Turkish Journal of Pediatric Disease

Original Article

ChatGPT-4o's performance in detecting diffusion restriction on brain MRI of neonates with hypoxic-ischemic encephalopathy: Does clinical information influence diagnostic interpretation?

Cemre Özenbaş[1], Burcin İşcan[2]

Department of Radiology, Private Buca Hospital, İzmir Tınaztepe University, İzmir, Türkiye
Department of Pediatrics, Private Buca Hospital, İzmir Tınaztepe University, İzmir, Türkiye

ABSTRACT

Objective: Hypoxic-ischemic encephalopathy is a major cause of neonatal morbidity and mortality. Diffusion-weighted imaging plays a key role in early diagnosis. With the increasing interest in large language models like ChatGPT, it is important to evaluate their potential in radiological interpretation. The aim of this study was to assess the diagnostic performance of ChatGPT-4o in identifying diffusion restriction on neonatal brain Diffusion-weighted imaging (DWI) and to determine whether clinical information (Thompson score) influences its diagnostic responses.

Material and Methods: This retrospective study included 36 neonates (18 with and 18 without diffusion restriction) who underwent brain DWI MRI between postnatal days 4 and 7. For each case, representative DWI and Apparent diffusion coefficient (ADC) images were uploaded to ChatGPT-4o in five separate sessions. The same process was repeated after adding the Thompson score. Performance was evaluated using sensitivity, specificity, Positive predictive value (PPV), Negative predictive value (NPV), Odds ratio (OR), Fleiss Kappa, and McNemar test.

Results: Without clinical information, sensitivity was 56.7%, specificity 90%, PPV 85%, and NPV 67.5% (OR; 11.77). With the Thompson score, sensitivity increased to 72.2%, specificity to 91.1%, PPV to 89%, and NPV to 76.6% (OR; 26.65). Intra-observer agreement was very high (without vs. with Thompson score; κ= 0.825 vs. κ= 0.920). McNemar test showed a statistically significant difference (p=0.045) after clinical data were included.

Conclusion: ChatGPT-4o showed high specificity and moderate sensitivity in detecting diffusion restriction on DWI. Clinical information significantly influenced diagnostic responses, highlighting both the potential and limitations of large language models (LLMs) in radiology.

Keywords: Artificial intelligence, diffusion-Weighted imaging, hypoxic-ischemic encephalopathy, large language models, neonatal MRI

INTRODUCTION

Hypoxic-ischemic encephalopathy (HIE), resulting from perinatal asphyxia, is a significant cause of morbidity and mortality in neonatal intensive care units. The incidence of HIE is approximately 2.5 per 1000 in term births without abnormalities and around 7 per 1000 in preterm births (1,2). Neuroimaging plays a critical role in the diagnosis, prognosis, and management of HIE. While USG, CT, and MRI are all part of the diagnostic imaging, MRI provides superior sensitivity and specificity compared to the others (3). Diffusion-weighted imaging (DWI), in particular, is more sensitive in detecting acute brain injury and enables early diagnosis and timely intervention (4).

Recent progress in large language models (LLMs) such as ChatGPT has generated interest in their potential application to radiology, a field that heavily relies on visual data. LLMs have been the subject of numerous studies in radiology practice, particularly in areas such as clinical decision-making, workflow optimization, structured report generation, and enhancing patient communication (5-9). Although various studies have evaluated the performance of different LLMs, the majority predominantly focus on different versions of ChatGPT, which is widely used in daily practice. Among these models, ChatGPT-4o (GPT-4 Omni) was developed by OpenAI (San Francisco, California, USA) and publicly released on May 13, 2024. While ChatGPT-4o has been investigated for its potential in interpreting radiological images across various domains, no study to date

Correspondence Address:
Cemre Özenbaş
e-mail: cozenbas@hotmail.com

Received : 29.07.2025
Accepted : 06.11.2025
DOI: 10.12956/TJPD.2025.1169

has assessed its performance in detecting diffusion restriction on DWI MRI for the diagnosis of HIE. However, prior research has demonstrated that the inclusion of clinical information can significantly influence the image interpretation of LLMs (10-12).

The aim of our study was to evaluate the performance of ChatGPT-4o in detecting diffusion restriction on neonatal brain DWI MRI, as well as to assess how the inclusion of clinical information influences its image interpretation. In this context, the study also explored the potential impact of integrating LLMs into the radiology workflow and investigated how future developments in this field might affect diagnostic processes.

