

ỨNG DỤNG TRÍ TUỆ NHÂN TẠO TRONG ĐÁNH GIÁ NĂNG LỰC TƯ DUY NGÔN NGỮ QUA CÁC BÀI TẬP TỰ LUẬN CỦA SINH VIÊN

NGUYỄN NGỌC ANH¹

VŨ THỊ KIM HOA²

NGUYỄN XUÂN THẮNG³

Abstract: In the context of the increasing application of artificial intelligence (AI) in education, the use of artificial intelligence in essay assessment opens up new approaches to educational evaluation. This study aims to propose a framework for assessing students' linguistic thinking competence and to examine the applicability of AI in evaluating its components through essay writing tasks, drawing on selected courses at Hanoi University. The paper focuses on clarifying the theoretical foundations of linguistic thinking competence assessment and analyzing the role of essay-based tasks in evaluating this competence. It also presents the theoretical basis for AI-assisted essay scoring and reports an experimental implementation involving task development, rubric design, and prompt construction for - AI-assisted assessment.

Keywords: artificial intelligence; linguistic thinking competence; essay writing tasks; rubric; prompt design.

1. Dẫn nhập

Trong bối cảnh đổi mới giáo dục theo định hướng phát triển năng lực, hoạt động đánh giá trong giáo dục đại học đang chuyển từ việc đo lường khả năng tái hiện kiến thức sang chú trọng năng lực vận dụng, phân tích và kiến tạo tri thức của người học. Trong đó, năng lực tư duy ngôn ngữ thể hiện qua khả năng tổ chức ý tưởng, triển khai lập luận, sử dụng phương tiện diễn đạt phù hợp và tạo lập văn bản mạch lạc đang trở thành một trong những năng lực cốt lõi cần được nhận diện và đánh giá một cách hệ thống.

Trong xu hướng đó, bài tập tự luận đóng vai trò quan trọng, như một công cụ đánh giá có khả năng phản ánh toàn diện năng lực tư duy của sinh viên, vượt ra ngoài phạm vi tái hiện kiến thức đơn thuần. Thông qua việc phân tích, lập luận và diễn đạt, người học bộc lộ rõ năng lực tư duy ngôn ngữ của cá nhân. Tuy nhiên, trên thực tế, việc chấm và phản hồi các bài tự luận đòi hỏi giảng viên (GV) phải đầu tư đáng kể về thời gian và công sức, đặc biệt khi cần đưa ra nhận xét chi tiết, mang tính cá nhân hoá nhằm chỉ ra mức độ phát triển năng lực của từng sinh viên. Trong các lớp học đông, áp lực này trở nên rõ rệt khiến cho việc nhận xét khó chi tiết và dễ thiếu thống nhất trong đánh giá. Vì vậy, vấn đề ứng dụng công nghệ, đặc biệt là trí tuệ nhân tạo để hỗ trợ hoạt động đánh giá năng lực trở thành nhu cầu tất yếu.

Hiện nay, các hệ thống chấm điểm tự động bài tự luận (Automated Essay Scoring System - AES) không chỉ giúp rút ngắn thời gian chấm bài, nâng cao tính nhất quán trong đánh giá mà còn hỗ trợ phản hồi về cấu trúc bài viết, lập luận và cách sử dụng ngôn ngữ. Đặc biệt, sự xuất hiện của các mô hình ngôn ngữ lớn (Large Language Model - LLM) tiếp tục mở rộng năng lực của AES, cho phép hệ thống tiếp cận sâu hơn tới các biểu hiện tư duy của người học như tổ chức ý tưởng, lập luận và khả năng diễn đạt, qua đó góp phần nâng cao hiệu quả đánh giá năng lực tư duy ngôn ngữ. Tuy nhiên, khảo sát các nghiên cứu hiện nay cho thấy phần lớn các công trình về AES vẫn chủ yếu tập trung vào việc xác định điểm số tổng thể của bài viết, trong khi việc tiếp cận theo hướng phân tích cấu trúc nội tại của năng lực tư duy ngôn ngữ còn chưa được quan tâm đúng mức. Khoảng trống này đặt ra yêu cầu cần thiết phải xây dựng những khung lý thuyết và công cụ đánh giá có khả năng phân tách và đo lường các thành tố cụ thể của năng lực, thay vì chỉ dừng lại ở đánh giá kết quả đầu ra.

Xuất phát từ những vấn đề nêu trên, bài báo hướng đến hai mục tiêu: (1) đề xuất khung đánh giá năng lực tư duy ngôn ngữ của sinh viên; (2) khảo sát mức độ tương hợp giữa kết quả chấm điểm của

¹ Trường Đại học Hà Nội; Email: anhnn@hanu.edu.vn

² Trường Đại học Hà Nội

³ Trường Đại học Hà Nội

AI và giảng viên đối với điểm tổng bài tự luận khi cùng vận dụng khung đánh giá này trong các học phần Nhập môn Việt ngữ học, Văn bản tiếng Việt, Cơ sở văn hoá Việt Nam tại Trường Đại học Hà Nội. Qua đó, nghiên cứu góp phần hoàn thiện cơ sở lý luận và thực tiễn cho việc vận dụng khung đánh giá năng lực tư duy ngôn ngữ trong môi trường đánh giá có sự hỗ trợ của AI.

2. Cơ sở lý thuyết

2.1. Lý thuyết về đánh giá năng lực ngôn ngữ trong dạy học

2.1.1. Đánh giá năng lực ngôn ngữ: khái niệm và khung tiếp cận

(1) Khái niệm năng lực

Năng lực là một khái niệm trung tâm trong giáo dục hiện đại, được tiếp cận từ nhiều góc độ khác nhau. Trong các nghiên cứu của nước ngoài, năng lực thường được quy vào phạm trù khả năng hành động hiệu quả trong bối cảnh cụ thể. Chẳng hạn, OECD (2005) cho rằng năng lực là “khả năng đáp ứng một cách hiệu quả những yêu cầu phức hợp trong một bối cảnh cụ thể”; Denyse Tremblay (2002) cho rằng năng lực là “khả năng hành động, thành công và tiến bộ dựa vào việc huy động và sử dụng hiệu quả tổng hợp các nguồn lực để đối mặt với các tình huống trong cuộc sống”. Điểm chung của các quan niệm này là đều nhấn mạnh đến tính hành động, tính hiệu quả và khả năng huy động tổng hợp các nguồn lực của năng lực.

Trong khi đó, các nhà giáo dục học Việt Nam có xu hướng tiếp cận năng lực như một thuộc tính hoặc phẩm chất của cá nhân gắn với hoạt động. Theo Từ điển tiếng Việt của Hoàng Phê (2020), năng lực là “khả năng, điều kiện chủ quan hoặc tự nhiên sẵn có để thực hiện một hoạt động nào đó” hoặc “phẩm chất tâm sinh lý và trình độ chuyên môn tạo cho con người khả năng hoàn thành một loại hoạt động nào đó với chất lượng cao”. Tác giả Hoàng Hòa Bình (2015) cho rằng năng lực là thuộc tính cá nhân được hình thành và phát triển thông qua học tập, rèn luyện, được bộc lộ qua hoạt động và đảm bảo hoạt động đạt hiệu quả.

Từ các quan điểm trên, có thể thấy dù cách diễn đạt khác nhau, các nghiên cứu đều thống nhất những đặc trưng cốt lõi của năng lực như: (i) gắn với hoạt động và được bộc lộ qua hoạt động; (ii) thể hiện khả năng thực hiện hoạt động một cách hiệu quả; (iii) là sự tổng hợp của nhiều thành tố như tri thức, kỹ năng và thái độ. Đây cũng là cơ sở quan trọng để chuyển từ tiếp cận giáo dục dựa trên nội dung sang tiếp cận phát triển năng lực, đồng thời định hướng cho việc xây dựng các phương thức đánh giá năng lực người học trong giáo dục hiện nay.

(2) Khái niệm đánh giá năng lực

Trong giáo dục hiện đại, đánh giá năng lực (Competency-based Assessment) trở thành xu hướng tất yếu, thay thế cho cách tiếp cận đánh giá dựa trên tái hiện kiến thức. Đánh giá năng lực không chỉ nhằm xác định mức độ nắm vững tri thức mà quan trọng hơn là đo lường khả năng vận dụng tri thức đó trong các tình huống học tập và thực tiễn. Điều này gắn liền với sự chuyển dịch của giáo dục từ mô hình “truyền thụ kiến thức” sang mô hình “phát triển năng lực”, trong đó người học được kì vọng có thể làm chủ tri thức, kỹ năng và đáp ứng yêu cầu của xã hội và thị trường lao động. Vì vậy, đánh giá năng lực cần được thiết kế theo hướng tích hợp, chú trọng vào quá trình thực hiện nhiệm vụ và sản phẩm học tập, qua đó phản ánh một cách toàn diện năng lực của người học. Theo Nguyễn Công Khanh (2015), đánh giá theo cách tiếp cận năng lực là đánh giá theo chuẩn về sản phẩm đầu ra... nhưng sản phẩm đó không chỉ là kiến thức, kỹ năng mà chủ yếu là khả năng vận dụng kiến thức, kỹ năng và thái độ cần có để thực hiện nhiệm vụ học tập đạt tới một chuẩn nào đó.

