Tạp chí Khoa học Đại học Cần Thơ
Phần C: Khoa học Xã hội, Nhân văn và Giáo dục
ISSN 1859-2333 | e-ISSN 2815-5599

DOI:10.22144/ctujos.2026.239

ĐÁNH GIÁ KHẢ NĂNG XỬ LÝ PHẦN ĐỌC HIỂU TIẾNG NHẬT TRÌNH ĐỘ N3 CỦA CHATGPT QUA CÁC TÀI LIỆU LUYỆN THI JLPT

Nguyễn Thị Phi*, Nguyễn Minh Thọ, Tô Lê Đức Anh và Bùi Nguyễn Thị Thanh Nhàn
Bộ môn tiếng Nhật, Trường Đại học FPT Cần Thơ, Việt Nam
**Tác giả liên hệ (Corresponding author): phint10@fe.edu.vn*

Thông tin chung (Article Information)

Nhận bài (Received): 19/12/2025
Sửa bài (Revised): 03/02/2026
Duyệt đăng (Accepted): 23/08/2026

Title: *Evaluating ChatGPT's ability to solve N3-Level Japanese reading comprehension tasks using JLPT test preparation materials*

Author(s): *Nguyen Thi Phi, Nguyen Minh Tho, To Le Duc Anh and Bui Nguyen Thi Thanh Nhan*

Affiliation(s): *Department of Japanese Language, FPT University Can Tho, Viet Nam*

TÓM TẮT

Nghiên cứu này được thực hiênh nhằm tập trung đánh giá khả năng xử lý phần đọc hiểu tiếng Nhật N3 của ChatGPT thông qua 9 tài liệu luyện thi và được lặp lại ba lần. Phương pháp nghiên cứu kết hợp định lượng và định tính. Kết quả định lượng cho thấy ChatGPT đạt tỷ lệ trả lời đúng trung bình trên 60% ở cả 9 tài liệu, trong đó ba tài liệu trên 80% và cao nhất là 89,9%. Kiểm định thống kê cho thấy không có sự khác biệt có ý nghĩa giữa ba lần chạy, phản ánh tính ổn định và độ tin cậy của mô hình. Về mặt định tính, 71,8% lời giải được đánh giá là "đúng đầy đủ ý chính", 27,3% "đúng một phần ý chính" và chỉ 0,9% "không đúng ý chính". Các lỗi chủ yếu là hiểu sai ý hoặc bỏ sót thông tin, cho thấy hạn chế của ChatGPT trong việc xử lý văn bản dài và phức tạp. Kết quả nghiên cứu giúp khẳng định tiềm năng ứng dụng trí tuệ nhân tạo trong việc học tiếng Nhật, đồng thời gợi ý cách sử dụng công cụ này một cách kiểm soát và hiệu quả.

Từ khóa: *ChatGPT, đọc hiểu tiếng Nhật, JLPT, trí tuệ nhân tạo*

ABSTRACT

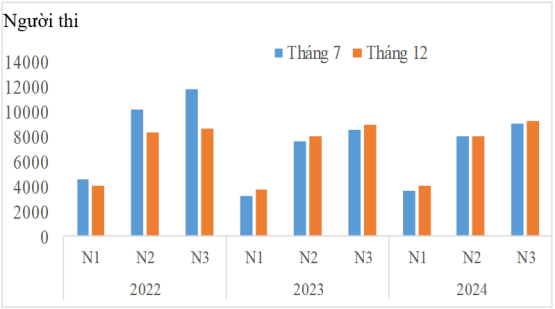
This study evaluates ChatGPT's ability to handle Japanese N3 reading comprehension through a dataset of nine JLPT practice materials, each repeated three times to ensure objectivity. A mixed-method approach combining quantitative and qualitative analysis was employed. Quantitatively, ChatGPT achieved an accuracy rate of over 60% across all materials, with three exceeding 80% and the highest reaching 89.9%. Statistical testing indicated no significant differences across the three runs, reflecting the model's stability and reliability. Qualitatively, 71.8% of ChatGPT's responses were rated as "fully correct" in terms of accurately conveying the key explanations from the book, 27.3% were rated as "partially correct", and only 0.9% were deemed "incorrect" when compared to the explanations provided in the book. Errors mainly stemmed from misinterpretation or omission of information, highlighting ChatGPT's limitations in processing long and complex texts. The findings provide empirical evidence of the potential application of artificial intelligence in Japanese language learning, while also suggesting the need for controlled, effective, and context-appropriate use of such tools in modern education.

Keywords: *Artificial intelligence, ChatGPT, JLPT, Japanese reading comprehension*

1. GIỚI THIỆU

Trong những năm gần đây, sự phát triển nhanh chóng của trí tuệ nhân tạo (AI), đặc biệt là các mô hình ngôn ngữ lớn (LLMs), đã mở ra nhiều cơ hội mới cho lĩnh vực giáo dục. Các công cụ tạo sinh như ChatGPT của OpenAI được đánh giá cao nhờ khả năng xử lý ngôn ngữ linh hoạt và tự nhiên. Công cụ này có thể hỗ trợ nhiều nhiệm vụ trong giáo dục, bao gồm giải thích học thuật, xây dựng bài giảng và cá nhân hóa việc học (Quân, 2023; Anh & Ghi, 2025). Các nghiên cứu trong và ngoài nước đã chứng minh ChatGPT có tiềm năng ứng dụng trong giảng dạy ngoại ngữ, bao gồm cung cấp nghĩa từ vựng, giải thích ngữ pháp và hỗ trợ luyện tập kỹ năng ngôn ngữ cơ bản. Đặc biệt, nghiên cứu gần đây đã chỉ ra khả năng đáng chú ý của ChatGPT trong các kỳ thi chuẩn hóa: ChatGPT đã vượt qua kỳ thi cấp quốc gia ngành dược tại Nhật Bản với tỷ lệ trả lời đúng 80% (Sato & Ogasawara, 2024). Tại Việt Nam, một số nghiên cứu đã khám phá tiềm năng của ChatGPT trong giáo dục phổ thông và đại học. Nghiên cứu của Vinh và ctv. (2023) chỉ ra rằng ChatGPT có thể thực hiện các nhiệm vụ trong bài kiểm tra môn Toán và Ngữ văn cấp trung học phổ thông, đồng thời nghiên cứu của Nga (2024) cũng đã khẳng định vai trò hỗ trợ của ChatGPT trong xây dựng bài giảng, thiết kế học liệu và cá nhân hóa quá trình học tập.

