ỨNG DỤNG AI TRONG HỌC TẬP VÀ THÁCH THỨC TỪ CÔNG CỤ PHÁT HIỆN: GÓC NHÌN TỪ GIẢNG DẠY TIẾNG TRUNG QUỐC VÀ ĐỀ XUẤT QUY TRÌNH ĐÁNH GIÁ MỚI

***Nguyễn Tá Nam*[1]**
Email: ntnam@daihocthudo.edu.vn

Ngày tòa soạn nhận được bài báo: 26/07/2025
Ngày phản biện đánh giá: 12/09/2025
Ngày bài báo được duyệt đăng: 15/10/2025

DOI: 10.59266/houjs.2025.981

Tóm tắt: *Sự phổ biến của các công cụ AI như Chat GPT đặt ra thách thức về tính nguyên bản trong học thuật, dẫn đến sự lạm dụng các phần mềm phát hiện văn bản AI. Nghiên cứu này phân tích những hạn chế cố hữu của các công cụ này và hệ quả tiêu cực của chúng đối với chất lượng nghiên cứu. Bằng phương pháp định tính kết hợp phân tích tài liệu và một trường hợp điển hình, nghiên cứu chỉ ra ba điểm yếu chính: (1) tỷ lệ dương tính giả cao với văn bản học thuật chất lượng, (2) kết quả thiếu ổn định, tạo ra "vòng xoáy chỉnh sửa" bất tận, và (3) sự thừa nhận về độ tin cậy thấp từ chính các nhà phát triển. Từ đó, nghiên cứu kết luận việc sử dụng chúng như một thước đo duy nhất để đánh giá đạo văn là phản tác dụng, vô tình khuyến khích hành vi "tự kiểm duyệt" và làm nghèo nàn nội dung học thuật. Là giải pháp thay thế, nghiên cứu đề xuất Mô hình Đánh giá Toàn diện Dựa trên Quá trình và Năng lực Thực (CPCAM), một khung đánh giá lấy con người làm trung tâm, nhằm bảo vệ tính toàn vẹn học thuật thực chất trong kỷ nguyên số.*

Từ khoá: *Phát hiện AI, dương tính giả, đánh giá học thuật, đạo văn AI, Turnitin*

I. Đặt vấn đề

Sự phát triển của các mô hình ngôn ngữ lớn (LLMs) như ChatGPT và Wenxin Yiyan đã mang lại những cơ hội đột phá trong giáo dục ngôn ngữ, đặc biệt là giảng dạy tiếng Trung như một ngoại ngữ. Các công cụ này trở thành trợ lý đắc lực, hỗ trợ toàn diện quá trình viết từ ý tưởng đến hoàn thiện văn bản (Kasneci và cộng sự, 2023).

Tuy nhiên, cùng với tiềm năng này xuất hiện những thách thức về đạo đức học thuật và phương pháp đánh giá. Nhiều cơ sở giáo dục đã sử dụng các công cụ phát hiện văn bản AI như một giải pháp kiểm soát, nhưng chính sự phụ thuộc này lại tạo

[1] Trường Đại học Thủ Đô Hà Nội

ra vấn đề mới. Các nghiên cứu gần đây (Liang và cộng sự, 2023; Weber-Wulff và cộng sự, 2023) chỉ ra tỷ lệ dương tính giả đáng kể, dẫn đến cáo buộc oan sai đặc biệt với những sinh viên có năng lực ngôn ngữ tốt. Thực trạng này làm xói mòn niềm tin và tạo ra môi trường học tập căng thẳng.