Như vậy, xét một cách khái quát, có thể coi đánh giá năng lực là quá trình thu thập thông tin và chứng cứ về khả năng vận dụng kiến thức, kỹ năng của học sinh để giải quyết các vấn đề trong học tập và đời sống thực tiễn. Khác với đánh giá truyền thống vốn tập trung vào việc ghi nhớ lý thuyết, phương pháp này chú trọng vào việc người học “làm được gì” từ những điều đã học.

(3) Đánh giá năng lực tư duy ngôn ngữ

- Khái niệm năng lực tư duy ngôn ngữ

Trong nghiên cứu ngôn ngữ học, ngôn ngữ không chỉ là phương tiện giao tiếp mà còn là công cụ của tư duy. Vì vậy, năng lực tư duy ngôn ngữ của người học cần được xem xét thông qua các sản phẩm học thuật mà người học sử dụng ngôn ngữ để biểu đạt tư duy.

Trong phạm vi nghiên cứu này, năng lực tư duy ngôn ngữ được xác định là khả năng của người học trong việc huy động tri thức ngôn ngữ và tri thức nền tảng liên quan để hiểu, phân tích, lí giải và chứng minh các vấn đề, đồng thời biểu đạt các kết quả tư duy đó bằng ngôn ngữ một cách rõ ràng, logic và có hệ thống.

Xét về cấu trúc, năng lực tư duy ngôn ngữ có thể được xem là sự tích hợp của ba thành tố cơ bản. Thứ nhất, tri thức nền tảng, thể hiện ở khả năng xác định đúng các nội dung, luận điểm và khái niệm liên quan đến vấn đề. Thứ hai, kĩ năng lập luận, thể hiện ở khả năng phân tích, giải thích, chứng minh và triển khai các luận điểm thông qua hệ thống lập luận và dẫn chứng phù hợp. Thứ ba, kĩ năng diễn đạt, thể hiện ở khả năng tổ chức bài viết theo cấu trúc rõ ràng và trình bày nội dung bằng ngôn ngữ học thuật chính xác, mạch lạc.

- Khung đánh giá năng lực tư duy ngôn ngữ

Từ cách tiếp cận trên, năng lực tư duy ngôn ngữ có thể được cụ thể hoá thành một khung đánh giá với các tiêu chí phản ánh những biểu hiện có thể quan sát được trong sản phẩm học thuật của người học. Khung đánh giá này được xây dựng theo hướng tiếp cận năng lực, nhấn mạnh không chỉ kết quả tiếp nhận tri thức mà còn chú trọng quá trình huy động, tổ chức và diễn đạt tri thức thông qua ngôn ngữ.

Việc đánh giá năng lực tư duy ngôn ngữ tập trung vào ba phương diện chính. Thứ nhất, năng lực huy động tri thức nền tảng, thể hiện qua khả năng xác định đúng vấn đề, nắm vững nội dung cốt lõi, hiểu và giải thích được các khái niệm, hiện tượng hoặc luận điểm liên quan đến yêu cầu của bài viết. Đây là cơ sở để người học triển khai các thao tác tư duy ở mức độ cao hơn.

Thứ hai, năng lực lập luận, thể hiện qua khả năng phân tích, so sánh, khái quát hoá và triển khai luận điểm theo trình tự logic. Phương diện này còn được biểu hiện ở việc người học biết lựa chọn và sử dụng dẫn chứng, ví dụ hoặc dữ liệu phù hợp nhằm làm rõ và tăng sức thuyết phục cho lập luận. Đây được xem là thành tố trung tâm của năng lực tư duy ngôn ngữ bởi nó phản ánh khả năng tổ chức và vận hành tư duy thông qua ngôn ngữ.

Thứ ba, năng lực diễn đạt học thuật, thể hiện qua khả năng tổ chức bài viết theo cấu trúc rõ ràng, liên kết chặt chẽ giữa các ý, đồng thời sử dụng ngôn ngữ chính xác, phù hợp với phong cách học thuật. Phương diện này không chỉ phản ánh khả năng sử dụng ngôn ngữ mà còn cho thấy mức độ mạch lạc và hiệu quả trong quá trình biểu đạt tư duy của người học.

Thứ tư, năng lực tư duy sáng tạo, thể hiện qua khả năng tiếp cận vấn đề từ những góc nhìn riêng, biết kết nối và mở rộng vấn đề trong những bối cảnh khác nhau, đồng thời đưa ra những nhận xét, đánh giá hoặc đề xuất mang dấu ấn cá nhân trên cơ sở tri thức học thuật. Phương diện này phản ánh mức độ chủ động, linh hoạt và sáng tạo của người học trong quá trình kiến tạo và vận dụng tri thức thông qua ngôn ngữ.

Từ bốn phương diện trên, hệ thống tiêu chí đánh giá được thiết kế nhằm đo lường cụ thể từng biểu hiện của năng lực tư duy ngôn ngữ trong bài viết tự luận. Các tiêu chí không tồn tại tách biệt mà có mối liên hệ chặt chẽ, bổ sung cho nhau trong việc phản ánh toàn diện năng lực của người học. Việc xây dựng khung đánh giá theo hướng phân tách các thành tố cụ thể giúp quá trình đánh giá trở nên minh bạch, có hệ thống và nhất quán hơn, đồng thời tạo cơ sở cho việc ứng dụng trí tuệ nhân tạo trong nhận diện và phân tích các biểu hiện của tư duy ngôn ngữ thông qua bài viết học thuật.

- Vai trò của bảng tiêu chí đánh giá (rubric) trong đánh giá năng lực tư duy ngôn ngữ

Các tiêu chí đánh giá nêu trên cần được cụ thể hoá thông qua rubric nhằm chuyển hoá các yêu cầu mang tính định tính thành các mức độ đánh giá rõ ràng và có thể đo lường. Việc sử dụng rubric không chỉ giúp đảm bảo tính nhất quán và độ tin cậy trong quá trình chấm điểm mà còn góp phần minh bạch hoá tiêu chuẩn đánh giá, từ đó hỗ trợ người học định hướng cách thức thực hiện bài làm.

Bên cạnh đó, rubric còn đóng vai trò như một công cụ kết nối giữa lí thuyết về năng lực và hoạt động đánh giá trong thực tiễn, cho phép các tiêu chí năng lực được cụ thể hoá thành những biểu hiện quan sát được trong sản phẩm học tập của người học. Nhờ đó, việc đánh giá không chỉ dừng lại ở nhận định cảm tính mà được thực hiện trên cơ sở các tiêu chí rõ ràng và có hệ thống.

2.1.2. Đánh giá năng lực tư duy ngôn ngữ qua các bài tập tự luận

Trong các hình thức đánh giá hiện nay, bài tập tự luận giữ vai trò quan trọng trong việc đo lường những khía cạnh phức tạp của năng lực ngôn ngữ, đặc biệt là năng lực tư duy ngôn ngữ. Khác với các dạng bài kiểm tra lựa chọn có sẵn phương án trả lời, bài tập tự luận yêu cầu người học phải huy động

tri thức, tổ chức ý tưởng và kiến tạo câu trả lời một cách độc lập. Nhờ đó, hình thức này không chỉ phản ánh mức độ ghi nhớ kiến thức mà còn cho phép đánh giá khả năng hiểu, vận dụng và triển khai tri thức trong những ngữ cảnh cụ thể.

So với trải nghiệm khách quan, vốn chủ yếu kiểm tra khả năng nhận diện và lựa chọn đáp án đúng, bài tập tự luận có ưu thế trong việc đánh giá các năng lực tư duy bậc cao. Người học không chỉ cần nắm vững kiến thức mà còn phải biết phân tích, lập luận và trình bày quan điểm một cách logic, mạch lạc. Đây cũng là hình thức cho phép đánh giá năng lực lập luận và tư duy phản biện thông qua cách tổ chức và phát triển ý tưởng trong bài viết. Bên cạnh đó, bài tập tự luận còn tạo điều kiện để người học thể hiện đặc điểm cá nhân và phong cách sử dụng ngôn ngữ, qua đó phản ánh rõ nét hơn năng lực tư duy ngôn ngữ của từng cá nhân.

Trong bài làm tự luận, năng lực tư duy ngôn ngữ của người học được biểu hiện qua nhiều phương diện. Trước hết là khả năng phân tích và giải thích các khái niệm ngôn ngữ học, thể hiện mức độ nắm vững lý thuyết. Tiếp theo là khả năng xử lý ngữ liệu và triển khai luận điểm, phản ánh năng lực tư duy và lập luận. Cuối cùng là khả năng diễn đạt bằng ngôn ngữ học thuật, thể hiện qua tính chính xác, mạch lạc và phù hợp của cách trình bày. Những biểu hiện này tạo cơ sở để xây dựng các tiêu chí đánh giá cụ thể trong việc đo lường năng lực tư duy ngôn ngữ của sinh viên.