Bên cạnh đó, Việt Nam xếp thứ 3 trong số các quốc gia có số lượng người học tiếng Nhật cao nhất tại Đông Nam Á, với 169.582 người vào năm 2021, chỉ sau Indonesia và Thái Lan (Japan Foundation, 2021). Cụ thể, tính theo cơ cấu người học, có tới 45.752 người học tiếng Nhật ở chương trình THPT, con số này cao nhất trong khu vực; ngoài ra, có hơn 89.000 người học ngoài hệ thống giáo dục chính quy như trung tâm ngoại ngữ, trường nghề hoặc chương trình đào tạo thực hành. Trong đó, JLPT (Japanese-Language Proficiency Test) là kỳ thi chuẩn hóa quốc tế, được công nhận rộng rãi trên toàn cầu để đánh giá trình độ tiếng Nhật, bao gồm bốn kỹ năng: từ vựng, ngữ pháp, đọc hiểu và nghe hiểu. Được tổ chức dưới dạng trắc nghiệm, kỳ thi này đã chứng minh vai trò trung tâm trong việc đánh giá năng lực tiếng Nhật toàn cầu thông qua sự gia tăng số lượng thí sinh, sự đa dạng về độ tuổi và mục đích tham gia. Với năm cấp độ từ N5 (cơ bản) đến N1 (nâng cao), JLPT không chỉ là công cụ đánh giá ngôn ngữ mà còn trở thành chuẩn mực quốc tế trong giáo dục và phát triển nghề nghiệp. Tại Việt Nam, số lượng thí sinh đăng ký thi JLPT liên tục tăng qua các năm, với số lượng cao nhất thuộc về kỳ thi JLPT cấp độ N3 (Hình 1).



Hình 1. Số lượng thí sinh đăng ký thi JLPT tại Việt Nam năm 2022 – 2024

Nguồn: Japanese Language Proficiency Test (2024)

Tuy nhiên, kỹ năng đọc hiểu tiếng Nhật vẫn là một thách thức đối với người học. Nghiên cứu của Dirgandini et al. (2025) cho thấy sinh viên Indonesia ở trình độ tiếng Nhật sơ – trung cấp gặp khó khăn trong việc hiểu văn bản do hạn chế về từ vựng, tốc độ đọc chậm và thiếu kỹ năng suy luận. Yang et al. (2023) nhấn mạnh rằng đọc hiểu hiệu quả không chỉ đòi hỏi kiến thức ngôn ngữ cơ bản, mà còn cần các kỹ năng nhận thức cao cấp như suy luận và phân tích. Kintsch (1998) cũng cho rằng năng lực đọc hiểu không chỉ liên quan đến từ vựng và cú pháp mà còn đòi hỏi xây dựng mô hình ngữ nghĩa để suy luận và tích hợp thông tin. Nishizawa et al. (2022) cũng chỉ ra rằng phần đọc hiểu trong kỳ thi JLPT, đặc biệt từ cấp độ N3 trở lên, yêu cầu người học tổng hợp, phân tích và diễn giải thông tin trong ngữ cảnh đa dạng. Những yếu tố trên cho thấy khó khăn trong đọc hiểu tiếng Nhật không chỉ xuất phát từ rào cản ngôn ngữ. Người học còn cần kỹ năng xử lý thông tin và tư duy chiến lược. Đồng thời, trong xử lý ngôn ngữ tự nhiên, đọc hiểu cũng là công cụ đánh giá khả năng hiểu ngôn ngữ của AI (Rajpurkar et al., 2016).

Vì vậy, việc lựa chọn JLPT N3 trong nghiên cứu này dựa trên tính phổ biến và mức độ yêu cầu đọc hiểu đặc thù của cấp độ này. N3 là cấp độ có số lượng thí sinh đăng ký cao nhất tại Việt Nam, đồng thời phần đọc hiểu từ cấp độ này trở lên đòi hỏi người học phải tổng hợp và suy luận thông tin trong ngữ cảnh, thay vì chỉ nhận diện kiến thức ngôn ngữ cơ bản như ở các cấp độ thấp hơn. Do đó, N3 được xem là cấp độ phù hợp để đánh giá năng lực đọc hiểu trong bối cảnh học tập thực tế cũng như khả năng xử lý ngôn ngữ của ChatGPT.

Bên cạnh đó, mặc dù nhiều nghiên cứu đã ghi nhận tiềm năng của ChatGPT trong hỗ trợ học tập và xử lý ngôn ngữ, nhưng các công trình hiện nay chủ yếu đánh giá hiệu suất tổng quát, chưa phân tích một cách hệ thống khả năng đọc hiểu trong bối cảnh

kỳ thi chuẩn hóa tiếng Nhật như JLPT. Đặc biệt, vẫn thiếu các nghiên cứu kết hợp đánh giá định lượng và định tính về chất lượng lập luận, độ ổn định và loại hình lỗi của ChatGPT ở trình độ N3, cũng như bằng chứng thực nghiệm trong bối cảnh người học tại Việt Nam.

Xuất phát từ thực tế đó, việc khảo sát hiệu quả của ChatGPT trong xử lý các bài đọc hiểu tiếng Nhật trình độ N3 không chỉ giúp đánh giá năng lực của công cụ trong lĩnh vực ngôn ngữ học ứng dụng, mà còn góp phần xem xét khả năng tích hợp AI vào giảng dạy và hỗ trợ người học tại Việt Nam một cách thiết thực. Vì vậy, nghiên cứu này được thực hiện với mục đích đánh giá năng lực đọc hiểu tiếng Nhật của ChatGPT trong bối cảnh kỳ thi JLPT N3, thông qua bộ dữ liệu gồm các câu hỏi đọc hiểu kèm đáp án và phần giải thích chính thức từ tài liệu luyện thi. Các mục tiêu cụ thể gồm:

– Đánh giá độ chính xác của câu trả lời: Phương án lựa chọn của ChatGPT được so sánh với đáp án chính thức nhằm xác định tỷ lệ đúng sai và mức độ ổn định.

– Phân tích chất lượng phần giải thích: Cách lý giải của ChatGPT được đối chiếu với phần giải thích trong sách, từ đó xem xét khả năng suy luận theo ngữ cảnh, ngữ nghĩa và ngữ pháp.

– Khảo sát đặc điểm phản hồi của ChatGPT: Xu hướng, điểm mạnh và hạn chế trong quá trình xử lý bài đọc được ghi nhận, từ đó chiến lược sử dụng hiệu quả trong dạy – học ngoại ngữ đã được đề xuất.

Trên cơ sở đó, nghiên cứu hướng đến mục tiêu tổng thể là đánh giá mức độ phù hợp, độ tin cậy và tiềm năng ứng dụng của ChatGPT trong hỗ trợ người học tiếng Nhật, đặc biệt trong giai đoạn ôn luyện kỹ năng đọc hiểu ở trình độ N3. Đồng thời, kết quả của nghiên cứu cũng gợi mở các khuyến nghị về việc tích hợp có kiểm soát các công cụ AI vào chương trình giảng dạy ngoại ngữ hiện đại.

