Prompt Engineering Strategies Improve the Diagnostic Accuracy of GPT-4 Turbo in Neuroradiology Cases

Akihiko Wada, MD[1], Toshiaki Akashi, MD[1], George Shih, MD[2], Akifumi Hagiwara, MD[1], Mitsuo Nishizawa, MD[1], Yayoi Hayakawa, MD[1], Junko Kikuta, MD[1], Keigo Shimoji, MD[1], Katsuhiro Sano, MD[1], Koji Kamagata, MD[1], Atsushi Nakanishi, MD[1] and Shigeki Aoki, MD[1]

[1]Department of Radiology Juntendo University School of Medicine

[2]Department of Clinical Radiology Cornell Medical College

Corresponding author: Akihiko Wada, MD (a-wada@juntendo.ac.jp)

Abstract

Background: Large language models (LLMs) like GPT-4 demonstrate promising capabilities in medical image analysis, but their practical utility is hindered by substantial misdiagnosis rates ranging from 30-50%.

Purpose: To improve the diagnostic accuracy of GPT-4 Turbo in neuroradiology cases using prompt engineering strategies, thereby reducing misdiagnosis rates.

Materials and Methods: We employed 751 publicly available neuroradiology cases from the American Journal of Neuroradiology Case of the Week Archives. Prompt instructions guided GPT-4 Turbo to analyze clinical and imaging data, generating a list of five candidate diagnoses with confidence levels. Strategies included role adoption as an imaging expert, step-by-step reasoning, and confidence assessment.

Results: Without any adjustments, the baseline accuracy of GPT-4 Turbo was 55.1% to correctly identify the top diagnosis, with a misdiagnosis rate of 29.4%. Considering the five candidates' improved applicability, it is 70.6%. Applying a 90% confidence threshold increased the accuracy of the top diagnosis to 72.9% and the applicability of the five candidates to 85.9%, while reducing misdiagnoses to 14.1%, but limited the analysis to half of cases.

Conclusion: Prompt engineering strategies with confidence level thresholds demonstrated the potential to reduce misdiagnosis rates in neuroradiology cases analyzed by GPT-4 Turbo. This research paves the way for enhancing the feasibility of AI-assisted diagnostic imaging, where AI suggestions can contribute to human decision-making processes. However, the study lacks analysis of real-world clinical data. This highlights the need for further investigation in various specialties and medical modalities to optimize thresholds that balance diagnostic accuracy and practical utility.

Keywords: Neuroradiology, Diagnostic Accuracy, GPT-4 Turbo, Prompt Engineering, AI in Medical Imaging

Introduction

Large language models (LLMs) have shown considerable promise in the processing of textual information, achieving performance levels that approximate human expertise (1–8). These models have been applied in various fields, including medicine, where they offer the potential to assist in the interpretation of complex medical data. In medical imaging diagnostics, LLMs, such as GPT-3.5 and GPT-4, have been explored for their ability to analyze radiological texts. Current reports suggest that these models achieve approximately 50-69% accuracy in identifying correct diagnoses from imaging findings (9–13). However, the 30-50% misdiagnosis rate is relatively high, which may be insufficient for assisting healthcare professionals in making accurate clinical decisions. Physicians generally seek diagnostic tools that minimize misdiagnosis risks to ensure patient safety and effective treatment. This potential and limitations of LLM in medical applications and the need to improve accuracy and reduce misdiagnosis rates in clinical settings (14).

Prompt engineering, a method of giving precise instructions to LLMs, has shown effectiveness in eliciting desired responses and is reported to improve LLMs' performance in various applications (15). In this study, we explore how prompt engineering can improve the diagnostic accuracy of LLM in medical imaging to reduce misdiagnosis rates. In this study, we aim to utilize the latest LLM, GPT-4 Turbo, along with prompt engineering techniques to improve diagnostic abilities in medical imaging, particularly focusing on reducing misdiagnosis rates. By adopting these advanced technologies, we seek to improve the precision of diagnostic suggestions, using the capabilities of GPT-4 Turbo to address the challenges of accurately interpreting medical images.

Materials and Methods

This study was carried out using the checklist for artificial intelligence in medical imaging. It was exempted from institutional review board oversight because it used publicly available data.

Data Collection

Our methodology examined 751 publicly available neuroradiological cases from 2012 to 2023 from the American Journal of Neuroradiology (AJNR) Case of the Week Archives (https://www.ajnr.org/cow/by/diagnosis) (16). The AJNR Case of the Week site separates clinical information, images, and diagnoses/explanations into separate tabs. GPT-4 Turbo accessed the indicated URL to retrieve clinical information and image findings as textual information without knowing the diagnosis name.

