

# XÂY DỰNG HỆ THỐNG CHẤM BÀI TỰ ĐỘNG DỰA TRÊN OFFICE OPEN XML, RUBRIC JSON VÀ AI HỖ TRỢ PHẢN HỒI CHO HỌC PHẦN TIN HỌC ĐẠI CƯƠNG

Phan Thị Đài Trang<sup>1</sup>, Nguyễn Thị Như<sup>1</sup>, Tống Thị Lan Chi<sup>2</sup>

Ngày nhận bài: 24/06/2026; Ngày phản biện thông qua: 07/08/2026; Ngày duyệt đăng: 07/08/2026

## TÓM TẮT

Nghiên cứu trình bày việc xây dựng và đánh giá hệ thống chấm bài tự động AGS cho các bài thực hành Microsoft Word, Excel và PowerPoint trong học phần Tin học đại cương. Hệ thống phân tích cấu trúc Office Open XML của các tệp .docx, .xlsx và .pptx, chuẩn hóa dữ liệu thành JSON AST, định vị đối tượng bằng Anchor Locator và thực thi các tiêu chí được mô hình hóa dưới dạng Rubric JSON thông qua Rule Engine. Quy trình được điều phối trên nền tảng n8n; mô hình ngôn ngữ lớn chỉ hỗ trợ diễn giải kết quả và tạo phản hồi, không trực tiếp quyết định điểm số. Dữ liệu ban đầu gồm 1.374 lượt bài; sau khi loại tệp không hợp lệ và trùng lặp, 1.257 lượt được xử lý, trong đó 1.247 cặp điểm đầy đủ được sử dụng để phân tích. Kết quả cho thấy MAE toàn hệ thống là 0,618 điểm trên thang điểm 10; 84,4% lượt chấm có sai lệch không quá một điểm và độ chệch trung bình là -0,051 điểm. Excel có sai lệch tuyệt đối thấp nhất. Phân tích 204 trường hợp cần rà soát cho thấy sai lệch chủ yếu liên quan đến việc chưa nhận diện đầy đủ các phương án thực hiện hợp lệ hoặc một số luật chấm chưa kiểm tra đầy đủ kỹ thuật. Kết quả cho thấy tính khả thi của kiến trúc chấm thống nhất và có khả năng truy vết, đồng thời khẳng định sự cần thiết của giảng viên trong kiểm soát các trường hợp ngoại lệ.

**Từ khóa:** Hệ thống chấm bài tự động; Office Open XML; Rubric JSON; Rule Engine; Phản hồi học tập bằng trí tuệ nhân tạo; Tin học đại cương.

## 1. MỞ ĐẦU

Đánh giá kết quả học tập giữ vai trò quan trọng trong quá trình đào tạo, cung cấp căn cứ để xác định mức độ đạt chuẩn đầu ra, phản hồi cho người học và điều chỉnh hoạt động dạy học. Trong các học phần thực hành tin học, đặc biệt là học phần Tin học đại cương, người học thường được yêu cầu tạo và hoàn thiện các sản phẩm số bằng Microsoft Word, Excel và PowerPoint. Việc đánh giá các sản phẩm này không chỉ căn cứ vào nội dung hiển thị mà còn phải xem xét cấu trúc tài liệu, định dạng, bố cục, công thức, tham chiếu dữ liệu và kỹ thuật sử dụng công cụ. Khi số lượng bài nộp lớn, việc chấm thủ công theo từng tiêu chí làm gia tăng đáng kể khối lượng công việc của giảng viên, kéo dài thời gian phản hồi và có thể phát sinh khác biệt trong cách áp dụng rubric giữa các lần chấm.

Sự phát triển của trí tuệ nhân tạo, xử lý ngôn ngữ tự nhiên và các công nghệ phân tích dữ liệu đã thúc đẩy mạnh mẽ nghiên cứu về đánh giá tự động trong giáo dục. Tổng quan của Gao et al. (2024) cho thấy các hệ thống đánh giá tự động ở giáo dục sau phổ thông đã được phát triển với nhiều loại dữ liệu đầu vào, phương pháp xử lý và hình thức đầu ra khác nhau. Tuy nhiên, các nghiên cứu vẫn còn khác biệt đáng kể về thiết kế thực nghiệm, cách xác lập điểm chuẩn và chỉ số đánh giá, làm hạn chế khả năng so sánh giữa các hệ thống. Tương tự, Sembey et al. (2024) nhận định rằng các công nghệ mới như trí tuệ nhân tạo, phân tích học tập và tự động hóa đang ngày càng được ứng dụng trong đánh giá và phản hồi ở giáo dục đại học, nhưng vẫn thiếu các khung thiết kế và

đánh giá thống nhất, đặc biệt đối với những hệ thống xử lý nhiều loại sản phẩm học tập.

Trong giáo dục lập trình, các công cụ chấm tự động thường dựa trên phân tích tĩnh, thực thi chương trình, kiểm thử và so sánh đầu ra. Các công cụ này có khả năng rút ngắn thời gian chấm và cung cấp phản hồi nhanh, nhưng độ tin cậy vẫn phụ thuộc vào mức độ đầy đủ của bộ tiêu chí, bộ kiểm thử và khả năng nhận diện những lời giải hợp lệ khác với đáp án mẫu (Messer et al., 2024). Đối với bài trả lời ngắn, các nghiên cứu cũng chỉ ra rằng hệ thống chấm tự động phải giải quyết đồng thời các vấn đề về biến thể ngôn ngữ, sự tương đương ngữ nghĩa và tính nhất quán với đánh giá của con người (Padó et al., 2024). Những kết quả này cho thấy tự động hóa đánh giá không chỉ là bài toán tính toán điểm, mà còn là bài toán mô hình hóa tiêu chí và nhận diện nhiều phương án thực hiện hợp lệ.

Khác với bài trắc nghiệm, câu trả lời ngắn hoặc bài lập trình, sản phẩm Word, Excel và PowerPoint được lưu dưới dạng tài liệu có cấu trúc phức hợp. Một bài làm có thể hiển thị đúng về mặt trực quan nhưng không sử dụng đúng kỹ thuật được yêu cầu. Ngược lại, nhiều cách thực hiện khác nhau có thể tạo ra kết quả tương đương và đều cần được công nhận. Trong Word, định dạng có thể được thiết lập trực tiếp hoặc kế thừa từ Style; nội dung có thể được bố trí bằng Tab, bảng hoặc các đối tượng khác. Trong Excel, các công thức khác nhau có thể cho cùng một kết quả, trong khi một giá trị hiển thị đúng có thể được nhập thủ công thay vì được tính bằng công thức. Trong PowerPoint, đối tượng có thể được đặt trực tiếp trên slide, kế thừa từ Slide Layout hoặc Slide Master, hoặc được tổ chức

<sup>1</sup>Khoa Khoa học Tự nhiên và Công nghệ, Trường Đại học Tây Nguyên;

<sup>2</sup>Phòng Khoa học, Hợp tác và Đổi mới sáng tạo, Trường Đại học Tây Nguyên;

Tác giả liên hệ: Nguyễn Thị Như; Email: ntnhu@ttn.edu.vn.

thành các nhóm hình. Vì vậy, việc chấm tự động chỉ dựa trên nội dung hiển thị, chuỗi văn bản hoặc sự tồn tại của đối tượng chưa đủ để phản ánh chính xác mức độ hoàn thành bài tập.

Một số nghiên cứu đã đề xuất các giải pháp tự động hóa việc đánh giá kỹ năng sử dụng Microsoft Office. Kovačić và Green (2012) xây dựng phương pháp chấm tự động kỹ năng bảng tính và cơ sở dữ liệu, qua đó cho thấy khả năng tích hợp đánh giá thực hành với môi trường quản lý học tập. Hekman (2019) phát triển công cụ chấm bảng tính Microsoft Excel nhằm kiểm tra các thành phần của bài làm và giảm công sức chấm thủ công. Iwata và Matsui (2022) đề xuất hệ thống chấm tự động cho tệp Excel và Word, hướng tới giảm thời gian chấm, cung cấp phản hồi tức thời và hỗ trợ quản lý bài nộp. Bên cạnh đó, Hunnicutt (2016) sử dụng bảng tính để hỗ trợ tự động hóa việc tính điểm báo cáo hoặc bài thuyết trình dựa trên rubric. Tuy nhiên, các công trình hiện có chủ yếu tập trung vào một hoặc hai định dạng, sử dụng các luật được xây dựng riêng cho từng dạng bài và chưa hình thành một cơ chế thống nhất để phân tích, định vị và chấm các thành phần trong Word, Excel và PowerPoint bằng cùng một kiến trúc.

Về mặt kỹ thuật, các tệp .docx, .xlsx và .pptx được xây dựng trên chuẩn Office Open XML. Các đặc tả ECMA-376 và ISO/IEC 29500 quy định hệ thống từ vựng XML, cấu trúc đóng gói, các thành phần tài liệu và quan hệ giữa chúng (Ecma International, 2021; International Organization for Standardization, 2021). Theo đó, một tài liệu Office là một gói chứa nhiều phần XML, tài nguyên và tệp quan hệ. Cấu trúc này cho phép hệ thống truy cập trực tiếp các thuộc tính như đoạn văn, font chữ, thiết lập trang, bảng, công thức, biểu đồ, slide, shape, placeholder và các quan hệ với Slide Layout hoặc Slide Master mà không phụ thuộc hoàn toàn vào giao diện hiển thị của Microsoft Office. Tuy nhiên, WordprocessingML, SpreadsheetML và PresentationML sử dụng các mô hình cấu trúc khác nhau. Vì vậy, để áp dụng cùng một cơ chế chấm cho ba định dạng, cần xây dựng một lớp biểu diễn trung gian có khả năng chuẩn hóa các thành phần tài liệu trước khi thực thi luật đánh giá.

Rubric là nền tảng để chuyển yêu cầu của bài tập thành các tiêu chí có thể quan sát và chấm điểm. Việc sử dụng rubric có thể làm rõ kỳ vọng của nhiệm vụ, tăng tính minh bạch của đánh giá và hỗ trợ người học tự điều chỉnh quá trình thực hiện (Panadero & Jonsson, 2013). Jonsson và Svingby (2007) cho rằng rubric có khả năng cải thiện độ tin cậy của đánh giá, đặc biệt khi tiêu chí được mô tả cụ thể, gắn với nhiệm vụ và được sử dụng nhất quán. Tuy nhiên, rubric không tự động bảo đảm độ tin cậy và độ giá trị của điểm số. Hiệu quả của rubric còn phụ thuộc vào chất lượng mô tả tiêu chí, sự phân định giữa các mức độ đạt và cách người chấm áp dụng rubric trong thực tế (Rezaei & Lovorn, 2010).

