AI-BASED FRAMEWORK FOR QUESTION QUALITY EVALUATION

Vu Xuan Hanh*[1], Do Duy Trinh[2], Nguyen Quang Anh[3]

* Corresponding author
Email: hanhvx@hou.edu.vn

[2] Email: trinhdd@hou.edu.vn
[3] Email: anhnq@hou.edu.vn
[1,2,3] Hanoi Open University
101 Nguyen Hien street, Bach Mai ward, Hanoi, Vietnam

Received: 27/3/2026
Revised: 07/7/2026
Accepted: 10/8/2026
Published: 20/8/2026

Abstract: Assessing question quality is an essential step in ensuring the validity and reliability of assessment, yet existing methods remain heavily reliant on experts, are qualitative in nature, and difficult to scale. This paper proposes AF2QE (AI-based Framework for Question Quality Evaluation), a pre-validation framework integrating three tiers to provide quantitative, transparent, and repeatable quality indicators. The semantic analysis tier uses Sentence-BERT to measure the alignment of questions with course content and course learning outcomes (CLO). The pedagogical validation tier classifies questions according to Bloom's taxonomy using a model that combines TF-IDF and LinearSVC, and calculates content and CLO coverage indices. The statistical validation tier estimates the difficulty and discrimination indices using Bloom weights and semantic distance. Experiments on 650 questions from the Object-Oriented Programming course show that the Bloom classification model achieved an accuracy of 92.50%, and 84% of the questions simultaneously met all three criteria regarding content alignment, cognitive level, and optimal difficulty range. Initial results indicate that AF2QE has potential as a useful pre-screening tool to support experts in validating the quality of question banks.

Keywords: *Item quality evaluation, automated item generation, Bloom's taxonomy, pre-deployment validation, course learning outcomes, classical test theory.*

KHUNG ĐÁNH GIÁ CHẤT LƯỢNG CÂU HỎI DỰA TRÊN TRÍ TUỆ NHÂN TẠO

Vũ Xuân Hạnh*[1], Đỗ Duy Trinh[2], Nguyễn Quang Ánh[3]

* Tác giả liên hệ
Email: hanhvx@hou.edu.vn

[2] Email: trinhdd@hou.edu.vn
[3] Email: anhnq@hou.edu.vn
[1,2,3] Trường Đại học Mở Hà Nội
101 Nguyễn Hiền, phường Bạch Mai, Hà Nội, Việt Nam

Nhận bài: 27/3/2026
Chỉnh sửa xong: 07/7/2026
Chấp nhận đăng: 10/8/2026
Xuất bản: 20/8/2026

Tóm tắt: Đánh giá chất lượng câu hỏi là khâu thiết yếu trong đảm bảo giá trị và độ tin cậy của khảo thí, song các phương pháp hiện hành vẫn phụ thuộc nhiều vào chuyên gia, mang tính định tính và khó mở rộng quy mô. Bài viết này đề xuất AF2QE (AI-based Framework for Question Quality Evaluation), một khung tiền kiểm định tích hợp ba tầng nhằm cung cấp các chỉ số chất lượng định lượng, minh bạch và có thể lặp lại. Tầng phân tích ngữ nghĩa sử dụng Sentence-BERT để đo mức độ liên kết của câu hỏi với nội dung giáo trình và chuẩn đầu ra học phần (CLO). Tầng xác thực sư phạm phân loại câu hỏi theo thang Bloom bằng mô hình kết hợp TF-IDF, LinearSVC, tính toán chỉ số bao phủ nội dung và CLO. Tầng kiểm định thống kê ước lượng độ khó và độ phân hóa dựa trên trọng số Bloom và khoảng cách ngữ nghĩa. Thực nghiệm trên 650 câu hỏi thuộc học phần Lập trình Hướng đối tượng cho thấy mô hình phân loại Bloom đạt độ chính xác 92,50% và 84% câu hỏi đồng thời đáp ứng ba tiêu chí về liên kết nội dung, mức nhận thức và vùng độ khó tối ưu. Kết quả bước đầu cho thấy, AF2QE có tiềm năng sàng lọc tiền kiểm định hữu ích, hỗ trợ chuyên gia trong kiểm định chất lượng ngân hàng câu hỏi.

Từ khóa: *Đánh giá chất lượng câu hỏi, tạo sinh câu hỏi tự động, thang phân loại Bloom, tiền kiểm định, chuẩn đầu ra học phần, lí thuyết kiểm thử cổ điển.*

1. Đặt vấn đề

Đánh giá chất lượng ngân hàng câu hỏi là thành tố cốt lõi trong khảo thí và đảm bảo chất lượng. Tan và cộng sự (2025) nhận định, một đề thi là tập hợp các câu hỏi đạt chuẩn không chỉ bao phủ nội dung mà còn phải bảo đảm giá trị nội dung và độ tin cậy, tức phản ánh đúng chuẩn năng lực và cho kết quả ổn định, nhất quán giữa các lần đánh giá. Bên cạnh đó, Makhlouf và cộng sự (2020) cho rằng, đề thi cần có sự phân hóa để phân loại người học theo mức độ đạt chuẩn đầu ra. Tam và cộng sự (2024) cho rằng, sự liên kết chặt chẽ giữa nội dung, chuẩn đầu ra và phân bố hợp lí theo thang nhận thức Bloom giúp đánh giá toàn diện. Tuy nhiên, Circi và cộng sự (2023) phát biểu, thực tế đánh giá đề thi vẫn chủ yếu dựa vào chuyên gia và kinh nghiệm cá nhân, mang tính định tính và khó mở rộng khi ngân hàng câu hỏi lớn.

Khojasteh và cộng sự (2025) nhận định, sự phát triển mạnh mẽ của trí tuệ nhân tạo (AI), đặc biệt là các mô hình ngôn ngữ lớn (LLMs), mở ra khả năng hỗ trợ đánh giá đề thi theo nội dung, thang Bloom, CLO và ngữ nghĩa, góp phần nâng cao tính khách quan và nhất quán. Opait và cộng sự (2025) cho rằng, với khả năng xử lí ngôn ngữ ở mức ngữ nghĩa sâu, AI không chỉ tạo sinh mà còn có thể phân tích cấu trúc, mục tiêu nhận thức, tính nhất quán, phát hiện yếu tố mơ hồ, thuật ngữ đa nghĩa hoặc điều kiện chưa đầy đủ của câu hỏi. Tan và cộng sự (2025) xây dựng hệ thống hỗ trợ phân loại câu hỏi theo thang Bloom, góp phần giảm khác biệt trong diễn giải giữa các chuyên gia và tăng tính nhất quán thông qua nhận diện các khái niệm và ngữ cảnh.

Trong thực tế, phương pháp đánh giá chất lượng đề thi dựa hoàn toàn vào chuyên gia (SME) vẫn được xem là "chuẩn mực vàng" trong nhiều hệ thống giáo dục. Tan và cộng sự (2025) chỉ ra, SME bộc lộ nhiều hạn chế về nguồn lực và tính hệ thống. Circi và cộng sự (2023) cho rằng, phương pháp này tốn kém thời gian, chi phí và đòi hỏi nhiều vòng phản biện, đặc biệt He và cộng sự (2024) nhấn mạnh thách thức ở các lĩnh vực yêu cầu cập nhật. Nghiên cứu của Opait và cộng sự (2025) cho thấy mức độ không đồng thuận có thể dao động từ 20% đến 30%.