MATERIALS and METHODS

This retrospective descriptive study included neonates who were admitted to Izmir Tinaztepe University Private Buca Hospital tertiary-level neonatal intensive care unit between January 2023 and December 2024, received wholebody therapeutic hypothermia due to moderate-to-severe HIE, and had diffusion-weighted MRI scans performed between postnatal days 4 and 7. All MRI scans were acquired using a 1.5 Tesla system (GE Healthcare SIGNA HDe). The Thompson scores obtained from all patients were recorded within the first four hours postnatally.

Total 42 brain diffusion MRI scans from patients diagnosed with moderate-to-severe HIE were initially reviewed. Due to image artifacts that adversely affected image quality, 6 scans were excluded. The final sample consisted of 18 patients with normal MRI findings and 18 patients with diffusion restriction consistent with HIE. To minimize the effect of the small sample size, each image was assessed at five different time points, resulting in a total of 90 diffusion MRI images with normal findings and 90 with diffusion restriction being analyzed.The flow chart of the study is shown in Figure 1.

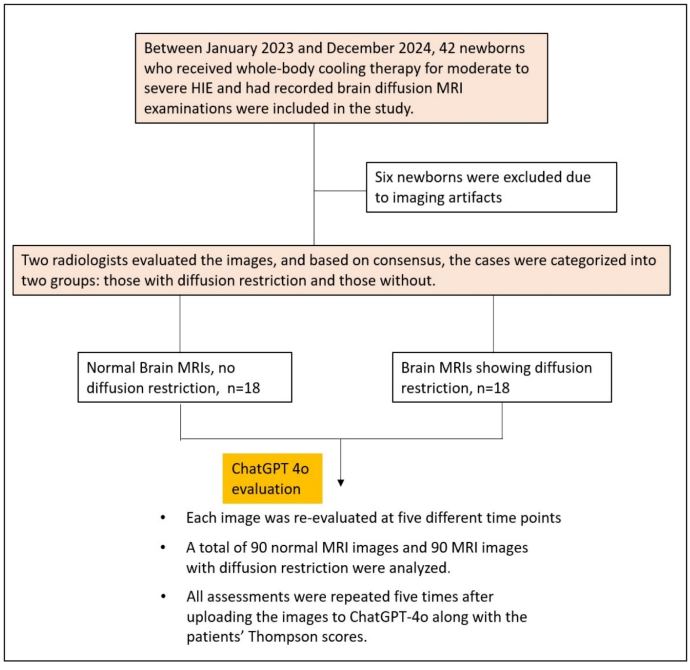

Figure 1: The flowchart of the study.

MRI scans were independently reviewed by two radiologists to determine the presence of diffusion restriction. As both radiologists obtained identical results for all patients, an interobserver agreement analysis was not performed. Based on these evaluations, the infants were divided into two groups: those with diffusion restriction and those without diffusion restriction. Since all patients with diffusion restriction demonstrated lesions in the periventricular white matter, the most representative slices from this region were selected for the diffusion restriction group. To ensure methodological consistency, slices including both lateral ventricles and the periventricular white matter were also selected in the group without diffusion restriction. All images were automatically saved in JPEG format by the hospital's PACS system.

First, the prepared DWI trace images were uploaded to ChatGPT-4o with the question "Can you tell me which MR sequence is this?" to evaluate its ability to recognize the sequence (Figure 2A). Then, the ADC map of the same patient was uploaded with the question "I am sending the ADC map of the same patient. Are there any diffusion restrictions in the images? If so, in which location of the brain?" to assess whether it could detect diffusion restriction (Figure 2B). The following prompt was entered into ChatGPT-4o for the evaluation conducted after providing the clinical information: 'I will send the Thompson score, which is an indicator of the neurological status, along with the diffusion MRI and ADC maps of the newborn patient. Could you please indicate whether there is diffusion restriction?' (Figure 2C, D). For each patient, a separate session was initiated to avoid any influence from previous image assessments. The images of each patient were evaluated five times at one-week intervals. All assessments were repeated five times after uploading the images to ChatGPT-4o along with the patients' Thompson scores. The results obtained with the clinical information included were recorded.