Rubric quá tổng quát có thể tạo ra nhiều cách diễn giải, trong khi rubric quá cứng nhắc có thể loại bỏ những phương án thực hiện khác nhưng vẫn hợp lệ. Panadero (2020) lưu ý rằng rubric có thể đơn giản hóa quá mức những năng lực phức tạp hoặc tạo cảm giác khách quan nếu các tiêu chí chưa phản ánh đầy đủ bản chất của nhiệm vụ. Đối với hệ thống chấm tự động, hạn chế này trở nên rõ ràng hơn vì mỗi yêu cầu phải được chuyển thành điều kiện có thể thực thi. Do đó, rubric cần được biểu diễn dưới dạng có cấu trúc, trong đó xác định rõ đối tượng cần kiểm tra, phạm vi tìm kiếm, thuộc tính, toán tử so sánh, giá trị kỳ vọng, dung sai, trọng số và nguyên tắc cho điểm thành phần.

Sự xuất hiện của các mô hình ngôn ngữ lớn mở rộng khả năng đánh giá câu trả lời mở và sinh phản hồi bằng ngôn ngữ tự nhiên. Pecuchova et al. (2025) cho thấy các mô hình GenAI có tiềm năng hỗ trợ chấm câu hỏi mở trong giáo dục đại học, nhưng mức độ phù hợp với người chấm phụ thuộc đáng kể vào mô hình, thiết kế lời nhắc, dữ liệu tham chiếu và phương pháp đánh giá. Đối với sản phẩm Office, phần lớn tiêu chí liên quan đến thuộc tính kỹ thuật có thể được xác định trực tiếp từ cấu trúc OOXML. Vì vậy, sử dụng mô hình ngôn ngữ lớn để quyết định toàn bộ điểm số có thể làm giảm tính ổn định, khả năng giải thích và khả năng kiểm toán. Một kiến trúc phù hợp hơn là sử dụng Rule Engine để tính điểm dựa trên dữ liệu cấu trúc và rubric đã xác định, trong khi mô hình ngôn ngữ lớn chỉ hỗ trợ diễn giải lỗi và tạo phản hồi cho người học.

Từ tổng quan nghiên cứu có thể xác định ba khoảng trống chính. Thứ nhất, còn thiếu một kiến trúc chấm thống nhất cho cả Word, Excel và PowerPoint dựa trên cấu trúc bên trong của tài liệu. Thứ hai, các hệ thống hiện có chưa làm rõ đầy đủ cơ chế định vị đối tượng cần chấm khi bài làm của sinh viên thay đổi vị trí, nội dung hoặc cấu trúc so với tệp mẫu. Thứ ba, nhiều nghiên cứu tập trung vào khả năng triển khai kỹ thuật nhưng chưa phân tích đầy đủ mức độ phù hợp giữa điểm máy và điểm giảng viên, nguyên nhân của các trường hợp sai lệch và cơ chế chuyển các bài có độ tin cậy thấp trở lại cho giảng viên rà soát.

Nghiên cứu này đề xuất hệ thống chấm bài tự động AGS cho các bài thực hành Word, Excel và PowerPoint trong học phần Tin học đại cương. Hệ thống phân tích tài liệu Office Open XML và chuyển đổi các thành phần của tài liệu thành cấu trúc JSON AST thống nhất. Anchor Locator được sử dụng để xác định đối tượng cần đánh giá, trong khi các tiêu chí được biểu diễn bằng Rubric JSON và thực thi bởi Rule Engine. Quy trình chấm được điều phối bằng nền tảng n8n; mô hình ngôn ngữ lớn được sử dụng ở lớp hỗ trợ phản hồi và không trực tiếp quyết định điểm số.

Kết quả chấm tự động được đối sánh với điểm tham chiếu do một giảng viên xác định trên tập dữ liệu đã được mã hóa. Sai lệch được phân tích riêng

cho Word, PowerPoint và Excel. Các trường hợp có sai lệch tuyệt đối vượt quá một điểm trên thang điểm 10 hoặc thiếu dữ liệu được nhận diện trong phân tích hậu nghiệm và đưa vào quy trình rà soát. Do việc xác định vượt ngưỡng cần có điểm giảng viên, quy trình này được sử dụng để đánh giá và cải tiến hệ thống, không được xem là cơ chế phân luồng độc lập trong vận hành thực tế. Sau rà soát, các nguyên nhân chẩn đoán được ghi nhận theo nhóm liên quan đến rubric, Anchor Locator, parser, Rule Engine và phương án thực hiện tương đương.

Trên cơ sở đó, nghiên cứu tập trung trả lời các câu hỏi sau:

- Mức độ phù hợp giữa điểm do AGS xác định và điểm do giảng viên chấm đối với các bài Word, Excel và PowerPoint như thế nào?
- Đặc điểm sai lệch giữa điểm AGS và điểm tham chiếu của giảng viên khác nhau như thế nào giữa ba loại bài thực hành?
- Những nguyên nhân kỹ thuật và đánh giá nào dẫn đến các trường hợp sai lệch vượt quá một điểm?
- Việc nhận diện hậu nghiệm các trường hợp vượt ngưỡng hoặc thiếu dữ liệu cung cấp thông tin gì cho thiết kế cơ chế rà soát có sự tham gia của giảng viên?

Đóng góp chính của nghiên cứu gồm: xây dựng kiến trúc chấm thống nhất cho ba định dạng Microsoft Office; đề xuất lớp biểu diễn JSON AST và Rubric JSON nhằm tách biệt dữ liệu tài liệu khỏi logic đánh giá; thiết kế và triển khai Anchor Locator để định vị đối tượng dựa trên nội dung, cấu trúc và quan hệ trong tài liệu; sử dụng Rule Engine để bảo đảm quá trình tính điểm có tính xác định và khả năng truy vết; đồng thời đánh giá thực nghiệm mức độ phù hợp giữa điểm AGS và điểm tham chiếu của giảng viên. Nghiên cứu cũng xây dựng quy trình rà soát hậu nghiệm để phân tích các trường hợp sai lệch. Hiệu quả định lượng riêng của Anchor Locator chưa được đánh giá độc lập và được xác định là một hướng nghiên cứu tiếp theo.

## 2. VẬT LIỆU VÀ PHƯƠNG PHÁP

### 2.1. Thiết kế nghiên cứu

Nghiên cứu được thực hiện theo thiết kế phát triển và đánh giá hệ thống, gồm hai giai đoạn chính. Giai đoạn thứ nhất tập trung xây dựng hệ thống chấm bài tự động AGS (*Automated Grading System*) cho các bài thực hành Microsoft Word, Excel và PowerPoint. Giai đoạn thứ hai đánh giá mức độ phù hợp giữa điểm do AGS xác định và điểm chấm của giảng viên trên dữ liệu bài kiểm tra thực tế.

Điểm do một giảng viên xác định theo cùng phiên bản rubric được sử dụng làm điểm tham chiếu thực hành. Điểm AGS được tạo độc lập từ tệp bài làm. Hai bộ điểm được đối sánh ở cấp từng lượt chấm và từng loại sản phẩm. Các trường hợp có sai lệch tuyệt đối vượt quá 1,0 điểm hoặc thiếu dữ liệu được nhận diện hậu nghiệm để giảng viên rà soát và phân tích nguyên nhân. Thiết kế này đánh giá mức phù hợp giữa AGS

và điểm tham chiếu, nhưng không thiết lập chuẩn vàng từ sự đồng thuận của nhiều giám khảo.

Nghiên cứu không sử dụng mô hình ngôn ngữ lớn để trực tiếp quyết định điểm số. Điểm được xác định bằng các luật có cấu trúc trong Rule Engine; mô hình ngôn ngữ lớn chỉ được sử dụng để diễn giải kết quả và hỗ trợ tạo phản hồi bằng ngôn ngữ tự nhiên.

### 2.2. Dữ liệu nghiên cứu

Dữ liệu nghiên cứu gồm các bài kiểm tra thực hành Word, Excel và PowerPoint của học phần Tin học đại cương. Mỗi người học thực hiện ba phần bài. Tập thu nhận ban đầu gồm 420 tệp Word, 426 tệp PowerPoint và 528 tệp Excel, tổng cộng 1.374 lượt tệp. Sau khi loại tệp trùng lặp, tệp không đúng định dạng và chỉ giữ một tệp hợp lệ cuối cùng cho mỗi người học ở từng loại bài, tập phân tích còn 1.257 lượt, tương ứng 419 người học  $\times$  3 loại bài. Việc giảm mạnh số tệp Excel chủ yếu do dữ liệu ban đầu chứa nhiều lần nộp và tệp trùng; các tệp bị loại không được đưa vào đối sánh điểm.

Trước khi phân tích, mã người học được thay thế bằng mã định danh ẩn dạng SVxxxx. Các thông tin có khả năng truy vết trực tiếp đến người học, như họ tên, mã số sinh viên chính thức, và tên tệp gốc, không được sử dụng trong tập dữ liệu phân tích.

Sau quá trình kiểm tra dữ liệu, số lượt đối sánh có đủ điểm của cả giảng viên và AGS gồm: 416 bài Word; 418 bài PowerPoint; 413 bài Excel.

Tổng cộng có 1.247 cặp điểm hợp lệ được sử dụng trong phân tích định lượng. Mười lượt chấm còn lại chưa đủ dữ liệu đối sánh, gồm chín trường hợp thiếu kết quả AGS và một trường hợp thiếu điểm giảng viên. Các trường hợp này được phân tích riêng như lỗi dữ liệu hoặc lỗi xử lý và không được đưa vào tính toán các chỉ số sai số giữa hai phương pháp.

Điểm của giảng viên và AGS được biểu diễn trên thang điểm từ 0 đến 10. Các giá trị được giữ ở độ chính xác ban đầu và không làm tròn trước khi tính sai lệch.

### 2.3. Hệ thống chấm bài tự động AGS

#### 2.3.1. Kiến trúc tổng thể

AGS được thiết kế theo kiến trúc module gồm các thành phần chính: Module tiếp nhận và kiểm tra tệp; Bộ phân tích tài liệu Office Open XML; Bộ chuyển đổi sang cấu trúc JSON AST; Anchor Locator; Rubric JSON; Rule Engine; Module tổng hợp điểm và tạo báo cáo; Module hỗ trợ sinh phản hồi; Quy trình điều phối bằng n8n.

Các tệp .docx, .xlsx và .pptx được xử lý dựa trên cấu trúc Office Open XML theo ECMA-376 và ISO/IEC 29500 (Ecma International, 2021; International Organization for Standardization, 2021). Sau khi kiểm tra định dạng, hệ thống giải nén gói OOXML, đọc các thành phần XML và quan hệ giữa các phần trong tài liệu.

Quá trình chấm được điều phối bằng n8n. Nền tảng này thực hiện các nhiệm vụ tiếp nhận tệp, gọi

module phân tích tương ứng, chuyển dữ liệu sang Rule Engine, lưu kết quả và tạo báo cáo. n8n không thực hiện logic tính điểm mà chỉ đóng vai trò điều phối quy trình.

### 2.3.2. Phân tích tài liệu Office Open XML

Ba bộ phân tích riêng được xây dựng cho Word, Excel và PowerPoint do mỗi định dạng sử dụng mô hình dữ liệu OOXML khác nhau.

