức còn tồn tại khi ứng dụng AI vào đánh giá đại học - đúng tinh thần chủ đề hội thảo.

II. Cơ sở lý thuyết

2.1. Chuẩn đầu ra học phần và nguyên tắc đánh giá nhất quán

Chuẩn đầu ra học phần (CLO) mô tả kiến thức, kỹ năng mà người học phải đạt được sau khi hoàn thành học phần; là thành phần bắt buộc trong đề cương chi tiết theo Thông tư 17/2021/TT-BGDĐT (Bộ Giáo dục và Đào tạo, 2021). Trong

thực tiễn ra đề thi, CLO được thể hiện qua ma trận đề thi: mỗi câu hỏi gán với một CLO và một mức nhận thức - đây chính là cấu trúc đầu vào tự nhiên cho hệ thống sinh câu hỏi tự động. Nguyên tắc đánh giá nhất quán (constructive alignment) của Biggs (1996) khẳng định đề thi phải đo đúng những gì CLO yêu cầu; một câu hỏi đúng nội dung nhưng sai mức nhận thức so với CLO vẫn không đạt. Đây là lý do nghiên cứu đưa “phù hợp CLO” và “đúng mức Bloom” thành hai tiêu chí đánh giá riêng biệt.

2.2. Thang Bloom cải tiến và nguyên tắc soạn câu hỏi TNKQ

Thang Bloom cải tiến (Anderson và cộng sự, 2001) gồm sáu mức: Nhớ, Hiểu, Vận dụng, Phân tích, Đánh giá, Sáng tạo; với TNKQ, bốn mức đầu là khả thi nhất. Haladyna và cộng sự (2002) tổng hợp 31 nguyên tắc soạn câu hỏi đã kiểm chứng; Rodriguez (2005) phân tích tổng hợp 80 năm nghiên cứu cho thấy ba phương án là tối ưu về hiệu quả đo lường, dù thực tiễn Việt Nam phổ biến bốn phương án. Các nguyên tắc này được mã hóa trực tiếp vào chỉ dẫn cho LLM.

2.3. Sinh câu hỏi bằng LLM và đánh giá bằng chấm chéo

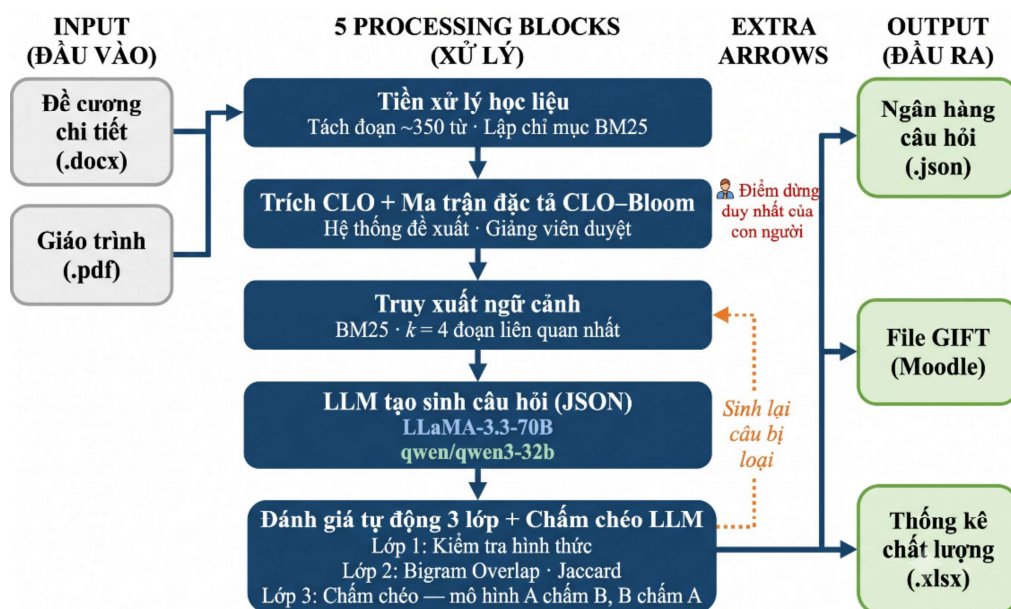
Các hệ thống AQG truyền thống phụ thuộc tài nguyên ngôn ngữ (WordNet, ontology) vốn nghèo nàn với tiếng Việt và chỉ sinh được câu mức thấp (Dhawaleswar Rao và Saha, 2020; Kurdi và cộng sự, 2020). LLM cho phép sinh câu hỏi bằng chỉ dẫn ngôn ngữ tự nhiên mà không cần huấn luyện lại; kết hợp kỹ thuật truy xuất tăng cường (RAG) giúp neo nội dung vào giáo trình, tránh LLM “bịa” thông tin. Trong y khoa, Kiyak và cộng sự (2024)