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Đối tượng và phạm vi nghiên cứu

Trong nghiên cứu này, tác giả sử dụng ChatGPT Plus dựa trên mô hình GPT-4.0 của OpenAI trong suốt quá trình thu thập và phân tích dữ liệu. ChatGPT Plus là phiên bản trả phí của ChatGPT, cho phép truy cập các mô hình ngôn ngữ lớn tiên tiến hơn so với phiên bản miễn phí; trong đó, GPT-4.0 được thiết kế nhằm nâng cao khả năng hiểu ngữ cảnh, suy luận và xử lý ngôn ngữ đa ngữ (OpenAI, 2025).

Đối tượng nghiên cứu là các câu trả lời do ChatGPT Plus (phiên bản GPT-4.0 trong giai đoạn nghiên cứu) tạo ra khi xử lý bộ câu hỏi đọc hiểu của kỳ thi năng lực tiếng Nhật JLPT cấp độ N3. Các đáp án và phần giải thích trong sách luyện thi JLPT N3 được sử dụng làm chuẩn đối chiếu nhằm đánh giá mức độ chính xác và chất lượng phản hồi của mô hình.

Nghiên cứu được tiến hành trong khoảng thời gian từ ngày 12/5/2025 đến ngày 10/8/2025. Toàn bộ dữ liệu đọc hiểu được trích từ 9 đầu sách luyện thi JLPT N3 phổ biến trên thị trường (Bảng 1). Mỗi đầu sách được gán mã quyển 1-9 và được sử dụng thống nhất trong toàn bộ nghiên cứu nhằm đảm bảo tính nhất quán và ngắn gọn trong phân tích.

2.2. Phương pháp thu thập dữ liệu

Trong quá trình nghiên cứu, câu lệnh (prompt) được thiết kế bám sát nguyên văn yêu cầu của từng dạng bài đọc hiểu trong chín tài liệu luyện thi JLPT N3. Do mỗi dạng bài có chỉ dẫn khác nhau, prompt được điều chỉnh tương ứng nhằm giữ nguyên mục tiêu và cấu trúc của đề bài gốc. Đối với mỗi câu hỏi, ChatGPT được yêu cầu đọc đoạn văn và lựa chọn đáp án phù hợp theo đúng yêu cầu của đề, không bổ sung hướng dẫn phân tích hay gợi ý suy luận. Cách tiếp cận này nhằm mô phỏng điều kiện làm bài thực tế và hạn chế ảnh hưởng của kỹ thuật thiết kế prompt đến kết quả.

Nguồn dữ liệu chính bao gồm: 9 đầu sách luyện thi JLPT N3 có phần đáp án và lời giải thích chi tiết. Công cụ được sử dụng để xử lý và tạo dữ liệu là ChatGPT Plus và việc phân tích số liệu được thực hiện bằng phần mềm IBM SPSS Statistics 25. Trong suốt quá trình nghiên cứu, toàn bộ các câu hỏi đọc hiểu từ từng đầu sách được nhập nguyên văn vào ChatGPT Plus trong ba lần độc lập, bảo đảm giữ nguyên cấu trúc và yêu cầu của đề bài gốc. Kết quả đầu ra gồm đáp án và lời giải thích của ChatGPT được ghi nhận để phân tích bằng cách đối chiếu với đáp án và lời giải thích trong 9 tài liệu sách luyện thi nêu trên.

2.3. Phương pháp phân tích dữ liệu

2.3.1. Đánh giá định lượng

Câu trả lời của ChatGPT được tổng hợp và tính tỷ lệ chính xác theo công thức:

Tỷ lệ chính xác (%) = (Số câu trả lời đúng/ Tổng số câu hỏi) x 100

Việc kiểm định thống kê được thực hiện trên từng đầu sách bằng phần mềm IBM SPSS Statistics 25, sử dụng kiểm định phi tham số Friedman Test

để xác định sự khác biệt có ý nghĩa thống kê giữa ba lần xử lý dữ liệu.

2.3.2. Đánh giá định tính

Dữ liệu đầu ra được phân thành hai nhóm. Nhóm 1 gồm các trường hợp câu trả lời của ChatGPT trong cả ba lần đều trùng khớp với đáp án trong sách. Với nhóm này, việc đánh giá "mức độ đúng ý chính" được thực hiện dựa trên mức độ tương đồng về nội dung cốt lõi giữa phần giải thích do ChatGPT tạo ra và lời giải chính thức trong các tài liệu luyện thi JLPT N3. Cụ thể, ba mức đánh giá được xác định như sau: (i) mức 0 – "không đúng ý chính", khi lời giải của ChatGPT diễn giải sai nội dung, bỏ sót ý chính cốt lõi hoặc không liên quan đến lập luận trong lời giải chuẩn; (ii) mức 1 – "đúng một phần ý chính", khi lời giải chỉ bao quát một số ý quan trọng nhưng thiếu hoặc chưa làm rõ các thông tin then chốt có ảnh hưởng đến việc lựa chọn đáp án và (iii) mức 2 – "đúng đầy đủ ý chính", khi lời giải của ChatGPT phản ánh đầy đủ ý nghĩa trọng tâm và lập luận chính của lời giải trong sách, dù cách diễn đạt có thể khác. Việc đánh giá được thực hiện thủ công bởi nhóm nghiên cứu trên cơ sở đối chiếu trực tiếp với đáp án và lời giải trong sách.

Bảng 1. Danh sách 9 tài liệu luyện thi JLPT N3

Quyển	Tên sách	Số câu hỏi đọc hiểu	Nhà xuất bản	Năm xuất bản
1	ぜったい合格！日本語能力試験完全模試N3 (Zettai Goukaku! Nihongo Nouryoku Shiken Kanzen Moshi N3)	48	Jresearch	2013
2	新完全マスター読解N3 (Shin Kanzen Master Dokkai N3)	146	3A Corporation	2015
3	日本語能力試験模試と対策N3 (Nihongo Nouryoku Shiken Moshi to Taisaku N3)	32	Ask	2010
4	日本語能力試験模試と対策N3 Vol 2 (Nihongo Nouryoku Shiken Moshi to Taisaku N3 Vol. 2)	32	Ask	2012
5	ドリル＆ドリル聴解と読解 N3 (Drill & Drill Choukai to Dokkai N3)	50	Unicom Inc.	2014
6	日本語能力試験これ一冊 N3 (Nihongo Nouryoku Shiken Kore Issatsu N3)	79	Ask	2023
7	日本語能力試験対策問題＆要点整理N3 (Nihongo Nouryoku Shiken Taisaku Mondai & Youten Seiri N3)	82	JMA Management Center Inc.	2010
8	パターン別徹底ドリルN3 (Pattern-betsu Tettei Drill N3)	19	Alc	2010
9	日本語総まとめ読解N3 (Nihongo Sou Matome Dokkai N3)	97	NXB Trẻ	2018