Ngày nay, có thể tạo sinh câu hỏi tự động (Automated Item Generation - AIG) dựa trên các mô hình hình thức hóa tri thức. He và cộng sự (2024) cho rằng, mục tiêu của AIG là mở rộng ngân hàng câu hỏi, đảm bảo tính nhất quán và giảm chi phí so với biên soạn thủ công. AIG thường được chia thành ba hướng chính: Dựa trên khuôn mẫu, dựa trên bản thể học và dựa trên LLMs. Mỗi hướng có ưu và nhược điểm riêng: Hướng dựa trên khuôn mẫu đảm bảo tính nhất quán nhưng bị giới hạn bởi độ đa dạng thấp; hướng dựa trên bản thể học cho phép lí luận có cấu trúc song đòi hỏi công sức xây dựng bản thể lớn; hướng dựa trên LLMs linh hoạt và tự nhiên nhưng thiếu kiểm soát về mức nhận thức và độ tin cậy. So với các tiếp cận đã được công bố, hệ thống của Tan và cộng sự (2025) tập trung vào phân loại Bloom và căn chỉnh CLO nhưng chưa tích hợp ước lượng tâm trắc học tiền triển khai; Zhai và cộng sự (2022) đi sâu vào chấm điểm tự động và phân tích năng lực nhưng hoạt động ở giai đoạn hậu kiểm; Circi và cộng sự (2023) tập trung vào sinh câu hỏi tự động mà chưa có cơ chế kiểm định chất lượng tích hợp. Mặc dù đạt nhiều tiến bộ, phần lớn nghiên cứu AIG tập trung vào sinh câu hỏi hơn là kiểm định chất lượng. Circi và cộng sự (2023) nhận định, các thách thức còn tồn tại gồm đảm bảo giá trị nội dung, kiểm soát mức Bloom, phát hiện mơ hồ ngữ nghĩa và duy trì độ tin cậy khi mở rộng quy mô.

Khoảng trống nghiên cứu ở đây là chưa có một quy trình đồng thời để: 1) Định lượng mức độ liên kết ngữ nghĩa với CLO, 2) Kiểm soát phân bố nhận thức theo thang Bloom, 3) Ước lượng độ khó/phân hóa trước khi có dữ liệu thực nghiệm.

Từ những thách thức và khoảng trống nêu trên, nghiên cứu này đề xuất một khung đánh giá chất lượng câu hỏi dựa trên trí tuệ nhân tạo hướng đến cung cấp một công cụ đánh giá chất lượng câu hỏi thông qua các chỉ số định lượng minh bạch, có thể lặp lại và mở rộng nhằm hỗ trợ chuyên gia trong quá trình đánh giá. Câu hỏi nghiên cứu đặt ra là: Liệu một khung tích hợp dựa trên AI, kết hợp phân tích ngữ nghĩa, xác thực sư phạm và ước lượng thống kê, có thể đánh giá chất lượng câu hỏi một cách hệ thống, định lượng và có khả năng mở rộng trước khi triển khai thực nghiệm không?

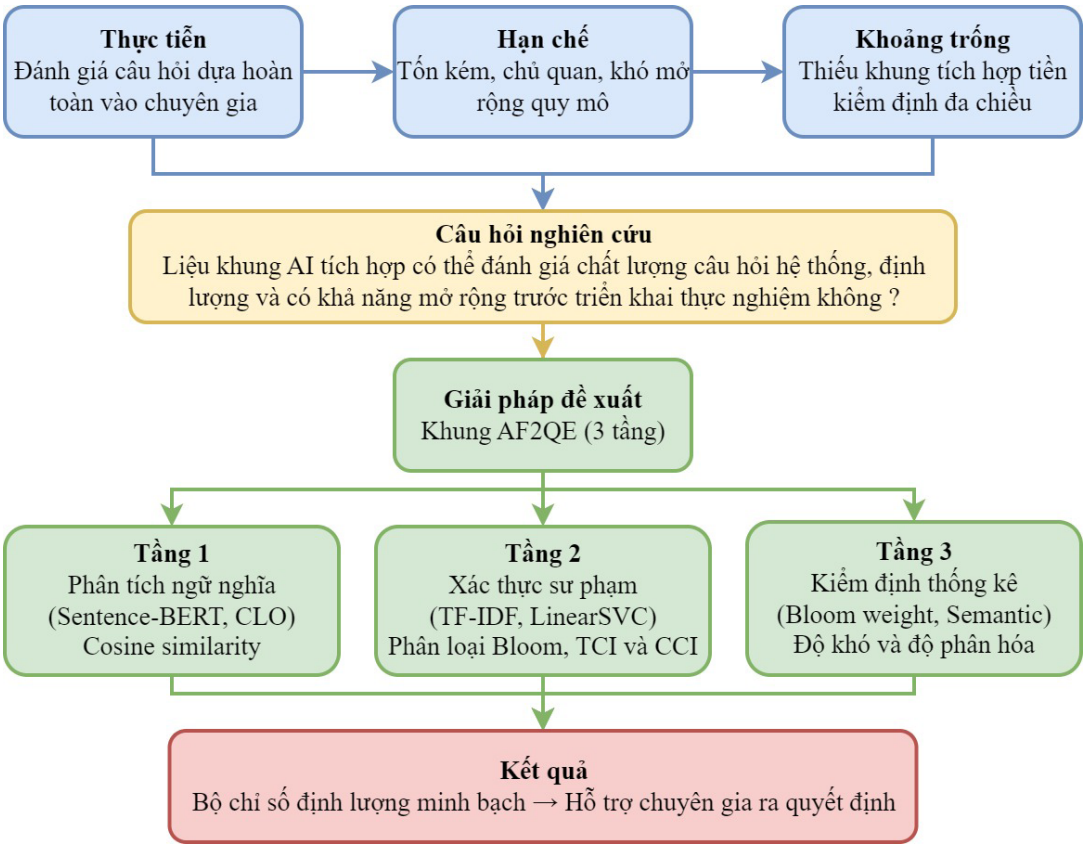
Hình 1 trình bày logic tổng thể của nghiên cứu. Xuất phát từ thực tiễn đánh giá câu hỏi phụ thuộc hoàn toàn vào chuyên gia, với các hạn chế về chi phí, tính chủ quan và khả năng mở rộng, nghiên cứu xác định khoảng trống về việc thiếu khung tiền kiểm định đa chiều. Từ đó, câu hỏi nghiên cứu được đặt ra và dẫn đến giải pháp AF2QE với ba tầng xử lí tuần tự: Phân tích ngữ nghĩa, xác thực sư phạm và kiểm định thống kê. Đầu ra cuối cùng là bộ chỉ số định lượng minh bạch, hỗ trợ chuyên gia ra quyết định.

