<u>Research Letter</u>

Impact of Attached File Formats on the Performance of ChatGPT-4 on the Japanese National Nursing Examination: Evaluation Study

Kazuya Taira[1], RN, PHN, PhD; Takahiro Itaya[2], RN, MPH, DrPH; Shuntaro Yada[3,4], PhD; Kirara Hiyama[2], RN, MPH; Ayame Hanada[2], RN, PHN, BHS

[1]Human Health Sciences, Graduate School of Medicine, Kyoto University, Kyoto, Japan

[2]Department of Healthcare Epidemiology, Graduate School of Medicine and Public Health, Kyoto University, Kyoto, Japan

[3]Graduate School of Science and Technology, Nara Institute of Science and Technology, Ikoma, Japan

[4]Faculty of Library, Information and Media Science, University of Tsukuba, Tsukuba, Japan

Corresponding Author:

Kazuya Taira, RN, PHN, PhD
Human Health Sciences, Graduate School of Medicine
Kyoto University
53, Shogoinkawara-cho, Sakyo-ku
Kyoto, 606-8507
Japan
Phone: 81 0757513927
Email: taira.kazuya.5m@kyoto-u.ac.jp

Abstract

This research letter discusses the impact of different file formats on ChatGPT-4's performance on the Japanese National Nursing Examination, highlighting the need for standardized reporting protocols to enhance the integration of artificial intelligence in nursing education and practice.

Introduction

Numerous generative artificial intelligences (AIs), exemplified by all versions of ChatGPT [1] and Llama [2], have been developed using large language models and evaluated in health care, particularly in nursing education [3,4], successfully passing national nursing examinations in several countries [5,6]. Generative AIs are evolving to handle multimodal information, including text and images [1]. However, previous evaluations have not assessed the impact of file formats [5,6].

Prompts, particularly long ones, can affect response accuracy owing to potential context loss or exceeded token limits [7-9]. In this study, we hypothesized that the file format attached to prompts could affect the results of nursing research that uses generative AI and aimed to evaluate its impact on ChatGPT-4's performance on the Japanese National Nursing Examination. The findings of this study would be useful for improving the quality of reports on future nursing research that uses generative AI.

Methods

Ethics Approval

This study did not require ethical approval or informed consent, as the data analyzed were obtained from a published database from the Ministry of Health, Labour and Welfare.

Generative AI Model

We used the original, unmodified GPT-4 (gpt-4-1106-preview, accessed March 2024) without additional training, tuning, or data. ChatGPT, launched by OpenAI in 2022, with GPT-4 released in March 2023, is currently widely used.

Input Data

The dataset included all 50 basic knowledge questions from the 2023 Japanese National Nursing Examination, along with 190 general questions. The passing standard for these basic knowledge questions was approximately 80%. ChatGPT-3.5 has consistently failed to meet this standard [4], leading us to consider whether performance might vary based on file format. Questions were prepared in TEXT (.txt), DOCX (.docx), PDF (.pdf), and IMAGE (.jpg) formats and in a format that directly described all questions in the prompt (PROMPT-ONLY format). Although other formats, including CSV, JSON, XML, and Markdown, could be used to present questions and choices, we excluded them to maintain consistency and focus on more common formats.

Prompt Engineering

The prompts for each file format are summarized in Textbox 1.

Textbox 1. Prompts provided to ChatGPT-4. The files (mentioned at the end of the prompt for TXT, DOCX, PDF, and JPG formats) were made viewable via OpenAI's application programming interface (API) function: ASSISTANT (type = retrieval).

\<Prompt for PROMPT-ONLY format\>
You are an expert in the field of nursing. Answer the given questions briefly and numerically. {Question number}. {Question}. Options: (1) {Option 1}, (2) {Option 2}, (3) {Option 3}, (4) {Option 4}

> *Example: 1. Which vessel sends blood from the fetus to the placenta in the fetal circulation? Options: (1) Common carotid artery, (2) Pulmonary artery, (3) Umbilical artery, and (4) Umbilical vein.*