cung cấp bằng chứng dùng ChatGPT sinh câu hỏi cho kỳ thi được lý. Về đánh giá, hướng dùng LLM làm giám khảo (LLM-as-a-judge) được Zheng và cộng sự (2023) đặt nền tảng phương pháp luận với MT-

Bench; tuy nhiên, cần dùng nhiều mô hình độc lập chấm chéo để tránh thiên kiến tự đánh giá - đây là khoảng trống mà nghiên cứu này góp phần lấp.

III. Phương pháp và vật liệu nghiên cứu

3.1. Kiến trúc quy trình



Hình 1: Sơ đồ tổng thể quy trình tạo sinh và đánh giá tự động câu hỏi TNKQ

Quy trình gồm năm khối nối tiếp, với một điểm dừng để giảng viên duyệt ma trận đặc tả trước khi hệ thống chạy sinh câu hỏi (Hình 1).

- **Khối 1 - Tiền xử lý:** giáo trình được tách thành các đoạn ~350 từ (gói đầu 60 từ), lập chỉ mục BM25.

- **Khối 2 - Ma trận đặc tả:** hệ thống tự trích CLO từ đề cương, đề xuất bản nháp; giảng viên duyệt.

- **Khối 3 - Truy xuất:** với mỗi dòng ma trận, tìm $k = 4$ đoạn giáo trình phù hợp nhất.

- **Khối 4 - Sinh câu hỏi:** LLM nhận chỉ dẫn mã hóa 31 nguyên tắc soạn câu hỏi và định nghĩa thao tác từng mức

Bloom, trả kết quả JSON gồm câu dẫn, bốn phương án, đáp án, giải thích và trích dẫn giáo trình.

• Khối 5 - Đánh giá 3 lớp.

3.2. Vật liệu và cấu hình thực nghiệm

Học phần: *Mạng và truyền thông* (7E1010.22), chương trình CNTT, giáo trình 237 trang. Đề cương có 4 CLO; hệ thống tự phát hiện CLO4 (kỹ năng mềm: làm việc nhóm, thuyết trình) không phù hợp TNKQ và tự động loại. Hai mô hình thuộc hai họ khác nhau: LLaMA-3.3-70B (Meta) và Qwen3-32B (Alibaba), qua nền tảng Groq. Mỗi mô hình lần lượt sinh và chấm chéo cho mô hình kia. Tham số: nhiệt độ 0,4; mỗi lượt 5 câu; $k = 4$. Tổng 40 câu (20/mô hình).

Ma trận đặc tả gồm 5 tổ hợp CLO-Bloom (4 câu/tổ hợp, cột “Duyệt = x”): CLO1×Hiểu (kiến thức mạng cơ bản), CLO1×Vận dụng (cấu hình thiết bị), CLO2×Vận dụng (giao thức tầng vận chuyển), CLO3×Vận dụng (định tuyến liên mạng) và CLO3×Phân tích (chẩn đoán sự cố mạng). Việc đưa mức Phân tích vào ma trận nhằm kiểm chứng khả năng kiểm soát mức nhận thức cao của LLM.

3.3. Khung đánh giá tự động và các chỉ số

Chất lượng câu hỏi được đánh giá hoàn toàn tự động qua ba lớp, với bốn chỉ số toán học:

Lớp 1 - Kiểm tra hình thức: loại câu thiếu trường dữ liệu, phương án trùng lặp, độ dài chênh lệch quá 3,5 lần, hoặc dùng cụm bị cấm.