Xuất phát từ thực tế đó, nghiên cứu này tập trung giải quyết câu hỏi: "Làm thế nào thiết lập khung đánh giá bài tập viết tiếng Trung công bằng và hiệu quả trong kỷ nguyên AI, vượt ra khỏi sự phụ thuộc vào các công cụ phát hiện tự động không hoàn hảo?" Nghiên cứu sẽ: (i) phân tích hạn chế khoa học của công cụ phát hiện hiện tại, (ii) đề xuất mô hình đánh giá toàn diện dựa trên quá trình và năng lực thực, và (iii) thảo luận hàm ý sư phạm cho dạy viết trong bối cảnh chuyển đổi số.

II. Cơ sở lý luận

Cơ sở lý luận của nghiên cứu được xây dựng trên ba trụ cột chính, làm rõ tính không phù hợp của công cụ phát hiện AI và sự cần thiết của một phương pháp đánh giá thay thế trong bối cảnh dạy viết tiếng Trung.

2.1. Công cụ phát hiện văn bản AI: Cơ chế hoạt động và những hạn chế nền tảng

Các công cụ phát hiện AI hoạt động dựa trên các chỉ số thống kê như "độ bất định" (perplexity) và "sự biến đổi" (burstiness) để phân biệt văn bản của người và AI (Mitchell và cộng sự, 2023; Gao và cộng sự, 2022). Tuy nhiên, chúng tồn tại những hạn chế nghiêm trọng: (1) độ tin cậy thấp với tiếng Trung do sự khác biệt về ngữ pháp, chữ viết và sự đa dạng phương ngữ (Zhang & Chen, 2013); (2) tỷ lệ dương tính giả cao, đe dọa sự công bằng khi dễ nhầm lẫn các bài luận học thuật chặt chẽ của người là do AI tạo ra (Weber-Wulff và cộng sự, 2023); và (3) dễ dàng bị đánh lừa thông qua các chỉnh sửa đơn giản nhằm tăng tính tự nhiên cho văn bản (Krishna và cộng sự, 2023).

2.2. Lý thuyết về đánh giá giáo dục: Từ đánh giá sản phẩm sang đánh giá quá trình

Dựa trên lý thuyết "đánh giá vì sự học" (Black & Wiliam, 1998), nghiên cứu ủng hộ sự chuyển dịch từ đánh giá sản phẩm sang đánh giá quá trình. Các phương pháp thay thế bao gồm: đ*ánh giá thực* (Authentic Assessment), sử dụng các bài tập đòi hỏi phân tích và trải nghiệm cá nhân mà AI khó bắt chước (Wiggins, 1990), và đánh giá tích hợp (Integrative Assessment), kết hợp portfolio, tự đánh giá, đặc biệt là phỏng vấn hoặc thuyết trình để xác thực quyền sở hữu trí tuệ (Carless, 2015).

2.3. Đặc thù của việc dạy và học kỹ năng viết tiếng Trung Quốc

Bối cảnh cụ thể của tiếng Trung làm trầm trọng thêm vấn đề. Ở trình độ cao, người học giỏi thường sử dụng các thành ngữ (成语) và cấu trúc câu rõ ràng, mạch lạc - những đặc điểm trùng lặp với văn phong của AI. Điều này khiến họ trở thành đối tượng có nguy cơ bị cáo buộc oan cao hơn (Li, 2022). Do đó, việc phát triển một khung đánh giá công bằng không chỉ là nhu cầu sư phạm chung mà còn là một yêu cầu cấp thiết mang tính đặc thù ngành.

Tóm lại, cơ sở lý luận khẳng định: (1) Công cụ phát hiện AI là không đáng tin cậy, (2) Lý thuyết giáo dục ủng hộ các phương pháp đánh giá quá trình, và (3) tính cấp thiết của vấn đề càng được nhấn mạnh trong bối cảnh giảng dạy tiếng Trung, tạo tiền đề cho việc đề xuất một khung đánh giá mới.