Hệ thống được phát triển bằng Python và triển khai dưới dạng dịch vụ web sử dụng Flask 3.0.3 cùng Werkzeug 3.0.1. Ba thư viện python-docx 1.2.0, openpyxl và python-pptx lần lượt được sử dụng để thao tác ở cấp đối tượng đối với tài liệu Word, Excel và PowerPoint. Các thư viện này hỗ trợ truy xuất các thành phần phổ biến như đoạn văn, bảng, worksheet, ô dữ liệu, công thức, slide, shape, placeholder và thuộc tính định dạng cơ bản. Tuy nhiên, một số thuộc tính

kỹ thuật của Office Open XML, đặc biệt là style kế thừa, quan hệ giữa các phần, Slide Layout, Slide Master và các cấu trúc XML chuyên biệt, không được các thư viện cấp đối tượng biểu diễn đầy đủ. Vì vậy, hệ thống kết hợp thư viện lxml với module chuẩn zipfile để giải nén gói OOXML, phân tích trực tiếp XML, xử lý namespace và thực hiện truy vấn XPath. Dữ liệu sau phân tích được chuẩn hóa thành JSON AST bằng module json, trong khi re hỗ trợ nhận diện các mẫu chuỗi và mã bài. Hệ thống sử dụng Jinja2 để kết xuất báo cáo từ mẫu HTML, WeasyPrint để chuyển báo cáo sang PDF, PyPDF2 3.0.1 để xử lý tệp PDF và requests để trao đổi dữ liệu qua HTTP giữa các dịch vụ Python, workflow n8n và các API liên quan. Danh sách các thư viện chính được trình bày trong Bảng 1.

**Bảng 1. Các thư viện và module Python chính được sử dụng trong hệ thống AGS**

| Thư viện /module | Chức năng trong hệ thống                                                                                                                       |
|------------------|------------------------------------------------------------------------------------------------------------------------------------------------|
| Flask            | Xây dựng dịch vụ web/API để tiếp nhận yêu cầu từ giao diện hoặc workflow n8n, gọi các mô-đun phân tích và trả kết quả chấm.                    |
| Werkzeug         | Hỗ trợ ứng dụng Flask trong xử lý yêu cầu HTTP, tải tệp, kiểm tra tên tệp và quản lý các tiện ích web.                                         |
| python-docx      | Đọc và xử lý tài liệu Word .docx ở cấp đối tượng, gồm đoạn văn, run, bảng, style và các thuộc tính định dạng cơ bản.                           |
| openpyxl         | Đọc và xử lý tệp Excel .xlsx, gồm worksheet, ô dữ liệu, công thức, kiểu dữ liệu, định dạng số, font, màu nền, đường viền và vùng gộp.          |
| python-pptx      | Đọc và xử lý tệp PowerPoint .pptx, gồm slide, shape, placeholder, text box, nội dung và định dạng văn bản, kích thước và vị trí đối tượng.     |
| lxml             | Phân tích trực tiếp các tệp XML bên trong gói OOXML, xử lý namespace, XPath và các thuộc tính không được thư viện cấp đối tượng hỗ trợ đầy đủ. |
| PyPDF2           | Đọc, kiểm tra và thao tác với các tệp PDF được sử dụng trong quy trình báo cáo hoặc tài liệu đầu vào.                                          |
| Jinja2           | Kết xuất báo cáo động từ mẫu HTML bằng cách chèn điểm số, trạng thái từng tiêu chí, cảnh báo và nội dung phản hồi.                             |
| zipfile          | Mở và giải nén các tệp .docx, .xlsx và .pptx để truy cập các thành phần Office Open XML.                                                       |

Đối với Word, hệ thống trích xuất các thành phần như: Thiết lập trang và section; Đoạn văn và run; Font chữ và định dạng ký tự; Căn lề, khoảng cách và giãn dòng; Style; Tab stop và tab leader; Danh sách đánh số; Bảng, hàng, cột và ô; Hình ảnh; Header, footer và đánh số trang; Liên kết và các đối tượng liên quan.

Đối với Excel, hệ thống trích xuất: Worksheet và thứ tự sheet; Giá trị ô; Công thức; Kiểu dữ liệu; Định dạng số; Font, màu nền và đường viền; Ô gộp; Tham chiếu ô và vùng; Bảng, biểu đồ; Sắp xếp, lọc và xác thực dữ liệu.

Đối với PowerPoint, hệ thống phân tích danh sách và thứ tự slide; Slide Layout và Slide Master; Placeholder; Text box; Shape; Nội dung và định dạng văn bản; Hình ảnh; Kích thước và vị trí đối tượng; Theme và một số thuộc tính trình bày.

Các thư viện Python chuyên dụng được sử dụng để hỗ trợ thao tác ở cấp đối tượng. Trong những trường hợp thư viện không cung cấp đầy đủ thuộc tính cần thiết, hệ thống truy cập trực tiếp các thành phần XML

trong gói OOXML.

### 2.3.3. Biểu diễn trung gian JSON AST

Dữ liệu trích xuất từ ba loại tệp được chuyển thành cấu trúc trung gian JSON AST. Mục tiêu của lớp biểu diễn này là chuẩn hóa các đối tượng tài liệu để Rule Engine không phải xử lý trực tiếp sự khác biệt giữa WordprocessingML, SpreadsheetML và PresentationML.

Mỗi nút trong JSON AST gồm các thành phần cơ bản: Loại đối tượng; Mã định danh; Quan hệ cha-con; Nội dung văn bản; Thuộc tính định dạng; Thuộc tính bố cục; Vị trí trong tài liệu; Các quan hệ với đối tượng khác; Thông tin nguồn trong cấu trúc OOXML.

Ví dụ, một đoạn văn Word có thể được chuẩn hóa thành đối tượng gồm nội dung, căn lề, style, font, cỡ chữ, khoảng cách đoạn và các tab stop. Một ô Excel được biểu diễn bằng địa chỉ ô, giá trị, công thức, kiểu dữ liệu và định dạng. Một đối tượng PowerPoint được biểu diễn bằng loại shape, nội dung, tọa độ, kích thước, placeholder và quan hệ với layout.

JSON AST cho phép tách biệt hai lớp của hệ thống: lớp phân tích tài liệu và lớp logic chấm. Nhờ đó, việc cập nhật parser không làm thay đổi trực tiếp các luật chấm nếu cấu trúc trung gian vẫn được duy trì nhất quán.

## 2.4. Xây dựng rubric và luật chấm

### 2.4.1. Rubric JSON

Rubric của từng bài kiểm tra được xây dựng từ đề bài, đáp án và hướng dẫn chấm của giảng viên. Mỗi tiêu chí được chuyển thành một đối tượng JSON có cấu trúc, bao gồm: Mã tiêu chí; Tên và mô tả tiêu chí; Nhóm tiêu chí; Điểm tối đa; Đối tượng cần kiểm tra; Mô neo dùng để định vị đối tượng; Phạm vi tìm kiếm; Đường dẫn thuộc tính; Toán tử so sánh; Giá trị kỳ

vọng; Dung sai; Quy tắc cho điểm; Nội dung phản hồi khi đạt hoặc không đạt.

Một tiêu chí có thể sử dụng một hoặc nhiều luật. Các luật được kết hợp bằng toán tử logic AND, OR hoặc được tính điểm theo từng thành phần. Các tiêu chí phức hợp được tách thành tiêu chí con để hạn chế tình trạng một lỗi nhỏ làm mất toàn bộ điểm của nhóm yêu cầu.

Đối với các thuộc tính số như kích thước trang, lề, tọa độ hoặc kích thước đối tượng, rubric có thể khai báo dung sai. Đối với công thức Excel, rubric có thể khai báo nhiều công thức hoặc nhóm hàm tương đương thay vì chỉ so sánh chuỗi công thức tuyệt đối.

Ví dụ Rubric JSON

```
{
  "criteria_name": "Định dạng Slide Master",
  "allocated_points": 2.0000000000000004,
  "anchor_locator": {
    "type": "slide_master",
    "tag": "p:sldMaster"
  },
  "rules": [
    {
      "description": "Sử dụng đúng Theme",
      "property_to_check": "properties.app_properties.Template",
      "expected_value": "Facet",
      "action": "VERIFY_PROPERTY",
      "match_type": "EXACT",
      "points": 0.12
    }
  ],
  "sub_criteria": [
    {
      "criteria_name": "Tiêu đề (Cấp 1)",
      "allocated_points": 0.12,
      "anchor_locator": {
        "type": "master_style",
        "style_type": "titleStyle",
        "level": 1
      },
      "rules": [
        {
          "description": "Kiểm tra Font chữ",
          "property_to_check": "properties.text_styles.titleStyle.lvl1Pr.font_name",
          "expected_value": "Times New Roman",
          "action": "VERIFY_PROPERTY",
          "match_type": "EXACT",
          "points": 0.06,
          "level": 1
        }
      ],
      {
        "description": "Kiểm tra Cỡ chữ",
        "property_to_check": "properties.text_styles.titleStyle.lvl1Pr.sz",
        "expected_value": "4000",
        "action": "VERIFY_PROPERTY",
        "match_type": "EXACT",
        "points": 0.06,

```

```
"level": 1
```

```
}
]
}
```

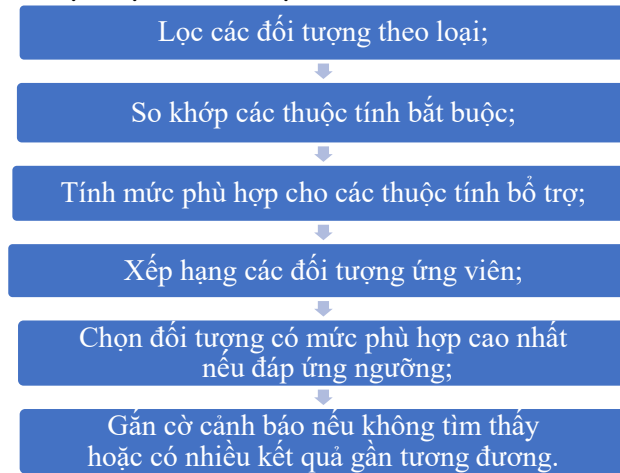
#### 2.4.2. Anchor Locator

Anchor Locator được sử dụng để xác định đối tượng cần chấm trong JSON AST. Cơ chế này cần thiết vì vị trí của đoạn văn, bảng, ô, worksheet hoặc shape có thể thay đổi giữa các bài làm.

Mỏ neo được xác định từ một hoặc nhiều thuộc

tính: Loại đối tượng; Nội dung hoặc từ khóa; Tên style; Tên worksheet; Vị trí tương đối; Quan hệ cha-con; Số thứ tự; Đặc trưng bố cục; Các thuộc tính định dạng có tính phân biệt.

Quá trình định vị gồm các bước:



**Hình 1. Quy trình định vị đối tượng**

Mỗi đối tượng ứng viên được chấm mức phù hợp theo tổng có trọng số của các thuộc tính bắt buộc và bổ trợ. Thuộc tính bắt buộc phải đồng thời thỏa mãn; các thuộc tính bổ trợ được dùng để xếp hạng ứng viên. Nếu không có ứng viên đạt ngưỡng hoặc có nhiều ứng viên có mức phù hợp gần tương đương, hệ thống không tự động kết luận tiêu chí đạt mà gắn cờ không chắc chắn để phục vụ rà soát. Trong nghiên cứu này, các tham số trọng số và ngưỡng được cấu hình theo từng nhóm tiêu chí trong Rubric JSON; đây là thành phần cần được chuẩn hóa và công bố chi tiết hơn trong các nghiên cứu tiếp theo.