2. Phương pháp nghiên cứu

2.1. Chỉ số đo lường câu hỏi

Trong đo lường giáo dục, chất lượng câu hỏi và

 https://doi.org/10.15625/2615-8957/12610807



Hình 1: *Logic tổng thể của nghiên cứu*

đề thi truyền thống thường được đánh giá dựa trên Thuyết kiểm thử cổ điển (Classical Test Theory - CTT) và Lí thuyết ứng đáp câu hỏi (Item Response Theory - IRT). Các chỉ số phổ biến gồm Độ khó (p), Độ phân hóa (D), hệ số tương quan point-biserial và Cronbach's alpha, đóng vai trò trung tâm trong phân tích đặc tính tâm trắc sau khi kì thi được triển khai trên mẫu thực tế (Kim và cộng sự, 2023; Toraman và cộng sự, 2022).

Theo Kim và cộng sự (2023), Độ khó (p) là tỉ lệ trả lời đúng, phản ánh mức độ dễ - khó của câu hỏi. Độ phân hóa (D) được tính qua nhóm trên - nhóm dưới hoặc tương quan điểm - tổng. Cronbach's alpha dao động 0.872 - 0.978 cho thấy độ tin cậy nội tại cao. Ở mức sâu hơn, IRT cho phép ước lượng tham số câu hỏi độc lập với mẫu và hàm phân tích thông tin. Toraman và cộng sự (2022) nhấn mạnh khả năng đánh giá giá trị và sai số chuẩn. Ngoài ra, Opait và cộng sự (2025) giới thiệu các chỉ số như Krippendorff's Alpha và Cohen's Kappa được dùng để đo mức độ đồng thuận giữa người chấm và hệ thống AI. Điểm chung của các chỉ số này là phụ thuộc vào dữ liệu sau kì thi. Do yêu cầu phản hồi thực tế để ước lượng tham số ổn định, chúng chủ yếu phục vụ phân tích hậu kiểm và chưa hỗ trợ hiệu quả cho đánh giá trước triển khai.

2.2. Ứng dụng trí tuệ nhân tạo trong đánh giá

Zhai và cộng sự (2022) cho rằng, các ứng dụng AI trong đánh giá giáo dục đang chuyển từ đo lường hiệu suất đơn thuần sang phân tích chi tiết đặc trưng nhận thức và năng lực người học. Nổi bật là tự động chấm bài tự luận với mức độ đồng thuận cao so với giám khảo, đo bằng Cohen's Kappa, đồng thời có thể phân tích thành phần kĩ năng như lập luận và sử dụng bằng chứng. AI hỗ trợ phân loại Bloom thông qua NLP và các mô hình học máy, góp phần kiểm tra phân bố nhận thức và căn chỉnh chuẩn đầu ra (Mohammed và Omar, 2018).

Bên cạnh đó, AI cung cấp phản hồi tự động gần thời gian thực, xác định sai lệch khái niệm và đề xuất cải thiện, thậm chí suy luận đặc điểm nhận thức để cá nhân hóa hỗ trợ (Zhai và cộng sự, 2022). Tuy nhiên, Tan và cộng sự (2025) chỉ ra rằng, phần lớn nghiên cứu tập trung vào các nhiệm vụ đơn lẻ như tạo sinh, phân loại hoặc chấm điểm, thiếu tích hợp và liên kết với các khung đo lường truyền thống như CTT hoặc IRT.

2.3. Khung đánh giá chất lượng câu hỏi đề xuất (AF2QE)

Khung AF2QE được đề xuất nhằm đảm bảo tính

nhất quán về nhận thức và độ tin cậy sư phạm, minh họa tại Hình 2. Hệ thống được thiết kế theo kiến trúc đa tầng, tích hợp cả phân tích ngữ nghĩa, xác thực sư phạm và kiểm định thống kê, nhằm cung cấp một quy trình đánh giá toàn diện từ trước đến sau khi sử dụng câu hỏi. AF2QE được xây dựng với ba tầng đánh giá theo cấu trúc hình thức, mỗi thành phần đại diện cho một phép biến đổi tuần tự tác động lên câu hỏi nhằm đảm bảo tính hợp lệ ở các khía cạnh khác nhau trong quá trình đánh giá.

$$AF2QE = (S, P, T) \qquad (1)$$

Để đảm bảo sự tương đồng ngữ nghĩa và phân loại chính xác theo thang Bloom, trước hết câu hỏi được đánh giá ở mức biểu diễn khái niệm bằng cách ánh xạ nội dung vào không gian vector ngữ nghĩa đa chiều, nơi các quan hệ khái niệm được lượng hóa thông qua các thước đo tương đồng phân bố. Cách tiếp cận này giảm sự phụ thuộc trùng khớp từ khóa bề mặt và cho phép so sánh định lượng với các CLO.

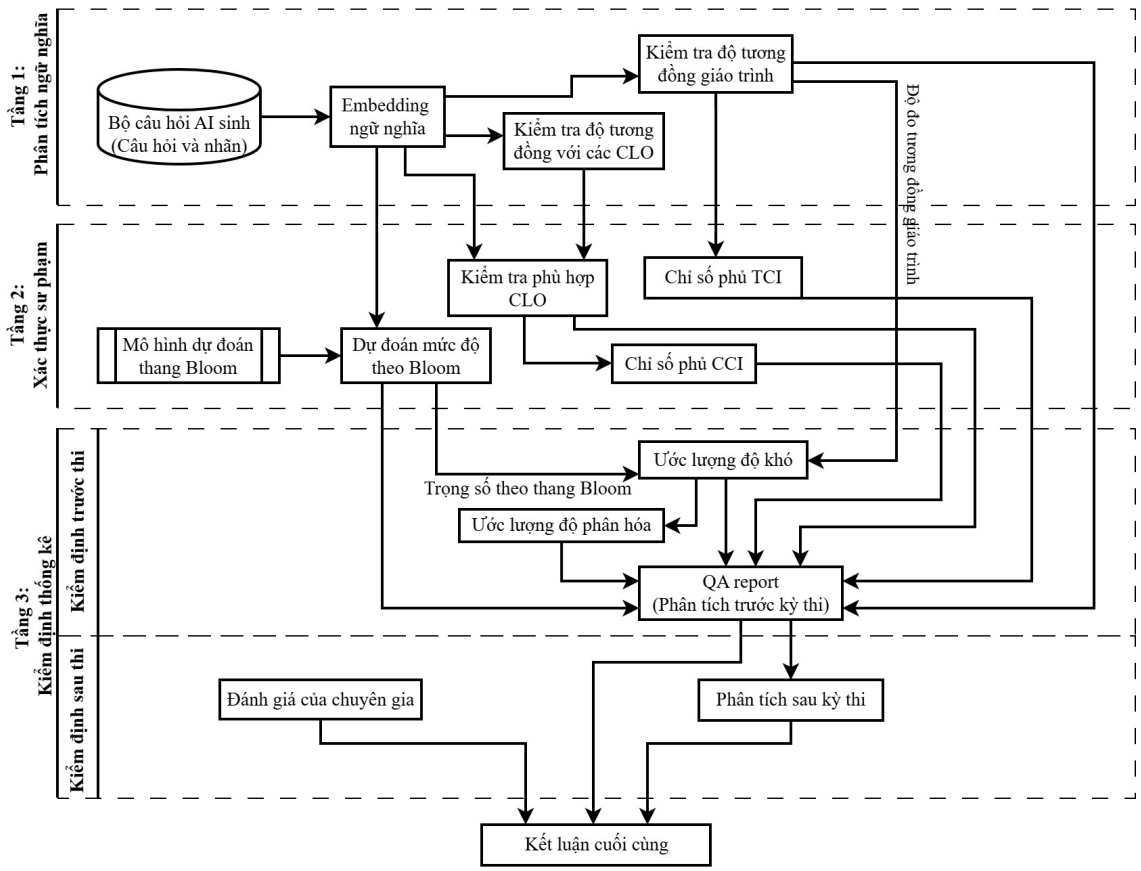
Tiếp theo, câu hỏi được đánh giá dưới góc độ xác thực sư phạm, tập trung vào thao tác nhận thức mà người học phải thực hiện thay vì chỉ nội dung bề mặt. Sự phù hợp giữa động từ hành động, ngữ cảnh và mức độ tư duy được kiểm tra thông qua cơ chế ánh xạ nhận thức, qua đó hạn chế sai lệch do mơ hồ ngôn ngữ và đảm bảo nguyên tắc căn chỉnh giữa mục tiêu, nội dung và đánh giá.