III. Phương pháp nghiên cứu

3.1. Thiết kế nghiên cứu

Nghiên cứu này được thiết kế theo phương pháp định tính, sử dụng cách tiếp cận phân tích tài liệu (document analysis) và nghiên cứu trường hợp (case study). Cách tiếp cận này phù hợp để đi sâu khám phá và minh họa cho những hạn chế phức tạp và đầy tính bối rối của các công cụ phát hiện văn bản AI trong bối cảnh thực tế (Creswell & Poth, 2018). Trọng tâm là phân tích một trường hợp cụ thể nhằm cung cấp bằng chứng thực nghiệm sâu sắc, làm cơ sở cho các lập luận và khuyến nghị về mặt lý thuyết.

3.2. Thu thập dữ liệu

Tài liệu hiện trường (Artifacts) và Trải nghiệm cá nhân: Nguồn dữ liệu chính và mang tính quyết định cho nghiên cứu này là một tài liệu hiện trường cụ thể: một bài nghiên cứu khoa học gốc có tiêu đề: "Barriers to Implementing CLIL in Tourism English Courses: A Survey of Teachers' Perspectives in Several Vietnamese Universities" do chính tác giả viết hoàn toàn bằng tiếng Anh. Bài viết này trải qua quy trình kiểm tra nghiêm ngặt bằng phần mềm Turnitin. Báo cáo kết quả kiểm tra từ phần mềm này (có chỉ số phần trăm hoặc cảnh báo về khả năng do AI tạo ra) được lưu lại và sử dụng làm bằng chứng để phân tích. Trải nghiệm này đóng vai trò như một trường hợp điển hình (typical case) minh họa cho vấn đề chung mà nhiều nhà nghiên cứu và sinh viên có thể gặp phải.

IV. Kết quả nghiên cứu

Phân tích dữ liệu từ nguồn tài liệu thứ cấp và kết quả thực nghiệm từ trường hợp cụ thể đã làm sáng tỏ hai nhóm kết quả chính, cung cấp bằng chứng thực tế cho những hạn chế mang tính hệ thống của các công cụ phát hiện văn bản AI.

4.1. Phân tích định tính: Xu hướng dương tính giả trong đánh giá văn bản học thuật

Kết quả tổng hợp từ các nghiên cứu trước (Weber-Wulff và cộng sự, 2023) chỉ ra một xu hướng nổi bật: các công cụ phát hiện AI có tỷ lệ dương tính giả (false positive) đáng kể. Xu hướng này đặc biệt nghiêm trọng với các văn bản học thuật - vốn có đặc điểm là cấu trúc chặt chẽ, ngôn ngữ trang trọng và sử dụng từ vựng chuyên ngành chuẩn mực. Những đặc điểm này, vốn là thước đo của một bài viết chất lượng cao, lại trùng khớp một cách ngẫu nhiên với mẫu hình văn bản do AI tạo ra. Sự trùng lặp này dẫn đến hệ quả tất yếu là các bài viết của con người, đặc biệt là của các researchers và sinh viên xuất sắc, dễ bị gắn nhãn sai một cách oan uổng. Điều này không chỉ làm giảm lòng tin vào công nghệ mà còn tiềm ẩn nguy cơ gây tổn hại đến uy tín và động lực học tập của người học.

4.2. Phân tích định tính trường hợp: Minh chứng về tính không ổn định và thiếu nhất quán

Nghiên cứu tiến hành phân tích một trường hợp cụ thể để kiểm chứng độ tin cậy của công cụ. Một bài báo khoa học gốc do tác giả tự viết hoàn toàn được đưa vào kiểm tra lặp lại nhiều lần bằng các phần mềm khác nhau (Turnitin, GPTZero) và ở các thời điểm chỉnh sửa khác nhau. Kết quả thu được thể hiện qua Bảng 1 cho thấy sự biến động đáng kể và thiếu nhất quán trong tỷ lệ phần trăm bị nghi ngờ.