Lớp 2 - Kiểm tra ngữ nghĩa. Gồm hai chỉ số. Chỉ số thứ nhất đo độ bám sát giáo trình (Bigram Overlap - BO): tỷ lệ cặp từ liên tiếp (bigram) trong phần trích dẫn căn cứ do LLM cung cấp thực sự xuất hiện trong đoạn giáo trình:

$$BO(e, C) = \frac{|BG(e) \cap BG(C)|}{|BG(e)|} \quad (1)$$

trong đó e là căn cứ trích dẫn (evidence) do LLM cung cấp, C là đoạn giáo trình đã truy xuất, $BG(\cdot)$ là tập bigram. Nghiên cứu chọn bigram thay từ đơn vì tiếng Việt là ngôn ngữ đơn âm tiết, khiến độ trùng từ đơn cao giả tạo. Câu có $BO < 0,20$ bị gắn cờ nghi bịa nội dung (ngưỡng đặt theo thực hành, cần hiệu chỉnh trong nghiên cứu tiếp theo). Do BO đo mức trùng lặp giữa căn cứ trích dẫn với chính đoạn giáo trình đã cấp cho LLM, chỉ số này có tính vòng tròn: giá trị BO cao gần như tất yếu và

chỉ chứng minh câu hỏi bám sát đoạn tư liệu đã cấp, chưa đủ để khẳng định hoàn toàn không có ảo giác nội dung.

Chỉ số thứ hai đo chất lượng phương án nhiều bằng độ tương đồng Jaccard giữa từng nhiều d và đáp án đúng k :

$$JS(d, k) = \frac{|T(d) \cap T(k)|}{|T(d) \cup T(k)|} \quad (2)$$

với $T(\cdot)$ là tập từ (token). Nhiều hợp lệ khi $0,01 \leq JS \leq 0,90$; ngoài khoảng này bị gắn cờ.

Lớp 3 - Chấm chéo LLM. Mô hình giám khảo (khác họ với mô hình sinh) tự trả lời câu hỏi chỉ dựa trên giáo trình, không biết đáp án, đồng thời tự gán mức Bloom. Mức đồng thuận đo bằng hệ số Cohen's kappa:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (3)$$

với P_o là tỷ lệ đồng ý quan sát được, P_e là tỷ lệ đồng ý kỳ vọng ngẫu nhiên. Theo thang phân loại của Landis và Koch (1977), $\kappa \geq 0,60$ được coi là mức tốt. Cần lưu ý κ đo mức đồng thuận giữa hai mô hình giám khảo, không phải độ chính xác so với đáp án chuẩn do giảng viên xác nhận. Song song với κ , nghiên cứu còn dùng chỉ số khớp mức Bloom (B_{match}) là tỷ lệ đồng ý quan sát được (P_o) riêng cho nhãn Bloom giữa mô hình sinh và mô hình mẫu, chưa hiệu chỉnh theo cơ hội ngẫu nhiên - đóng vai trò tương tự P_o trong công thức (3) nhưng áp dụng cho mức nhận thức thay vì đáp án. Tỷ lệ câu đạt cả ba lớp (Pass Rate):

$$P = \frac{N_{\text{đạt}}}{N_{\text{tổng}}} \times 100\% \quad (4)$$

Giả thuyết: H1 - $P \geq 70\%$; H2 - κ đáp án $\geq 0,60$; H3 - κ Bloom $\geq 0,60$.

IV. Kết quả và thảo luận

4.1. Thống kê mô tả

Hệ thống sinh 40 câu hỏi (20/mô hình) theo 5 tổ hợp CLO-Bloom. Sau ba lớp đánh giá và khử trùng lặp, ngân hàng cuối gồm 7 câu (Bảng 1). Đáng lưu ý: hai mô hình tạo ra tập câu hỏi đạt không trùng

lặp về CLO - LLaMA có câu CLO3 trong khi Qwen không có; ngược lại Qwen có câu CLO2 trong khi LLaMA không có. Toàn bộ 8 câu mức Phân tích bị loại hoàn toàn dù ma trận yêu cầu, cho thấy mức nhận thức cao là thách thức chung của cả hai mô hình.