Nhóm 2 gồm các trường hợp có ít nhất một câu trả lời sai so với đáp án trong sách; đối với nhóm này, tiến hành phân loại lỗi, ghi nhận tần suất cũng như quan sát đặc điểm phản hồi của ChatGPT. Các lỗi sai của ChatGPT được phân thành năm nhóm: (1) Lỗi không hiểu đúng yêu cầu: đề là trường hợp mô hình xác định sai mục tiêu câu hỏi. (2) Lỗi hiểu sai ý: phản ánh việc diễn giải nội dung văn bản không chính xác. (3) Lỗi bỏ sót ý: xảy ra khi mô hình chỉ xử lý một phần thông tin và bỏ qua chi tiết then chốt. (4) Lỗi suy diễn là việc đưa ra kết luận vượt quá phạm vi thông tin được cung cấp trong văn bản. (5) Lỗi lựa chọn đáp án là trường hợp mô hình hiểu đúng nội dung nhưng lựa chọn sai phương án trả lời.

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Phân tích kết quả định lượng

Kết quả cho thấy tất cả 9 tài liệu đều đạt tỷ lệ trả lời đúng trung bình trên 60%, trong đó quyển 2, quyển 6 và quyển 9 vượt 80%. Kết quả này phản ánh mức độ chính xác tương đối cao của ChatGPT trong xử lý các bài đọc hiểu trình độ N3. Mô hình cho thấy khả năng tiếp cận văn bản, nhận diện thông tin trọng

yếu và lựa chọn đáp án phù hợp (Bảng 2). Kết quả này tương đồng với nghiên cứu của Taira et al. (2023), khi công cụ này đạt tỷ lệ đúng trung bình 75,1% đối với câu hỏi kiến thức cơ bản và 64,5% đối với câu hỏi tổng quát trong kỳ thi cấp phép điều dưỡng quốc gia Nhật Bản – một bài thi chuẩn hóa, phức tạp và sử dụng tiếng Nhật bản địa.

Đáng chú ý, quyển 2 và quyển 6 đạt tỷ lệ đúng vượt trội trên 84% ở cả 3 lần, thể hiện không chỉ hiệu suất xử lý ngôn ngữ cao mà còn sự ổn định đầu ra của ChatGPT giữa các lần thử nghiệm. Ngược lại, kết quả ở quyển 3 cho thấy sự dao động rõ rệt, với tỷ lệ đúng ở lần thứ 2 chỉ đạt 53,3% – thấp nhất trong toàn bộ tập hợp khảo sát. Phần lớn lỗi ở quyển sách này tập trung ở dạng câu xác định nghĩa của từ, cụm từ hoặc câu được gạch chân. Đây có thể xem là dạng câu hỏi đòi hỏi khả năng duy trì mạch logic xuyên suốt và là những yếu tố gây thách thức đối với các mô hình LLMs như ChatGPT. Theo Biswas (2023), các lỗi trong đọc hiểu phức tạp có thể được phân loại thành lỗi ngữ nghĩa và lỗi ngữ cảnh. Những lỗi này bao gồm diễn giải sai từ vựng đa nghĩa, không nhận diện chính xác các biểu đạt quy chiếu hoặc mất mạch khi xử lý thông tin trải dài qua nhiều câu.

Bảng 2. Tỷ lệ chính xác của ChatGPT qua 3 lần lặp lại (mức ý nghĩa α = 0,05)

Quyển	L1 (%)	L2 (%)	L3 (%)	Trung bình (%)	p-value
1	81,3	83,3	75,0	79,9	0,27
2	87,0	85,6	84,3	85,6	0,69
3	65,6	53,3	65,6	61,5	0,41
4	71,9	71,9	65,6	69,8	0,74
5	78,0	68,0	62,0	69,3	0,06
6	84,8	92,4	92,4	89,9	0,06
7	78,1	73,2	81,7	77,6	0,16
8	79,0	68,4	73,7	73,7	0,69
9	78,4	80,4	84,5	81,1	0,28

Ghi chú: L1 là lần 1, L2 là lần 2 và L3 là lần 3

Về mặt thống kê, mức ý nghĩa α = 0,05 được sử dụng để đánh giá sự khác biệt giữa ba lần chạy của ChatGPT. Kết quả kiểm định cho thấy, mặc dù có sự chênh lệch nhất định giữa các lần lặp, không có sự khác biệt nào đạt ý nghĩa thống kê giữa ba lần thử nghiệm ở bất kỳ đầu sách nào (p > 0,05). Giá trị p thấp nhất được ghi nhận ở quyển 5 và quyển 6, với p = 0,06 (Bảng 4), tuy nhiên các giá trị này vẫn lớn hơn mức ý nghĩa đã chọn (α = 0,05). Điều này khẳng định ChatGPT thể hiện tính ổn định trong việc xử lý các câu hỏi đọc hiểu JLPT N3. Phát hiện này hoàn toàn phù hợp với kết luận từ Wan et al. (2025), rằng mặc dù LLMs như ChatGPT có tính ngẫu nhiên nội tại trong quá trình tạo sinh văn bản, các mô hình vẫn có thể đạt được độ ổn định đáng kể nếu sử dụng cấu hình lặp lại hoặc tổng hợp đầu ra.

Sự khác biệt về kết quả giữa các tài liệu luyện thi JLPT N3 có liên quan chặt chẽ đến độ khó, cấu trúc câu hỏi và loại văn bản của từng đầu sách. Cụ thể, các tài liệu mô phỏng sát đề thi thật như quyển 3, quyển 4 và quyển 5 bao gồm nhiều dạng yêu cầu phức hợp (xác định ý gạch chân, ý trong dấu ngoặc, đối chiếu bảng thông tin), đòi hỏi xử lý chi tiết và tích hợp thông tin, từ đó làm tăng tỷ lệ phản hồi chỉ đúng một phần. Ngược lại, quyển 6 có cấu trúc đơn giản hơn, chủ yếu sử dụng đoạn văn ngắn và dài với yêu cầu tìm ý đúng hoặc thông tin trực tiếp trong bài, nên ghi nhận kết quả ổn định hơn. Quyển 7 tập trung vào dạng câu hỏi hình và biểu đồ, song hình ảnh đơn giản và có minh họa rõ ràng, giúp giảm độ phức tạp trong xử lý. Quyển 8 có số lượng câu ít, chủ yếu là bài tìm ý đúng và xác định ý gạch chân, trong khi quyển 9 được biên soạn theo lộ trình từ cơ bản đến nâng cao, tập trung vào các dạng tìm thông tin trên hình và xác định ý gạch chân. Những khác biệt này cho thấy hiệu suất của ChatGPT chịu ảnh hưởng đáng kể từ đặc điểm ngữ liệu và hình thức đánh giá, chứ không chỉ từ năng lực nội tại của mô hình.