Bảng 1. Kết quả kiểm tra văn bản bằng các công cụ phát hiện AI

Lần	Công cụ	Tỷ lệ (%)	Nhận diện chính	Ghi chú
1	Turnitin	32	Introduction; 2.2.; 2.5; Conclusion	-
2	Turnitin	31	Các mục tương tự lần 1	Đã viết lại toàn bộ các phần bị nghi ngờ nhưng kết quả không thay đổi.
3	Turnitin	26	Xuất hiện thêm các câu mới ở phần Keywords và 2.3.1	Phần đã sửa bị nghi ngờ, phần mới thêm vào cũng ngay lập tức bị nghi ngờ.
4	GPTZero	< 20	-	Đạt yêu cầu theo tiêu chí< 20%.
5	Turnitin	26	-	Gửi bài cho Ban tổ chức (BTC), kết quả không khớp với công cụ khác.
6	Turnitin	20	-	Tiếp tục chỉnh sửa.
7	Bên thứ 3	10	-	Kết quả từ một công cụ kiểm tra khác.
8	Turnitin	20	-	Tác giả lựa chọn phương án xoá bớt phần bị cho là do AI viết đi

Kết quả trên cho thấy:

(1) Tính không ổn định: Cùng một văn bản, qua các lần chỉnh sửa, cho kết quả biến động liên tục (32% -> 31% -> 26% -> 20%) mà không theo một quy luật rõ ràng.

(2) Sự thiếu nhất quán giữa các công cụ: Một công cụ (GPTZero) cho kết quả dưới 20%, trong khi công cụ khác (Turnitin) trên cùng một bài lại cho kết quả 26%, dẫn đến các quyết định trái ngược.

(3) Tính không minh bạch: Cơ chế nhận diện không rõ ràng. Các phần đã được viết lại hoàn toàn vẫn tiếp tục bị gắn cờ, và các phần mới thêm vào ngay lập tức bị nghi ngờ, tạo thành một "vòng xoáy" không lối thoát cho người viết.

4.3. Khẳng định từ nhà phát triển về tính không hoàn hảo

Đáng chú ý, ngay trong báo cáo kết quả kiểm tra của Turnitin, nhà phát triển đã có một tuyên bố miễn trừ trách nhiệm (disclaimer) rõ ràng (Bảng 2). Tuyên bố này thừa nhận rằng công cụ "có thể không phải lúc nào cũng chính xác" và khuyến nghị "không nên được sử dụng làm căn cứ duy nhất" cho các hành động kỷ luật, đồng thời nhấn mạnh sự cần thiết của "sự xem xét kỹ lưỡng và phán đoán của con người".

Bảng 2. Tuyên bố miễn trừ trách nhiệm của Turnitin

Nội dung gốc (English)	Tạm dịch (Tiếng Việt)
Disclaimer: Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positives or false negatives), so it should not be used as the sole basis for adverse actions against a student. It requires further scrutiny and human judgment, in conjunction with an organization's application of its specific academic policies, to determine whether academic misconduct has occurred.	Tuyên bố miễn trừ trách nhiệm: Công cụ đánh giá bài viết AI của chúng tôi được thiết kế để hỗ trợ các nhà giáo dục xác định văn bản có thể được tạo bởi công cụ AI. Công cụ đánh giá bài viết AI của chúng tôi có thể không phải lúc nào cũng chính xác (ví dụ: mô hình AI của chúng tôi có thể tạo ra kết quả dương tính giả hoặc âm tính giả), do đó không nên được sử dụng làm căn cứ duy nhất để đưa ra các biện pháp kỷ luật đối với sinh viên. Cần có sự xem xét kỹ lưỡng hơn và phán quyết của con người, kết hợp với việc áp dụng các chính sách học thuật cụ thể của tổ chức, để xác định xem có hành vi vi phạm học thuật nào xảy ra hay không.