Với những đặc trưng trên, bài tập tự luận là hình thức phù hợp để đánh giá năng lực tư duy ngôn ngữ của sinh viên trong nghiên cứu này.

2.2. Ứng dụng AI trong đánh giá bài tự luận

2.2.1. Tổng quan về AES và ứng dụng AI trong chấm bài tự luận

Trong những năm gần đây, sự phát triển vượt bậc của các mô hình trí tuệ nhân tạo mà tiêu biểu là mô hình ngôn ngữ lớn đã mang lại những lợi ích đáng kể trong lĩnh vực giáo dục - đào tạo. LLM là các mô hình xác suất có khả năng hiểu và sinh ngôn ngữ tự nhiên dựa trên kiến thức được thu thập từ các tập dữ liệu cực lớn, vì vậy được ứng dụng rộng rãi trong dịch thuật, chatbot, trợ lý ảo và các hệ thống hỗ trợ học tập.

Trong lĩnh vực đánh giá, hệ thống chấm điểm tự động bài luận (Automated Essay Scoring - AES) sử dụng các mô hình thống kê hoặc AI để gán điểm cho bài viết (Rodriguez et al., 2019). Ban đầu, AES được phát triển nhằm giảm chi phí và thời gian chấm bài, đồng thời tăng tính nhất quán trong đánh giá. Với sự phát triển của AI, AES không chỉ hỗ trợ cho điểm mà còn có thể phân tích bài tự luận theo thời gian thực, cung cấp phản hồi về ngữ pháp, cách dùng từ, phong cách, tổ chức và phát triển ý (Mansour et al., 2024).

Lợi ích của ứng dụng AI trong chấm tự luận là rõ ràng: tăng tốc độ chấm, bảo đảm tính nhất quán theo thời gian, hỗ trợ kiểm tra theo yêu cầu, cho phép phản hồi nhanh và mở đường cho cá thể hoá học tập. Tuy nhiên, AES vẫn có giới hạn khi đánh giá các phương diện chuyên sâu của năng lực tư duy ngôn ngữ như hàm ý, chất lượng lập luận, ý nghĩa xã hội của ngôn ngữ hoặc những khác biệt tinh tế giữa các nhóm người học. Vì vậy, xu hướng hiện nay không xem AI là công cụ thay thế người chấm, mà là hệ thống hỗ trợ quyết định đánh giá và phản hồi (Wei et al., 2023).

2.2.2. Khung đánh giá bài tự luận với sự hỗ trợ của AI

Khi được yêu cầu đánh giá bài tự luận, AI không “chấm điểm” theo cách hiểu và diễn giải bài viết giống như con người. Hệ thống tự động chấm điểm vận hành bằng cách nhận diện các đặc trưng của văn bản, hoặc quan hệ giữa các đặc trưng đó với điểm số mẫu được chấm bởi người chấm, rồi áp dụng mô hình cho bài viết mới. Độ tin cậy khi chấm điểm của AI chỉ đạt được khi AI được huấn luyện kỹ theo rubric chấm, được cung cấp bài mẫu đã gán mức điểm và phải vượt qua bước hiệu chuẩn trước khi được đưa vào triển khai thực tế. Do đó, với AI, vai trò của rubric và câu lệnh (prompt) chấm không phải phụ trợ mà là phần lõi quan trọng tạo nên tính chính xác và độ tin cậy của hệ thống. Nếu không có rubric rõ ràng và prompt chấm phù hợp, đầu ra của AI chỉ là một phán đoán ngôn ngữ có vẻ hợp lý chứ chưa phải một quyết định đánh giá có giá trị đo lường (Pack et al., 2024).

Một khung đánh giá bài tự luận với sự hỗ trợ AI nên được tổ chức theo chuỗi logic. Trước hết, cần xác định rõ năng lực cần đo: chấm để cho điểm tổng kết, để phản hồi quá trình học, hay để đánh

giá năng lực tư duy ngôn ngữ. Tiếp theo, cần xây rubric với các tiêu chí quan sát được, mô tả được mức độ và bằng chứng điển hình cho từng mức điểm. Sau đó, rubric cần được chuyển hoá thành prompt chấm điểm, yêu cầu AI đọc đề, đối chiếu từng tiêu chí, trích dẫn các bằng chứng trong bài luận, cho điểm sơ bộ theo từng tiêu chí, giải thích lí do đánh giá, chỉ ra những điểm chưa đáp ứng yêu cầu và bảo đảm phản hồi chỉ dừng ở mức phân tích, nhận xét thay vì can thiệp viết lại nội dung cho người học. Bước tiếp theo là hiệu chuẩn prompt chấm bằng các bài mẫu chuẩn hoặc các ví dụ đã được người chấm xác nhận. Cuối cùng, cần có sự kiểm tra lại của GV đối với các trường hợp điểm thấp bất thường, hoặc điểm chấm được sử dụng trong các quyết định có rủi ro cao. Đồng thời phải giám sát liên tục kết quả chấm của AI để kịp thời điều chỉnh khi phát hiện kết quả chấm có thiên lệch, thiếu nhất quán, hoặc khi có sự thay đổi về trình độ và năng lực của người học. Như vậy, AI chỉ hỗ trợ đánh giá bài tự luận một cách có hệ thống khi nó được đặt trong một hệ quy chiếu đánh giá đã được con người thiết kế trước (Stahl et al., 2024).

Các nghiên cứu gần đây cho thấy việc thêm tiêu chí rubric, ví dụ mẫu và cấu trúc nhiệm vụ phù hợp có thể giúp cải thiện chất lượng chấm của AI. Ở góc độ kỹ thuật, prompt chấm tốt không nhất thiết phải càng dài càng tốt. Quan trọng hơn là prompt chấm phải đặc tả rõ tiêu chí và đầu ra cần có. Một nghiên cứu so sánh nhiều mô hình AI chấm bài luận cho thấy ở ba trong bốn mô hình, prompt chấm rút gọn tuy sử dụng ít đơn vị xử lý ngôn ngữ (token) hơn nhưng vẫn cho kết quả gần tương đương với prompt chi tiết (Yoshida, 2025). Nghiên cứu cũng khẳng định prompt chấm có kèm rubric cho kết quả tốt hơn đáng kể so với prompt không có rubric. Một hướng nghiên cứu khác cho thấy việc thay đổi mẫu prompt chấm về trật tự thao tác chấm hoặc số lượng ví dụ mẫu có thể làm thay đổi đáng kể kết quả chấm và phản hồi. Vì thế, trong thực hành sư phạm, thay vì đòi hỏi prompt chấm quá dài và khó tái sử dụng, nên ưu tiên prompt chấm ngắn gọn nhưng bắt buộc AI thực hiện đúng quy trình: đọc đề, xác định tiêu chí, phân tích bằng chứng, cho điểm sơ bộ, tạo phản hồi và nêu mức độ chắc chắn. Rubric càng rõ ràng, prompt chấm càng nhất quán thì đánh giá của AI càng đáng tin cậy (Mughal et al., 2026).

2.2.3. Vai trò của AI trong đánh giá năng lực tư duy ngôn ngữ qua bài tự luận

Vai trò đầu tiên của AI là hỗ trợ phân tích cấu trúc và nội dung của bài tự luận. Các hệ thống AES hiện nay có thể nhận diện những phương diện mà người chấm thường quan tâm như: tổ chức đoạn văn, phát triển luận điểm, mức độ liên quan nội dung, từ vựng, phong cách, mạch lạc và liên kết. Nhiều hướng dẫn về chấm điểm tự động cũng nhấn mạnh rằng các tiêu chí chấm bài luận không được dừng ở ngữ pháp và từ vựng, mà phải bao gồm cả các đặc trưng diễn ngôn như tổ chức, mức độ liên quan nội dung và phát triển chủ đề. Điều này phù hợp với các mô tả về năng lực viết trong các khung chuẩn, trong đó “tư duy viết”, “mạch lạc và liên kết”, “phạm vi từ vựng”, “phạm vi sử dụng ngôn ngữ” được xem là các thành phần cốt lõi của năng lực viết. Vì vậy, AI có thể đóng vai trò như một bộ phân tích sơ cấp giúp GV nhìn cấu trúc và nội dung của bài viết một cách có hệ thống hơn.