Cuối cùng, kiểm định thống kê trước thi sử dụng các tham số như độ khó và khả năng phân hóa được ước lượng qua mô hình đo lường và đối chiếu. Bước này xác định câu hỏi không chỉ phù hợp về nội dung và nhận thức mà còn đạt yêu cầu về độ tin cậy và giá trị đo lường để triển khai thực tế.

Như vậy, khung AF2QE thể hiện một chuỗi biến đổi liên hoàn: Từ biểu diễn ý nghĩa, đến căn chỉnh nhận thức và cuối cùng là kiểm định thống kê. Cách tiếp cận này cho phép quá trình xác thực câu hỏi được thực hiện một cách hệ thống, định lượng và có khả năng lặp lại, hỗ trợ chuyên gia ra quyết định.

2.3.1. Tầng Phân tích ngữ nghĩa

Tầng này được thiết kế nhằm đánh giá mức độ tương đồng ngữ nghĩa của câu hỏi với giáo trình, CLO. Sau khi tiền xử lí và phân đoạn tài liệu để loại bỏ nhiễu và đảm bảo mật độ thông tin học thuật, hệ thống sử dụng mô hình Sentence-BERT để ánh xạ mỗi đơn vị nội dung vào vector đa chiều, sau đó



Hình 2: *Khung AF2QE đề xuất*

 https://doi.org/10.15625/2615-8957/12610807

được chuẩn hóa và lưu trữ như một cơ sở tri thức ngữ nghĩa của học phần (Reimers và Gurevych, 2019).

Biểu diễn vector

Đầu vào của hệ thống gồm:

Câu hỏi: q

Tập đoạn giáo trình: $D = \{d_1, d_2, \ldots, d_n\}$

Tập chuẩn đầu ra: $C = \{c_1, c_2, \ldots, c_m\}$

Các thành phần này được ánh xạ vào vector đa chiều thông qua hàm embedding:

$$v_x = f(x) \qquad (2)$$

trong đó $f(\cdot)$ là mô hình Sentence-BERT và $x \in \{q, d_i, c_j\}$.

Để đảm bảo tính nhất quán trong tính toán độ tương đồng, các vector được chuẩn hóa về dạng L2:

$$\hat{v}_x = \frac{v_x}{\|v_x\|_2} \qquad (3)$$

Tính toán độ tương đồng ngữ nghĩa

Sau khi chuẩn hóa, độ tương đồng giữa câu hỏi và từng đoạn giáo trình hoặc từng CLO được tính bằng cosine similarity:

$$\text{sim}(q, d_i) = \hat{v}_q \cdot \hat{v}_{d_i} \qquad (4)$$

$$\text{sim}(q, c_j) = \hat{v}_q \cdot \hat{v}_{c_j} \qquad (5)$$

Giá trị này nằm trong khoảng [–1;1], phản ánh mức độ tương đồng ngữ nghĩa giữa các biểu diễn.

Các chỉ số đánh giá ngữ nghĩa

Từ các giá trị tương đồng, trích xuất hai chỉ số chính:

Chỉ số tương đồng tối đa giữa câu hỏi và một đoạn trong giáo trình:

$$SR_{max}(q) = \max_i \text{sim}(q, d_i) \qquad (6)$$

Chỉ số tương đồng trung bình:

$$SR_{avg}(q) = \frac{1}{n} \sum_{i=1}^{n} \text{sim}(q, d_i) \qquad (7)$$

2.3.2. Tầng Xác thực sư phạm

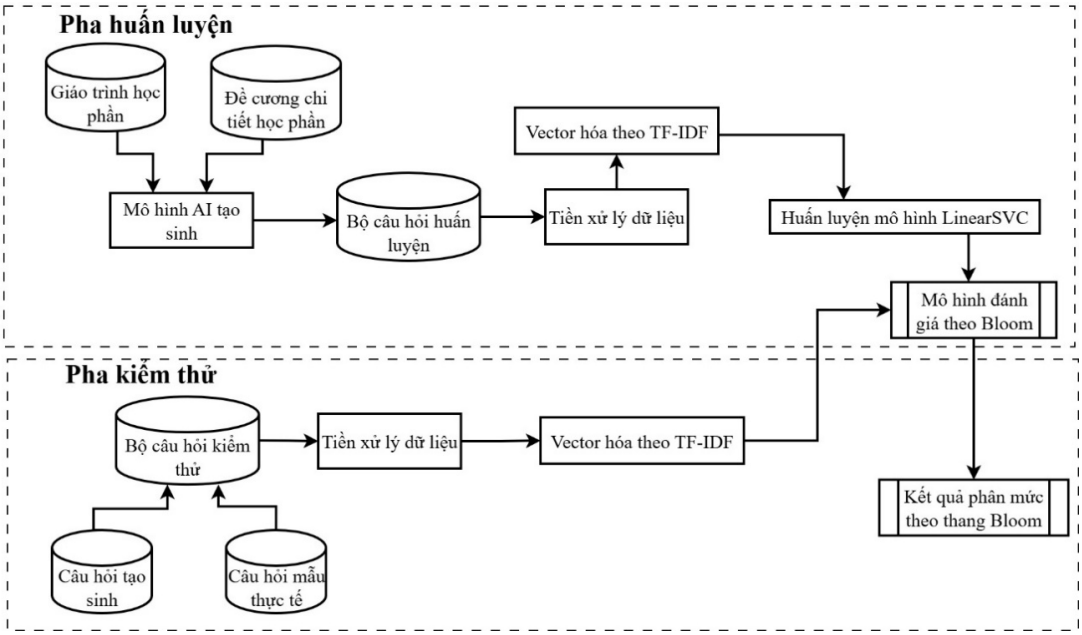
Kiểm tra tính chuẩn xác và phù hợp với mức độ theo thang Bloom thông qua mô hình học máy, độ phủ của câu hỏi và các chỉ số phù hợp với CLO (xem Hình 3).