Tuyên bố này của chính nhà cung cấp công cụ đã củng cố mạnh mẽ cho luận điểm chính của nghiên cứu: các công cụ phát hiện AI hiện tại là không hoàn hảo và việc sử dụng chúng như một thẩm phán duy nhất là một thực hành thiếu khoa học và tiềm ẩn nhiều rủi ro. Điều này đặt ra một câu hỏi lớn về tính hợp lý của việc thiết lập các ngưỡng phần trăm cố định (ví dụ: <20%) để đánh giá tính nguyên bản trong học thuật, đặc biệt lại càng khó khăn và thiếu chính xác hơn khi áp dụng cho các văn bản bằng ngôn ngữ không phải tiếng Anh như tiếng Trung Quốc.

Rõ ràng rằng, ngay cả trong kết quả kiểm tra sử dụng AI mà Ban tổ chức đưa ra ở phía dưới cũng có dòng chữ tuyên bố miễn trừ trách nhiệm và có nói kết quả kiểm tra có thể là chưa chính xác (Bảng 2).

V. Thảo Luận

5.1. Nghịch lý của công cụ phát hiện AI và tác động đến tính toàn vẹn học thuật

Nghiên cứu phát hiện một nghịch lý: việc tuân thủ các công cụ phát hiện AI (vốn có độ tin cậy thấp) có thể dẫn đến những hành vi làm suy giảm chính tính toàn vẹn và chất lượng học thuật. Để đáp ứng các ngưỡng phần trăm cứng nhắc, người viết buộc phải cắt bỏ hoặc làm nghèo nàn những nội dung chất lượng cao thay vì cải thiện chúng, dẫn đến ba hệ quả chính:

(1) Phá vỡ tính mạch lạc: Việc loại bỏ các phần then chốt (như tổng quan lý thuyết, tóm tắt kết quả) làm giảm tính logic và sự trọn vẹn của bài nghiên cứu.

(2) Suy giảm chất lượng ngôn ngữ: Người viết tránh dùng các thuật ngữ chuyên môn và cấu trúc phức tạp - vốn là dấu hiệu của văn phong học thuật cao cấp - để không bị hệ thống đánh dấu.

(3) Lệch lạc mục tiêu: Trọng tâm chuyển từ sáng tạo tri thức sang "tối ưu hóa để vượt qua kiểm tra", biến công cụ bảo vệ tính nguyên bản thành tác nhân kìm hãm sáng tạo.

Như vậy, công cụ được thiết kế để bảo vệ tính nguyên bản lại trở thành tác nhân gián tiếp kìm hãm sự sáng tạo và làm xói mòn chất lượng học thuật. Bài nghiên cứu có thể đạt ngưỡng kỹ thuật về "độ nguyên bản", nhưng giá trị học thuật thực sự của nó lại bị suy giảm.

5.2. Hàm ý cho chính sách và thực tiễn giảng dạy

Phát hiện này phủ nhận tính khoa học của việc sử dụng các ngưỡng phần trăm cố định (ví dụ: <15% nội dung AI) làm tiêu chí đánh giá. Thay vào đó, cần:

(1) Định vị lại vai trò công cụ: Chỉ sử dụng kết quả từ máy dò AI như một "tín hiệu cảnh báo" để mở ra đối thoại, không phải bằng chứng kết tội.

(2) Ưu tiên đánh giá của con người: Đánh giá dựa trên tính logic, chiều sâu học thuật và sự phù hợp với quá trình của người học.

(3) Chuyển đổi sang đánh giá quá trình: Áp dụng các phương pháp như hồ sơ học tập (portfolio) và phỏng vấn bảo vệ (viva voce) để đánh giá năng lực thực.

5.3. Kết luận và khuyến nghị

Nghiên cứu khẳng định: việc áp dụng cứng nhắc công cụ phát hiện AI là phản tác dụng, tạo ra "môi trường học thuật phòng thủ" nơi người viết tự kiểm duyệt và làm nghèo nội dung. Để chuyển đổi sang mô hình thích ứng chủ động, nghiên cứu đề xuất:

(1) Với cơ sở đào tạo:

a) Xây dựng chính sách đánh giá dựa trên quá trình.

b) Phát triển khung đánh giá đa chiều, tích hợp portfolio, phỏng vấn bảo vệ và bài tập tại lớp.