Vai trò thứ hai là hỗ trợ phân tích tư duy ngôn ngữ và kĩ năng viết - một cấp độ sâu hơn so với việc phát hiện các lỗi ngôn ngữ bề mặt. Một nghiên cứu gần đây về đánh giá viết bậc cao bằng AI cho thấy mô hình có thể xác định và phản hồi khá chính xác chất lượng của bài luận theo bốn phương diện: đáp ứng yêu cầu đề bài, mạch lạc và liên kết, phạm vi từ vựng và độ chính xác ngữ pháp; tuy nhiên mức độ cụ thể của phản hồi không đồng đều (Saricaoglu et al., 2025). AI thường hoạt động hiệu quả hơn trong việc phát hiện lỗi ngữ pháp và từ vựng, nhưng với tiêu chí đáp ứng yêu cầu đề bài và mạch lạc chỉ AI chỉ đánh giá ở mức khái quát.

Trên cơ sở xác định ưu nhược điểm đó của AI, nhóm nghiên cứu đề xuất hướng tiếp cận đánh giá dựa trên đặc trưng, trong đó mỗi tiêu chí trong rubric được chia thành các tiêu tiêu chí cụ thể kèm theo hệ thống câu hỏi nhằm xác định mức độ đáp ứng của bài viết. Cách tiếp cận này giúp chuyển những biểu hiện vốn khó định lượng của kỹ năng viết và tư duy ngôn ngữ thành các dấu hiệu có thể quan sát và diễn giải rõ ràng hơn. Các nghiên cứu về độ giá trị của hệ thống chấm bài tự động cũng cho thấy phản hồi của AI chỉ thực sự có ý nghĩa khi bám sát cấu trúc năng lực cần đánh giá, được người học hiểu đúng và hỗ trợ hiệu quả cho quá trình chỉnh sửa bài viết. Điểm đáng chú ý là độ giá trị của phản hồi còn được nâng cao hơn khi có sự trao đổi giữa GV và người học về cách hiểu đúng các phản hồi này. Nói cách khác, AI có thể hỗ trợ nhận diện cách người học tư duy và biểu đạt tư duy thông qua hệ thống lập luận, cách tổ chức ý, sử dụng dẫn chứng và lựa chọn phương tiện ngôn ngữ; tuy nhiên, để

những phân tích này thực sự góp phần nâng cao kết quả học tập vẫn cần có sự định hướng và điều chỉnh từ phía GV.

Vai trò thứ ba là tự động hoá chấm điểm và phản hồi. Ở góc độ sư phạm, đây là ưu thế mạnh nhất của AI. Một thử nghiệm ngẫu nhiên có đối chứng với 190 người học ngoại ngữ cho thấy AES có tác động tích cực giúp người học cải thiện kỹ năng viết luận (Wei et al., 2023). Nghiên cứu so sánh việc nhận phản hồi từ AI với nhận phản hồi từ GV trong khoá học về kỹ năng viết tranh luận, kết quả cho thấy cả hai nhóm người học đều tiến bộ đáng kể. Khác biệt giữa hai nguồn phản hồi không có ý nghĩa thống kê khi cùng được dẫn hướng bởi rubric phân tích và prompt chấm có cấu trúc. Tuy nhiên, các nghiên cứu này cũng chỉ ra AI có thể gần với người chấm ở điểm tổng quát nhưng vẫn yếu hơn trong việc đưa ra phản hồi chi tiết và những gợi ý cụ thể giúp người học cải thiện bài viết. Vì vậy, trong đánh giá năng lực tư duy ngôn ngữ, AI đặc biệt phù hợp với đánh giá quá trình, là hình thức đánh giá có mục tiêu giúp người học cải thiện, yêu cầu phản hồi nhanh, cá thể hoá và lặp lại thường xuyên. Đối với đánh giá định kỳ (kiểm tra, thi ...), mô hình kết hợp giữa GV và AI vẫn là lựa chọn an toàn hơn vì lí do độ tin cậy, độ bao quát cấu trúc năng lực và tính công bằng của hoạt động đánh giá.

Tóm lại, các kết quả nghiên cứu gần đây cho thấy AI đã trở thành một công cụ có tính khả dụng cao trong đánh giá bài tự luận, nhưng hiệu quả của nó không nằm ở khả năng “tự chấm điểm” độc lập mà ở năng lực hỗ trợ đánh giá có cấu trúc. Khi được gắn với rubric rõ ràng, prompt chấm chặt chẽ, bài mẫu chuẩn và cơ chế giám sát phù hợp, AI có thể giúp giáo viên và cơ sở giáo dục phân tích bài viết nhanh hơn, nhất quán hơn và tạo phản hồi kịp thời hơn. Đối với đánh giá năng lực tư duy ngôn ngữ qua bài tự luận, vai trò phù hợp nhất của AI là: hỗ trợ đọc cấu trúc và nội dung, đánh giá điểm mạnh - điểm yếu theo tiêu chí, tạo phản hồi giúp người học cải thiện kỹ năng. Do đó, khi ứng dụng AI trong tự động chấm điểm bài tự luận, câu hỏi quan trọng nhất không phải là “AI chấm bài tự luận tốt đến đâu”, mà là “AI được thiết kế theo khung đo lường nào, bằng rubric nào, bằng prompt nào và dưới sự giám sát của con người ra sao”.

3. Ứng dụng AI trong đánh giá năng lực tư duy ngôn ngữ qua các bài tập tự luận

3.1. Phương pháp nghiên cứu

Nghiên cứu được triển khai theo hướng thực nghiệm ứng dụng, nhằm khảo sát mức độ tương hợp giữa kết quả chấm điểm của hệ thống AI và giảng viên đối với bài tập tự luận ở một số học phần được lựa chọn. Nghiên cứu chủ yếu sử dụng phương pháp phân tích định lượng để đo lường mức độ tương hợp về điểm số giữa hai nguồn đánh giá thông qua các chỉ số thống kê phù hợp. Bên cạnh đó, phương pháp phân tích định tính được vận dụng như một công cụ bổ trợ nhằm xem xét nội dung phản hồi của AI và giảng viên, qua đó nhận diện một số điểm tương đồng và khác biệt trong cách tiếp cận chất lượng bài viết.

Trên cơ sở đối sánh hai nguồn đánh giá, nghiên cứu tập trung trước hết vào mức độ tương hợp giữa điểm tổng do AI và giảng viên chấm theo cùng một khung đánh giá năng lực tư duy ngôn ngữ được cụ thể hoá bằng các rubric chấm điểm. Đồng thời, việc phân tích nội dung phản hồi được thực hiện thông qua việc nhận diện các chủ đề nổi bật trong nhận xét và xem xét một số trường hợp điển hình có sự khác biệt về điểm số hoặc cách đánh giá. Kết quả phân tích định tính chủ yếu mang ý nghĩa minh họa và hỗ trợ diễn giải cho các kết quả định lượng, góp phần làm rõ tính phù hợp cũng như những giới hạn của AI trong hỗ trợ chấm bài tự luận ở bậc đại học.

3.2. Mô hình AI và ngữ liệu nghiên cứu

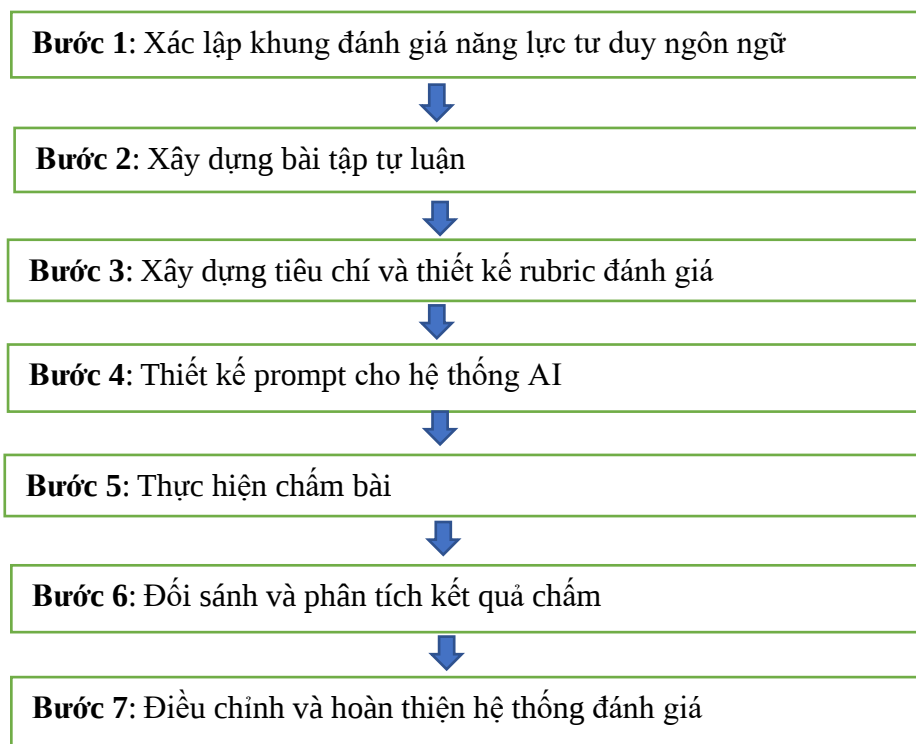
Mô hình AI sử dụng để chấm điểm là Gemini 3 Flash, là mô hình ngôn ngữ lớn phiên bản thương mại do Google phát hành. Phần mềm chấm điểm sử dụng trong nghiên cứu được phát triển bằng ngôn ngữ lập trình Python và triển khai tại Trường Đại học Hà Nội. Phần mềm có các tính năng cơ bản của hệ thống chấm điểm tự động như: tạo tài khoản và quản lí người dùng; quản lí bài tập; quản lí prompt chấm bài; quản lí phiên chấm bài ... Phần mềm giao tiếp với Gemini 3 Flash thông qua các thư viện Python API tiêu chuẩn do Google cung cấp. Hệ thống được thiết kế dễ dàng mở rộng để làm việc với các mô hình ngôn ngữ lớn nguồn mở đã được huấn luyện/ tinh chỉnh chuyên dụng cho tác vụ chấm điểm và được triển khai nội bộ tại Trường Đại học Hà Nội.

