Cureus
Part of SPRINGER NATURE

Open Access Original Article

"This Is a Quiz" Premise Input: A Key to Unlocking Higher Diagnostic Accuracy in Large Language Models

Review began 10/05/2024
Review ended 10/20/2024
Published 10/25/2024

DOI: 10.7759/cureus.72383

Yusuke Asari [1], Ryo Kurokawa [1], Yuki Sonoda [1], Akifumi Hagiwara [1], Jun Kamohara [1], Takahiro Fukushima [1], Wataru Gonoi [1], Osamu Abe [1]

1. Radiology, The University of Tokyo, Tokyo, JPN

Corresponding author: Ryo Kurokawa, kuroro63@gmail.com

Abstract

Purpose

Large language models (LLMs) are neural network models that are trained on large amounts of textual data, showing promising performance in various fields. In radiology, studies have demonstrated the strong performance of LLMs in diagnostic imaging quiz cases. However, the inherent differences in prior probabilities of a final diagnosis between clinical and quiz cases pose challenges for LLMs, as LLMs had not been informed about the quiz nature in previous literature, while human physicians can optimize the diagnosis, consciously or unconsciously, depending on the situation. The present study aimed to test the hypothesis that notifying LLMs about the quiz nature might improve diagnostic accuracy.

Methods

One hundred and fifty consecutive cases from the "Case of the Week" radiological diagnostic quiz case series on the American Journal of Neuroradiology website were analyzed. GPT-4o and Claude 3.5 Sonnet were used to generate the top three differential diagnoses based on the textual clinical history and figure legends. The prompts included or excluded information about the quiz nature for both models. Two radiologists evaluated the accuracy of the diagnoses. McNemar's test assessed differences in correct response rates.

Results

Informing the quiz nature improved the diagnostic performance of both models. Specifically, Claude 3.5 Sonnet's primary diagnosis and GPT-4o's top 3 differential diagnoses significantly improved when the quiz nature was informed.

Conclusion

Informing the quiz nature of cases significantly enhances LLMs' diagnostic performances. This insight into LLMs' capabilities could inform future research and applications, highlighting the importance of context in optimizing LLM-based diagnostics.

Categories: Radiology
Keywords: bayes' theorem, claude 3.5 sonnet, gpt-4o, large language model, prompt engineering

Introduction

Large language models (LLMs) are neural network models that are trained on massive amounts of textual data, demonstrating excellent performance in various natural language processing tasks, including those in the medical field [1-3]. Recently, there have been attempts to apply LLMs to radiological diagnostics. In order to apply LLM to radiological diagnosis, it is necessary to be familiar with the characteristics of LLM. Numerous studies have reported the use of LLMs in quiz cases for diagnostic radiologists. Reports indicate that by inputting textual data such as clinical history and image findings, ChatGPT's GPT-4 model (OpenAI, San Francisco, CA, USA) performed well in diagnostic imaging quizzes from radiology journals [4,5]. Other studies have shown that LLMs achieve good results in Radiology Board examinations across different countries [6-8]. Comparative studies among vendors have also been conducted, with Sonoda et al. [9] reporting that in text-based "Diagnosis Please" assessments, Claude 3 Opus (Anthropic, San Francisco, CA, USA), GPT-4o (OpenAI, San Francisco, CA, USA), and Gemini 1.5 Pro (Google, Mountain View, CA, USA) scored significantly higher in this respective order. Efforts to use vision-language models that directly input radiological images are underway, and studies have reported improved accuracy in LLMs through key image input [10,11].

However, compared to human radiologists, Suthar et al. [12] reported that LLMs still perform lower than human radiologists in challenging neuroradiology cases. One notable issue is that the prior probability of a

final diagnosis differs between clinical cases and quiz cases. Human radiologists, when presented with a quiz, tend to provide answers focusing on rare diseases, recognizing that these are more likely to be the intended answers. Conversely, they are inclined to avoid suggesting common diseases (such as cerebral infarction and subarachnoid haemorrhage), as these are less likely to be the quiz's target. In contrast, the aforementioned studies did not inform the LLMs that they were dealing with quizzes, thus putting them at a disadvantage compared to humans.

We hypothesized that notifying LLMs that they are dealing with quiz cases for diagnostic radiologists might improve their diagnostic accuracy. The present study aimed to assess the impact of revealing the quiz nature of cases on the diagnostic performance of LLMs.

This article was previously posted to the medRxiv preprint server on September 23, 2024 (https://www.medrxiv.org/content/10.1101/2024.09.20.24314101v1).

Materials And Methods

Figure *1* shows the overview of this study.