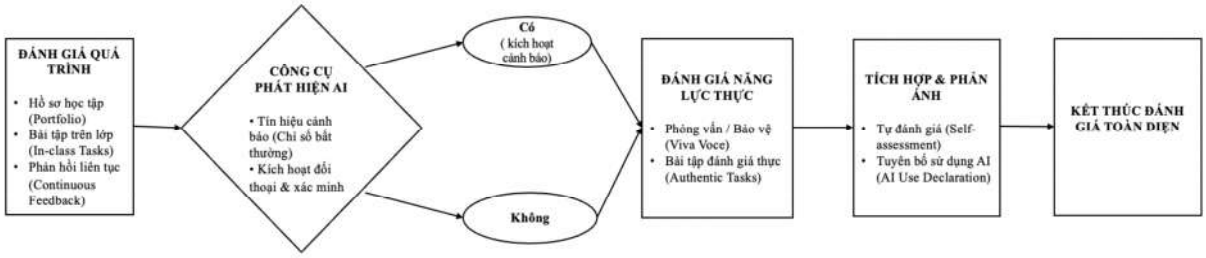
(2) Với giảng viên:

a) Chuyển từ vai trò kiểm tra sang hướng dẫn.

b) Giáo dục về sử dụng AI có đạo đức.

Đặc biệt trong giảng dạy ngôn ngữ, nơi sự tinh tế trong diễn đạt dễ bị công cụ AI quy kết sai, các khuyến nghị này càng trở nên cấp thiết.

Là giải pháp tổng thể, nghiên cứu đề xuất Mô hình Đánh giá Toàn diện Dựa trên Quá trình và Năng lực Thực (CPCAM - Comprehensive Process and Competency-based Assessment Model) (Hình 1).



Hình 1. Mô hình đánh giá toàn diện CPCAM (Comprehensive Process and Competency-based Assessment Model)

(Nguồn: Mô hình được đề xuất bởi tác giả)

Tương lai của đánh giá học thuật phụ thuộc vào việc xây dựng một "hệ sinh thái đánh giá thông minh", nơi công nghệ hỗ trợ và con người giữ vai trò trung tâm trong việc đưa ra các phán quyết công bằng và am hiểu chuyên môn.

VI. Kết luận

Nghiên cứu đã chỉ ra thực trạng đáng quan ngại: sự phụ thuộc thiếu phê phán vào các công cụ phát hiện văn bản AI như "cảnh sát" đánh giá tính nguyên bản trong giáo dục đại học và nghiên cứu khoa học. Bằng phương pháp phân tích tài liệu và trường hợp điển hình, nghiên cứu chứng minh các công cụ này tồn tại nhiều điểm yếu nghiêm trọng về độ chính xác, tính ổn định và minh bạch.

Hệ quả sâu xa không dừng ở các cáo buộc oan sai, mà tạo ra nghịch lý phản học thuật: nhà nghiên cứu buộc phải lựa chọn giữa việc giữ lại nội dung chất lượng cao (với nguy cơ bị từ chối) hoặc tự cắt bỏ chúng để đáp ứng tiêu chuẩn kỹ thuật. Lựa chọn thứ hai trực tiếp làm suy giảm giá trị học thuật - điều mà các công cụ này vốn được tạo ra để bảo vệ.

Để giải quyết nghịch lý này, nghiên cứu đề xuất Mô hình Đánh giá Toàn diện Dựa trên Quá trình và Năng lực Thực (CPCAM) như giải pháp thay thế toàn diện. Mô hình chuyển trọng tâm từ "bắt gian lận" sang "xác thực và phát triển năng lực" thông qua ba trụ cột: (1) Đánh giá quá trình liên tục, (2) Đánh giá năng lực thực, và (3) Tích hợp phản ánh và minh bạch trong sử dụng AI.