Mẫu nghiên cứu được lựa chọn từ nhóm sinh viên chính quy khoá 2024 và 2025 đang theo học học phần Nhập môn Việt ngữ học, Văn bản tiếng Việt và Cơ sở văn hoá Việt Nam, học kì 1 năm học 2025-2026 tại Trường Đại học Hà Nội. Việc chọn mẫu dựa trên một số tiêu chí cụ thể: (i) sinh viên có bài tập tự luận hợp lệ trong học phần; (ii) dữ liệu thu thập bảo đảm đầy đủ và có thể xử lý; (iii) mẫu phản ánh tính đại diện cho các lớp thuộc cùng học phần; (iv) phù hợp với phạm vi cũng như điều kiện triển khai của nghiên cứu.

Trên cơ sở đó, quy mô thực nghiệm được xác định gồm 106 bài viết tự luận của sinh viên chính quy khoá 2024 và 2025 thuộc các khoa ngoại ngữ, Trường Đại học Hà Nội. Các bài tập tự luận được thu thập qua Google Form, chuyển thành tệp Excel và nhập vào hệ thống để chấm điểm tự động⁴. Hệ thống thực hiện đánh giá sơ bộ từng bài viết theo hướng dẫn chấm đã tích hợp, sau đó xuất phản hồi chi tiết cho người học. Để kiểm tra độ tin cậy và độ phù hợp trong kết quả chấm của AI, nhóm tiến hành rà soát xác suất kết quả, đối chiếu giữa kết quả chấm của AI với GV 106/106 bài (100%). Việc đối chiếu tập trung vào hai tiêu chí: mức độ tương đồng điểm số (định lượng) và tương đồng về nhận xét (định tính).

3.3. Quy trình thực hiện

Nhằm triển khai hiệu quả giải pháp “Ứng dụng trí tuệ nhân tạo trong đánh giá năng lực tư duy ngôn ngữ qua các bài tập tự luận của sinh viên”, nhóm nghiên cứu đã thiết kế và tổ chức thực hiện một quy trình gồm bảy bước theo hướng hệ thống, bảo đảm tính khoa học và khả năng triển khai trong thực tiễn, cụ thể như sau:



Sơ đồ 1. Quy trình thực hiện nghiên cứu

Bước 1: Xác lập khung đánh giá năng lực tư duy ngôn ngữ

Trên cơ sở lí thuyết về khung năng lực tư duy ngôn ngữ đã phân tích ở trên, nhóm nghiên cứu đề xuất khung đo năng lực tư duy ngôn ngữ như sau:

⁴ Ngữ liệu nghiên cứu: https://docs.google.com/spreadsheets/d/1aAWsdFMLHyfEx0QVhD-bQMsgPrMuRLWs/edit?usp=drive_link&ouid=100307688903414450192&rtopof=true&sd=true

Bảng 1. Khung đánh giá năng lực tư duy ngôn ngữ

Trục năng lực tư duy	Chỉ số hành vi cụ thể
1. Huy động tri thức	- Xác định đúng vấn đề và yêu cầu của đề bài - Tổng hợp được các khái niệm, thuật ngữ và tri thức liên quan đến chủ đề - Giải thích được nội dung, hiện tượng hoặc luận điểm bằng cơ sở lí thuyết phù hợp - Thể hiện mức độ hiểu và vận dụng tri thức vào ngữ cảnh cụ thể của bài viết
2. Năng lực lập luận	- Xây dựng được hệ thống luận điểm rõ ràng, phù hợp với vấn đề đề bài nêu ra - Triển khai lập luận theo trình tự logic - Biết phân tích, so sánh, chứng minh hoặc phản biện để làm rõ quan điểm - Sử dụng được dẫn chứng phù hợp nhằm tăng tính thuyết phục cho lập luận
3. Năng lực diễn đạt học thuật	- Tổ chức bài viết theo bố cục rõ ràng, hợp lí - Sử dụng ngôn ngữ chính xác, phù hợp với phong cách học thuật - Diễn đạt mạch lạc, có sự liên kết giữa các câu và đoạn văn - Sử dụng thuật ngữ, phương tiện liên kết và thao tác diễn đạt phù hợp với mục đích giao tiếp học thuật
4. Năng lực tư duy sáng tạo	- Thể hiện cách tiếp cận hoặc góc nhìn riêng đối với vấn đề - Biết kết nối, mở rộng hoặc liên hệ vấn đề với các vấn đề khác - Đưa ra được nhận xét, đánh giá hoặc đề xuất cá nhân có giá trị học thuật

Khung năng lực này là cơ sở để xây dựng bài tập tự luận, thiết kế tiêu chí đánh giá và triển khai hệ thống rubric trong các bước tiếp theo của nghiên cứu.

Bước 2: Xây dựng bài tập tự luận

Căn cứ khung năng lực tư duy ngôn ngữ đã xác lập, bám sát mục tiêu học phần và định hướng đánh giá năng lực, nhóm nghiên cứu thiết kế các bài tập tự luận phù hợp. Nội dung bài tập được xây dựng bảo đảm đáp ứng chuẩn đầu ra của các học phần Nhập môn Việt ngữ học, Văn bản tiếng Việt và Cơ sở văn hoá Việt Nam, đồng thời hướng tới các thao tác tư duy như giải thích, phân tích, chứng minh và vận dụng. Đề bài cũng tạo điều kiện để người học sử dụng dẫn chứng hoặc ví dụ minh họa và diễn đạt bằng ngôn ngữ học thuật. Ba bài tập tự luận của ba học phần được thiết kế với mức độ chuẩn hoá học thuật cao và đa dạng về loại hình nhiệm vụ. Ở học phần Nhập môn Việt ngữ học, bài tập thiên về trình bày kiến thức có cấu trúc rõ ràng, với các nội dung được phân tách cụ thể và gắn trực tiếp với thang điểm. Bài tập Cơ sở văn hoá Việt Nam lại thuộc dạng nghị luận dài, yêu cầu kết hợp tri thức, lập luận, dẫn chứng và tổ chức văn bản hoàn chỉnh. Trong khi đó, bài tập Văn bản tiếng Việt là dạng trung gian, vừa yêu cầu lập luận ngắn gọn vừa đảm bảo có dẫn chứng minh họa. Sự kết hợp này tạo nên một hệ bài tập có khả năng bao quát nhiều dạng năng lực viết khác nhau, đồng thời cung cấp các đơn vị đánh giá tương đối rõ ràng, thuận lợi cho việc triển khai và kiểm chứng hiệu quả của hệ thống chấm điểm tự động. Các yêu cầu này đồng thời phản ánh các tiêu chí đánh giá năng lực tư duy ngôn ngữ, bao gồm: mức độ nắm vững tri thức, khả năng lập luận và chứng minh, khả năng sử dụng dẫn chứng và khả năng diễn đạt. Nhờ đó, tiêu chí đánh giá được định hình ngay từ khâu xây dựng đề bài, làm cơ sở cho các bước thiết kế rubric và triển khai đánh giá sau đó.