\<Prompt for TXT, DOCX, PDF, and JPG formats\>
You are an expert in the field of nursing. Answer briefly and numerically all questions given by the file.
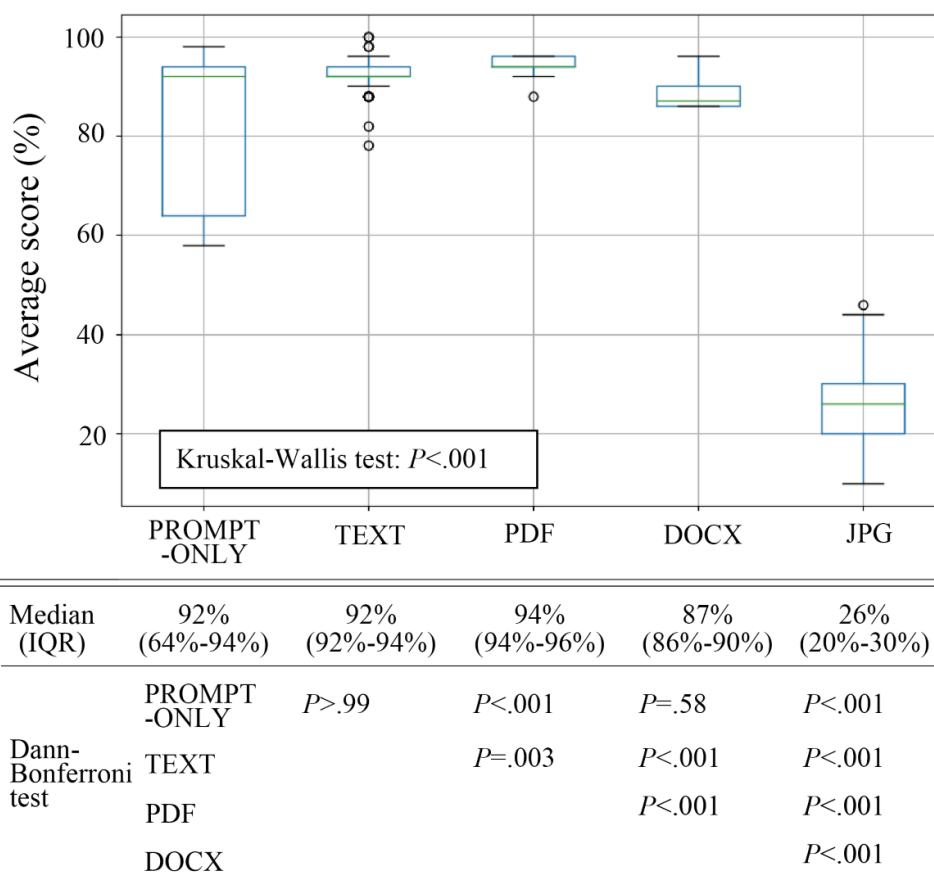

Data Analyses

Prompts for all formats were processed for 100 iterations each; the median and IQR of the percentage of correct answers were calculated. Differences among the percentages of correct answers by the attached file format were compared using the Kruskal-Wallis test and Dann-Bonferroni test. Statistical analyses were performed using Python (version 3.11.4) with the *pandas* (version 1.5.3) and *matplotlib* (version 3.7.1) libraries.

Results

The median percentages of correct answers were 92% (IQR 64%-94%), 92% (IQR 92%-94%), 94% (IQR 94%-96%), 87% (IQR 86%-90%), and 26% (IQR 20%-30%) for PROMPT-ONLY, TEXT, PDF, DOCX, and JPG formats, respectively. The differences between the attached formats were statistically significant in all pairs (P<.01) except for the PROMPT-ONLY versus TEXT and PROMPT-ONLY versus DOCX pairs (Figure 1).

Figure 1. Performance evaluation of ChatGPT-4 on the Japanese National Nursing Examination by the attached file format. Outliers, shown as dots, are values below Q1 – 1.5 or above Q3 + 1.5 in the IQR.