Based on the collected data, sensitivity, specificity, positive predictive value, negative predictive value, and odds ratio were calculated to evaluate the performance of ChatGPT-4o in detecting diffusion restriction. Intra-observer agreement across the five repeated responses provided by ChatGPT-4o for each case was assessed using the Fleiss Kappa method. The McNemar test was used to determine whether there was a significant difference between the responses given before and after the inclusion of the Thompson score. All statistical analyses were performed using IBM SPSS Statistics version 26.

RESULTS

In the first task, which involved identifying the uploaded MR sequence, ChatGPT-4o correctly recognized all sequences. For the evaluation of diffusion restriction, ChatGPT-4o identified 51 out of 90 images with diffusion restriction as true positives, and 81 out of 90 normal images as true negatives. However, 9 normal images were classified as false positives, and 39 images with diffusion restriction were classified as false negatives. Based on the evaluation results, the sensitivity of ChatGPT-4o was found to be 56.7% (CI: 46%–67%). The specificity was determined as 90% (CI: 84%–96%). The positive predictive value (PPV) was 85% (CI: 76%–94%), and the negative predictive value (NPV) was 67.5% (CI: 59%–76%). The calculated odds ratio (OR) was 11.77.

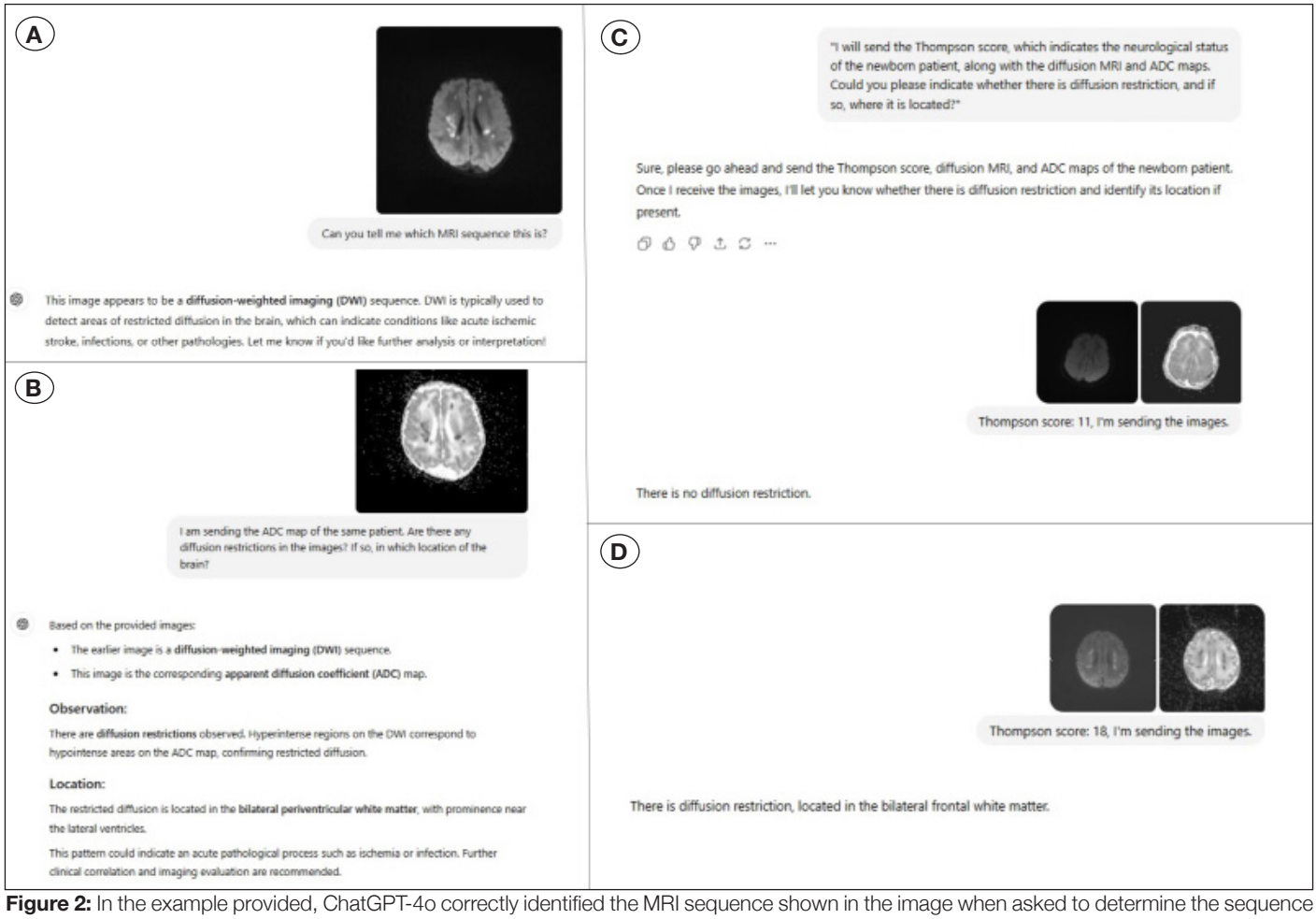

Figure 2: In the example provided, ChatGPT-4o correctly identified the MRI sequence shown in the image when asked to determine the sequence type **(A)**. In the second question directed to ChatGPT-4o, the model correctly identified diffusion restriction in the periventricular white matter **(B)**. Example case without ***(C)*** and with **(D)** diffusion restriction evaluated by ChatGPT-4o using the prompt provided with the Thompson score.

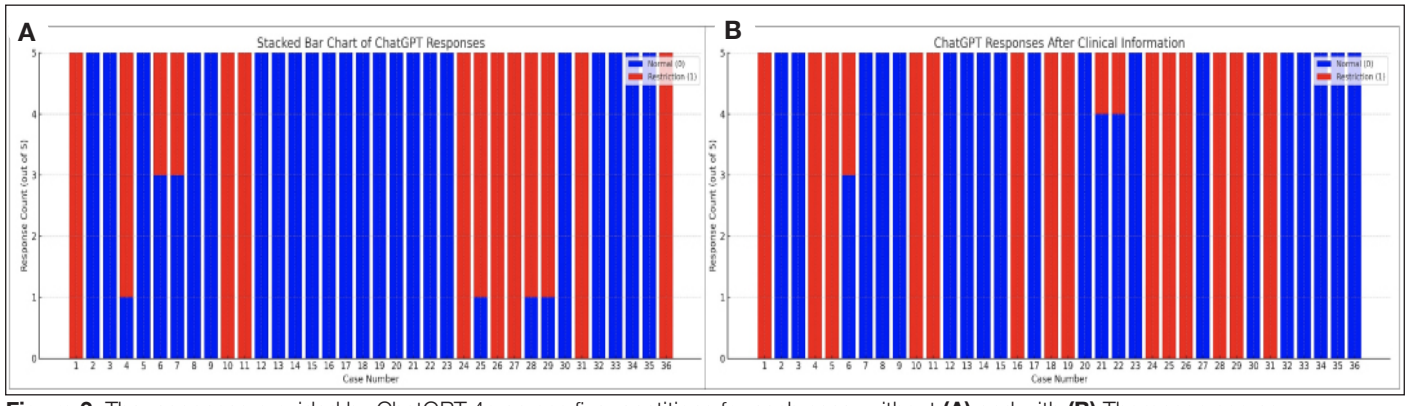

Figure 3: The responses provided by ChatGPT-4o across five repetitions for each case, without **(A)** and with **(B)** Thompson scores.

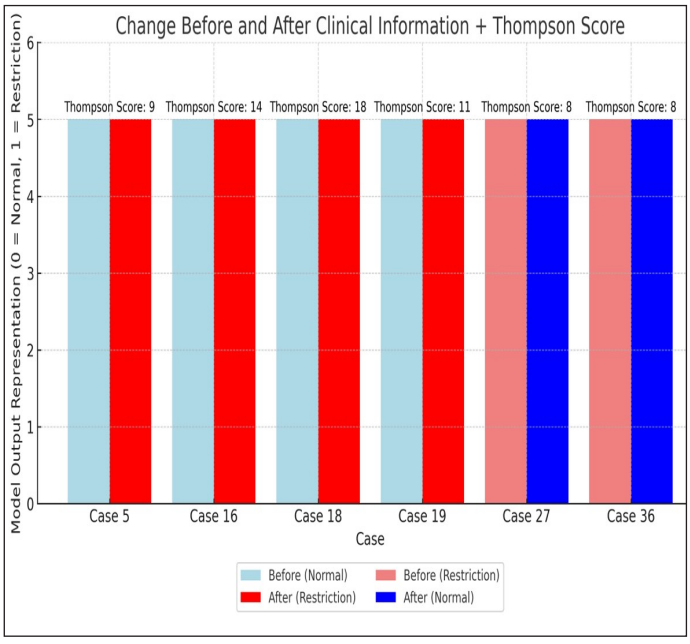

Figure 4: Cases where ChatGPT-4o's interpretation consistently changed in all five repeated evaluations after the inclusion of Thompson scores.