Dưới đây là một câu hỏi tự luận thuộc học phần Văn bản tiếng Việt:

Bảng lập luận và dẫn chứng phù hợp, hãy chứng minh nhận định: “Từ ngữ trong phong cách khoa học là lớp từ ngữ đa phong cách, trung hòa về màu sắc cảm xúc, chủ yếu được dùng trong ý nghĩa khái quát” (Đình Trọng Lạc, Nguyễn Thái Hòa, *Phong cách học tiếng*

Bước 3. Xây dựng tiêu chí và thiết kế rubric đánh giá

Căn cứ yêu cầu của câu hỏi tự luận và các thành tố năng lực cần đánh giá, nhóm nghiên cứu xây dựng rubric theo hướng chuẩn hoá tiêu chí đánh giá từ nội dung bài làm. Trước hết, từ câu hỏi và đáp án gợi ý, xác định các đơn vị tri thức cốt lõi (luận điểm/nội dung bắt buộc). Mỗi đơn vị được chuyên hoá thành tiêu chí với các mức độ đạt được khác nhau, nhằm đánh giá mức độ nắm vững kiến thức. Tiếp theo, thiết kế thang điểm theo mức độ hoàn thành (đúng/sai, đủ/thiếu) kèm mô tả cụ thể, đảm bảo khả năng phân hoá và tính nhất quán khi chấm. Đồng thời, bổ sung các tiêu chí về lập luận, diễn đạt học thuật và tư duy sáng tạo nhằm đánh giá khả năng vận dụng, tổ chức và kiến tạo tri thức của người

học. Cụ thể, tiêu chí lập luận và diễn đạt tập trung vào cấu trúc bài viết, tổ chức luận điểm, tính mạch lạc, cách sử dụng dẫn chứng và phương tiện ngôn ngữ học thuật; trong khi tiêu chí tư duy sáng tạo được thể hiện thông qua khả năng tiếp cận vấn đề từ góc nhìn riêng, cách triển khai lập luận linh hoạt, khả năng khái quát hoá, liên hệ hoặc mở rộng vấn đề trên cơ sở tri thức chuyên môn. Các yếu tố hình thức và điểm thưởng cũng được tích hợp vào rubric nhằm khuyến khích những biểu hiện sáng tạo phù hợp với yêu cầu học thuật. Từ đó, rubric được chuẩn hoá thành bốn nhóm tiêu chí: huy động tri thức, lập luận, diễn đạt học thuật và tư duy sáng tạo, làm cơ sở cho việc chấm điểm bằng AI. Dưới đây là một ví dụ về một tiêu chí đã cụ thể hoá bằng các chỉ số đo lường được trong rubric.

Bảng 2. Ví dụ về rubric đánh giá năng lực tư duy ngôn ngữ

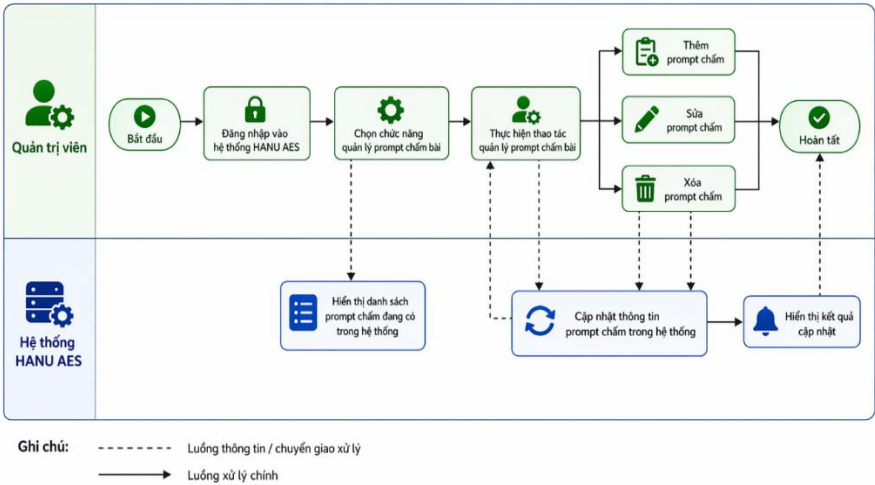
Tiêu chí	Mức độ đạt yêu cầu	Điểm
Điểm thưởng cho sự sáng tạo (Bonus)	- Cộng 0,5 điểm sáng tạo đối với bài làm có ít nhất 1 trong 2 biểu hiện sau: - Cách triển khai lập luận linh hoạt, không mang tính liệt kê, có câu phân tích để làm rõ luận điểm. - Có câu kết mang tính tổng hợp, không lặp nguyên văn đề, thể hiện tư duy khái quát. - Không cộng điểm sáng tạo nếu: - Tổng điểm đã đạt tối đa (10 điểm) - Nội dung lí thuyết đạt dưới 4,0/6; - Dẫn chứng sai/không liên quan.	0.5

Bước 4. Thiết kế prompt cho hệ thống AI

Prompt chấm bài là câu lệnh mà phần mềm dùng để tương tác với mô hình ngôn ngữ lớn trong quá trình chấm bài. Mỗi một bài tập sẽ có một prompt chấm bài tương ứng. Prompt chấm bài được soạn thảo bằng ngôn ngữ thông thường nhưng tuân thủ một số quy tắc nhất định để ánh xạ các yêu cầu trong rubrics chấm điểm sang câu lệnh giao tiếp với mô hình ngôn ngữ lớn. Với mỗi bài tập, căn cứ vào đề bài, đáp án, rubrics chấm, quản trị viên sẽ tạo ra prompt chấm bài phù hợp cho bài tập đấy. Phần mềm cho phép quản lí prompt chấm bài độc lập với quản lí bài tập để hỗ trợ tái sử dụng và tinh chỉnh prompt chấm bài. Quy trình quản lí prompt chấm bài như sau:

Quy trình quản lý prompt chấm bài – HANU AES

Mô tả luồng quản lý prompt chấm bài trên hệ thống HANU AES



Sơ đồ 2. Quy trình quản lý prompt chấm bài

Bước 5. Thực hiện chấm bài

Các bài viết tự luận của sinh viên được chấm song song bởi hệ thống chấm điểm bằng AI và GV nhằm phục vụ đối sánh kết quả. Nhóm nghiên cứu thu thập được 106 bài luận thuộc ba học phần, gồm: 34 bài Nhập môn Việt ngữ học, 37 bài Cơ sở văn hoá Việt Nam và 35 bài Văn bản tiếng Việt.

Trên cơ sở đó, hệ thống AI thực hiện chấm toàn bộ 106 bài luận, tạo ra kết quả cho từng bài gồm: điểm theo từng tiêu chí kèm nhận xét chi tiết, điểm tổng và nhận xét chung. Song song với đó, ba GV có chuyên môn sâu ở các học phần tương ứng được mời chấm thủ công cùng tập bài này. Cả hai hình thức chấm đều sử dụng chung bộ câu hỏi, đáp án, rubric và bài viết mẫu, nhằm đảm bảo tính thống nhất về chuẩn đánh giá. Quá trình chấm điểm của hệ thống AI và của GV được triển khai độc lập, không có sự trao đổi thông tin kết quả trong quá trình chấm, đảm bảo tính khách quan cho bước đối sánh tiếp theo.

Bước 6. Đối sánh và phân tích kết quả chấm

Ở bước này, nghiên cứu tiến hành đối sánh kết quả chấm của hệ thống AI và GV trên 106 bài luận, xét trên cả điểm số và nội dung nhận xét. Với dữ liệu nhận xét (69/106 bài), quá trình phân tích tập trung vào ba phương diện: (i) mức độ chi tiết và cụ thể trong phản hồi của AI; (ii) sự tương đồng về nội dung đánh giá chính giữa AI và GV; (iii) mối liên hệ giữa độ giống nhau về câu chữ và mức độ tương đồng trong đánh giá. Kết quả phân tích được sử dụng để nhận diện những điểm phù hợp và khác biệt trong cách chấm giữa hai nguồn, làm cơ sở cho bước điều chỉnh và hoàn thiện hệ thống ở giai đoạn tiếp theo.

Bước 7. Điều chỉnh và hoàn thiện hệ thống đánh giá

Trên cơ sở kết quả đối sánh và phân tích dữ liệu thực nghiệm, nhóm nghiên cứu tiến hành rà soát và điều chỉnh hệ thống đánh giá nhằm nâng cao mức độ phù hợp giữa kết quả chấm của AI với yêu cầu đánh giá năng lực tư duy ngôn ngữ. Các nội dung được điều chỉnh bao gồm tiêu chí trong rubric, cách diễn đạt prompt và cơ chế phản hồi của hệ thống. Quá trình này góp phần hoàn thiện hệ thống đánh giá theo hướng gia tăng tính nhất quán, khả năng nhận diện các biểu hiện năng lực và nâng cao hiệu quả hỗ trợ giảng viên trong hoạt động đánh giá bài tập tự luận.