AI Model and Platform

We leveraged GPT-4 Turbo, specifically the "gpt-4-1106-preview" model, an advanced version of LLM developed by OpenAI (17). This version, notable for its enhanced capabilities and a vast 128k context window, allows for comprehensive analysis within a single prompt. We use the MD.ai Reporting/Chat application (https://chat.md.ai?) for direct URL access to extract relevant clinical and imaging information, omitting diagnoses to test the diagnostic accuracy of GPT-4 Turbo.

Prompt Instruction

In this study, we used three strategy prompt designs to improve the precision of GPT -4 turbo diagnostic suggestions in neuroradiology cases. These strategies include role adoption, step-by-step thinking, and confidence level assessment (Figure 1). We introduced these strategies based on the knowledge from recent engineering guides and reports (15,18).

Role Adoption

Adopting roles involves directing the LLM to act as an expert in the diagnosis of medical imaging. This strategy aims to shift the LLM's process from data processing to decision-making with domain knowledge. By acting as an expert, the LLM can prioritize relevant information for disease diagnosis from medical images such as CT, MRI, and X-rays. This approach is based on the assumption that by narrowing the LLM's focus to a specific domain, the quality and relevance of its diagnostic suggestions will improve. Role-specific data extraction will lead to higher response precision as the LLM tailors its output to reflect the experience expected of a medical imaging diagnostician.

Step-by-Step Thinking

Step-by-step thinking requires the LLM to take a systematic approach to the diagnostic task, mirroring the logical progression a human expert would take in analyzing clinical information and imaging findings. Our prompt instructs the LLM to suggest an initial differential diagnosis from clinical information alone, then receive imaging findings, and modify the differential diagnosis. This approach encourages the LLM to gradually integrate and refine candidate diagnoses with input data. Such a structured approach could make LLM diagnostic reasoning more transparent and more accessible to interpret, potentially increasing the accuracy and reliability of the suggestions provided.

Confidence assessment

Instructions that ask LLMs to analyze their responses and evaluate their confidence level are called "self-monitoring."(18) LLMs self-assess their confidence based on task difficulty, match with training data, and consistency of their responses. Users can obtain quantitative information about the LLM's list of candidate diagnoses.

The specific text of the instructions for GPT-4 Turbo

The following prompts enabled GPT-4 Turbo to extract textual data of clinical information and imaging findings from the AJNR case of the week site and generate a list of diagnostic candidates and their confidence levels.

 # Role

You are an expert in medical imaging diagnosis with extensive experience interpreting various medical images, including CT, MRI, and X-rays. Your expertise includes identifying pathologies, understanding radiology clinical report contexts, and correlating to imaging findings with potential diagnoses proofread.

 # Request

Along with the following Regulation prompt, present a refined list of five differential diagnoses, including the most probable diagnosis and four alternatives.
Each diagnosis should have a corresponding confidence level based on your comprehensive analysis.

 # Regulation

Using the clinical information provided: {# URL of clinical information}, list five initial differential diagnoses. Then, review the imaging findings: {# URL of image findings}, and update your diagnoses accordingly. Reflect on how the new data alters your assessment. For each diagnosis in your updated list, assign a confidence level between 0% and 100%, considering the task's complexity and the extent to which clinical and imaging data support each diagnosis.

Evaluation of GPT-4 Turbo's Diagnostic Accuracy

Two board-certified neuroradiologists with 15 and 28 years of clinical experience (T.A. and A.W) evaluated the diagnostic suggestions from GPT-4 Turbo. We evaluated the accuracy of GPT-4 Turbo's responses using a three-tier scale: "Excellent" for instances where the top diagnostic suggestion matched the correct diagnosis, "Good" when the correct diagnosis was among the suggested candidates, and "Insufficient" if the correct diagnosis was not listed among the suggested candidates.

Results

Table 1 presents the breakdown of the diagnostic performance of GPT-4 Turbo in 751 cases of neuroradiology. The 'Excellent' category, where the top predicted diagnosis matched the ground truth,

was achieved in 55.1% of the cases. The 'Good' category, where the correct diagnosis was included among the top five predictions, covered an additional 15.5% of cases, totaling 70.6%. However, in 29.4% of cases categorized as 'Insufficient,' the correct diagnosis was not present within GPT-4 Turbo's predictions.