Trong nghiên cứu này, Anchor Locator được đánh giá gián tiếp thông qua log thực thi, cảnh báo và quá trình rà soát các bài có sai lệch. Tập dữ liệu chưa được gắn nhãn độc lập ở cấp đối tượng để xác định đối tượng chuẩn cho từng tiêu chí. Vì vậy, nghiên cứu chưa tính precision, recall, F1-score hoặc tỷ lệ định vị đúng riêng cho Anchor Locator. Các kết quả liên quan đến cơ chế này chỉ được sử dụng để mô tả khả năng vận hành và chẩn đoán dạng lỗi, không được diễn giải như bằng chứng định lượng về độ chính xác độc lập.

#### 2.4.3. Rule Engine và cách tính điểm

Sau khi Anchor Locator xác định được đối tượng, Rule Engine truy xuất thuộc tính tương ứng và so sánh với giá trị trong rubric. Kết quả của mỗi luật gồm: Giá trị thực tế; Giá trị kỳ vọng; Trạng thái thực thi; Kết luận đạt hoặc không đạt; Điểm được cấp; Mức tin cậy; Mã cảnh báo, nếu có.

Điểm của tiêu chí được tính theo điểm thành phần: mỗi luật con có một trọng số;

Điểm toàn bài là tổng điểm của các tiêu chí sau khi

giới hạn trong thang điểm 10. Rule Engine là thành phần duy nhất quyết định điểm tự động.

#### 2.5. Quy trình chấm và thu thập kết quả

Quy trình xử lý một bài làm được thực hiện theo các bước:

- Tiếp nhận tệp và xác định loại tài liệu;
- Kiểm tra định dạng, tính toàn vẹn và khả năng đọc tệp;
- Phân tích cấu trúc OOXML;
- Chuyển dữ liệu sang JSON AST;
- Nạp rubric tương ứng với bài;
- Xác định mỏ neo cho từng tiêu chí;
- Thực thi các luật chấm;
- Tổng hợp điểm;
- Ghi nhận trạng thái, cảnh báo và thời gian xử lý;
- Tạo báo cáo kết quả;
- Sinh phản hồi mô tả từ kết quả Rule Engine.

Đối với những tệp không thể phân tích, hệ thống không gán điểm 0. Kết quả được để thiếu và gán mã lỗi tương ứng, chẳng hạn lỗi tệp, lỗi parser, không hỗ trợ định dạng hoặc quy trình xử lý không hoàn thành.

Điểm của giảng viên và điểm AGS được lưu trong cùng một bảng dữ liệu theo mã bài ẩn danh. Sai lệch có dấu được xác định theo công thức:

$$PD_i = S_{AGS,i} - S_{GV,i}$$

Trong đó  $S_{AGS,i}$  là điểm do AGS xác định và  $S_{GV,i}$  là điểm do giảng viên chấm cho bài thứ (i). Giá trị dương cho biết AGS chấm cao hơn giảng viên; giá trị âm cho biết AGS chấm thấp hơn giảng viên.

Sai lệch tuyệt đối được tính như sau:

$$AE_i = |S_{AGS,i} - S_{GV,i}|$$

## 2.6. Phân loại trạng thái và quy trình giảng viên rà soát

Mỗi lượt đối sánh được phân loại theo ba trạng thái:

- ĐẠT: có đủ hai điểm và sai lệch tuyệt đối không vượt quá 1,0 điểm;
- GV CHẤM LẠI: có đủ hai điểm nhưng sai lệch tuyệt đối lớn hơn 1,0 điểm;
- THIẾU DỮ LIỆU: thiếu điểm AGS hoặc thiếu điểm giảng viên.

Ngưỡng 1,0 điểm được sử dụng như mức chấp nhận thực hành ban đầu trên thang điểm 10 và được giảng viên thống nhất trước khi đối sánh. Chỉ số này được gọi là tỷ lệ đồng thuận trong ngưỡng  $\pm 1$  điểm, không phải độ chính xác phân loại. Vì ngưỡng phụ thuộc điểm tham chiếu của giảng viên, nó chỉ phục vụ đánh giá hậu nghiệm và xác định trường hợp cần rà soát trong nghiên cứu này.

Các bài thuộc nhóm GV CHẤM LẠI được giảng viên mở lại và đối chiếu theo từng tiêu chí rubric. Trong quá trình rà soát, giảng viên xem xét: Tập bài làm gốc; Đề và rubric; Điểm ban đầu; Điểm AGS; Kết quả từng lượt; Mô neo được hệ thống lựa chọn; Giá trị thực tế và giá trị kỳ vọng; Cảnh báo của parser hoặc Rule Engine.

Sau rà soát, nguyên nhân sai lệch được phân loại thành các nhóm:

- AGS chưa nhận diện cách thực hiện hợp lệ;
- Công thức hoặc tham chiếu Excel tương đương chưa được công nhận;
- Parser chưa trích xuất đầy đủ thuộc tính;
- Anchor Locator không tìm thấy hoặc nhận nhầm đối tượng;
- Luật chấm quá chặt;
- Luật chấm quá rộng;
- Khác biệt trong cách cho điểm thành phần;
- Điểm chấm thủ công ban đầu chưa chính xác;
- Thiếu kết quả do lỗi tập, parser hoặc workflow.

**Hình 2. Danh sách các bài cần GV xử lý lại**

| MÃ BÀI | LOẠI BÀI | ĐIỂM GV | ĐIỂM AGS | PD (AGS-GV) | PD  | STATUS         | XỬ LÝ BAN ĐẦU                                                       | LÝ DO SAU RÀ SOÁT                                                                                     | HÀNH ĐỘNG ĐỀ XUẤT                                                                                       |
|--------|----------|---------|----------|-------------|-----|----------------|---------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| SV0005 | PPT      | 8,5     | 5,5      | -3,0        | 3,0 | GV CHẤM LẠI    | GV mở bài, đối soát theo rubric và xác nhận điểm cuối cùng.         | AGS chấm thấp hơn rõ rệt; bài PowerPoint sử dụng layout tùy chỉnh, nhóm shape hoặc đối tượng kẻ thừa  | Giữ điểm GV sau chấm lại; cập nhật parser cho grouped shape, placeholder và Slide Master, điều          |
| SV0007 | WORD     | 9,0     |          |             |     | THIẾU ĐIỂM AGS | Kiểm tra tập parser và chạy lại AGS; nếu vẫn lỗi, GV chấm thủ công. | AGS không trả kết quả; khả năng tập Word không được parser đọc đầy đủ, tập lỗi hoặc workflow chưa     | Kiểm tra tập và log xử lý, chạy lại AGS. Nếu vẫn không có kết quả, giữ điểm GV sau rà soát và loại lượt |
| SV0008 | EXCEL    | 5,0     |          |             |     | THIẾU ĐIỂM AGS | Kiểm tra tập parser và chạy lại AGS; nếu vẫn lỗi, GV chấm thủ công. | AGS không trả kết quả; khả năng workbook có công thức, sheet hoặc cấu trúc OOXN/L khiến parser thất   | Kiểm tra tập và log xử lý, chạy lại AGS. Nếu vẫn không có kết quả, giữ điểm GV sau rà soát và loại lượt |
| SV0009 | EXCEL    | 6,0     | 3,5      | -2,5        | 2,5 | GV CHẤM LẠI    | GV mở bài, đối soát theo rubric và xác nhận điểm cuối cùng.         | AGS chấm thấp hơn; một số công thức tương đương hoặc cách tham chiếu hợp lệ chưa được Rule Engine     | Giữ điểm GV; bổ sung nhóm công thức tương đương, tên sheet linh hoạt và cho điểm một phần.              |
| SV0010 | WORD     | 7,0     |          |             |     | THIẾU ĐIỂM AGS | Kiểm tra tập parser và chạy lại AGS; nếu vẫn lỗi, GV chấm thủ công. | AGS không trả kết quả; khả năng tập Word không được parser đọc đầy đủ, tập lỗi hoặc workflow chưa     | Kiểm tra tập và log xử lý, chạy lại AGS. Nếu vẫn không có kết quả, giữ điểm GV sau rà soát và loại lượt |
| SV0011 | EXCEL    | 7,0     | 8,5      | 1,5         | 1,5 | GV CHẤM LẠI    | GV mở bài, đối soát theo rubric và xác nhận điểm cuối cùng.         | AGS chấm cao hơn nhẹ; khác biệt do làm tròn, dùng sai hoặc điểm một phần của một số tiêu chí Excel.   | Giữ điểm GV sau rà soát; hiệu chỉnh tolerance và tròn số.                                               |
| SV0014 | EXCEL    | 5,0     |          |             |     | THIẾU ĐIỂM AGS | Kiểm tra tập parser và chạy lại AGS; nếu vẫn lỗi, GV chấm thủ công. | AGS không trả kết quả; khả năng workbook có công thức, sheet hoặc cấu trúc OOXN/L khiến parser thất   | Kiểm tra tập và log xử lý, chạy lại AGS. Nếu vẫn không có kết quả, giữ điểm GV sau rà soát và loại lượt |
| SV0016 | PPT      | 9,0     | 6,5      | -2,5        | 2,5 | GV CHẤM LẠI    | GV mở bài, đối soát theo rubric và xác nhận điểm cuối cùng.         | AGS chấm thấp hơn; một số bộ cục placeholder hoặc cách trình bày hợp lệ chưa được rubric và parser    | Giữ điểm GV; bổ sung nhiều cấu trúc hợp lệ cho layout placeholder, cho điểm một phần và cập nhật        |
| SV0016 | EXCEL    | 6,0     | 4,5      | -1,5        | 1,5 | GV CHẤM LẠI    | GV mở bài, đối soát theo rubric và xác nhận điểm cuối cùng.         | AGS chấm thấp hơn nhẹ; chênh lệch do công thức tương đương, làm tròn hoặc điểm thành phần.            | Giữ điểm GV sau rà soát; điều chỉnh tolerance số học và quy tắc partial score.                          |
| SV0019 | WORD     | 6,0     | 3,0      | -3,0        | 3,0 | GV CHẤM LẠI    | GV mở bài, đối soát theo rubric và xác nhận điểm cuối cùng.         | AGS không trả kết quả; qua rà soát, một nhóm tiêu chí Word có thể chưa được nhận diện đúng do định    | Giữ điểm GV; bổ sung cách định dạng hợp lệ vào rubric, đượ                                              |
| SV0020 | WORD     | 9,0     | 7,0      | -2,0        | 2,0 | GV CHẤM LẠI    | GV mở bài, đối soát theo rubric và xác nhận điểm cuối cùng.         | AGS chấm thấp hơn; một số tiêu chí định dạng hoặc bộ cục Word được GV chấp nhận nhưng rubric từ       | Giữ điểm GV sau rà soát; tách tiêu chí lớn thành tiêu chí con, bổ sung partial score và điều chỉnh m    |
| SV0020 | EXCEL    | 7,0     | 5,5      | -1,5        | 1,5 | GV CHẤM LẠI    | GV mở bài, đối soát theo rubric và xác nhận điểm cuối cùng.         | AGS chấm thấp hơn; chênh lệch do công thức tương đương, làm tròn hoặc điểm thành phần.                | Giữ điểm GV sau rà soát; điều chỉnh tolerance số học và quy tắc partial score.                          |
| SV0021 | PPT      | 9,0     | 6,0      | -3,0        | 3,0 | GV CHẤM LẠI    | GV mở bài, đối soát theo rubric và xác nhận điểm cuối cùng.         | AGS chấm thấp hơn rõ rệt; bài PowerPoint sử dụng layout tùy chỉnh, nhóm shape hoặc đối tượng kẻ thừa  | Giữ điểm GV sau chấm lại; cập nhật parser cho grouped shape, placeholder và Slide Master, điều          |
| SV0023 | WORD     | 9,5     | 6,5      | -3,0        | 3,0 | GV CHẤM LẠI    | GV mở bài, đối soát theo rubric và xác nhận điểm cuối cùng.         | AGS chấm thấp hơn rõ rệt; qua rà soát, một nhóm tiêu chí Word có thể chưa được nhận diện đúng do định | Giữ điểm GV sau chấm lại; rà soát log anchor và parser, bổ sung cách định dạng hợp lệ vào rubric, đượ   |
| SV0023 | EXCEL    | 6,0     | 7,5      | 1,5         | 1,5 | GV CHẤM LẠI    | GV mở bài, đối soát theo rubric và xác nhận điểm cuối cùng.         | AGS chấm cao hơn nhẹ; khác biệt do làm tròn, dùng sai hoặc điểm một phần của một số tiêu chí Excel.   | Giữ điểm GV sau rà soát; hiệu chỉnh tolerance và tròn số.                                               |
| SV0025 | EXCEL    | 6,0     |          |             |     | THIẾU ĐIỂM AGS | Kiểm tra tập parser và chạy lại AGS; nếu vẫn lỗi, GV chấm thủ công. | AGS không trả kết quả; khả năng workbook có công thức, sheet hoặc cấu trúc OOXN/L khiến parser thất   | Kiểm tra tập và log xử lý, chạy lại AGS. Nếu vẫn không có kết quả, giữ điểm GV sau rà soát và loại lượt |
| SV0027 | PPT      | 4,0     | 2,0      | -2,0        | 2,0 | GV CHẤM LẠI    | GV mở bài, đối soát theo rubric và xác nhận điểm cuối cùng.         | AGS chấm thấp hơn; một số bộ cục placeholder hoặc cách trình bày hợp lệ chưa được rubric và parser    | Giữ điểm GV; bổ sung nhiều cấu trúc hợp lệ cho layout placeholder, cho điểm một phần và cập nhật        |
| SV0028 | WORD     | 8,5     | 7,0      | -1,5        | 1,5 | GV CHẤM LẠI    | GV mở bài, đối soát theo rubric và xác nhận điểm cuối cùng.         | AGS chấm thấp hơn nhẹ; chênh lệch chủ yếu do cách cho điểm một phần đối với font, căn lề, khoảng cách | Giữ điểm GV sau rà soát; hiệu chỉnh trong số hoặc quy tắc điểm thành phần nếu lỗi lần lại.              |