Bảng 1: Thống kê mô tả bộ câu hỏi sinh được

Chỉ số	LLaMA-3.3	Qwen3-32B	Tổng	Ngưỡng H1
Tổng sinh	20	20	40	—
Câu đạt	3	4	7	—
Pass Rate - (4)	15,0%	20,0%	17,5%	$\geq 70\%$
Bloom	Hiệu 1; VD 2	Hiệu 2; VD 2	Hiệu 3; VD 4	—

4.2. Xu hướng tích cực

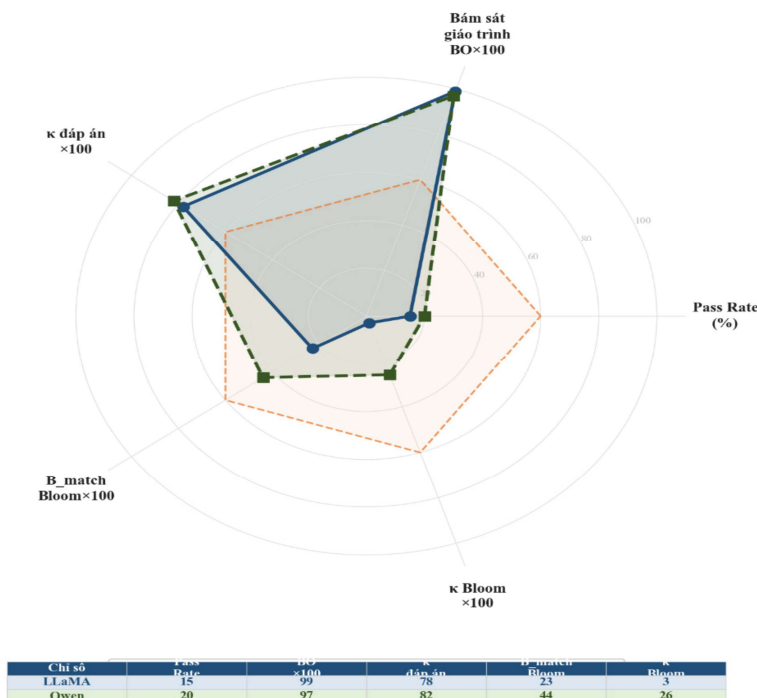
Bảng 2: Chỉ số chất lượng và đồng thuận chéo

Chỉ số	LLaMA	Qwen	Ngưỡng	Kết quả
BO - (1), TB $\pm$ SD	0,99 $\pm$ 0,03	0,97 $\pm$ 0,04	$\geq 0,20$	✓
JS - (2), TB $\pm$ SD	0,21 $\pm$ 0,23	0,18 $\pm$ 0,21	0,01 - 0,90	✓
P _o đáp án	0,85	0,89	—	—
κ đáp án - (3)	0,78	0,82	$\geq 0,60$	✓ H2
B _{match} Bloom	0,23	0,44	$\geq 0,65$	× H3
κ Bloom - (3)	0,03	0,26	$\geq 0,60$	× H3

Cohen's κ đáp án đạt 0,78 (LLaMA) và 0,82 (Qwen) - mức “tốt” và “rất tốt” - vượt ngưỡng H2 ($\geq 0,60$), cho thấy mức đồng thuận giữa hai mô hình giám khảo về đáp án là cao. Độc lập với chỉ số này, độ bám sát giáo trình theo BO cũng đạt mức rất cao (0,97-0,99); tuy nhiên do BO chỉ đo mức trùng lặp với chính đoạn giáo trình đã cấp cho LLM, chỉ số này khẳng định câu hỏi bám sát đoạn tư liệu đã cho, chưa đủ để kết luận hoàn toàn không có ảo giác nếu thiếu thẩm định của giảng viên. Hai chỉ số này là xu hướng tích cực tương đồng giữa hai mô hình. Tuy nhiên, phân bố CLO đạt lại rất khác nhau: LLaMA có câu CLO3 (2 câu) nhưng thiếu CLO2; Qwen