Table 2 illustrates how adjusting the confidence threshold affected diagnostic accuracy and adoption rates. At the default 60% threshold covering all cases, the rate of 'Excellent' predictions remained at 55.1%, while 'Excellent + Good' constituted 70.6% of cases. Crucially, the false positive rate (misdiagnosis) was 29.4%. As shown in Figure 2, increasing the confidence threshold led to improved precision, but a decrease in adoption rates. Setting a strict 90% threshold reduced the misdiagnosis rate to 14.1%, effectively halving diagnostic errors. Concurrently, the 'Excellent' rate rose to 72.9%, though this came at the cost of analyzing only 47% of cases meeting this high-confidence criteria.

In summary, prompt engineering strategies coupled with confidence thresholds demonstrated a trade-off between diagnostic precision and practical utility for GPT-4 Turbo in neuroradiology cases. Higher confidence filters minimized misdiagnosis risks, but limited the proportion of cases the AI system could confidently assess.

Discussion

Improvement in Diagnostic Accuracy Through Prompt Engineering

The baseline performance of GPT-4 Turbo, marked with a 55.1% accuracy rate, is consistent with the results of previous studies in this domain, underscoring the ongoing challenge of improving diagnostic precision in complex medical fields (12). A strategic innovation in this study involved the incorporation of regulation in the prompt design to increase the utility of LLM responses. By shifting from presenting a single diagnostic candidate to offering a set of five, researchers increased the likelihood of including the correct diagnosis within LLM suggestions to 70%, consequently reducing the rate of insufficient responses to 30%.

Risk Reduction through Confidence Threshold

We introduced a novel approach to using confidence thresholds to filter LLM responses, adding an analytical layer. By establishing a high confidence threshold (90% or higher), the probability that LLM provided an inaccurate diagnosis was effectively halved from 29.4% to 14.1%. This highlights the potential of prompt engineering to increase the utility of LLM in clinical settings. The quantitative impact of employing confidence thresholds is substantial, and the number of insufficient responses from LLMs in a dataset of 751 cases is expected to decrease from 221 to 50. However, this strategy reveals a significant trade-off that requires human decision-making in the absence of

LLM suggestions in 53% of cases.

Consideration and Limitations

The efficacy of diagnostic efforts is fundamentally based on human experience. The performance of LLMs in diagnostic tasks is contingent upon several factors: the quality and quantity of clinical information and imaging findings provided, the capability to extract pertinent insights from images, and selecting the most probable diagnosis from the LLM's suggestions. Although setting high confidence thresholds significantly mitigates the risk of incorrect diagnoses, it concurrently limits the LLM's applicability by reducing the number of cases it can address. Additionally, asking medical professionals to consider multiple diagnostic proposals could potentially improve the precision of the diagnosis. However, this approach introduces an additional cognitive load, which can affect clinical workflows and decision-making processes. A limitation of this study is its focus on evaluating precision in scenarios where sufficient information is provided for diagnosis. There is a need for further research on the LLM's performance with real-world data, which may present incomplete information or atypical cases, and in domains beyond neuroradiology.

Future Perspectives

The results of this study suggest avenues for enhancing the accuracy of LLM suggestions:

* Expanding the learning data, especially for tasks where confidence levels were low.
* Implementing few-shot learning as part of prompt engineering to refine the LLM reasoning process by providing it with examples of human thought processes.

Discussing the potential to improve these methods to increase the proportion of high-confidence cases is crucial. Future studies could explore optimizing these strategies to balance diagnostic precision with usability, possibly integrating LLM suggestions more effectively into clinical decision support systems.

Conclusion

Prompt engineering strategies and confidence level thresholds applied to GPT-4 Turbo improve the diagnostic accuracy in neuroradiology, offering a promising avenue for AI-assisted clinical decision-making.

ACKNOWLEDGMENTS

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant [22K07674]. The authors would like to express their gratitude for the financial support provided, which has been instrumental in the advancement of this research.