3.2. Phân tích kết quả định tính

3.2.1. Phân tích mức độ đúng ý chính của ChatGPT

Đối với dữ liệu đầu ra thuộc nhóm 1, gồm những câu có đáp án đúng qua 3 lần lặp lại, kết quả thống kê tần suất mức độ đúng ý chính qua các lời giải thích của ChatGPT so với lời giải của các tài liệu luyện thi JLPT N3 được nêu ở Bảng 3.

Bảng 3. Tần suất mức độ đúng ý chính của lời giải thích ChatGPT

Quyển	Mức độ 0	Mức độ 1	Mức độ 2
1	0	0	33
2	0	2	106
3	0	4	6
4	1	5	10
5	0	4	24
6	0	43	20
7	2	25	23
8	0	4	6
9	0	3	9
Tổng hợp	3	90	237
Tỷ lệ (%)	0,9	27,3	71,8

Ghi chú: mức độ 0 = không đúng ý chính, 1 = đúng một phần ý chính, 2 = đúng đầy đủ ý chính.

Kết quả cho thấy phần lớn lời giải thích của ChatGPT đạt mức độ "đúng đầy đủ ý chính", với tỷ lệ 71,8%. Điều này cho thấy công cụ này có khả năng diễn giải tương đối đầy đủ nội dung văn bản trong đa số trường hợp. Tuy nhiên, vẫn còn tới 27,3% phản hồi chỉ "đúng một phần ý chính" gây ảnh hưởng đến độ chính xác khi lựa chọn đáp án. Đáng chú ý, tỷ lệ phản hồi "không đúng ý chính" là rất thấp, chỉ 0,9%.

Cụ thể, mức độ đúng ý chính ở mức 2 của quyển 1 đạt 100% và quyển 2 đạt 98,2%. Kết quả này cho thấy nội dung của hai tài liệu này phù hợp với năng lực xử lý ngôn ngữ hiện tại và phản hồi của ChatGPT gần như trùng khớp với lời giải trong sách. Tuy nhiên, trường hợp ở quyển 6 có tới hơn 68% câu chỉ đạt mức độ 1, vượt xa số lượng phản hồi ở mức độ 2 (32%). Phần lớn các phản hồi ở mức độ 1 thuộc nhóm câu hỏi yêu cầu xác định nội dung phù hợp với văn bản gốc. Nhiều câu trong số này liên quan đến cụm từ hoặc câu được gạch chân trong đoạn trích. Điều này cho thấy mô hình gặp khó khăn trong việc diễn giải sát với ý chính của sách, dù thể hiện được một phần hiểu đúng ý chính của bài đọc.

Trong nghiên cứu của Dentella et al. (2024), nhóm tác giả nhấn mạnh rằng các mô hình LLMs, mặc dù đạt độ lưu loát và chính xác cao ở bề mặt nhưng vẫn thiếu sự nhạy cảm ngữ nghĩa thực sự và thường mắc lỗi khi yêu cầu phân tích ngữ nghĩa có tính kết nối giữa các thành phần văn bản.

3.2.2. Phân tích lỗi sai cụ thể của ChatGPT

Dựa trên các kết quả thu được đối với dữ liệu nhóm 2 gồm các câu có lỗi sai xuất hiện trong các lần lặp lại, việc chọn đáp án sai của ChatGPT do 5 nhóm lỗi chính sau được thể hiện tại Bảng 4.

Các nhóm lỗi trên đã được ghi nhận trong nhiều nghiên cứu gần đây về hiệu suất của các mô hình LLMs. Cụ thể, Saka et al. (2024) đã xây dựng một kho dữ liệu phân loại lỗi ChatGPT, trong đó nhấn mạnh lỗi "so khớp bề mặt" (surface-level matching) và "suy luận dựa trên ảo tưởng" (hallucination-based inference) là những lỗi phổ biến trong các nhiệm vụ đọc hiểu. Ngoài ra, Biswas (2023) cũng đề cập rằng việc phân loại lỗi một cách hệ thống theo từng nhóm chức năng đóng vai trò nền tảng trong việc nâng cao hiệu quả vận hành của các mô hình LLMs khi ứng dụng vào các bối cảnh học thuật chuyên biệt.



Hình 2. Tỷ lệ các lỗi sai xuất hiện trong phản hồi của ChatGPT

Bảng 4. Thống kê tần suất lỗi sai của ChatGPT trong 9 tài liệu

Quyển	Không hiểu đúng yêu cầu đề	Hiểu sai ý	Bỏ sót ý	Suy diễn	Lựa chọn đáp án
1	7	5	8	9	0
2	3	20	26	12	2
3	1	10	17	6	3
4	1	6	13	6	3
5	4	17	12	7	4
6	1	17	4	1	1
7	4	23	16	9	3
8	0	4	8	0	3
9	0	1	3	0	0
Tổng lỗi	21	103	107	50	19

Kết quả cho thấy các nhóm lỗi phổ biến nhất là "Hiểu sai ý" (34,3%) và "Bỏ sót ý" (35,7%), chiếm tỷ lệ lớn nhất trong tổng số 300 lỗi ghi nhận được (Hình 2). Lỗi "Hiểu sai ý" xuất hiện rõ nhất trong quyển 7 với 23 lỗi trên tổng số 103 lỗi. Điều này phản ánh sự thiếu chính xác trong quá trình diễn giải

nội dung văn bản, dẫn đến xu hướng lựa chọn luận điểm không phù hợp hoặc diễn giải sai ý chính. Itabashi and Matsubayashi (2024) nhận định rằng ChatGPT có xu hướng thực hiện tốt các câu hỏi yêu cầu tổng quát hóa nội dung toàn văn, nhưng lại gặp khó khăn rõ rệt ở các câu hỏi đòi hỏi đối chiếu chi tiết hoặc diễn giải chính xác từng đoạn nhỏ.