có CLO2 (1 câu) nhưng thiếu CLO3. Nếu chỉ dùng một mô hình, ngân hàng câu hỏi sẽ thiếu hụt một số CLO nhất định, đây là rủi ro thực tiễn quan trọng khi triển khai AI hỗ trợ ra đề thi. Lựa chọn mô hình ảnh hưởng trực tiếp đến mức bao phủ CLO - bản thân điều này là lý do cần có giảng viên rà soát câu hỏi được tạo sinh.

Lưu ý: κ được tính trên $n = 13$ (LLaMA) và $n = 9$ (Qwen) câu được chấm chéo; với cỡ mẫu nhỏ này, khoảng tin cậy (CI) 95% ước tính khá rộng ($\pm 0,15-0,20$). Kết quả mang tính khám phá (exploratory), chưa đủ để kết luận chắc chắn (conclusive).



Hình 2: So sánh đa chiều chất lượng hai mô hình

4.3. Thách thức

Pass Rate 17,5% thấp hơn nhiều so với ngưỡng H1. Đáng lưu ý: toàn bộ 8 câu mức Phân tích bị loại hoàn toàn dù ma trận yêu cầu, do các phương án dạng kịch bản có Jaccard quá thấp hoặc bất đồng mức Bloom - khẳng định đây là mức nhận thức khó kiểm soát nhất. Phân tích hai nhóm nguyên nhân chính: nhóm 1 (~50% câu bị loại): “nhiều quá xa đáp án” theo Jaccard, thực chất là hạn chế của thước đo, không phải lỗi câu hỏi: hai chuỗi kỹ thuật dạng số (ví dụ “255.255.255.192” và “255.255.255.128”) có Jaccard rất thấp dù đều là subnet mask hợp lý. Nhóm 2 (~35%): bất đồng mức Bloom (κ Bloom chỉ 0,03-0,26) - đây là thách thức cốt lõi. Do phần lớn số câu bị loại ở Nhóm 1 xuất phát từ hạn chế của thước đo Jaccard chứ không phải chất lượng nội dung; Pass Rate 17,5% cần được diễn giải thận trọng và không nên dùng làm căn cứ duy nhất để kết luận về giá

thuyết H1; việc thay Jaccard bằng độ tương đồng ngữ nghĩa và tính lại Pass Rate là cần thiết trước khi đưa ra kết luận cuối cùng.

Ví dụ minh họa: Câu đạt (CLO1, mức Hiểu): “Khi một thiết bị trong mạng LAN gửi dữ liệu, vai trò của switch là gì?” - A. Định tuyến ra Internet; B. Xác định cổng đích dựa trên địa chỉ MAC; C. Chia nhỏ dữ liệu thành gói tin; D. Cung cấp tín hiệu xung nhịp. Giám khảo đồng ý đáp án B, gán mức Hiểu - khớp với mô hình sinh. Câu bị loại (CLO2, mức Vận dụng): câu về tính subnet bị gán cờ vì JS giữa các phương án nhiều dạng số quá thấp (JS < 0,01), dù nội dung chính xác.

Tổng hợp từ kết quả thực nghiệm, Bảng 3 xác định rõ những gì AI có thể đảm nhận đáng tin cậy và những gì giảng viên vẫn phải giữ vai trò quyết định. Mô hình triển khai phù hợp: “AI sinh câu hỏi - giảng viên rà soát mức nhận thức và CLO - AI định dạng và tạo sinh”.