Bước 1: Xây dựng mô hình và dự đoán nhãn theo thang Bloom cho câu hỏi

Theo Mohammed và cộng sự (2018), phương pháp sử dụng TF-IDF để vector hóa đặc trưng ngôn ngữ, sau đó áp dụng LinearSVC nhằm thực hiện phân loại trong không gian đặc trưng chiều cao. Cấu hình TF-IDF kết hợp LinearSVC được ưa chuộng trong phân loại văn bản học thuật nhờ tính đơn giản, ổn định và khả năng tổng quát hóa tốt.

Bước 2: Tính toán chỉ số bao phủ giáo trình (Textbook Coverage Index – TCI)

Bên cạnh việc tính mức độ liên quan ngữ nghĩa



Hình 3: *Xây dựng mô hình dự đoán theo thang Bloom*

tối đa cho từng câu hỏi, hệ thống còn mở rộng phân tích theo hướng bao phủ nội dung ở cả mức câu hỏi và mức tập câu hỏi. Khi sinh một số câu hỏi cho đề kiểm tra, các câu hỏi đó phải bao phủ được nội dung giáo trình và mỗi câu hỏi phải bao phủ được một phần của giáo trình (Singhania và cộng sự, 2022).

Chỉ số bao phủ ở mức câu hỏi

Với một câu hỏi **q** được xem là "bao phủ" nếu:

$$\text{sim}(q, d_i) \geq \tau \quad (8)$$

trong đó, τ là ngưỡng tương đồng.

Số đoạn được bao phủ:

$$N_{cov}(q) = \sum_{i=1}^{n} \text{sim}(q, d_i) \geq \tau \quad (9)$$

Chỉ số bao phủ của câu hỏi:

$$TCI(q) = \frac{N_{cov}(q)}{n} \quad (10)$$

$TCI_i \in [0,1]$ phản ánh tỉ lệ nội dung giáo trình mà câu hỏi đó có liên quan. Nếu giá trị thấp thì câu hỏi tập trung hẹp vào một phần kiến thức, còn nếu giá trị cao thì câu hỏi có phạm vi rộng hoặc mô tả khái quát.

Chỉ số bao phủ ở mức toàn bộ các câu hỏi

Để có thể đánh giá mức độ bao phủ tổng thể của tập câu hỏi đối với giáo trình, xác định với mỗi đoạn d_i xem có ít nhất một câu hỏi bao phủ hay không, thể hiện qua Z_i:

$$Z_i = \begin{cases} 1, \text{ nếu } \exists i: \text{sim}(q, d_i) \geq \tau \\ 0, \text{ ngược lại} \end{cases} \quad (11)$$

Chỉ số bao phủ toàn giáo trình:

$$TCI_{corpus} = \frac{1}{n} \sum_{i=1}^{n} Z_i \quad (12)$$

Chỉ số này cho biết tỉ lệ nội dung giáo trình được "chạm tới" bởi ít nhất một câu hỏi.

Bước 3: Xác định CLO phù hợp nhất

Để đánh giá được câu hỏi gần nhất với CLO nào, dựa vào chỉ số tương đồng giữa câu hỏi và CLO sim(q,c_j). Trong đó, độ tương đồng cao nhất được xác định bởi:

$$c^* = \underset{j}{\text{argmax}}\ \text{sim}(q, c_j) \quad (13)$$

Giá trị tương đồng tương ứng với:

$$SR_{CLO}(q) = \max_j (\text{sim}(q, c_j)) \quad (14)$$

Chỉ số này phản ánh mức độ gắn kết mạnh nhất giữa câu hỏi và một chuẩn đầu ra cụ thể.

Bước 4: Tính toán độ bao phủ chuẩn đầu ra (CLO Coverage Index – CCI)

Ngoài việc xác định CLO chính, hệ thống còn đánh giá mức độ bao phủ tổng thể thông qua chỉ số CCI. Cho ngưỡng tương đồng τ, tập các CLO được xem là có liên quan:

$$S = \{c_j | \text{sim}(q, c_j) \geq \tau\} \quad (15)$$

Chỉ số CCI được định nghĩa:

$$CCI = \frac{|S|}{m}, CCI \in [0,1] \quad (16)$$

Trong đó: |S| là số CLO đạt ngưỡng, m là tổng số CLO của học phần.

2.3.3. Tầng Kiểm định thống kê

Tại tầng này, hệ thống sẽ chia thành hai phần, phần Kiểm định trước thi và Kiểm định sau thi. Phần kiểm định trước thi sử dụng thang đo độ khó dựa trên Bloom kết hợp với độ tương đồng với giáo trình, nếu ở mức độ quá tương đồng với giáo trình kết hợp cùng Bloom ở mức thấp, độ khó của câu hỏi được sinh nằm ở mức đơn giản mà đa số người thi có thể thực hiện được.

Bước 1: Ước lượng độ khó

Theo Moghadamzadeh và cộng sự (2011), với bộ câu hỏi chưa có tính thực nghiệm, để kiểm tra độ khó trước khi thực nghiệm có thể dựa theo hai thành phần tiêu chí:

Thứ nhất, mức độ liên quan đến tài liệu ngữ nghĩa với giáo trình. Tại bước tính toán mức độ liên quan ngữ nghĩa tại tầng 1, chỉ số bao phủ đã được chuẩn hóa về $s \in [0,1]$, để mô hình hóa quan hệ phi tuyến giữa độ bám sát nội dung và độ khó, hàm logistic được sử dụng:

$$\text{semantic\_component} = \frac{1}{1 + e^{-k(0.5 - s)}} \quad (17)$$

trong đó, hệ số k nhằm điều chỉnh độ dốc và max(s) = 0.5 là điểm trung tâm.

Với hàm logistic này, s càng thấp thì càng yêu cầu sự suy luận vượt qua tài liệu, tương đương với độ khó tăng. Ngược lại, s càng cao thì càng bám sát tài liệu, độ khó giảm.

Thứ hai, thành phần nhận thức dựa trên thang Bloom. Mỗi mức Bloom $b \in \{BT1, BT2, ..., BT6\}$ được ánh xạ sang một trọng số nhận thức dựa theo thang điểm [0, 10] theo chương trình đào tạo và chuẩn hóa về $w_b \in [0,1]$ phản ánh giả định rằng, mức tư duy cao hơn tương ứng với độ khó lớn hơn. Tổng hợp lại, độ khó cuối cùng sẽ được tính bằng trung bình

 https://doi.org/10.15625/2615-8957/12610807

có trọng số:

$$dif = \alpha \times semantic\_component + (1-\alpha) \times w_b \quad (18)$$

Trong đó $\alpha \in [0,1]$ là hệ số cân bằng nhận thức và ngữ nghĩa.

Với dif ∈ [0,1] thể hiện rằng: Giá trị gần 0 biểu thị câu hỏi dễ, giá trị gần 1 biểu thị câu hỏi khó.