Lỗi "Bỏ sót ý" xảy ra chủ yếu ở quyển 2 với 26 lỗi trên tổng số 107 lỗi. Lỗi này xuất hiện khi ChatGPT chỉ tập trung vào một phần nội dung văn bản hoặc bỏ qua những thông tin then chốt có ảnh hưởng đến việc xác định đáp án chính xác. Theo Zhou et al. (2025), các mô hình LLMs thường có xu hướng chọn lọc thông tin dựa trên các cụm từ nổi bật thay vì xử lý toàn diện nội dung văn bản. Điều này dẫn đến việc bỏ sót các thông tin quan trọng nhưng không hiển hiện rõ ràng về mặt hình thức, ảnh hưởng trực tiếp đến độ chính xác của đáp án.

Kế đến, lỗi "suy diễn" chiếm 16,7% tổng số lỗi, là nhóm lỗi phổ biến thứ ba. Lỗi này xảy ra khi ChatGPT đưa ra suy luận vượt quá thông tin trong văn bản, dẫn đến việc tạo kết luận không được hỗ trợ bởi nội dung gốc. Các lỗi này xuất hiện nhiều nhất ở quyển 1 (9 lỗi), quyển 2 (12 lỗi) và quyển 7 (9 lỗi). Farquhar et al. (2024) nhấn mạnh rằng khả năng "đoán" của các mô hình LLMs có thể gây ra lỗi nếu việc suy luận không bám sát văn bản. Khi không kiểm soát tốt ranh giới giữa thông tin trong văn bản và tri thức nền, ChatGPT dễ tạo ra các suy luận không hợp lý. Hiện tượng này đặc biệt rõ trong các bài kiểm tra đọc hiểu yêu cầu đối chiếu chặt chẽ giữa câu hỏi và nội dung văn bản.

Ngược lại, các lỗi ít gặp nhất là "Không hiểu đúng yêu cầu đề" (7%) và "Lựa chọn đáp án" (6,3%). Lỗi "Không hiểu đúng yêu cầu đề" xuất hiện nhiều nhất trong quyển 1 với 7 lỗi và thường xảy ra khi ChatGPT không hiểu đúng mục tiêu của câu hỏi, dẫn đến việc phân tích sai hướng thông tin. Theo Borji (2023), một trong những nguyên nhân phổ biến dẫn đến sai sót của ChatGPT là do hệ thống không có khả năng xây dựng một "mô hình thế giới" (world model) đầy đủ, khiến nó không thể hiểu rõ bối cảnh ngầm định hay ý đồ của câu hỏi, đặc biệt trong các bài đọc hiểu có cấu trúc đa tầng. Tuy nhiên, lỗi này không xuất hiện trong quyển 8 và quyển 9, cho thấy ChatGPT có khả năng xác định đúng yêu cầu của đề trong hầu hết các trường hợp ở hai tài liệu này. Nghiên cứu của Thủy (2025) về chất lượng phản hồi của ChatGPT trong sửa đề viết tiếng Pháp cũng đề cập đến "Lỗi do không hiểu ý đồ của tác giả" và "Lỗi do không hiểu bối cảnh bài viết", lần lượt là 4% và 19%.

Tương tự, lỗi "Lựa chọn đáp án" xuất hiện ít, với quyển 5 ghi nhận 4 lỗi. Đây là lỗi khi ChatGPT hiểu đúng dữ kiện nhưng lựa chọn sai phương án trả lời chuẩn. Đây là một trong những dạng sai sót điển hình khi đánh giá khả năng xử lý bài đọc hiểu của ChatGPT. Mặc dù thuật ngữ này không được tác giả Borji (2023) sử dụng trực tiếp, song hiện tượng tương ứng đã được ghi nhận trong phân tích về các lỗi thuộc nhóm "reasoning" và "logic". Theo đó, ChatGPT có thể nắm được ngữ cảnh và dữ kiện chính xác, nhưng vẫn đưa ra câu trả lời sai do thất bại trong bước đối chiếu thông tin với yêu cầu câu hỏi hoặc lựa chọn đáp án không phù hợp với lập luận đã hình thành.

Ngoài ra, mặc dù quyển 9 có số lượng câu ít nhất với chỉ 15 câu có lời giải thích nhưng cũng ghi nhận 4 lỗi, trong đó tất cả đều thuộc nhóm "Hiểu sai ý" và "Bỏ sót ý", trong khi ba nhóm lỗi còn lại không xuất hiện. Những phát hiện này cho thấy sự khác biệt về cấu trúc ngữ liệu và độ phức tạp của câu hỏi giữa các đầu sách. Các yếu tố này có thể ảnh hưởng trực tiếp đến hiệu suất của ChatGPT. Việc xác định các nhóm lỗi này không chỉ giúp làm rõ các giới hạn hiện tại của ChatGPT trong việc xử lý ngôn ngữ mà còn cung cấp cơ sở để cải tiến công cụ trong các ứng dụng giảng dạy và đánh giá năng lực đọc hiểu tiếng Nhật.

3.2.2. Đặc điểm phản hồi của ChatGPT

Kết quả quan sát đặc điểm phản hồi của ChatGPT cho thấy các ưu điểm và nhược điểm trong quá trình giải các đề đọc của các tài liệu luyện thi JLPT N3 như sau:

Ưu điểm

Kết quả phân tích cho thấy ChatGPT sở hữu nhiều ưu điểm nổi bật trong quá trình xử lý bài đọc hiểu JLPT N3. (1) Duy trì định dạng ổn định: Kết quả cho thấy ChatGPT duy trì cấu trúc phản hồi nhất quán trong các phiên làm việc dài mà không cần nhắc lại hướng dẫn. Điều này phản ánh hiệu quả của các kỹ thuật có "mẫu câu lệnh hướng dẫn" (Instructional Prompt Patterns) và "hướng dẫn từng bước" (Step-by-Step Guidelines) tốt (Chen et al., 2025). (2) Tổ chức thông tin chủ động: ChatGPT tự động đề xuất cách trình bày dưới dạng bảng hoặc danh sách sau vài lượt truy vấn, giúp tăng tính trực quan và hỗ trợ người học theo dõi phân tích. Chen et al. (2025) và Saka et al. (2024) cũng đề cập rằng đây là kết quả từ quá trình tự điều chỉnh phản hồi, tương tự như kỹ thuật "gợi ý tự hiệu chỉnh" (Self-Refine Prompting). (3) Tóm tắt văn bản trước phân tích: Một số phản hồi cho thấy ChatGPT tự động

tóm lược nội dung trước khi thực hiện phân tích, thể hiện việc áp dụng chiến lược "gợi ý tóm tắt" (Summarization Prompting), giúp mô hình nắm bắt thông tin cốt lõi trước khi suy luận (Chen et al., 2025). (4) Tự điều chỉnh lỗi: ChatGPT có khả năng tự rà soát và chỉnh sửa lỗi, cho thấy sự áp dụng hiệu quả của các kỹ thuật "tự nhất quán" (Self-Consistency) và "tự hiệu chỉnh" (Self-Refine) để cải thiện độ chính xác qua nhiều bước như được Chen et al. (2025) chỉ ra.