	PROMPT-ONLY	TEXT	PDF	DOCX	JPG
Median (IQR)	92% (64%-94%)	92% (92%-94%)	94% (94%-96%)	87% (86%-90%)	26% (20%-30%)

Dann-Bonferroni test		PROMPT-ONLY	TEXT	PDF	DOCX	JPG
	PROMPT-ONLY		$P>.99$	$P<.001$	$P=.58$	$P<.001$
	TEXT			$P=.003$	$P<.001$	$P<.001$
	PDF				$P<.001$	$P<.001$
	DOCX					$P<.001$

Discussion

ChatGPT-4's performance on the Japanese National Nursing Examination varied significantly across file formats. The best performance was observed with PROMPT-ONLY, TEXT, and PDF formats (median scores >92%), followed by DOCX (87%), and the worst performance was with JPG (26%). The PROMPT-ONLY format exhibited a larger IQR and more variability than TEXT, PDF, and DOCX formats. JPG's poor performance highlights a significant limitation of generative AI, which excels at processing digital text but struggles with interpreting text from images. This "visual comprehension" gap has critical implications for AI applications involving nondigital text sources. The variability in PROMPT-ONLY performance may reflect reduced accuracy with longer prompts [7,8].

Therefore, to prepare for a future where generative AI is integrated into nursing practice and education [10], it is crucial to understand the interaction between humans and generative AI, including the impact of input file formats.

Additionally, it is essential to report the following aspects in a standardized manner:

- Name and version of the generative AI model
- Presence of additional training, tuning, or knowledge transfer
- Prompt design and attached file formats
- Response generation parameters, including the number of iterations, temperature settings, and maximum token count
- Execution environment (if applicable)

However, as we only examined ChatGPT-4's performance on the Japanese National Nursing Examination and the impact of major file formats, investigations on other formats and AI models are warranted. Particularly, evaluating the performance of AI that specializes in image processing and image formats other than JPG and expanding the evaluations to include national nursing examinations in other countries and clinical questions in practice will be important in future research.

Acknowledgments

interpretations. After using these tools, the authors reviewed and edited the content as necessary and take full responsibility for the content of the published article.

Conflicts of Interest

None declared.

References

1. OpenAI, Achiam J, Adler S, et al. GPT-4 technical report. arXiv. Preprint posted online on Mar 15, 2023. [doi: 10.48550/arXiv.2303.08774]
2. Topaz M, Peltonen LM, Michalowski M, et al. The ChatGPT effect: nursing education and generative artificial intelligence. J Nurs Educ. Feb 5, 2024:1-4. [doi: 10.3928/01484834-20240126-01] [Medline: 38302101]
3. Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. arXiv. Preprint posted online on Feb 27, 2023. [doi: 10.48550/arXiv.2302.1397]
4. Jin HK, Lee HE, Kim E. Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: a systematic review and meta-analysis. BMC Med Educ. Sep 16, 2024;24(1):1013. [doi: 10.1186/s12909-024-05944-8] [Medline: 39285377]
5. Taira K, Itaya T, Hanada A. Performance of the large language model ChatGPT on the National Nurse Examinations in Japan: evaluation study. JMIR Nurs. Jun 27, 2023;6:e47305. [doi: 10.2196/47305] [Medline: 37368470]
6. Su MC, Lin LE, Lin LH, Chen YC. Assessing question characteristic influences on ChatGPT's performance and response-explanation consistency: insights from Taiwan's Nursing Licensing Exam. Int J Nurs Stud. May 2024;153:104717. [doi: 10.1016/j.ijnurstu.2024.104717] [Medline: 38401366]
7. Ratnayake H, Wang C. A prompting framework to enhance language model output. Presented at: AI 2023: Advances in Artificial Intelligence: 36th Australasian Joint Conference on Artificial Intelligence; Nov 28 to Dec 1, 2023; Brisbane, Australia. [doi: 10.1007/978-981-99-8391-9_6]
8. Levy M, Jacoby A, Goldberg Y. Same task, more tokens: the impact of input length on the reasoning performance of large language models. arXiv. Preprint posted online on Feb 19, 2024. [doi: 10.18653/v1/2024.acl-long.818]
9. Zhang ZY, Verma A, Doshi-Velez F, Low BKH. Understanding the relationship between prompts and response uncertainty in large language models. arXiv. Preprint posted online on Jul 20, 2024. [doi: 10.48550/arXiv.2407.14845]
10. Goldberg CB, Adams L, Blumenthal D, et al. To do no harm — and the most good — with AI in health care. NEJM AI. Feb 22, 2024;1(3). [doi: 10.1056/AIp2400036]

Abbreviations

AI: artificial intelligence