References

1. Bhayana R. Chatbots and Large Language Models in Radiology: A Practical Primer for Clinical and Research Applications. Radiology. 2024;310(1):e232756. doi: 10.1148/radiol.232756.
2. Jiang S-T, Xu Y-Y, Lu X. ChatGPT in Radiology: Evaluating Proficiencies, Addressing Shortcomings, and Proposing Integrative Approaches for the Future. Radiology. 2023;308(1):e231335. doi: 10.1148/radiol.231335.
3. Biswas S. ChatGPT and the Future of Medical Writing. Radiology. 2023;223312. doi: 10.1148/radiol.223312.
4. Bhayana R, Bleakney RR, Krishna S. GPT-4 in Radiology: Improvements in Advanced Reasoning. Radiology. 2023;307(5):e230987. doi: 10.1148/radiol.230987.
5. Gertz R, Bunck A, Lennartz S, et al. GPT-4 for Automated Determination of Radiological Study and Protocol based on Radiology Request Forms: A Feasibility Study. Radiology. 2023;307 5:e230877. doi: 10.1148/radiol.230877.
6. Fink MA, Bischoff A, Fink CA, et al. Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer. Radiology. 2023;308(3):e231362. doi: 10.1148/radiol.231362.
7. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT. Radiology. 2023;307(4):e230424. doi: 10.1148/radiol.230424.
8. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol. 2023;1–9. doi: 10.1007/s00330-023-10213-1.
9. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations. Radiology. 2023;307(5):e230582. doi: 10.1148/radiol.230582.
10. Ueda D, Mitsuyama Y, Takita H, et al. Diagnostic Performance of ChatGPT from Patient History and Imaging Findings on the Diagnosis Please Quizzes. Radiology. 2023;308(1):e231040. doi: 10.1148/radiol.231040.
11. Kottlors J, Bratke G, Rauen P, et al. Feasibility of Differential Diagnosis Based on Imaging Patterns Using a Large Language Model. Radiology. 2023;308(1):e231167. doi: 10.1148/radiol.231167.
12. 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(1):73–79. doi: 10.1007/s00234-023-03252-4.

13. 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(8):e43958. doi: 10.7759/cureus.43958.
14. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and Other Large Language Models Are Double-edged Swords. Radiology. 2023;307(2):e230163. doi: 10.1148/radiol.230163
15. Prompt engineering - OpenAI API. https://platform.openai.com/docs/guides/prompt-engineering. Accessed 28 March 2024.
16. Case of the Week Diagnoses | American Journal of Neuroradiology. https://www.ajnr.org/cow/by/diagnosis. Accessed March 18, 2024.
17. GPT-4 Turbo in the OpenAI API. https://help.openai.com/en/articles/8555510-gpt-4-turbo. 2023, Accessed 28 March 2024.
18. Ickes W, Holloway R, Stinson LL, Hoodenpyle TG. Self□Monitoring in Social Interaction: The Centrality of Self□Affect. J Pers. 2006;74(3):659–84. doi: 10.1111/j.1467-6494.2006.00388.x

Disease Category	Excellent	Good	Insufficient	Proportion of Total Cases
Tumor	0.442	0.191	0.367	0.29 (215)
Demyelinating Disease	0.727	0.000	0.273	0.01 (11)
Infections and Inflammatory Disease	0.535	0.153	0.313	0.19 (144)
Vascular Disease	0.688	0.150	0.163	0.11 (80)
Genetic/Degenerative Disease	0.515	0.134	0.351	0.13 (97)
Trauma	0.667	0.267	0.067	0.02 (15)
Metabolic Disease	0.735	0.088	0.176	0.09 (68)
Malformation	0.563	0.229	0.208	0.06 (48)
Neurological Disease	0.571	0.000	0.429	0.01 (7)
Other	0.576	0.106	0.318	0.09 (66)
Total	0.551	0.154	0.294	1.00 (751)

Table 1. Breakdown of diagnostic tasks of neuroradiological imaging.

The 'Excellent' category represents cases where the top diagnostic prediction matches the correct diagnosis. 'Good' indicates cases where the correct diagnosis was included within the top five diagnostic predictions. 'Insufficient' marks cases where the correct diagnosis was not included in the diagnostic predictions.

Confidence Threshold	Excellent (%)	Good (%)	Insufficient (%)	Adoption Rate
≥60%	55.1	15.5	29.4	100% (751/751)
≥70%	55.3	15.5	29.2	99% (746/751)
≥80%	57.9	15.8	26.3	92% (689/751)
≥90%	72.9	13.0	14.1	47% (354/751)
=100%	87.5	12.5	0.0	1% (8/751)

Table 2. Diagnostic Accuracy of GPT-4 Turbo at Varying Confidence Thresholds

This table shows the GPT-4 Turbo's diagnostic accuracy and adoption rate at various confidence levels.