Nhược điểm

Bên cạnh những ưu điểm, kết quả phân tích cũng chỉ ra một số nhược điểm đáng chú ý trong phản hồi của ChatGPT. (1) Lặp lại nội dung: Khi xử lý văn bản dài, ChatGPT dễ lặp lại câu hỏi hoặc phân tích không liên quan, gây nhiễu cho người dùng. Cengiz et al. (2025) cũng nhấn mạnh rằng học viên thường giảm động lực và gặp khó khăn khi đưa ra phản hồi quá khó, dài dòng, vượt quá trình độ, thiếu rõ ràng hoặc chồng chéo giữa các nội dung. Saka et al. (2024) cũng đề cập rằng ChatGPT vẫn còn xuất hiện các phản hồi sai lệch hoặc không liên quan, cho thấy sự cần thiết phải cải thiện và tối ưu hóa quy trình phát triển prompt nhằm nâng cao độ chính xác của kết quả. (2) Giới hạn trong xử lý hình ảnh: Khi chỉ có hình ảnh mà không có văn bản kèm theo, ChatGPT thường trả về phản hồi mặc định như "It seems you've uploaded several images..." (tạm dịch: "Có vẻ như bạn đã tải lên vài hình ảnh..."), cho thấy giới hạn trong khả năng liên kết thị giác-ngôn ngữ. OpenAI (2025) cũng có đề cập rằng việc xử lý thông tin trong ảnh sẽ bị giảm đi đối với những ngôn ngữ không phải chữ Latin, như tiếng Hàn hay tiếng Nhật. (3) Phụ thuộc vào định dạng truy vấn: Mô hình dễ hiểu sai yêu cầu khi prompt không rõ ràng, dẫn đến phản hồi thiếu chính xác. Điều này trùng khớp với khẳng định của Ando et al. (2024) rằng hiệu suất của ChatGPT có thể thay đổi tùy thuộc vào cách đặt câu hỏi, và chất lượng phản hồi bằng tiếng Anh thường vượt trội so với các ngôn ngữ khác. Đồng thời, Saka et al. (2024) cũng nhấn mạnh rằng mặc dù GPT được huấn luyện trên dữ liệu lớn, khả năng hiểu kiến thức chuyên ngành còn hạn chế, do đó cần bổ sung ngữ cảnh và diễn giải rõ ràng để nâng cao hiệu quả sử dụng.

4. ĐỀ XUẤT

Từ kết quả nghiên cứu, một số đề xuất được đưa ra nhằm cải thiện cả phương diện nghiên cứu lẫn ứng dụng thực tiễn của ChatGPT trong luyện thi JLPT N3.

Về phương diện nghiên cứu, các nghiên cứu tiếp theo cần mở rộng phạm vi khảo sát sang các cấp độ JLPT cao hơn như N2 và N1, đồng thời đa dạng hóa nguồn tư liệu thay vì chỉ giới hạn ở chín đầu sách luyện thi. Bên cạnh đó, việc tăng số lần thử nghiệm cần được tiến hành để bao quát hơn các biến thiên ngẫu nhiên của mô hình, qua đó nâng cao độ tin cậy trong kiểm chứng kết quả.

Về phương diện ứng dụng, kết quả cho thấy ChatGPT hoạt động hiệu quả hơn đối với các dạng văn bản liền mạch (đoạn ngắn, trung bình, dài) và câu hỏi yêu cầu tóm tắt, xác định ý chính hoặc quan hệ nhân quả. Ngược lại, mô hình dễ mắc lỗi ở dạng "tìm kiếm thông tin" vốn có cấu trúc phi tuyến tính và nhiều yếu tố phi văn bản. Do đó, người học nên sử dụng ChatGPT như công cụ hỗ trợ để tóm tắt nội dung, gợi ý ý chính và định hướng phân tích, đồng thời cần đối chiếu với tài liệu gốc để tránh bị phụ thuộc hoàn toàn và nâng cao năng lực đọc hiểu độc lập.

5. KẾT LUẬN

Kết quả nghiên cứu cho thấy ChatGPT đạt hiệu suất tương đối cao trong xử lý bài đọc hiểu JLPT N3. Tỷ lệ trả lời đúng trung bình vượt 60% ở tất cả các tài liệu, trong đó ba tài liệu (quyển 2, quyển 6 và quyển 9) vượt mức 80% và quyển 6 đạt tỷ lệ cao nhất là 89,9%. Đồng thời, kết quả phân tích thống kê cho thấy không có sự khác biệt có ý nghĩa thống kê giữa ba lần thử nghiệm ($p > 0,05$), qua đó khẳng định tính ổn định của mô hình trong các phiên truy vấn lặp lại. Về kết quả định tính, 71,8% phản hồi của ChatGPT được đánh giá đúng đầy đủ ý chính của lời giải trong sách, 27,3% đúng một phần ý chính và chỉ 0,9% không đúng ý chính. Các lỗi sai phổ biến nhất tập trung vào hai nhóm "bỏ sót ý" (35,7%) và "hiểu sai ý" (34,3%), phản ánh hạn chế của mô hình trong việc duy trì mạch văn và xử lý các câu hỏi yêu cầu tích hợp thông tin đa tầng hoặc suy luận gián tiếp. Bên cạnh đó, nghiên cứu cũng ghi nhận một số đặc điểm tích cực trong phản hồi của ChatGPT như khả năng duy trì định dạng ổn định, tóm tắt nội dung trước khi phân tích và tự điều chỉnh sai sót, cho thấy tiềm năng của mô hình như một công cụ hỗ trợ hữu ích trong luyện thi JLPT nếu được sử dụng đúng cách.

Kết quả của nghiên cứu này cung cấp bằng chứng thực nghiệm về khả năng và giới hạn của ChatGPT trong xử lý bài đọc hiểu tiếng Nhật, qua đó đóng góp cho lĩnh vực ứng dụng trí tuệ nhân tạo trong giáo dục ngôn ngữ. Kết quả không chỉ có ý nghĩa học thuật trong việc đánh giá năng lực của mô

hình ngôn ngữ lớn mà còn định hướng việc sử dụng ChatGPT như công cụ hỗ trợ học tập.

Tuy nhiên, nghiên cứu còn hạn chế do chỉ tập trung vào JLPT N3; số lượng tài liệu và số lần thử nghiệm còn hạn chế, làm giảm khả năng khái quát hóa kết quả. Các nghiên cứu tiếp theo nên mở rộng sang các cấp độ cao hơn (N2, N1), tăng quy mô dữ liệu và so sánh với người học hoặc các mô hình ngôn ngữ khác nhằm làm rõ hơn vai trò của ChatGPT trong dạy và học ngoại ngữ.