Bảng 1: *Trọng số theo thang Bloom*

Mức nhận thức	Số điểm đạt	Trọng số tương ứng
Ghi nhớ	1, 2	0.2
Thông hiểu	3, 4	0.3
Áp dụng	5, 6	0.5
Phân tích	7	0.7
Đánh giá	8	0.85
Sáng tạo	9, 10	0.95

Công thức này phản ánh giả định rằng, độ khó chịu ảnh hưởng đồng thời bởi hai yếu tố, đó là độ phức tạp nhận thức và khoảng cách ngữ nghĩa so với tài liệu học tập (xem Bảng 1).

Bước 2: Ước lượng độ phân hóa

Trong khi độ khó thể hiện cho từng câu sẽ khiến việc phân biệt năng lực giữa các thí sinh ở mức ổn định thì dựa trên độ phân hóa là thể hiện năng lực cao, thấp của từng thí sinh đó. Nhưng khi chưa có dữ liệu thực tế, hệ thống mô phỏng ước lượng độ phân hóa như một hàm của độ khó (Hillsley, 2025).

Dựa trên nguyên lí trong lí thuyết kiểm tra rằng các câu hỏi có độ khó trung bình thường phân hóa tốt nhất, sử dụng hàm parabol Bernoulli:

$$D = 4dif(1-dif) \quad (19)$$

Với D ∈ [0,1], giá trị 4 được nhân vào hàm parabol để chuẩn hóa hàm về [0, 1] nhằm mô hình hóa dễ hàng hơn, khi D=0 tương đương với việc độ khó đạt 0 hoặc 1, lúc đó phân hóa nằm ở việc câu quá dễ khiến đại đa số sẽ trả lời đúng, câu quá khó thì đại đa số sẽ trả lời sai, thể hiện sự không phân hóa. Tại D cực đại khi độ khó dif=0.5, câu thể hiện độ khó trung bình tạo ra khả năng phân biệt tốt hơn.

Bước 3: Kiểm định sau thi

Theo Kim và cộng sự (2023), sau khi kì thi được tổ chức và thu thập dữ liệu làm bài của thí sinh, hệ thống tiến hành kiểm định thực nghiệm nhằm xác thực các ước lượng độ khó và độ phân hóa đã được tính toán ở giai đoạn tiền kiểm định. Ở bước này, các chỉ số tâm trắc học cổ điển được sử dụng để đo lường đặc tính thực tế của từng câu hỏi, bao gồm tỉ lệ trả lời đúng để xác định độ khó cũng như các chỉ số tương quan giữa điểm câu hỏi và tổng điểm bài thi nhằm đánh giá khả năng phân biệt năng lực. Việc so sánh giữa giá trị ước lượng trước thi và dữ liệu thực tế cho phép hệ thống đánh giá mức độ chính xác của mô hình. Nếu xuất hiện sai lệch đáng kể thì các tham số như hệ số logistic hoặc trọng số Bloom có thể được điều chỉnh để nâng cao độ tin cậy cho các lần triển khai sau.

Bước 4: Kiểm định bởi chuyên gia

Bên cạnh dữ liệu thực nghiệm, hệ thống còn tích hợp đánh giá của chuyên gia nhằm đảm bảo tính đúng đắn nội dung của câu hỏi (Kim và cộng sự, 2023). Các chuyên gia có thể xem xét mức độ phù hợp của câu hỏi với chuẩn đầu ra học phần, độ chính xác của phân loại Bloom cũng như nhận định chủ quan về độ khó và khả năng phân hóa. Đánh giá này được thực hiện thông qua thang đo định lượng hoặc rubric phân tầng. Sự so sánh giữa nhận định chuyên gia và giá trị ước lượng của hệ thống giúp xác định mức độ đồng thuận giữa mô hình tự động và chuyên môn sư phạm. Trong trường hợp xuất hiện khác biệt đáng kể, hệ thống có thể được hiệu chỉnh để cải thiện độ phù hợp với thực tế. Do đó, kiểm định chuyên gia đóng vai trò bảo đảm tính giá trị nội dung và tăng cường độ tin cậy học thuật cho toàn bộ quy trình đánh giá.

2.4. Thiết lập huấn luyện và kiểm thử

2.4.1. Dữ liệu

Tập dữ liệu gồm: Giáo trình học phần Lập trình hướng đối tượng, chuẩn đầu ra CLO của học phần, 600 câu hỏi huấn luyện được sinh từ 03 mô hình ChatGPT, Gemini, Deepseek và được gán nhãn theo thang Bloom 6 mức, với phân bố tương đối đồng đều (khoảng 100 câu/mức). 50 câu hỏi kiểm thử được sinh từ 05 mô hình gồm ChatGPT, Gemini, Deepseek, Perplexity, Claude, có kiểm tra tránh leakage. Trích xuất đặc trưng sử dụng TF-IDF với n-gram (1-gram, 2-gram) và LinearSVC được tối ưu tham số bằng phương pháp kiểm tra chéo. Dữ liệu huấn luyện được chia x_train, x_test với tỉ lệ 8:2.

2.4.2. Phương pháp đánh giá

Chúng tôi sử dụng bốn chỉ số để đánh giá mô hình, bao gồm: ACC (độ chính xác tổng thể); F1-Score của từng mức, FPR (tỉ lệ dương tính giả từng

mức), FNR (tỉ lệ âm tính giả từng mức). Các chỉ số này được tính theo công thức sau:

$$ACC = \frac{\sum_{k=1}^{6} C_{kk}}{\sum_{i=1}^{6}\sum_{j=1}^{6} C_{ij}} \quad (20)$$

$$F1-score_k = \frac{2TP_k}{2TP_k + FP_k + FN_k} \quad (21)$$

$$FPR_k = \frac{FP_k}{FP_k + TN_k} \quad (22)$$

$$FNR_k = \frac{FN_k}{FN_k + TP_k} \quad (23)$$

trong đó, TP, FP, FN và TN là các tham số của ma trận nhầm lẫn theo từng mức được trình bày trong Bảng 2.

Ngoài ra, sử dụng tỉ lệ phân lớp chính xác cho từng mức trong thang Bloom (DR) để đánh giá hiệu suất của mô hình đề xuất trong giai đoạn kiểm thử:

$$DR = \frac{NoCorrect}{NoTest} \quad (24)$$

trong đó, NoCorrect là số lượng câu hỏi phân loại chính xác và NoTest là tổng số câu hỏi được sử dụng để phân loại.

Bảng 2: *Ma trận nhầm lẫn cho lớp k*

		Thực tế	
		Mức k	Mức khác k
Phán đoán	Mức k	$TP_k = C_{kk}$	$FP_k = \sum_{i \neq k} C_{ik}$
	Mức khác k	$FN_k = \sum_{j \neq k} C_{kj}$	$TN_k = \sum_{i \neq k, j \neq k} C_{ij}$

3. Kết quả nghiên cứu

3.1. Kết quả

3.1.1. Phân loại theo thang Bloom

Độ chính xác tổng thể ACC đạt 92.5%. Tỉ lệ âm tính giả, tỉ lệ dương tính giả và F1-score của từng lớp hiển thị tại Bảng 3.