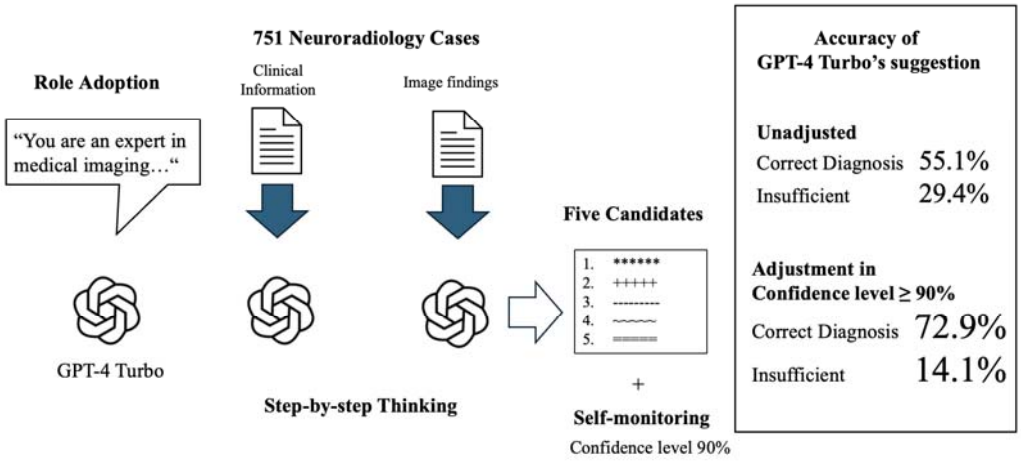

Figure 1: Methodology for prompt engineering to enhance GPT-4 Turbo's utility and reduce misdiagnoses in interpreting 751 neuroradiology cases from the American Journal of Neuroradiology Case of the Week Archives. A role-adoption prompt made GPT-4 Turbo an expert radiologist. GPT-4 Turbo extracted textual clinical information and imaging findings and then used step-by-step thinking to generate a list of 5 diagnostic candidates with confidence levels. The unadjusted performance showed a 55.1% correct diagnosis and 29.4% insufficient suggestions. Adjusting for a confidence level threshold of 90% improved the correct diagnosis to 72.9% but with 14.1% insufficient suggestions on the reduced subset meeting that threshold.

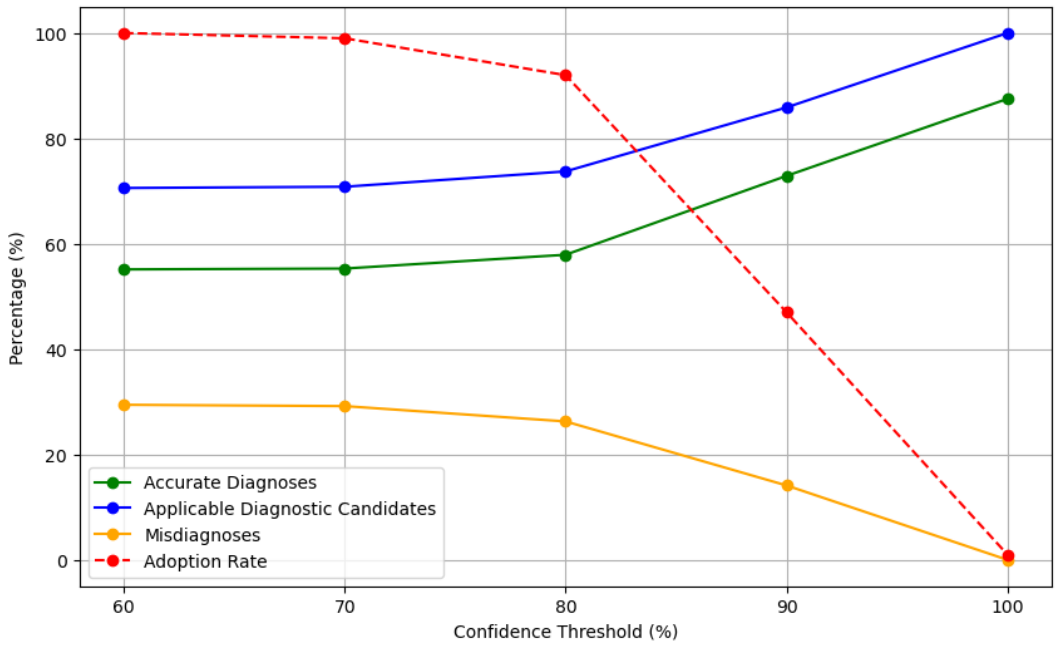

Figure 2. Impact of Confidence Thresholds on GPT-4 Turbo Diagnostic Performance

This graph shows GPT-4 Turbo's diagnostic accuracy across confidence thresholds. The green line shows accurate top suggestions that match the true diagnosis. The blue line indicates the cases with the correct diagnosis among the top five suggestions, suggesting usefulness. The orange line represents misdiagnoses, decreasing with higher confidence thresholds. The red dotted line shows the adoption rate, indicating how often the suggestions were suitable for clinical use. As confidence increases, so does accuracy, but adoption rates decline, highlighting a precision-practicality trade-off.