3.4. Kết quả nghiên cứu

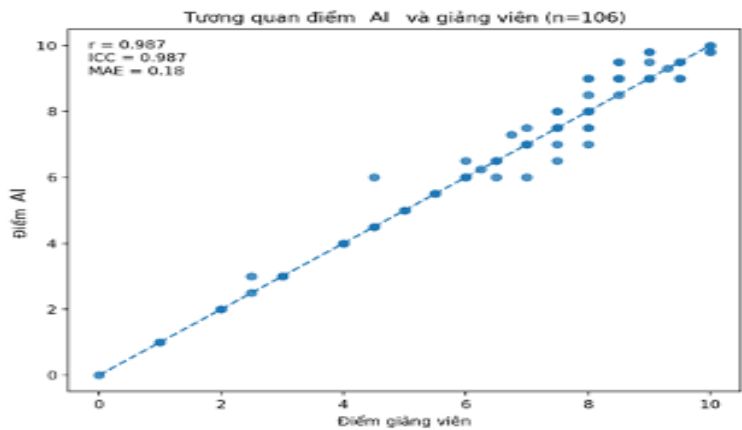
Phần này trình bày kết quả phân tích, đánh giá mức độ tương quan giữa điểm số và nội dung nhận xét do AI chấm và kết quả do GV chấm, từ đó xem xét độ tin cậy và độ giá trị của điểm do AI chấm. Các thang đo (metrics) được dùng trong phân tích định lượng bao gồm: hệ số tương quan Pearson (r), hệ số tương quan thứ tự xếp hạng Spearman (ρ), hệ số tương quan nội nhóm ICC, hệ số tương quan đồng thuận Lin CCC, các chỉ số thống kê như: điểm trung bình, sai số tuyệt đối trung bình (MAE), chênh lệch trung bình (Bias). Sau khi phân tích, tính toán, giá trị của các metrics này được trình bày trong bảng dưới đây:

Bảng 3. Kết quả phân tích tương quan giữa điểm chấm của phần mềm và của giảng viên

Metrics	Giá trị	Metrics	Giá trị
Số bài luận	106	Hệ số CCC	0,987
Điểm trung bình do AI chấm	6,647	Hệ số ICC	0,987
Điểm trung bình do GV chấm	6,583	MAE	0,184
Bias: Chênh lệch trung bình	0,064	Số bài có điểm trùng nhau	71,7%
Hệ số Pearson (r)	0,987	Số bài có điểm lệch dưới 0,5đ	22,6%
Spearman (ρ)	0,981	Số bài có điểm lệch từ 0,5đ đến 1,5đ	5,7%

Kết quả cho thấy mức độ tương hợp điểm chấm giữa AI và GV là rất cao. Trên toàn bộ 106 bài luận, hệ số Pearson đạt 0,987, Spearman đạt 0,981, ICC đạt 0,987 và CCC đạt 0,987. Các chỉ số này cho thấy AI không chỉ xếp hạng các bài luận tương đối giống GV, mà còn chấm rất sát về mặt trị số điểm. Về sai số tuyệt đối, MAE của bộ dữ liệu 106 bài luận chỉ là 0.18 điểm và độ lệch trung bình là 0,064 điểm. Có 76/106 bài luận (71.7%) được AI chấm đúng bằng điểm chấm của GV; 24/106 bài luận (22,6%) lệch không quá 0,5 điểm; và 6/106 bài luận (5,7%) có điểm chênh lệch từ 0,5 đến 1,5 điểm.

Hình dưới đây biểu diễn mức độ tương quan giữa điểm chấm của AI và của GV.



Hình 1. Tương quan điểm chấm giữa phần mềm và giảng viên

Để kiểm tra xem AI chấm điểm có hoạt động ổn định ở các mức chất lượng bài viết khác nhau hay không, các bài luận được chia theo dải điểm do GV chấm: yếu (dưới 5đ), trung bình - khá (từ 5đ đến dưới 7đ), khá - giỏi (từ 7đ đến dưới 8,5đ) và xuất sắc (từ 8,5 trở lên). Các chỉ số so sánh được trình bày trong Bảng 4:

Bảng 4. So sánh các chỉ số theo dải điểm

Dải điểm	Số bài luận	Điểm TB Giảng viên	Điểm TB AI	Bias	MAE	Trùng điểm (%)	Lệch dưới 0,5đ (%)
Yếu	24	3,08	3,17	0,08	0,08	91,7	4,1
Trung bình - khá	23	5,96	5,98	0,02	0,07	87,0	13,0
Khá - giỏi	33	7,56	7,55	-0,02	0,26	63,6	21,2
Xuất sắc	26	9,13	9,31	0,18	0,29	50,0	34,6

Bảng 4 cho thấy AI chấm bám rất sát GV ở tất cả các dải điểm, nhưng độ khít không hoàn toàn đồng đều. Ở dải yếu, trung bình và khá, mức trùng điểm tuyệt đối rất cao (xấp xỉ 87 - 92%) và MAE nhỏ. Ở nhóm bài giỏi và xuất sắc, kết quả chấm của AI vẫn gần với giảng viên nhưng độ lệch có xu hướng tăng nhẹ, cho thấy việc phân hoá giữa các bài viết tốt và rất tốt là một thách thức lớn hơn đối với hệ thống chấm tự động. Đáng chú ý, ở dải xuất sắc, điểm trung bình của AI là 9,31đ cao hơn điểm trung bình của GV (9,13đ) khoảng 0,18 điểm. Chênh lệch này chưa đủ lớn để tạo thành thiên lệch mang tính hệ thống trên toàn bộ dữ liệu, nhưng phản ánh xu hướng AI đánh giá nới lỏng hơn so với GV ở nhóm bài điểm cao.

Việc phân tích, so sánh các nhận xét do AI viết và do GV viết được thực hiện trên 69 bài luận, tập trung vào ba câu hỏi: (i) AI viết nhận xét dài và cụ thể đến mức nào; (ii) nhận xét của AI và GV có giống nhau về những nội dung chính hay không; (iii) mức độ giống nhau về câu chữ trong nhận xét có phản ánh đúng mức độ tương đồng về đánh giá hay không.

Thống kê cho thấy các nhận xét của AI có độ dài trung bình 80,2 từ (trung vị 77), trong khi nhận xét chung của GV có độ dài trung bình 53,6 từ (trung vị 58). Như vậy, AI có xu hướng viết dài hơn và giải thích cụ thể hơn.

Khi đo độ giống nhau về câu chữ, mức tương đồng lại chỉ ở mức thấp: chỉ số cosine similarity trung bình giữa từng cặp nhận xét là 0,258, trung vị là 0,183; tỷ lệ giao nhau từ vựng trung bình là 0,286; và chỉ có 1 cặp nhận xét trùng khớp gần như hoàn toàn. Nếu chỉ xét thước đo thống kê từ vựng thì nhận xét của AI và của GV lệch nhau khá xa. Nguyên nhân chủ yếu nằm ở sự khác biệt về phong cách diễn đạt: AI thường cung cấp phản hồi dài, chi tiết, đôi khi bám sát hoặc viện dẫn trực tiếp rubric; trong khi GV có xu hướng nhận xét ngắn gọn, mang tính chẩn đoán cao và tập trung vào những lỗi cốt lõi. Vì vậy, hai nhận xét có thể dùng từ rất khác nhau nhưng vẫn đồng thuận khi đánh giá bản chất của bài làm (vẫn chấm điểm số tương tự nhau).

Nhóm nghiên cứu sử dụng thuật toán TF-IDF để xác định các chủ đề chính của các nhận xét của AI và GV. Bảng 5 dưới đây trình bày kết quả phân tích và so sánh chủ đề nhận xét:

Bảng 5. So sánh từ khoá và chủ đề của nhận xét của phần mềm và giảng viên

Chủ đề	Phần mềm (%)	Giảng viên (%)	Cùng xuất hiện (%)
Thiếu dẫn chứng / ví dụ / hiện tượng ngôn ngữ	46 (66,7%)	40 (58,0%)	32 (46,4%)
Sơ sài / thiếu chiều sâu	15 (21,7%)	15 (21,7%)	9 (13,0%)
Lạc đề / chưa đúng trọng tâm	2 (2,9%)	3 (4,3%)	2 (2,9%)
Kiến thức / thuật ngữ chưa chính xác	20 (29,0%)	29 (42,0%)	13 (18,8%)
Bố cục / trình bày / mạch lạc, tốt	27 (39,1%)	33 (47,8%)	20 (29,0%)
Phân tích tốt / đáp ứng yêu cầu	10 (14,5%)	34 (49,3%)	8 (11,6%)

Các chủ đề và từ khoá nổi bật nhất ở cả hai nhóm nhận xét là tương đối nhất quán, các chủ đề/từ khoá chính gồm: “thiếu dẫn chứng/hiện tượng ngôn ngữ”, “bài còn sơ sài”, “kiến thức hoặc thuật ngữ chưa chính xác”, bình luận về như mức độ mạch lạc và khả năng đáp ứng yêu cầu đề bài. Điều này cho thấy AI không chỉ cho điểm gần với GV mà còn thường nhận diện đúng loại vấn đề mà GV cũng chú ý.

Tuy nhiên, cấu trúc chủ đề cho thấy một khác biệt đáng chú ý. GV dùng nhóm nhận xét tích cực “phân tích tốt/ đáp ứng yêu cầu” ở gần một nửa số bài, trong khi AI chỉ dùng nhóm này ở khoảng một phần bảy số bài. Ngược lại, AI có xu hướng nhấn mạnh lỗi và điểm còn thiếu sót của bài viết nhiều hơn, ngay cả với các bài được cho điểm cao. Đây là dấu hiệu cho thấy AI phù hợp hơn với vai trò công cụ phản hồi theo hướng giúp sửa bài, còn GV giữ được sự cân bằng tốt hơn giữa khen và góp ý.