Dữ liệu thử nghiệm được loại bỏ nhãn, sử dụng mô hình phân loại theo thang Bloom cho kết quả tại Bảng 4. Các câu hỏi ở mức Thông hiểu phân mức sai 02 câu, 01 câu sang mức Ghi nhớ, 01 câu sang mức Áp dụng.

Bảng 3: *Tỉ lệ F1-Score, FNR và FPR của các mức nhận thức*

Mức nhận thức	F1-score	FNR	FPR
Ghi nhớ	92.68%	5.00%	2.00%
Thông hiểu	82.93%	15.00%	4.00%
Áp dụng	97.44%	5.00%	0.00%
Phân tích	90.00%	10.00%	2.00%
Đánh giá	97.44%	5.00%	0.00%
Sáng tạo	95.00%	5.00%	1.00%

Bảng 4: *Kết quả phân lớp pha kiểm thử*

Lớp	Số lượng	Phân lớp chính xác	Tỉ lệ
Ghi nhớ	9	9	100.00%
Thông hiểu	8	6	75.00%
Áp dụng	8	8	100.00%
Phân tích	9	9	100.00%
Đánh giá	8	8	100.00%
Sáng tạo	8	8	100.00%

3.1.2. Tổng quan thống kê mô tả

Tập kiểm thử gồm 50 câu hỏi được đánh giá qua ba tầng của khung AF2QE. Kết quả đạt 0.742 về trung bình độ liên quan ngữ nghĩa (SR), 0.701 về sự tương đồng với chuẩn đầu ra CLO và 0.81 về độ phủ nội dung.

Các câu hỏi đạt ngưỡng SR ≥ 0.65 đạt 88%, cho thấy mức độ bám sát nội dung, có 6 câu có SR ≤ 0.6, chủ yếu thuộc nhóm suy luận mở rộng. Điều này cho thấy bộ câu hỏi có độ liên kết tốt với nội dung, đồng thời vẫn duy trì mức độ mở rộng hợp lí để đánh giá tư duy phân tích.

3.1.3. Phân loại theo thang Bloom dựa trên trọng số

Phân loại tự động cho 50 câu hỏi cho kết quả phân bố thể hiện tại Bảng 5:

Bảng 5 cho thấy, các câu hỏi ở mức Thông hiểu, Áp dụng và Phân tích được ước lượng nhiều hơn, các câu ở mức Ghi nhớ, Đánh giá, Sáng tạo được ước lượng ít hơn.

- Độ khó ước lượng trung bình: 0.53

 https://doi.org/10.15625/2615-8957/12610807

Bảng 5: *Phân loại câu hỏi theo thang Bloom dựa trên trọng số*

Mức nhận thức	Số câu theo thang Bloom	Số câu theo thang Bloom dựa trên trọng số	Tỉ lệ
Ghi nhớ	9	6	12%
Thông hiểu	8	11	22%
Áp dụng	8	14	28%
Phân tích	9	12	24%
Đánh giá	8	5	10%
Sáng tạo	8	2	4%

- Độ lệch chuẩn: 0.11

- Phân bố tập trung mạnh trong khoảng 0.45 – 0.60

Điều này cho thấy câu hỏi được thiết kế ở mức Trung bình Khá, phù hợp để tối ưu hóa khả năng phân hóa theo CTT. Khoảng 64% câu hỏi nằm trong vùng tối ưu ($0.4 \leq$ Độ khó ≤ 0.6).

3.1.4. Độ phân hóa ước lượng

Chỉ số độ phân hóa ước lượng được mô hình hóa theo hàm parabol đạt 0.68. Các câu có độ khó ước lượng gần 0.5 cho giá trị phân hóa cao nhất, phù hợp với kì vọng lí thuyết rằng, câu hỏi có độ khó trung bình thường phân biệt tốt nhất giữa nhóm học viên mạnh và yếu.

Qua đánh giá, 84% câu đạt chuẩn ở cả ba tiêu chí. Các ngưỡng này được chọn dựa trên tiêu chí của CTT: SR ≥ 0.65 phản ánh mức liên kết nội dung chấp nhận được; Bloom $\geq$ Áp dụng đảm bảo câu hỏi đánh giá tư duy bậc trung trở lên phù hợp đại học; độ khó 0.4–0.7 là vùng tối ưu theo CTT để phân hóa người học. 08 câu được khuyến nghị hiệu chỉnh, 03 câu do SR thấp và 05 câu chỉ ở mức độ Ghi nhớ.

Chỉ số Coverage Index 0.81 cho thấy nội dung đề thi bao phủ tốt nội dung giáo trình, không tập trung quá mức vào một chương cụ thể.

3.2. Thảo luận

Thực nghiệm trên 50 câu hỏi cho thấy khung AF2QE vận hành ổn định và nhất quán giữa ba tầng đánh giá. Ở tầng ngữ nghĩa, mức độ liên quan ngữ nghĩa trung bình đạt 0.742 (đa số ≥ 0.65), cho thấy mức độ bám sát giáo trình tốt; một số câu có SR thấp nhưng Bloom cao phản ánh nhóm câu hỏi yêu cầu suy luận bậc cao. Về sư phạm, hơn 50% câu hỏi tập trung ở mức Áp dụng và Phân tích, với trọng số Bloom trung bình 0.64. Tương quan dương giữa trọng số Bloom và ước lượng độ khó ($r \approx 0.61$) củng cố giả định rằng, mức độ nhận thức cao thường đi kèm độ khó lớn hơn. Đáng chú ý, mức Thông hiểu có F1-score thấp nhất (82.93%) và tỉ lệ phân lớp sai cao nhất trong pha kiểm thử (75%), phần lớn do ranh giới ngữ nghĩa giữa Ghi nhớ và Thông hiểu thường mờ trong ngôn ngữ tự nhiên - tương đồng với nhận định của Mohammed và cộng sự (2018) rằng hai mức nhận thức bậc thấp nhất là khó phân biệt nhất bằng đặc trưng từ vựng. So với hệ thống của Tan và cộng sự (2025) chỉ đạt khoảng 85% độ chính xác phân loại Bloom với LLM, kết quả 92.5% của AF2QE là tín hiệu khả quan, tuy cần thận trọng khi so sánh do khác biệt về ngôn ngữ và miền kiến thức.

Ở tầng thống kê, ước lượng độ khó trung bình đạt 0.53 (phần lớn trong khoảng 0.4–0.6) cho thấy đề thi tập trung trong vùng phân hóa tối ưu theo CTT; Độ phân hóa ước lượng trung bình 0.68 phù hợp với mô hình parabol lí thuyết, trong đó câu hỏi có độ khó trung bình đạt khả năng phân biệt cao nhất. Tương quan âm nhẹ giữa mức độ liên quan ngữ nghĩa và ước lượng độ khó ($r \approx -0.28$) gợi ý rằng, câu hỏi bám sát nội dung thường dễ hơn. Tổng thể, 84% câu hỏi đạt đồng thời ba tiêu chí, cho thấy AF2QE có tiềm năng như một công cụ tiền kiểm định hiệu quả.