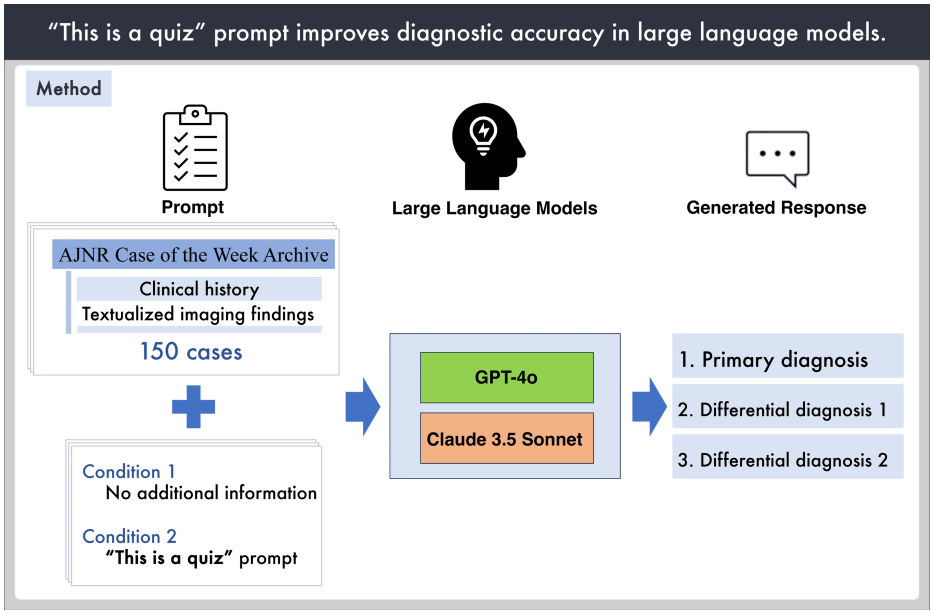

FIGURE 1: Overview of the study

No ethical approval was required, as this study only used publicly available information on the AJNR website [13].

We utilized the clinical history and figure legends from the "Case of the Week" series, a weekly quiz case compilation for diagnostic radiologists available on the American Journal of Neuroradiology website [13]. A total of 150 consecutive cases from August 2021 to June 2024 were analyzed.

We used GPT-4o (OpenAI, San Francisco, CA, USA; released on May 13, 2024) and Claude 3.5 Sonnet (Anthropic, San Francisco, CA, USA; released on June 27, 2024) to list the primary diagnoses and two differential diagnoses for the cases.

We used application programming interfaces to access each model (GPT-4o: gpt-4o-2024-05-13; Claude 3.5 Sonnet: claude-3-5-sonnet-20240620) on August 8, 2024. To ensure reproducibility, generation parameters for all models were set as temperature = 0.0. Each input was conducted in an independent session so as to prevent previous inputs from influencing subsequent ones. The prompt was as follows: "Assuming you are a physician, please respond with the most likely diagnosis and the next two most likely differential diagnoses based on the attached information," with or without the following sentences in the prompt to reveal the premise that they are quiz cases for experienced diagnostic radiologists ("additional prompt") for LLM. "They are quizzes of diagnostic imaging for doctors specializing in diagnostic radiology, and the purpose of the questions is to share knowledge on rare diseases and their imaging findings. Please keep this premise in mind and answer the questions, considering that common diseases are less likely to be asked." We submitted each prompt to the models only once and used the first response generated for evaluation.

One trainee radiologist with three years of experience and one board-certified diagnostic radiologist with 11 years of experience assessed the accuracy of the primary diagnosis and two differential diagnoses generated by the models by consensus.

McNemar's test assessed the difference in correct response rates for the overall accuracy under Conditions 1 (without additional prompt) and 2 (with additional prompt) for each model and between the models. Two-sided p-values < 0.05 were considered statistically significant. Statistical analyses were performed using R (version 4.1.1; R Foundation for Statistical Computing, Vienna, Austria).

Results

As shown in Table *1*, revealing the quiz context significantly improved the top three differential diagnoses for GPT-4o and the primary diagnosis for Claude 3.5 Sonnet (p = 0.020 and p = 0.0075, respectively). While improvements were observed in GPT-4o's primary diagnosis and Claude 3.5 Sonnet's top three differential diagnoses upon revealing the quiz context, these did not reach statistical significance. Claude 3.5 Sonnet exhibited significantly superior diagnostic performance compared to GPT-4o under all conditions (p < 0.0001, McNemar's test).

Diagnosis	LLM	Condition	Accuracy	p-value
Primary diagnosis	GPT-4o	1	33/150 (22.0%)	-
		2	40/150 (26.7%)	0.071
	Claude 3.5 Sonnet	1	62/150 (41.3%)	-
		2	72/150 (48.0%)	0.0075*
Top 3 differential diagnoses	GPT-4o	1	45/150 (30.0%)	-
		2	54/150 (36.0%)	0.020*
	Claude 3.5 Sonnet	1	73/150 (48.7%)	-
		2	80/150 (53.3%)	0.052

TABLE 1: Diagnostic performances of large language models

All P-values shown are the results of McNemar's test.