Bên cạnh phân tích định lượng, nhóm nghiên cứu cũng tiến hành phân tích định tính thông qua việc xác định nguyên nhân dẫn đến sự chênh lệch điểm lớn giữa kết quả chấm của hệ thống và GV ở một số bài làm của sinh viên. Cụ thể có 05 bài chênh lệch trong khoảng từ 1,0 đến 1,5 điểm như mô tả trong Bảng 6:

Bảng 6. Danh sách bài luận có mức chênh lệch điểm từ 1,0 đến 1,5 điểm

Mã SV	Điểm Hệ thống	Điểm Giảng viên	Chênh lệch
2427101724	6,0	4,5	1,5
2327060008	6,5	7,5	-1,0
2427010774	7,0	8,0	-1,0
2427010960	6,0	7,0	-1,0
2537041122	9,0	8,0	1,0

Trong 05 bài viết có sai lệch điểm được lựa chọn phân tích, có 01 trường hợp thuộc học phần Nhập môn Việt ngữ học, 03 trường hợp thuộc học phần Văn bản tiếng Việt và 01 trường hợp thuộc học phần Cơ sở văn hoá Việt Nam, như vậy sai lệch tập trung chủ yếu ở học phần Văn bản tiếng Việt. Nhìn chung, sự khác biệt giữa hệ thống AI và GV xuất phát từ những cách tiếp cận khác nhau trong vận dụng rubric. AI có xu hướng dựa vào các dấu hiệu bề mặt và ấn tượng tổng thể để đưa ra đánh giá, xử lý tiêu chí theo ngưỡng hoặc theo từng biểu hiện riêng lẻ, đồng thời còn hạn chế ở các tiêu chí định lượng và hình thức như độ dài, bố cục. Ngược lại, GV đánh giá theo hướng phân tích, linh hoạt, chú trọng độ chính xác khái niệm, mức độ đáp ứng đầy đủ yêu cầu và chất lượng hệ thống dẫn chứng.

Một số trường hợp tiêu biểu dưới đây là minh chứng cho sự khác biệt nói trên giữa cách chấm của AI và GV: Ở mã 2427101724, AI dựa vào việc bài viết có đề cập các khái niệm như “bậc phân đoạn”, “âm đầu - vần - thanh điệu” để suy ra mức độ hiểu biết và cho điểm cao, trong khi GV phát hiện lỗi sai bản chất (hiểu sai bậc 1, thiếu bậc 2) nên trừ điểm đáng kể. Ngược lại, với mã 2327060008, AI lại cho điểm thấp do bám vào các tiêu chí hình thức như thiếu tên tác giả và đánh giá sai độ dài văn bản, trong khi GV đánh giá tổng thể nội dung và kiểm tra trực tiếp nên ghi nhận bài đạt yêu cầu. Những trường hợp này cho thấy AI có thể vừa đánh giá “nói lỏng” ở nội dung, vừa “chặt” ở hình thức, trong khi GV ưu tiên đánh giá bản chất học thuật và mức độ hoàn thiện thực tế của bài viết.

Từ góc độ đánh giá chất lượng phản hồi, nhận xét của AI có ba ưu điểm chính. Thứ nhất, nhận xét thường đầy đủ, giải thích cụ thể vì sao bài được mức điểm hiện tại. Thứ hai, AI nêu khá rõ những phần còn thiếu/ hạn chế của bài làm so với yêu cầu trong rubrics chấm điểm. Thứ ba, với lớp học đông sinh viên, AI có thể hỗ trợ tạo nhận xét/ phản hồi nhanh và khá nhất quán. Dù vậy, nhận xét của AI cũng có ba hạn chế. Thứ nhất, văn phong nhận xét dễ mang tính khuôn mẫu và thường thể hiện sự mức độ khẳng định cao ngay cả trong những trường hợp suy luận chưa thật chính xác. Thứ hai, AI có xu

hướng viết dài, đôi khi nêu quá nhiều tiểu tiết so với những điều được GV coi là lỗi trọng tâm. Thứ ba, mức giống nhau về câu chữ cho thấy nếu dùng AI như công cụ thay thế hoàn toàn GV ở phần nhận xét, sinh viên có thể nhận được phản hồi chưa chuẩn phong cách sư phạm dù điểm số vẫn được chấm tương đối chính xác.

4. Kết luận

Kết quả nghiên cứu cho thấy hệ thống AI được xây dựng không chỉ có thể vận hành như một công cụ hỗ trợ đánh giá năng lực tư duy ngôn ngữ của sinh viên, mà còn cho thấy tính khả thi của việc ứng dụng trí tuệ nhân tạo trong việc đánh giá các thành tố của năng lực này thông qua bài tập tự luận. Trên cơ sở tích hợp rubric và thiết kế prompt, hệ thống đã bước đầu tái hiện được logic chấm điểm theo tiêu chí, đồng thời cho thấy mức độ tương thích nhất định khi đối sánh với kết quả của GV.

Về ý nghĩa thực tiễn, cách tiếp cận này góp phần nâng cao hiệu quả đánh giá trong dạy học, hỗ trợ GV trong việc chấm bài và cung cấp phản hồi, đồng thời hướng tới sự nhất quán và minh bạch trong quá trình đánh giá theo năng lực. Tuy nhiên, nghiên cứu vẫn còn những hạn chế như phạm vi ngữ liệu chưa lớn và sự phụ thuộc đáng kể vào chất lượng thiết kế prompt, điều này có thể ảnh hưởng đến độ ổn định của kết quả.

Trong thời gian tới, nhóm nghiên cứu kì vọng tiếp tục mở rộng ngữ liệu, đa dạng hoá dạng bài tập và hoàn thiện hệ thống đánh giá theo hướng tối ưu hoá prompt, nhằm nâng cao độ tin cậy và khả năng ứng dụng của mô hình trong thực tiễn giáo dục.

TÀI LIỆU THAM KHẢO

Tiếng Việt

1. Hoàng Hòa Bình (2015). *Năng lực và đánh giá theo năng lực*. Tạp chí Khoa học Đại học Sư phạm Thành phố Hồ Chí Minh. Số 6 (71).
2. Nguyễn Thu Hà (2014). *Giảng dạy theo năng lực và đánh giá theo năng lực trong giáo dục: Một số vấn đề lý luận cơ bản*. Tạp chí Khoa học Đại học Quốc gia Hà Nội: Nghiên cứu Giáo dục, Tập 30, Số 2.
3. Nguyễn Công Khanh (2015). *Đánh giá học sinh theo cách tiếp cận năng lực*. NXB Đại học Sư phạm.
4. Hoàng Phê (2020). *Từ điển tiếng Việt*. NXB Đà Nẵng.

Tiếng nước ngoài

5. DeSeCo (2002). *Education - Lifelong Learning and the Knowledge Economy*. Key Competencies for the Knowledge Society. In: *Proceedings of the DeSeCo Symposium, Stuttgart*.
6. Mughal, N., Imran, A. S., Daudpota, S. M., et al. (2026). *Exploring potential of large language models for automated essay scoring in education*. Discover Artificial Intelligence, 6, 166. <https://doi.org/10.1007/s44163-026-01002-y>
7. OECD (2002). *Definition and Selection of Competencies: Theoretical and Conceptual Foundation*.
8. Pack, A., Barrett, A., & Escalante, J. (2024). *Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability*. Computers and Education: Artificial Intelligence, 6, 100234. <https://doi.org/10.1016/j.caeai.2024.100234>
9. Tremblay Denyse (2002). *The Competency-Based Approach: Helping learners become autonomous*. In *Adult Education - A Lifelong Journey*.
10. Rodriguez, P. U., Jafari, A. H., & Ormerod, C. M. (2019). *Language models and Automated Essay Scoring*. arXiv (Cornell University). <https://doi.org/10.48550/arXiv.1909.09482>
11. Stahl, M., Biermann, L., Nehring, A., & Wachsmuth, H. (2024). *Exploring LLM prompting strategies for joint essay scoring and feedback generation*. In Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024) (pp. 283-298). Association for Computational Linguistics.
12. Watheq Ahmad Mansour, Salam Albatarni, Sohaila Eltanbouly, & Tamer Elsayed. *Can Large Language Models Automatically Score Proficiency of Written Essays?*. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 2777-2786, Torino, Italia. ELRA and ICCL.
13. Wei P, Wang X and Dong H. (2023). *The impact of automated writing evaluation on second language writing skills of Chinese EFL learners: a randomized controlled trial*. Front. Psychol. 14:1249991. doi: 10.3389/fpsyg.2023.1249991
14. Yoshida, L. (2025). *Do We Need a Detailed Rubric for Automated Essay Scoring Using Large Language Models?*. In: Cristea, A.I., Walker, E., Lu, Y., Santos, O.C., Isotani, S. (eds) *Artificial Intelligence in Education*. AIED. Lecture Notes in Computer Science(), vol 15882. Springer, Cham. https://doi.org/10.1007/978-3-031-98465-5_8