Tuy nhiên, nghiên cứu còn một số hạn chế cần được thừa nhận: Thứ nhất, thực nghiệm chỉ được tiến hành trên 50 câu hỏi từ một học phần duy nhất. Do đó, khả năng khái quát hóa kết quả sang các lĩnh vực và ngôn ngữ khác còn hạn chế và cần được kiểm chứng thêm. Thứ hai, các chỉ số độ khó và độ phân hóa mới ở mức ước lượng tiền triển khai, chưa được đối chiếu với dữ liệu phản hồi thực nghiệm từ người học, nên chưa thể khẳng định độ chính xác của mô hình đo lường. Thứ ba, việc gán trọng số Bloom theo thang tuyến tính là giả định đơn giản hóa, có thể chưa phản ánh đầy đủ sự khác biệt nhận thức thực tế giữa các mức tư duy và cần được hiệu chỉnh dựa trên dữ liệu thực nghiệm.

4. Kết luận

Nghiên cứu này đề xuất và kiểm chứng khung AF2QE như một giải pháp tiền kiểm định chất lượng câu hỏi trong bối cảnh giáo dục đại học có sự hỗ trợ của AI tạo sinh. Với cấu trúc ba tầng gồm phân tích ngữ nghĩa, xác thực sư phạm và ước lượng thống kê, AF2QE hướng tới chuẩn hóa quy trình xây dựng và rà soát ngân hàng câu hỏi theo hướng hệ thống, minh bạch và có thể lặp lại. Kết quả thực nghiệm trên 50

câu hỏi cho thấy những tín hiệu ban đầu tích cực: Đa số câu hỏi đạt mức liên kết nội dung chấp nhận được với giáo trình, phân bố tương đối hợp lí theo thang Bloom và tập trung trong vùng độ khó trung bình. Các mối quan hệ giữa trọng số Bloom, mức độ liên quan ngữ nghĩa và ước lượng độ khó nhìn chung phù hợp với nguyên lí đo lường giáo dục. Tuy nhiên, cần lưu ý rằng, quy mô mẫu nhỏ và phạm vi thực nghiệm hạn chế chưa cho phép đưa ra kết luận tổng quát. AF2QE đóng vai trò hỗ trợ chuyên gia ra quyết định, cung cấp các chỉ số định lượng giúp phát hiện sớm câu hỏi lệch nội dung, sai mức nhận thức hoặc phân bố độ khó chưa hợp lí. Tỉ lệ lớn câu hỏi đạt đồng thời cả ba tiêu chí đánh giá cho thấy tính khả thi của khung trong giai đoạn tiền kiểm định.

Tài liệu tham khảo

Ahmed, A., Kerr, E. & O'Malley, A. (2025). Quality assurance and validity of AI-generated single best answer questions. *BMC Medical Education, 25*.

Circi, R., Hicks, J. & Sikali, E. (2023). Automatic item generation: foundations and machine learning-based approaches for assessments. *Frontiers in Education, 8*. DOI: https://doi.org/10.3389/feduc.2023.858273.

Gunawan, D., Sembiring, C. A. & Budiman, M. A. (2018). The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents. *2nd International Conference on Computing and Applied Informatics (ICCAI 2017).* DOI: 10.1088/1742-6596/978/1/012120.

He, X., Lin, Z., Gong, Y., Jin, A.-L., Zhang, H., Lin, C., Chen, W. (2024). AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Industry Track*, pp.165-190. Mexico City, Mexico.

Hillsley, K. (2025). Question format is the best predictor of item discrimination: a multivariable analysis. *Journal of Microbiology & Biology Education, 26*(3).

Khojasteh, L., Kafipour, R., Pakdel, F. & Mukundan, J. (2025). Empowering medical students with AI writing co-pilots: design and validation of AI self-assessment toolkit. *BMC Medical Education, 25*. DOI :https://doi.org/10.1186/s12909-025-06753-3.

Kim, Y. H., Kim, B. H., Kim, J., Jung, B. & Bae, S. (2023). Item difficulty index, discrimination index, and reliability of the 26 health professions licensing examinations in 2022, Korea: a psychometric study. *Journal of Educational Evaluation for Health Professions, 20*. DOI: https://doi.org/10.3352/jeehp.2023.20.31.

Makhlouf, K., Amouri, L., Chaabane, N. & EL-Haggar, N. (2020). Exam Questions Classification Based on Bloom's Taxonomy: Approaches and Techniques. *2020 2nd International Conference on Computer and Information Sciences (ICCIS).* Sakaka, Saudi Arabia. DOI: 10.1109/ICCIS49240.2020.9257698.

Moghadamzadeh, A., Salehi, K. & Khodaie, E. (2011). A Comparison the Information Functions of the Item and Test in One, Two and Three Parametric Model of the Item Response Theory (IRT). *International Conference on Education and Educational Psychology (ICEEPSY 2011).* Istanbul, Turkey.

Mohammed, M. & Omar, N. (2018). Question Classification Based on Bloom's Taxonomy Using Enhanced TF-IDF. *International Journal on Advanced Science, Engineering and Information Technology, 8*(4-2), pp.1679–1685. DOI: https://doi.org/10.18517/ijaseit.8.4-2.6835.

Opait, E.-E., Duca, A. & Olariu, M.-E. (2025). Evaluating Generative AI and Human Performance in Question-Answer Validation Tasks. *29th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2025), 270*, pp.3973-3982. DOI: https://doi.org/10.1016/j.procs.2025.09.522.

Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.* Hong Kong, China.

Singhania, S., Razniewski, S. & Weikum, G. (2022). Predicting Document Coverage for Relation Extraction. *Transactions of the Association for Computational Linguistics (TACL), 10*, pp.207-223.

Taheri, E. & Li, N. (2023). L2 Norm-Based Control Regularization for Solving Optimal Control Problems. *IEEE Access, 11*. DOI: 10.1109/ACCESS.2023.3331382.

Tam, D. T. & Quynh, P. T. (2024). Alignment between Course Learning Outcomes and Assessments: An Analysis. *International Journal of TESOL & Education, 4*(2). DOI: https://doi.org/10.54855/ijte.24422.

Tan, B., Armoush, N., Mazzullo, E., Bulut, O. & Gierl, M. J. (2025). A review of automatic item generation techniques leveraging large language. *International Journal of Assessment Tools in Education, 12*(2), pp.317-340. DOI: https://doi.org/10.21449/ijate.1602294.

Toraman, Ç., Karadağ, E. & Polat, M. (2022). Validity and reliability evidence for the scale of distance education satisfaction of medical students based on item response theory (IRT). *BMC Medical Education, 22*. DOI: https://doi.org/10.1186/s12909-025-06753-3.

Zhai, X., Haudek, K. C. & Ma, W. (2022). Assessing Argumentation Using Machine Learning and Cognitive Diagnostic Modeling. *Research in Science Education, 53*, pp.405-424.

 https://doi.org/10.15625/2615-8957/12610807