Việc rà soát 204 trường hợp được thực hiện bởi chính giảng viên cung cấp điểm tham chiếu, dựa trên tập gốc, rubric, log từng lượt, đối tượng do Anchor Locator chọn và giá trị thực tế-kỳ vọng. Mỗi trường hợp được gán một nguyên nhân chẩn đoán chính theo nhóm lỗi chi phối sai lệch; nếu có nhiều lỗi, nguyên nhân có ảnh hưởng điểm lớn nhất được ưu tiên. Dữ liệu sau rà soát được ghi riêng, gồm nguyên nhân, hành động đề xuất và điểm cuối cùng nếu có; điểm gốc không bị ghi đè. Do chỉ có một người rà soát, nghiên cứu chưa đánh giá độ tin cậy liên mã hóa và các nhóm nguyên nhân cần được diễn giải như bằng chứng chẩn đoán trong bối cảnh nghiên cứu, không phải phân loại đã được xác nhận độc lập.

## 2.7. Phương pháp phân tích dữ liệu

Phân tích được thực hiện riêng cho Word, PowerPoint và Excel, sau đó tổng hợp trên toàn bộ cặp điểm hợp lệ. Các chỉ số được báo cáo gồm số lượt chấm, sai số tuyệt đối trung bình (MAE), độ chệch

trung bình, tỷ lệ đồng thuận trong ngưỡng  $\pm 1$  điểm và giới hạn đồng thuận Bland-Altman. MAE tổng được tính trên toàn bộ 1.247 lượt chấm hợp lệ, tức là trung bình có trọng số theo số lượt của từng loại bài.

Mức sai số trung bình tuyệt đối được xác định bằng:

$$MAE = \frac{1}{n} \sum_{i=1}^n |S_{AGS,i} - S_{GV,i}|$$

Sai số căn trung bình bình phương được tính theo:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (S_{AGS,i} - S_{GV,i})^2}$$

Độ chệch trung bình được tính bằng:

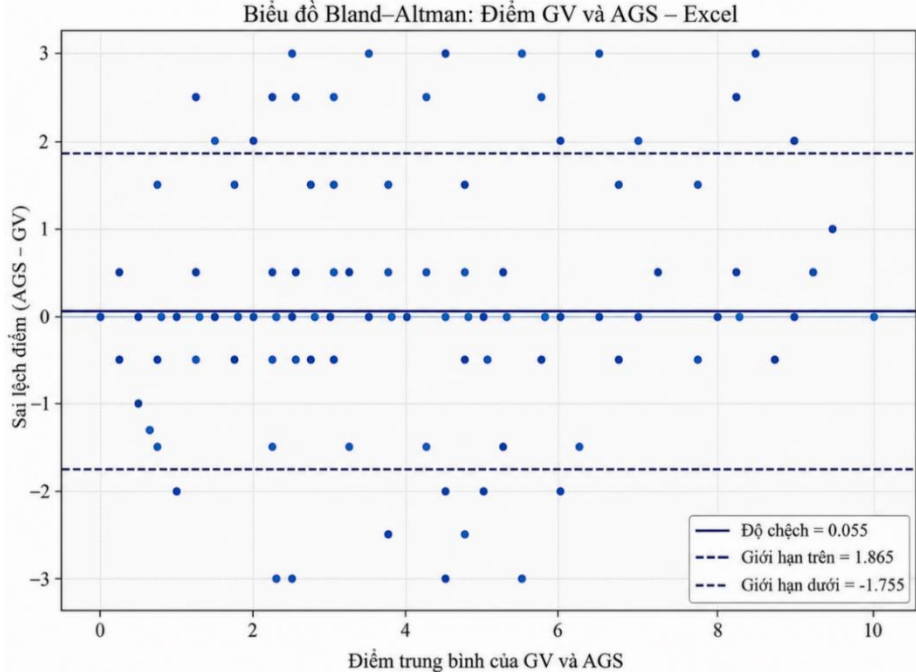
$$Bias = \frac{1}{n} \sum_{i=1}^n (S_{AGS,i} - S_{GV,i})$$

Trong đó, Bias âm cho thấy AGS có xu hướng chấm thấp hơn giảng viên và bias dương cho thấy xu hướng chấm cao hơn.

Nghiên cứu không sử dụng hệ số tương quan làm chỉ số kết luận chính vì tương quan không phản ánh trực tiếp mức đồng thuận giữa hai phương pháp chấm. Tỷ lệ phù hợp trong ngưỡng 1,0 điểm được xác định bằng:

$$P_{\leq 1} = \frac{\sum_{i=1}^n I(AE_i \leq 1)}{n} \times 100\%$$

Trong đó  $I(.)$  là hàm chỉ báo.  
Biểu đồ Bland–Altman được sử dụng để mô tả độ chênh trung bình và giới hạn đồng thuận 95% giữa AGS và điểm tham chiếu của giảng viên. Hình 3 minh hoạ biểu đồ Bland–Altman điểm Excel



Hình 3. Biểu đồ Bland–Altman điểm Excel

Sự khác nhau giữa ba loại bài được trình bày ở mức mô tả thông qua MAE, độ chênh và giới hạn đồng thuận. Nghiên cứu không thực hiện kiểm định suy luận cho sự khác biệt giữa ba nhóm; vì vậy, các nhận định so sánh không được diễn giải như bằng chứng về khác biệt có ý nghĩa thống kê.

Ngoài phân tích định lượng, các trường hợp được giảng viên rà soát được tổng hợp theo nhóm nguyên nhân. Tần suất và tỷ lệ của từng nhóm lỗi được thống kê cho Word, PowerPoint và Excel nhằm xác định những thành phần cần ưu tiên cải tiến trong parser, Anchor Locator, rubric và Rule Engine

Bảng 2. Thống kê nhóm lý do các bài cần xử lý

| Nhóm lý do                                                     | WORD | PPT | EXCEL | Tổng | Tỷ lệ |
|----------------------------------------------------------------|------|-----|-------|------|-------|
| AGS bỏ sót định dạng/bố cục Word hợp lệ                        | 45   | 0   | 0     | 45   | 22,1% |
| AGS bỏ sót layout/placeholder/shape PowerPoint hợp lệ          | 0    | 37  | 0     | 37   | 18,1% |
| Luật Excel chưa kiểm tra đầy đủ công thức/tham chiếu           | 0    | 0   | 37    | 37   | 18,1% |
| AGS chưa công nhận công thức hoặc tham chiếu Excel tương đương | 0    | 0   | 27    | 27   | 13,2% |
| Luật PowerPoint quá rộng hoặc nhận nhầm đối tượng              | 0    | 26  | 0     | 26   | 12,7% |
| Luật Word quá rộng hoặc chưa kiểm tra đúng kỹ thuật            | 22   | 0   | 0     | 22   | 10,8% |
| Thiếu kết quả AGS / lỗi tệp, parser hoặc workflow              | 3    | 1   | 5     | 9    | 4,4%  |
| Thiếu điểm GV                                                  | 0    | 0   | 1     | 1    | 0,5%  |

2.8. Kiểm soát chất lượng dữ liệu

- Trước khi phân tích, dữ liệu được kiểm tra theo các điều kiện sau:
- Mã bài không trùng lặp trong cùng loại sản phẩm;
  - Điểm nằm trong khoảng từ 0 đến 10;
  - Sai lệch PD đúng bằng điểm AGS trừ điểm giảng viên;
  - Dữ liệu thiếu không được tự động thay thế bằng 0;
  - Mỗi bài được gán đúng loại Word, PowerPoint

- hoặc Excel;
- Rubric và phiên bản hệ thống được xác định thống nhất;
  - Các trường hợp thiếu dữ liệu được tách khỏi tập tính sai số;
  - Điểm gốc và kết quả sau rà soát được lưu riêng.
- Trong quá trình kiểm tra, không phát hiện điểm nằm ngoài thang điểm 0–10. Các giá trị PD được tính lại bằng công thức để tránh trường hợp ô điểm trống bị phần mềm bảng tính hiểu là 0 và tạo ra sai lệch giả.
- 2.9. Bảo mật và đạo đức dữ liệu

Dữ liệu sử dụng trong nghiên cứu gồm các bài làm thực hành của học phần Tin học đại cương và đã được Bộ môn Công nghệ thông tin cho phép sử dụng cho mục đích nghiên cứu và cải tiến hệ thống. Dữ liệu được khai thác theo hình thức hồi cứu, không làm thay đổi quá trình kiểm tra hoặc kết quả học tập của sinh viên.