*Statistically significant.

Specifically, Claude 3.5 Sonnet, without the quiz context, incorrectly proposed sinonasal lymphoma as the primary diagnosis for the February 9, 2023 case [14] but correctly diagnosed the rare Rosai-Dorfman disease when informed of the quiz nature. Similarly, GPT-4o failed to include the rare L-2-hydroxyglutaric aciduria in its differential diagnoses for the October 20, 2022 case [15] without information of the quiz nature but correctly identified it as the primary diagnosis when informed of the quiz nature.

Conversely, there were several instances in which notifying the quiz nature of the case resulted in incorrect answers. For example, both LLMs were able to list perioperative ischaemic optic neuropathy as a differential diagnosis for the January 13, 2022 case [16] without the quiz nature. However, with the knowledge of the quiz nature, they provided incorrect responses. This suggests that common conditions, such as ischaemia, may lead to a higher likelihood of incorrect answers when the LLMs are informed that they are participating in a quiz.

Discussion

This study compared the diagnostic capabilities of different LLMs and explored the impact of informing them of the quiz nature of cases. Claude 3.5 Sonnet significantly outperformed GPT-4o in all conditions, and both models showed enhanced performance when aware of the quiz context.

Research involving LLMs solving quiz cases with definitive diagnoses is crucial for performance assessment of LLMs (inter-vendor comparisons, intra-vendor version comparisons), evaluating similarities and differences with human radiologists, and exploring future applications [9,11,12]. However, existing studies did not inform LLMs of the quiz nature, and in such situations, the LLMs generally exhibit difficulties with uncommon diagnoses compared to human radiologists [17].

According to Bayes' theorem, the pre-test probability (disease prevalence) is a critical determinant of post-test probability and subsequent diagnosis [18,19]. Physicians, including radiologists, understand that disease frequency varies depending on the situation [20,21]. Based on this premise - whether it is a quiz case, regional differences, or clinic setting versus third-party referral hospitals - they apply a gradient from high pre-test probability diagnoses to extremely low ones. We assumed that the same principle could apply to LLMs. We demonstrated that including the quiz context in the prompt significantly improved diagnostic performance.

Our study suggests future research directions: similar to how the quiz context enhanced diagnostic performance, presenting the clinical context might improve real-world diagnostic capabilities. It is known that appropriate clinical information enhances radiologists' diagnostic accuracy [22]. Providing LLMs with clinical information databases (age and gender distribution, region, facility size, disease prevalence, etc.) could yield optimized diagnostic results for individual patients. The differences in diagnostic performances from human radiologists should be further examined in the future.

This study has some limitations. We did not analyze disease- or category-specific performance due to the limited number of cases. The applicability of diagnostic imaging in areas other than neuroradiology was also not examined in the present study. Since the present study did not include the images themselves in the input data, future studies are needed to determine whether the combination of image data input and prompt engineering can improve the diagnostic performance of LLM. Additionally, the history and legends used are publicly available, possibly included in the training data of GPT-4o and Claude 3.5 Sonnet.

Conclusions

In conclusion, this study investigated whether notifying LLMs that they are dealing with quiz cases for diagnostic radiologists improves their diagnostic accuracy. By informing the LLM that the case is a quiz case, the diagnostic performance based on text data from the history and figure legend significantly improved. This study also demonstrated that Claude 3.5 Sonnet significantly outperformed GPT-4o in all conditions. LLMs, much like human radiologists, can benefit from contextual cues for better diagnosis. Understanding this feature of LLMs could be valuable for future research and clinical applications, eventually leading to more sophisticated uses of artificial intelligence in medical diagnosis and treatment.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Ryo Kurokawa, Yusuke Asari, Akifumi Hagiwara, Wataru Gonoi

Acquisition, analysis, or interpretation of data: Ryo Kurokawa, Yusuke Asari, Yuki Sonoda, Akifumi Hagiwara, Jun Kamohara, Takahiro Fukushima, Wataru Gonoi, Osamu Abe

Drafting of the manuscript: Ryo Kurokawa, Yusuke Asari

Critical review of the manuscript for important intellectual content: Ryo Kurokawa, Yuki Sonoda, Akifumi Hagiwara, Jun Kamohara, Takahiro Fukushima, Wataru Gonoi, Osamu Abe

Supervision: Osamu Abe

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

The large language models were used to generate diagnoses for the quiz cases. However, these models were not utilized in the research design, hypothesis formulation, data analysis, or manuscript writing in this study.