Do đó, kết luận then chốt của nghiên cứu này là các công cụ phát hiện AI không thể và không nên được sử dụng như thẩm phán độc nhất hay tiêu chí ngưỡng cứng. Tương lai của nền học thuật lành mạnh trong kỷ nguyên số phụ thuộc vào việc xây dựng các phương pháp đánh giá thông minh dựa trên con người, mà CPCAM là

một đề xuất cụ thể. Sự thay đổi này đòi hỏi chuyển dịch từ đánh giá sản phẩm sang quá trình, từ con số phần trăm sang đối thoại phản biện, và đặt niềm tin vào phán xét chuyên môn - chỉ như vậy mới thực sự khuyến khích sáng tạo và bảo vệ tính toàn vẹn tri thức bền vững.

Tài liệu tham khảo

[1]. Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7-74. https://doi.org/10.1080/0969595980050102

[2]. Carless, D. (2015). *Excellence in university assessment: Learning from award-winning practice*. Routledge. https://doi.org/10.4324/9781315740621

[3]. Creswell, J. W., & Poth, C. N. (2018). *Qualitative inquiry and research design: Choosing among five approaches* (4th ed.). SAGE Publications.

[4]. Gao, J., Ren, L., Yang, Y., Zhang, D., & Li, L. (2022). The impact of artificial intelligence technology stimuli on smart customer experience and the moderating effect of technology readiness. *International Journal of Emerging Markets*, *17*(4), 1123-1142.

[5]. Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Salehi, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

[6]. Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. arXiv preprint arXiv:2303.13408.

[7]. Li, X. (2022). *Teaching Chinese as a foreign language: Theory and practice*. Peking University Press.

[8]. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, *4*(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

[9]. Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. *Proceedings of the 40th International Conference on Machine Learning*. https://doi.org/10.48550/arXiv.2301.11305

[10]. OpenAI. (2023). *Educator considerations for ChatGPT*. https://platform.openai.com/docs/chatgpt-education

[11]. Turnitin. (2025). *AI writing detection: Disclaimer* [Screenshot]. Turnitin.

[12]. Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, *19*(1). https://doi.org/10.1007/s40979-023-00146-z

[13]. Wiggins, G. (1990). The case for authentic assessment. *Practical Assessment, Research, and Evaluation*, *2*(2). https://doi.org/10.7275/ffb1-mm19

[14]. Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). SAGE Publications.

[15]. Zhang, S. L., & Chen, S. (2013). A research on cross-cultural adaptation of foreign students in China. *Overseas English*, (20), 163-164.

THE APPLICATION OF AI IN LEARNING AND THE CHALLENGE FROM DETECTION TOOLS: A PERSPECTIVE FROM TEACHING CHINESE AND A PROPOSAL FOR A NEW ASSESSMENT FRAMEWORK

***Nguyen Ta Nam*[1]**

Abstract*: The growth of AI tools like ChatGPT poses challenges to academic originality, leading to the misuse of AI-generated text detection software. This study examines the inherent limitations of these detectors and their negative impacts on research quality. Using a qualitative approach that combines document analysis and a representative case study, the research highlights three main weaknesses: (1) a high false positive rate with high-quality academic writing, (2) unstable results that create an endless "editing loop" for authors, and (3) low reliability acknowledged by the developers themselves. The study concludes that relying solely on these tools for plagiarism detection is counterproductive, as it unintentionally promotes self-censorship and weakens academic content. As an alternative, the study proposes the Comprehensive Process and Competency-based Assessment Model (CPCAM), a human-centered evaluation framework designed to uphold substantive academic integrity in the digital age.*

Keywords: *AI detection, false positives, academic assessment, AI plagiarism, Turnitin*

[1] Hanoi Metropolitan University