Trước khi phân tích, họ tên, mã số sinh viên, tên tệp gốc và các thông tin định danh trực tiếp được loại bỏ. Mỗi bài làm được thay bằng một mã nghiên cứu không chứa thông tin nhận diện trực tiếp. Tệp bài làm gốc, JSON AST đầy đủ, điểm số và log chấm được lưu trong hạ tầng nội bộ, với quyền truy cập giới hạn cho các thành viên thực hiện nghiên cứu. Kết quả chỉ được báo cáo ở dạng tổng hợp.

Gemini chỉ được sử dụng ở bước tạo phản hồi bằng ngôn ngữ tự nhiên. Tệp Office gốc và JSON AST đầy đủ không được truyền đến dịch vụ bên ngoài. Trước khi gọi API, hệ thống tối thiểu hóa dữ liệu và chỉ gửi các trường kỹ thuật cần thiết ở cấp tiêu chí, gồm mã tiêu chí không định danh, mô tả yêu cầu, trạng thái thực hiện, giá trị kỳ vọng và lỗi do Rule Engine xác định. Họ tên, mã sinh viên, lớp học và tên tệp không được đưa vào yêu cầu. Điểm số được tính hoàn toàn trong hạ tầng nội bộ; phản hồi do Gemini sinh không làm thay đổi điểm.

Kết nối API được thực hiện qua HTTPS/TLS. Khóa API được lưu trong kho thông tin xác thực của n8n hoặc biến môi trường phía máy chủ, không được ghi trong mã nguồn hoặc cung cấp cho trình duyệt người dùng. Nội dung yêu cầu và phản hồi không được ghi đầy đủ vào log ứng dụng; log thực thi được giới hạn quyền truy cập và chỉ được lưu trong thời gian cần thiết cho kiểm tra và xử lý lỗi.

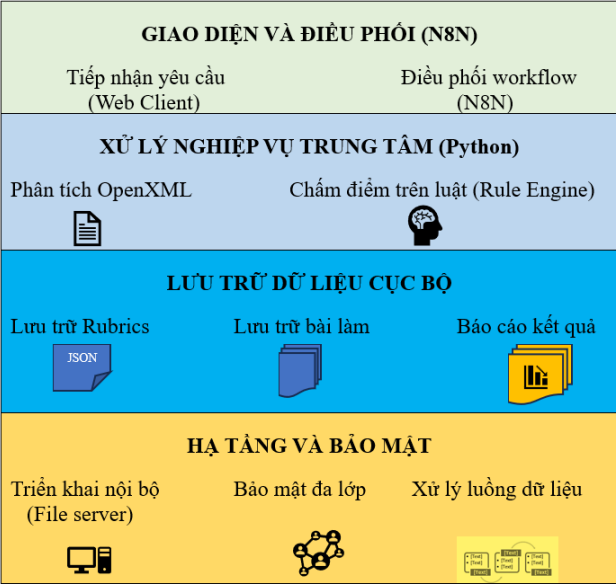
Hệ thống sử dụng Gemini API có trả phí. Chính sách xử lý dữ liệu của dịch vụ được xác định theo điều khoản áp dụng tại thời điểm triển khai. Nhóm tác giả không giả định rằng dịch vụ bên thứ ba hoàn toàn không lưu dữ liệu; do đó, nguyên tắc tối thiểu hóa và loại bỏ thông tin định danh được áp dụng trước mọi yêu cầu gửi ra ngoài.

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Kiến trúc tổng thể và hạ tầng triển khai

Nghiên cứu đã xây dựng và triển khai thành công hệ thống AGS theo kiến trúc module, gồm: tiếp nhận tệp, phân tích Office Open XML, chuyển đổi dữ liệu sang JSON AST, định vị đối tượng bằng Anchor Locator, thực thi Rubric JSON bằng Rule Engine, tổng

hợp điểm và sinh phản hồi. Quy trình được điều phối trên n8n; mô hình ngôn ngữ lớn chỉ hỗ trợ diễn giải kết quả, không tham gia trực tiếp vào quyết định điểm số. Các chức năng phân tích và chấm điểm được cung cấp dưới dạng dịch vụ Flask; n8n đảm nhiệm điều phối workflow, còn Jinja2 và WeasyPrint hỗ trợ kết xuất báo cáo kết quả.



Hình 4. Kiến trúc tổng thể của hệ thống

Điểm nổi bật của kiến trúc là khả năng xử lý thống nhất ba định dạng .docx, .xlsx và .pptx thông qua lớp biểu diễn trung gian JSON AST. Cách tổ chức này tách biệt dữ liệu tài liệu, logic định vị và luật chấm, nhờ đó có thể cập nhật parser hoặc rubric mà không phải thay đổi toàn bộ hệ thống. Đây là khác biệt quan trọng so với các giải pháp trước đây thường được xây dựng riêng cho một loại tài liệu hoặc một dạng bài cụ thể (Hekman, 2019; Iwata & Matsui, 2022; Kovačić & Green, 2012).

3.2. Khả năng chấm tự động trên ba loại bài

AGS được đánh giá trên 1.257 lượt chấm Word, PowerPoint và Excel. Sau khi kiểm tra dữ liệu, 1.247 cặp điểm có đầy đủ kết quả của cả giảng viên và AGS được đưa vào phân tích, gồm 416 bài Word, 418 bài PowerPoint và 413 bài Excel. Chín lượt thiếu kết quả AGS và một lượt thiếu điểm giảng viên được tách khỏi tập tính sai số.

Bảng 4 trình bày mức sai lệch giữa AGS và điểm tham chiếu của giảng viên. Tỷ lệ ở cột cuối được tính trên số cặp có đủ hai điểm của từng loại bài

Bảng 4. Mức sai lệch giữa điểm AGS và điểm giảng viên.

| Loại bài    | n            | MAE          | Độ chênh lệch trung bình | Tỷ lệ đồng thuận trong ngưỡng ±1 điểm |
|-------------|--------------|--------------|--------------------------|---------------------------------------|
| Word        | 416          | 0,692        | -0,154                   | 83,9%                                 |
| PowerPoint  | 418          | 0,714        | -0,054                   | 84,9%                                 |
| Excel       | 413          | 0,446        | 0,055                    | 84,5%                                 |
| <b>Tổng</b> | <b>1.247</b> | <b>0,618</b> | <b>-0,051</b>            | <b>84,4%</b>                          |

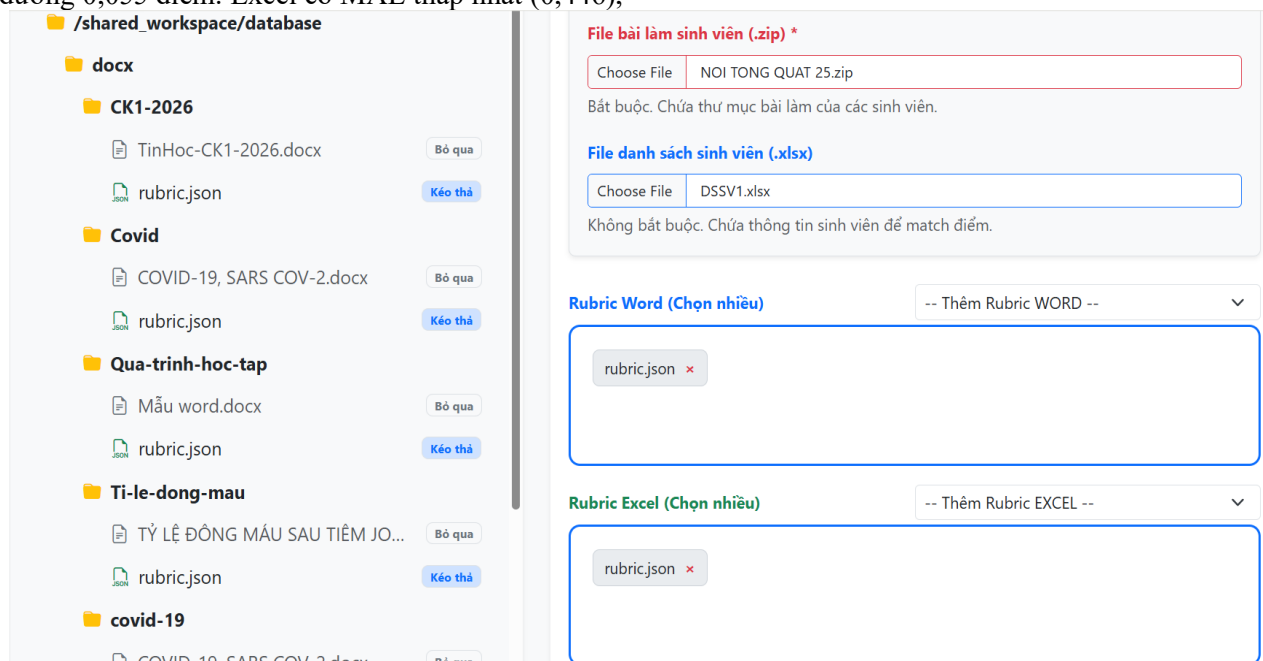
Kết quả cho thấy MAE gộp của 1.247 cặp điểm là 0,618 điểm trên thang điểm 10. Có 1.053/1.247 lượt

chấm, tương đương 84,4%, nằm trong ngưỡng sai lệch không quá một điểm. Tỷ lệ này được diễn giải là

mức đồng thuận trong ngưỡng  $\pm 1$  điểm, không phải độ chính xác tuyệt đối. Độ chệch chung -0,051 điểm cho thấy AGS không có xu hướng chệch lệch rõ rệt theo một chiều ở mức trung bình.

Ở cấp độ mô tả, AGS có độ chệch âm nhẹ đối với Word và PowerPoint, trong khi Excel có độ chệch dương 0,055 điểm. Excel có MAE thấp nhất (0,446),

còn Word và PowerPoint lần lượt là 0,692 và 0,714. Kết quả này gợi ý rằng các thuộc tính có cấu trúc xác định trong Excel thuận lợi hơn cho chấm dựa trên luật; tuy nhiên, nghiên cứu không thực hiện kiểm định suy luận nên không khẳng định sự khác biệt có ý nghĩa thống kê giữa ba loại bài.



**Hình 5. Giao diện chấm bài của hệ thống AGS**

Kết quả cung cấp bằng chứng ban đầu về tính khả thi của một kiến trúc chung cho ba loại sản phẩm Office trong bối cảnh khảo sát. So với các nghiên cứu trước chủ yếu tập trung vào bảng tính hoặc kết hợp Word và Excel, AGS mở rộng phạm vi xử lý sang PowerPoint và sử dụng cùng một lớp biểu diễn trung gian, cơ chế rubric và Rule Engine. Tuy nhiên, kết quả chưa cho phép khẳng định khả năng khái quát sang đề bài, học phần hoặc cơ sở đào tạo khác.