Bảng 3: Ranh giới năng lực AI và vai trò của giảng viên trong ra đề TNKQ theo CLO

Khía cạnh	AI làm được	Vấn cần giảng viên
Nội dung & đáp án	$\kappa = 0,78-0,82$ ✓ Đáng tin cậy	Xác nhận theo đáp án chính thức
Bám sát giáo trình	$BO = 0,97-0,99$ ✓ Rất tốt	Chọn đoạn giáo trình chuẩn
Mức nhận thức Bloom	$\kappa = 0,03-0,26$ ✓ Chưa tin cậy	Xác nhận đúng mức Bloom của CLO
Bao phủ CLO đều	Phụ thuộc mô hình ✗	Đảm bảo đủ tất cả CLO
Mức Phân tích trở lên	0/8 câu đạt ✗	Soạn hoặc rà soát câu mức cao

4.4. Thảo luận

Phát hiện quan trọng nhất: năng lực của LLM trong đánh giá đại học mang tính phân tầng. Ở tầng nội dung và đáp án (kiểm chứng được từ giáo trình), AI đạt mức tin cậy cao. Ở tầng phân loại mức nhận thức Bloom (mang tính diễn giải), AI kém tin cậy - nhất quán với Elkins và cộng sự (2024) ghi nhận ngay cả giáo viên cũng thường bất đồng khi phân loại Bloom. Hàm ý: các tiêu chí khách quan có thể tự động hóa; các tiêu chí diễn giải nên giữ vai trò giảng viên, AI chỉ là trợ lý gợi ý. Để cải thiện κ Bloom, các hướng kỹ thuật tiềm năng gồm: suy luận chuỗi (chain-of-thought) cho phân loại Bloom, dùng ví dụ mẫu (few-shot) với câu hỏi đã được giảng viên gán nhãn, hoặc tinh chỉnh mô hình trên dữ liệu Bloom tiếng Việt.

Hạn chế: (i) hai mô hình tuy khác họ nhưng cùng nền tảng suy luận; cần thêm họ thứ ba (Claude hoặc Gemini) hoàn toàn độc lập; (ii) cỡ mẫu nhỏ ($n = 40$, CI rộng), kết quả mang tính khám phá; (iii) chưa có thẩm định giảng viên làm tiêu chuẩn vàng; (iv) mới thử trên 1 học phần, cần mở rộng; (v) các ngưỡng $BO \geq 0,20$ và JS đặt theo kinh nghiệm, chưa được hiệu chỉnh bằng dữ liệu gán nhãn; (vi) mỗi tổ hợp CLO-Bloom chỉ sinh một lượt duy nhất, chưa đánh giá phương sai giữa các lần sinh do bản chất ngẫu nhiên của LLM.

V. Kết luận

Nghiên cứu đề xuất và thử nghiệm bước đầu quy trình tạo sinh câu hỏi TNKQ tiếng Việt bám sát CLO và thang Bloom, với cấu hình chấm chéo giữa hai LLM thuộc hai họ độc lập. Kết quả cho thấy: (1) hai mô hình tạo ra kết quả khác nhau về phân bố CLO đạt chuẩn đầu ra, nên lựa chọn mô hình ảnh hưởng trực tiếp đến mức bao phủ CLO; (2) cả hai đạt đồng thuận đáp án tốt ($\kappa = 0,78-0,82$) và bám sát giáo trình xuất sắc ($BO = 0,97-0,99$); (3) cả hai chưa tạo được câu mức Phân tích vượt qua bộ lọc tự động và cho κ Bloom rất thấp ($0,03-0,26$), khẳng định tạo sinh câu hỏi theo chuẩn đầu ra vẫn là thách thức chưa thể tự động hóa hoàn toàn, đòi hỏi vai trò không thể thiếu của giảng viên trong rà soát mức nhận thức và đảm bảo bao phủ CLO đầy đủ.

Về độ giá trị nội tại, kết quả có thể chịu ảnh hưởng bởi thiết kế prompt, ngưỡng lọc BO/JS và cách định nghĩa mức Bloom. Về độ giá trị ngoại tại, thực nghiệm mới thực hiện trên một học phần thuộc lĩnh vực mạng máy tính nên chưa thể khái quát sang các học phần khác. Về độ giá trị kết luận, cỡ mẫu nhỏ khiến các hệ số κ có khoảng tin cậy rộng. Về độ giá trị tiêu chuẩn, nghiên cứu chưa có bộ nhãn do giảng viên độc lập gán làm quy chuẩn.