TÀI LIỆU THAM KHẢO

Ando, K., Sato, M., Wakatsuki, S., Nagai, R., Chino, K., Kai, H., Sasaki, T., Kato, R., Nguyen, T. P., Guo, N., & Sultan, P. (2024). A comparative study of English and Japanese ChatGPT responses to anaesthesia-related medical questions. *BJA Open*, *10*(C), 100296. https://doi.org/10.1016/j.bjao.2024.100296

Anh, T. P. N., & Ghi, N. T. (2025). Ý định áp dụng ChatGPT và kết quả học tập của sinh viên tiếng Anh thương mại: Phân tích từ mô hình TAM và lý thuyết học tập kiến tạo. *Tạp chí Khoa học và Công nghệ – Trường Đại học Hùng Vương*, *11*(1), 62–73. https://doi.org/10.59775/1859-3968.254

Biswas, S. (2023). Evaluating Errors and Improving Performance of ChatGPT: A Research Paper. *Clinical Research and Studies*, *2*(3), 1-5. https://doi.org/10.31579/2835-2882/019

Borji, A. (2023). A Categorical Archive of ChatGPT Failures. *ArXiv*, abs/2302.03494. https://doi.org/10.48550/arXiv.2302.03494

Cengiz, B. C., Bilki, Z., Ataş, A. H., & Celik, B. (2025). Exploring second language writers' engagement with ChatGPT feedback: Revision behaviors and perceptions. *System*, *134*, 103837. https://doi.org/10.1016/j.system.2025.103837

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2025). Unleashing the potential of prompt engineering for large language models. *Patterns*, *6*(5), 1-44. https://doi.org/10.1016/j.patter.2025.101260

Dentella, V., Günther, F., Murphy, E., Marcus, G., & Leivada, E. (2024). Testing AI on language comprehension tasks reveals insensitivity to underlying meaning. *Scientific Reports*, *14*(1), 28083. https://doi.org/10.1038/s41598-024-79531-8

Dirgandini, M., Lewerissa, E. D., Alvin, & Setiana, S. M. (2025). Difficulties in Learning Japanese Reading Comprehension Reviewed from the Perceptions of Indonesian Students. *Forum for Linguistic Studies*, *7*(2), 1178–1190. https://doi.org/10.30564/fls.v7i2.8568

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, *630*, 625–630. https://doi.org/10.1038/s41586-024-07421-0

Itabashi, Y., & Matsubayashi, Y. (2024). An Analysis of Large Language Models' Reading Comprehension Ability for Narrative Texts. *Proceedings of the 30th Annual Meeting of the Association for Natural Language* (pp.309-313). Kobe International Conference Center. https://www.anlp.jp/proceedings/annual_meeting/2024/pdf_dir/A2-4.pdf?utm_source

Japan Foundation. (2021). *The current state of Japanese language education in Southeast Asia*. https://www.jpf.go.jp/j/project/japanese/survey/result/dl/survey2021/s_east_asia.pdf

Japanese Language Proficiency Test. (2024). *Past test data*. https://www.jlpt.jp/e/statistics/archive.html

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Nga, T. T. N. (2024). Tổng quan nghiên cứu về ứng dụng ChatGPT trong dạy học và đánh giá ứng dụng ChatGPT vào việc thiết kế bài giảng. *Tạp chí Giáo dục*, *24*(9), 21–26.

Nishizawa, H., Isbell, D. R., & Suzuki, Y. (2022). Review of the Japanese-Language Proficiency Test. *Language Testing*, *39*(3), 494–503. https://doi.org/10.1177/02655322221080898

OpenAI (2025). *ChatGPT Image Inputs FAQ*. https://help.openai.com/en/articles/8400551-chatgpt-image-inputs-faq?utm_source=chatgpt.com

Quân, P. N. (2023). ChatGPT hỗ trợ khả năng tự học và phát triển năng lực số cho thanh niên. *Kỷ yếu Hội thảo khoa học Cán bộ trẻ các Đại học Quốc gia, Đại học vùng mở rộng lần thứ II* (tr.232-237). Trường Đại học Ngoại ngữ – Đại học Đà Nẵng.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. *Conference on Empirical Methods in Natural Language Processing* (pp. 2383–2392). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1264

Saka, A., Taiwo, R., Saka, N., Salami, B. A., Ajayi, S., Akande, K., & Kazemi, H. (2024). GPT models in construction industry: Opportunities, limitations, and a use case validation. *Developments in the Built Environment*, *17*, 100300. https://doi.org/10.1016/j.dibe.2023.100300

Sato, H., & Ogasawara, K. (2024). ChatGPT (GPT-4) passed the Japanese National License Examination for Pharmacists in 2022, answering all items including those with diagrams: a descriptive study. *Journal of Educational Evaluation for Health Professions*, *21*(4), 1-8. https://doi.org/10.3352/jeehp.2024.21.4

Taira, K., Itaya, T., & Hanada, A. (2023). Performance of the large language model ChatGPT on the national nurse examinations in Japan: Evaluation study. *JMIR Nursing*, *6*, e47305. https://doi.org/10.2196/47305

Thủy, T. B. Đ. (2025). Chất lượng phản hồi của ChatGPT trong sửa bài viết tiếng Pháp. *Tạp chí Khoa học Đại học Sư phạm Hà Nội: Khoa học Giáo dục*, *70*(2), 33–44.

Vinh, A. L., Diễn, T. B., Quân, Q. L., & Luân, V. V. (2023). Khả năng thực hiện bài kiểm tra định kì môn Toán và môn Ngữ văn cấp trung học của công cụ ChatGPT: Kết quả nghiên cứu và một số khuyến nghị ban đầu. *Tạp chí Khoa học Giáo dục Việt Nam*, *19*(2), 1–10.

Wan, G., Wu, Y., Chen, J., & Li, S. (2025). Reasoning Aware Self-Consistency: Leveraging Reasoning Paths for Efficient LLM Sampling. *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3613–3635). Association for Computational Linguistics. https://aclanthology.org/2025.naacl-long.184/

Yang, L., Xiong, Y., & Chen, Q. (2023). The role of linguistic and cognitive skills in reading Chinese as a second language: A path analysis modeling approach. *Frontiers in Psychology, 14*, 1131913. https://doi.org/10.3389/fpsyg.2023.1131913

Zhou, T., Zhang, Y., Cao, S., Li, Y., Wang, T., & Yang, Y. (2025). Complementary advantages exhibited by ChatGPT and human readers in text reading inference. *Current Psychology*, *44*(12), 12679–12693. https://doi.org/10.1007/s12144-025-08010-8