### 3.3. Mức độ đồng thuận giữa AGS và giảng viên

Biểu đồ Bland–Altman được sử dụng để đánh giá độ chệch và giới hạn đồng thuận giữa hai phương pháp chấm.

**Bảng 5. Kết quả phân tích Bland–Altman**

| Loại bài   | Độ chệch | Giới hạn dưới | Giới hạn trên |
|------------|----------|---------------|---------------|
| Word       | -0,154   | -2,185        | 1,877         |
| PowerPoint | -0,054   | -2,090        | 1,982         |
| Excel      | 0,055    | -1,755        | 1,865         |

Độ chệch của cả ba loại bài đều gần 0, cho thấy không có xu hướng sai lệch hệ thống lớn theo một chiều ở mức trung bình. Tuy nhiên, giới hạn đồng thuận còn rộng: xấp xỉ -2,2 đến +1,9 điểm đối với Word, -2,1 đến +2,0 điểm đối với PowerPoint và -1,8 đến +1,9 điểm đối với Excel. Do đó, bias nhỏ không đồng nghĩa với đồng thuận cao ở cấp độ từng bài; vẫn có thể xuất hiện chênh lệch gần hai điểm trong phạm vi giới hạn đồng thuận 95%.

Phân tích Bland–Altman cũng cho thấy việc chỉ sử

dụng hệ số tương quan sẽ chưa đủ để đánh giá chất lượng hệ thống. Hai bộ điểm có thể tương quan cao nhưng vẫn tồn tại các trường hợp sai lệch lớn ở cấp độ từng bài. Vì vậy, nghiên cứu kết hợp MAE, độ chệch, giới hạn đồng thuận và tỷ lệ nằm trong ngưỡng để đánh giá đầy đủ hơn mức phù hợp giữa hai phương pháp.

### 3.4. Phân luồng bài cần giảng viên rà soát

Trong phân tích hậu nghiệm, ngưỡng sai lệch tuyệt đối một điểm được dùng để nhận diện các trường hợp cần xem xét lại. Trong 1.247 lượt có đủ dữ liệu, 194 lượt vượt ngưỡng, gồm:

- 67 bài Word;
- 63 bài PowerPoint;
- 64 bài Excel.

Ngoài ra, chín lượt thiếu kết quả AGS và một lượt thiếu điểm giảng viên được đưa vào nhóm xử lý dữ liệu. Tổng cộng có 204/1.257 lượt (16,2%) được nhận diện hậu nghiệm để rà soát.

Kết quả này cho thấy tiềm năng xây dựng cơ chế human-in-the-loop, nhưng chưa chứng minh AGS có thể tự phân luồng độc lập trong vận hành thực tế, vì việc xác định sai lệch vượt ngưỡng cần có điểm giảng viên. Một cơ chế triển khai thực tế cần dựa trên các tín hiệu nội tại như cảnh báo parser, Anchor không duy nhất, luật không thực thi được, mức tin cậy thấp hoặc lấy mẫu kiểm định ngẫu nhiên.

Việc phân luồng không đồng nghĩa với việc mặc định điểm AGS sai. Sau khi rà soát, giảng viên xác định nguyên nhân, giữ nguyên điểm gốc của cả hai

phương pháp và ghi nhận điểm cuối cùng. Cách làm này giúp bảo đảm khả năng truy vết, đồng thời tạo dữ liệu phản hồi để cải tiến rubric, parser, Anchor Locator và Rule Engine.

**Bảng 6. Phân bố nhóm nguyên nhân sai lệch và thiếu dữ liệu**

| Nhóm nguyên nhân chẩn đoán sau rà soát                         | Số lượng | Tỷ lệ |
|----------------------------------------------------------------|----------|-------|
| AGS bỏ sót định dạng hoặc bố cục Word hợp lệ                   | 45       | 22,1% |
| AGS bỏ sót layout, placeholder hoặc shape PowerPoint hợp lệ    | 37       | 18,1% |
| Luật Excel chưa kiểm tra đầy đủ công thức hoặc tham chiếu      | 37       | 18,1% |
| AGS chưa công nhận công thức hoặc tham chiếu Excel tương đương | 27       | 13,2% |
| Luật PowerPoint quá rộng hoặc nhận nhầm đối tượng              | 26       | 12,7% |
| Luật Word quá rộng hoặc chưa kiểm tra đúng kỹ thuật            | 22       | 10,8% |
| Thiếu kết quả AGS do lỗi tệp, parser hoặc workflow             | 9        | 4,4%  |
| Thiếu điểm giảng viên                                          | 1        | 0,5%  |

Có 109/204 trường hợp (53,4%) được chẩn đoán theo hướng AGS chưa nhận diện đầy đủ cách thực hiện hợp lệ, gồm định dạng Word kế thừa từ Style, bố cục PowerPoint sử dụng Slide Master hoặc grouped shape và công thức Excel tương đương nhưng khác biểu diễn trong rubric.

Có 85/204 trường hợp (41,7%) được chẩn đoán theo hướng luật chấm quá rộng hoặc chưa kiểm tra đầy đủ kỹ thuật, chẳng hạn chỉ kiểm tra sự tồn tại của đối tượng, nhận nhầm shape/placeholder hoặc xác nhận giá trị Excel mà chưa kiểm tra công thức và vùng tham chiếu. Mười trường hợp còn lại (4,9%) liên quan đến thiếu dữ liệu.

Kết quả này cho thấy sai lệch không xuất phát chủ yếu từ một lỗi đơn lẻ mà từ hai nhóm vấn đề đối nghịch:

- Luật quá chặt, làm bỏ sót cách làm hợp lệ;
- Luật quá rộng, làm công nhận bài chưa đáp ứng đầy đủ yêu cầu.

Phát hiện này có ý nghĩa đối với thiết kế rubric tự động. Rubric không nên chỉ định nghĩa một giá trị đúng duy nhất, mà cần hỗ trợ phương án tương đương, dung sai, điểm thành phần và quan hệ cấu trúc. Đồng thời, tiêu chí không nên chỉ kiểm tra sự tồn tại của đối tượng mà phải kiểm tra đúng đối tượng, đúng thuộc tính và đúng kỹ thuật thực hiện.

Kết quả nghiên cứu cung cấp bằng chứng liên quan đến ba khoảng trống đã xác định ở phần Mở đầu.

Thứ nhất, hệ thống triển khai một kiến trúc thống nhất cho Word, Excel và PowerPoint thông qua lớp JSON AST và một Rule Engine chung. Cách tổ chức này tách biệt quá trình phân tích tài liệu khỏi logic đánh giá và mở rộng phạm vi xử lý so với các nghiên cứu tập trung riêng vào bảng tính hoặc một số định dạng Office cụ thể (Hekman, 2019; Iwata & Matsui, 2022; Kovačić & Green, 2012).

Thứ hai, Anchor Locator được thiết kế để giảm sự phụ thuộc vào vị trí cố định của đối tượng và hỗ trợ cảnh báo khi kết quả định vị không chắc chắn. Quá trình rà soát ghi nhận một số trường hợp hệ thống không tìm thấy hoặc lựa chọn đối tượng chưa phù hợp; tuy nhiên, các trường hợp này không hình thành một mẫu đánh giá độc lập ở cấp đối tượng. Vì vậy,

### 3.5. Phân tích nguyên nhân sai lệch

Các trường hợp cần xử lý được phân loại theo nguyên nhân chẩn đoán chính sau khi một giảng viên rà soát tệp gốc, rubric và log hệ thống.

kết quả chỉ có ý nghĩa chẩn đoán dạng lỗi và chưa cho phép ước lượng precision, recall hoặc tỷ lệ định vị đúng của Anchor Locator trên toàn bộ dữ liệu.

Thứ ba, nghiên cứu không chỉ báo cáo điểm trung bình mà còn kết hợp MAE, độ chệch, giới hạn đồng thuận Bland–Altman và phân tích các trường hợp sai lệch. Điểm số được quyết định bởi Rubric JSON và Rule Engine, trong khi mô hình ngôn ngữ lớn chỉ hỗ trợ phản hồi. Quy trình rà soát hậu nghiệm cung cấp dữ liệu để cải tiến parser, Anchor Locator, rubric và Rule Engine; tuy nhiên, cơ chế phân luồng độc lập vẫn cần được đánh giá trong nghiên cứu tiếp theo.

Tỷ lệ 84,4% lượt chấm nằm trong ngưỡng  $\pm 1$  điểm và độ chệch chung gần 0 cho thấy kiến trúc có tính khả thi trong bối cảnh khảo sát. Tuy nhiên, giới hạn đồng thuận còn rộng và 194 lượt vượt ngưỡng cho thấy AGS chưa thể thay thế hoàn toàn giảng viên. Giá trị hiện tại của hệ thống nằm ở hỗ trợ chuẩn hóa, tự động hóa các tiêu chí rõ ràng và cung cấp tín hiệu cho việc rà soát trường hợp ngoại lệ.

### 3.6. Hạn chế và hướng cải tiến

Nghiên cứu vẫn còn một số hạn chế cần được xem xét khi diễn giải kết quả. Trước hết, dữ liệu thực nghiệm được thu thập trong phạm vi một học phần, tại một cơ sở đào tạo và trên một hệ thống đề kiểm tra cụ thể. Do đó, kết quả hiện tại mới phản ánh hiệu quả của AGS trong bối cảnh triển khai đã khảo sát và chưa đủ cơ sở để khái quát cho các học phần, nhóm người học hoặc cơ sở đào tạo khác.

Thứ hai, điểm tham chiếu được sử dụng để đối sánh với AGS chỉ dựa trên kết quả chấm của một giảng viên. Mặc dù giảng viên đã sử dụng rubric thống nhất và thực hiện rà soát lại các bài có sai lệch lớn, thiết kế này chưa cho phép đánh giá độ tin cậy liên giám khảo và chưa loại trừ hoàn toàn ảnh hưởng của khác biệt cá nhân trong cách diễn giải tiêu chí, cho điểm thành phần và xử lý các trường hợp ngoại lệ. Vì vậy, điểm của giảng viên trong nghiên cứu nên được xem là điểm tham chiếu thực hành, chưa phải là chuẩn vàng được thiết lập từ sự đồng thuận của nhiều chuyên gia.

Thứ ba, ngưỡng sai lệch tuyệt đối 1,0 điểm trên thang điểm 10 được sử dụng để phân luồng bài cần rà

soát là ngưỡng thực nghiệm do nhóm nghiên cứu lựa chọn. Ngưỡng này có ý nghĩa quản lý và kiểm soát chất lượng trong bối cảnh triển khai hiện tại, nhưng chưa được xác nhận là ngưỡng tối ưu cho mọi loại bài hoặc mọi mục tiêu đánh giá.

Thứ tư, phân tích chủ yếu được thực hiện ở cấp độ điểm tổng. Cách tiếp cận này cho phép đánh giá mức độ phù hợp chung giữa AGS và giảng viên nhưng chưa phản ánh đầy đủ khả năng nhận diện đúng hoặc sai của từng tiêu chí. Một bài có điểm tổng gần nhau vẫn có thể tồn tại sai lệch bù trừ giữa các tiêu chí, trong khi một sai lệch điểm tổng lớn có thể chỉ phát sinh từ một tiêu chí có trọng số cao.