Hướng phát triển: mở rộng sang ba họ mô hình kèm thẩm định giảng viên; thay Jaccard bằng độ tương đồng ngữ nghĩa; thực nghiệm đa học phần; và tinh chỉnh mô hình tiếng Việt chuyên biệt cho khảo thí.

Tài liệu tham khảo

- Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.
- Bộ Giáo dục và Đào tạo. (2021). *Thông tư số 17/2021/TT-BGDĐT ngày 22/6/2021 quy định về chuẩn chương trình đào tạo; xây dựng, thẩm định và ban hành chương trình đào tạo các trình độ của giáo dục đại học*. <https://vanban.chinhphu.vn/?pageid=27160&docid=203478>
- Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, 32, 347-364. <https://doi.org/10.1007/BF00138871>
- Dhawaleswar Rao, & Saha, S. K. (2020). Automatic multiple choice question generation from text: A survey. *IEEE Transactions on Learning Technologies*, 13(1), 14-25. <https://doi.org/10.1109/TLT.2018.2889100>
- Doughty, J., Wan, Z., Bompelli, A., Qayum, J., Wang, T., Zhang, J., Zheng, Y., Doyle, A., Sridhar, P., Agarwal, A., Bogart, C., Keylor, E., Kultur, C., Savelka, J., & Sakr, M. (2024). A comparative study of AI-generated (GPT-4) and human-crafted MCQs in programming education. In *Proceedings of the 26th Australasian Computing Education Conference* (pp. [cần bổ sung nếu có]). <https://doi.org/10.1145/3636243.3636256>
- Elkins, S., Kochmar, E., Cheung, J. C. K., & Serban, I. (2024). How teachers can use large language models and Bloom's taxonomy to create educational quizzes. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*. <https://doi.org/10.1609/aaai.v38i21.30353>
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309-333. https://doi.org/10.1207/S15324818AME1503_5
- Kıyak, Y. S., Coşkun, Ö., Budakoğlu, I., & Uluoğlu, C. (2024). ChatGPT for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. *European Journal of Clinical Pharmacology*, 80, 1-7. <https://doi.org/10.1007/s00228-024-03649-x>
- Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1), 121-204. <https://doi.org/10.1007/s40593-019-00186-y>
- Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159-174. <https://doi.org/10.2307/2529310>
- Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, 24(2), 3-13. <https://doi.org/10.1111/j.1745-3992.2005.00006.x>
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*.

AUTOMATIC GENERATION OF MULTIPLE-CHOICE QUESTIONS BASED ON COURSE LEARNING OUTCOMES, LARGE LANGUAGE MODELS, AND BLOOM'S TAXONOMY

Ngo Van Son¹, Bui Anh Tuan¹, Nguyen Quang Anh¹

Abstract: *This study proposes and pilot-tests a pipeline for the automatic generation of Vietnamese multiple-choice questions (MCQs) using large language models (LLMs). Each generated item is explicitly mapped to a course learning outcome (CLO) and a specific cognitive level of Bloom's Revised Taxonomy, and is anchored to course textbook content via BM25 retrieval to reduce hallucination. In contrast to prior approaches that rely on manual instructor evaluation or single-model self-assessment, this work introduces a cross-grading mechanism in which two LLMs from entirely independent model families (LLaMA-3.3-70B and qwen/qwen3-32b) evaluate each other's output as the automated quality assurance layer. Experiments on the Computer Networking course (237-page textbook, 3 viable CLOs, 40 generated items) show good-to-very-good answer agreement (Cohen's $\kappa = 0.78-0.82$) and very high content grounding ($BO = 0.97-0.99$), yet reveal a substantial challenge in Bloom-level classification ($\kappa = 0.03-0.26$). The results affirm the feasibility of AI-assisted MCQ generation in Vietnamese higher education, while delineating the boundary between what can be automated and what continues to require instructor judgment.*

Keywords: *automatic question generation, large language models, course learning outcomes, mô hình ngôn ngữ lớn cross-grading, higher education assessment*

¹ Hanoi Open University, Hanoi, Vietnam