References

1. GPT-4 technical report. (2023). Accessed: October 25, 2024: https://arxiv.org/abs/2303.08774.
2. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS: Large language models in medicine. Nat Med. 2023, 29:1930-40. 10.1038/s41591-023-02448-8
3. Singhal K, Azizi S, Tu T, et al.: Large language models encode clinical knowledge . Nature. 2023, 620:172-80. 10.1038/s41586-023-06291-2
4. Ueda D, Mitsuyama Y, Takita H, Horiuchi D, Walston SL, Tatekawa H, Miki Y: ChatGPT's diagnostic performance from patient history and imaging findings on the diagnosis please quizzes. Radiology. 2023, 308:e231040. 10.1148/radiol.231040
5. Horiuchi D, Tatekawa H, Shimono T, et al.: Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. Neuroradiology. 2024, 66:73-9. 10.1007/s00234-023-03252-4
6. Toyama Y, Harigai A, Abe M, Nagano M, Kawabata M, Seki Y, Takase K: Performance evaluation of ChatGPT, GPT-4, and Bard on the official board examination of the Japan Radiology Society. Jpn J Radiol. 2024, 42:201-7. 10.1007/s11604-023-01491-2
7. Bhayana R, Krishna S, Bleakney RR: Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology. 2023, 307:e230582. 10.1148/radiol.230582
8. Almeida LC, Farina EM, Kuriki PE, Abdala N, Kitamura FC: Performance of ChatGPT on the Brazilian radiology and diagnostic imaging and mammography board examinations. Radiol Artif Intell. 2024, 6:e230103. 10.1148/ryai.230103
9. Sonoda Y, Kurokawa R, Nakamura Y, et al.: Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in "Diagnosis Please" cases. Jpn J Radiol. 2024, 10.1007/s11604-024-01619-y
10. Oura T, Tatekawa H, Horiuchi D, et al.: Diagnostic accuracy of vision-language models on Japanese diagnostic radiology, nuclear medicine, and interventional radiology specialty board examinations. Jpn J Radiol. 2024, 10.1007/s11604-024-01633-0
11. Kurokawa R, Ohizumi Y, Kanzawa J, et al.: Diagnostic performances of Claude 3 Opus and Claude 3.5 Sonnet from patient history and key images in Radiology's "Diagnosis Please" cases. Jpn J Radiol. 2024, 10.1007/s11604-024-01634-z
12. Suthar PP, Kounsal A, Chhetri L, Saini D, Dua SG: Artificial intelligence (AI) in Radiology: a deep dive into ChatGPT 4.0's accuracy with the American Journal of neuroradiology's (AJNR) "case of the month". Cureus. 2023, 15:e43958. 10.7759/cureus.43958
13. The American Society of Neuroradiology: American Journal of Neuroradiology Case of the Week Archive . (2024). Accessed: October 25, 2024: https://www.ajnr.org/cow/by/year/.
14. The American Society of Neuroradiology: American Journal of Neuroradiology Case of the Week February 9, 2023. (2023). Accessed: October 25, 2024: https://www.ajnr.org/content/cow/02092023/.
15. The American Society of Neuroradiology: American Journal of Neuroradiology Case of the Week October 20, 2022. (2022). Accessed: October 25, 2024: https://www.ajnr.org/content/cow/10202022/.
16. The American Society of Neuroradiology: American Journal of Neuroradiology Case of the Week January 13, 2022. (2022). Accessed: October 25, 2024: https://www.ajnr.org/content/cow/01132022/.
17. Horiuchi D, Tatekawa H, Oura T, et al.: Comparing the diagnostic performance of GPT-4-based ChatGPT, GPT-4V-based ChatGPT, and radiologists in challenging neuroradiology cases. Clin Neuroradiol. 2024, 10.1007/s00062-024-01426-y
18. Bours MJ: Bayes' rule in diagnosis. J Clin Epidemiol. 2021, 131:158-60. 10.1016/j.jclinepi.2020.12.021
19. Burnside ES: Bayesian networks: computer-assisted diagnosis support in radiology . Acad Radiol. 2005, 12:422-30. 10.1016/j.acra.2004.11.030
20. Uy EJ: Key concepts in clinical epidemiology: estimating pre-test probability . J Clin Epidemiol. 2022, 144:198-202. 10.1016/j.jclinepi.2021.10.022
21. Attia JR, Nair BR, Sibbritt DW, et al.: Generating pre-test probabilities: a neglected area in clinical decision making. Med J Aust. 2004, 180:449-54. 10.5694/j.1326-5377.2004.tb06020.x
22. Castillo C, Steffens T, Sim L, Caffery L: The effect of clinical information on radiology reporting: a systematic review. J Med Radiat Sci. 2021, 68:60-74. 10.1002/jmrs.424