Thứ năm, một số thành phần phức tạp của tài liệu Office chưa được hệ thống xử lý đầy đủ, như style kẻ thừa, grouped shape, Slide Master, công thức động, các cách tham chiếu tương đương và những phương án trình bày khác nhưng vẫn đáp ứng yêu cầu. Ngoài ra, các tiêu chí mang tính thẩm mỹ, sáng tạo hoặc phụ thuộc vào phán đoán chuyên môn vẫn chưa phù hợp để tự động hóa hoàn toàn bằng Rule Engine.

Thứ sáu, nghiên cứu chưa đánh giá độc lập hiệu quả của Anchor Locator ở cấp đối tượng. Dữ liệu hiện tại lưu kết quả chấm, đối tượng được lựa chọn và cảnh báo thực thi nhưng chưa có tập nhãn chuẩn xác định đối tượng đúng cho từng tiêu chí. Do đó, nghiên cứu chưa tính được precision, recall, F1-score hoặc tỷ lệ định vị đúng riêng của cơ chế này. Nghiên cứu tiếp theo cần xây dựng tập dữ liệu được gán nhãn thủ công, bao gồm cả trường hợp thông thường, trường hợp có nhiều ứng viên và trường hợp không tồn tại đối tượng hợp lệ, đồng thời báo cáo kết quả riêng cho Word, Excel và PowerPoint.

Trong các nghiên cứu tiếp theo, cần mở rộng dữ liệu theo nhiều lớp học, nhiều đề kiểm tra, nhiều khóa học và nhiều cơ sở đào tạo để đánh giá khả năng khái quát của hệ thống. Đặc biệt, nên tổ chức ít nhất hai hoặc ba giảng viên chấm độc lập trên cùng một tập bài và sử dụng quy trình hội ý để xây dựng điểm đồng thuận. Khi đó, có thể tính hệ số tương quan nội lớp, hệ số kappa hoặc các chỉ số phù hợp khác nhằm đánh giá độ tin cậy liên giám khảo và thiết lập chuẩn tham chiếu vững chắc hơn cho AGS.

Bên cạnh đó, cần mở rộng phân tích xuống cấp độ từng tiêu chí và xây dựng tập nhãn chuẩn ở cấp đối tượng để đánh giá độc lập Anchor Locator bằng precision, recall, F1-score, tỷ lệ định vị đúng, tỷ lệ bỏ sót và tỷ lệ nhận nhầm. Dữ liệu từ các bài được giảng viên rà soát cũng nên được sử dụng để xây dựng bộ

kiểm thử hồi quy cho parser, rubric và Rule Engine.

Hệ thống cũng cần được hoàn thiện theo hướng hỗ trợ nhiều phương án hợp lệ, công thức tương đương, dung sai thích ứng và quan hệ cấu trúc phức tạp trong OOXML. Đồng thời, nên phát triển chỉ số độ tin cậy của từng tiêu chí và toàn bài để phân luồng dựa trên nhiều tín hiệu, thay vì chỉ dựa vào sai lệch điểm sau khi đã có điểm của giảng viên.

Cuối cùng, cần đánh giá AGS trong điều kiện vận hành thực tế thông qua tích hợp với hệ thống quản lý học tập, đo thời gian xử lý, khối lượng công việc giảng viên được giảm, chất lượng phản hồi và mức độ chấp nhận của người sử dụng. Những mở rộng này sẽ giúp xác định rõ hơn giá trị thực tiễn, độ tin cậy và khả năng triển khai AGS ở quy mô lớn.

#### 4. KẾT LUẬN

Nghiên cứu đã xây dựng và đánh giá hệ thống AGS cho các bài thực hành Word, Excel và PowerPoint dựa trên Office Open XML, JSON AST, Rubric JSON và Rule Engine. Kiến trúc module giúp tách biệt quá trình phân tích tài liệu, định vị đối tượng và thực thi luật chấm; mô hình ngôn ngữ lớn chỉ hỗ trợ tạo phản hồi và không trực tiếp quyết định điểm.

Trên 1.247 cặp điểm hợp lệ, MAE toàn hệ thống đạt 0,618 điểm, 84,4% lượt chấm nằm trong ngưỡng sai lệch  $\pm 1$  điểm và độ chệch chung là -0,051 điểm. Kết quả cho thấy hệ thống không có xu hướng chệch lệch rõ rệt theo một chiều ở mức trung bình; tuy nhiên, giới hạn đồng thuận còn rộng và 194 lượt vượt ngưỡng cho thấy AGS chưa thể thay thế hoàn toàn đánh giá của giảng viên.

Đóng góp chính của nghiên cứu nằm ở việc triển khai một kiến trúc chung cho ba định dạng Office, sử dụng lớp biểu diễn trung gian và luật chấm có khả năng truy vết. Anchor Locator hỗ trợ định vị đối tượng và phát hiện các trường hợp không chắc chắn, nhưng hiệu quả riêng của cơ chế này chưa được đánh giá bằng tập nhãn độc lập. Vì vậy, kết quả hiện tại được xem là bằng chứng về tính khả thi trong bối cảnh khảo sát, không phải bằng chứng xác nhận khả năng khái quát hoặc thay thế hoàn toàn việc chấm của con người.

Nghiên cứu tiếp theo cần mở rộng dữ liệu theo nhiều giảng viên, lớp học và cơ sở đào tạo; xây dựng tập nhãn để đánh giá precision, recall và F1-score của Anchor Locator; kiểm tra kết quả ở cấp tiêu chí; đồng thời đánh giá hiệu năng, chi phí và mức giảm khối lượng công việc trong điều kiện vận hành thực tế.

## TÀI LIỆU THAM KHẢO

- Ecma International. (2021). *Office Open XML file formats* (5th ed.; Standard ECMA-376). <https://ecma-international.org/publications-and-standards/standards/ecma-376/>
- Gao, R., Merzdorf, H. E., Anwar, S., Hipwell, M. C., & Srinivasa, A. R. (2024). Automatic assessment of text-based responses in post-secondary education: A systematic review. *Computers and Education: Artificial Intelligence*, 6, Article 100206. <https://doi.org/10.1016/j.caeai.2024.100206>
- Hekman, K. (2019, June). Automated grading of Microsoft Excel spreadsheets. In *2019 ASEE Annual Conference & Exposition*. American Society for Engineering Education. <https://doi.org/10.18260/1-2--32135>
- Hunnicutt, S. S. (2016). An automated grading spreadsheet for reports or presentations. *Journal of Chemical Education*, 93(1), 210–212. <https://doi.org/10.1021/acs.jchemed.5b00553>
- International Organization for Standardization. (2021). *Document description and processing languages—Office Open XML file formats—Part 2: Open packaging conventions* (ISO/IEC Standard No. 29500-2:2021). <https://www.iso.org/standard/77818.html>
- Iwata, K., & Matsui, Y. (2022). Implementation of an automated grading system for Microsoft Excel spreadsheets and Word documents. In T. Matsuo (Ed.), *Proceedings of the 11th International Congress on Advanced Applied Informatics* (EPIc Series in Computing, Vol. 81, pp. 289–302). EasyChair. <https://doi.org/10.29007/1zs6>
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130–144. <https://doi.org/10.1016/j.edurev.2007.05.002>
- Kovačić, Z. J., & Green, J. S. (2012). Automatic grading of spreadsheet and database skills. *Journal of Information Technology Education: Innovations in Practice*, 11, 53–70. <https://www.jite.org/documents/Vol11/JITEv11IIPp053-070Kovacic1051.pdf>
- Messer, M., Brown, N. C. C., Kölling, M., & Shi, M. (2024). Automated grading and feedback tools for programming education: A systematic review. *ACM Transactions on Computing Education*, 24(1), Article 10, 1–43. <https://doi.org/10.1145/3636515>
- Padó, U., Eryilmaz, Y., & Kirschner, L. (2024). Short-answer grading for German: Addressing the challenges. *International Journal of Artificial Intelligence in Education*, 34(4), 1321–1352. <https://doi.org/10.1007/s40593-023-00383-w>
- Panadero, E. (2020). A critical review of the arguments against the use of rubrics. *Educational Research Review*, 30, Article 100329. <https://doi.org/10.1016/j.edurev.2020.100329>
- Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, 9, 129–144. <https://doi.org/10.1016/j.edurev.2013.01.002>
- Pecuchova, J., Benko, L., & Drlik, M. (2025). Automated grading of open-ended questions in higher education using GenAI models. *International Journal of Artificial Intelligence in Education*, 35(6), 3813–3846. <https://doi.org/10.1007/s40593-025-00517-2>
- Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, 15(1), 18–39. <https://doi.org/10.1016/j.asw.2010.01.003>
- Sembey, R., Hoda, R., & Grundy, J. (2024). Emerging technologies in higher education assessment and feedback practices: A systematic literature review. *Journal of Systems and Software*, 211, Article 111988. <https://doi.org/10.1016/j.jss.2024.111988>

# DEVELOPMENT OF AN AUTOMATED GRADING SYSTEM BASED ON OFFICE OPEN XML, RUBRIC JSONS, AND AI-ASSISTED FEEDBACK FOR A GENERAL INFORMATICS COURSE

Phan Thi Dai Trang<sup>1</sup>, Nguyen Thi Nhu<sup>1</sup>, Tong Thi Lan Chi<sup>2</sup>

Received Date: 24/06/2026; Revised Date: 07/08/2026; Accepted for Publication: 07/08/2026

## ABSTRACT

This study presents the development and evaluation of an Automated Grading System (AGS) for Microsoft Word, Excel, and PowerPoint practical assignments in a General Informatics course. The system analyzes the Office Open XML structures of .docx, .xlsx, and .pptx files, normalizes the extracted data into a JSON-based abstract syntax tree (JSON AST), locates assessment objects using an Anchor Locator, and executes criteria represented as JSON Rubrics through a Rule Engine. The workflow is orchestrated on the n8n platform, while a large language model is used only to interpret results and generate feedback rather than to determine scores

<sup>1</sup>Faculty of Natural Sciences and Technology, Tay Nguyen University;

<sup>2</sup>Department of Science, Partnership and Innovation, Tay Nguyen University;

Corresponding author: Nguyen Thi Nhu; Email: [ntnhu@ttn.edu.vn](mailto:ntnhu@ttn.edu.vn).

directly. The initial dataset comprised 1,374 submissions. After invalid and duplicate files were removed, 1,257 submissions were processed, of which 1,247 complete instructor–AGS score pairs were included in the analysis. The overall mean absolute error was 0.618 points on a 10-point scale; 84.4% of the automated scores differed from the instructor’s scores by no more than one point, and the mean bias was  $-0.051$  points. Excel produced the lowest absolute discrepancy. Analysis of 204 cases requiring review showed that discrepancies were mainly associated with the system’s failure to recognize valid alternative implementations or with grading rules that did not fully verify the required techniques. The findings demonstrate the feasibility and traceability of a unified grading architecture while also highlighting the continued importance of instructor oversight for exceptional cases.

**Keywords:** *Automated grading system; Office Open XML; JSON AST; General Informatics.*