After re-evaluating all images with the addition of Thompson scores, the model's sensitivity was found to be 72.2% (CI: 62%–81%). Specificity was determined as 91.1% (CI: 84%–97%). PPV was 89.0% (CI: 81%–96%), and the NPV was 76.6% (CI: 68%–85%). The calculated OR was 26.65 (Table I).

The responses before and after the inclusion of the Thompson scores were compared using the McNemar test. The responses provided by ChatGPT-4o showed a statistically significant change after clinical information was added (p=0.045).

The intra-observer agreement of ChatGPT-4o was assessed using the Fleiss Kappa test. When evaluations were performed based solely on imaging, the agreement was calculated as 0.825 (95% CI: 0.765–0.885), indicating a very high level of consistency (Figure 3A). After the addition of clinical information in the form of Thompson scores, the agreement further increased to 0.920 (95% CI: 0.880–0.960), reflecting a high level of level of consistency (Figure 3B). Cases with responses that changed after the clinical information was provided are shown in Figure 4.

DISCUSSION

The main findings of this study indicate that ChatGPT-4o demonstrates moderate sensitivity (56.7%) and high specificity (90.0%) in detecting diffusion restriction on DWI MRI images. When clinical information (Thompson score) is incorporated, the model's sensitivity increases significantly to 72.2%, while specificity remains high at 91.1%. These results suggest that ChatGPT-4o is substantially influenced by the clinical context during image interpretation.

Although LLMs such as ChatGPT have only been publicly available for a few years, they have already become the focus of numerous studies in the field of radiology due to the wide variety of imaging modalities and the ease of access to data. In a study by Kuzan et al. (13), which evaluated diffusion restriction in the diagnosis of acute stroke, ChatGPT-4 achieved an accuracy rate of 79.5%. In a study conducted by Zhang et al. (14), aiming to evaluate intracranial hemorrhages on brain CT scans, ChatGPT-4 demonstrated an overall detection accuracy of 79.6% across all hemorrhage types, reaching up to 89% specifically in cases of epidural hematoma. In a study by Ozenbas et al. (15) which evaluated the performance of ChatGPT-4o in identifying MRI characteristics and making differential diagnoses of brain tumors the model achieved accuracy rates ranging from 79.5% to 88.3% for certain characteristic features including signal properties, perilesional edema and contrast enhancement. However its performance was found to be low in determining lesion localization and in distinguishing between intra-axial and extra-axial lesions. In a study conducted by Lacaita et al. (16), ChatGPT-4o detected pathologies with an accuracy of 72% on abdominal X-ray images and 66.1% on chest X-ray images. In recent studies, the relatively high diagnostic accuracy achieved by LLMs is considered promising for the future. The continual emergence of new applications and the ongoing updates of existing ones will inevitably further enhance these success rates. In addition to this positive outlook, there are also studies indicating that the integration of LLMs into clinical practice is not yet realistic. In a study by Hager et al. (17), which included 2400 real patient cases and a dataset covering four common abdominal pathologies, it was demonstrated that large language models performed significantly worse than physicians in making diagnoses, failed to follow diagnostic and treatment algorithms, and were unable to interpret laboratory results accurately. In addition, studies have shown that LLMs can generate incorrect yet convincing information and may produce potentially harmful outcomes such as ethical, legal, and privacy concerns, as well as hallucinations (18,19). However, it is important to remember that the LLMs evaluated in these studies were not originally designed for diagnostic purposes; rather, they were initially developed as text-based language models. In the field of radiology, in addition to AI programs specifically developed for diagnostic purposes, we believe that LLMs, given their wide accessibility and lower financial barriers, may also serve as valuable supportive tools within the diagnostic process.

There are several critical issues, such as bias and prejudice, that must be thoroughly investigated before adapting artificial intelligence and LLMs into medical diagnostic processes. In a study by Zack et al. (20), which evaluated biases in GPT-4's diagnostic decision-making, the authors demonstrated that the model's clinical decisions were influenced by race and gender information, and that it may propagate or even reinforce harmful societal biases. Schmidt et al. (21). demonstrated in their study that, when making diagnostic decisions, ChatGPT was more influenced by the patient's medical history compared

Tablo I: Comparative presentation of ChatGPT-4o's responses based on imaging alone and after the addition of clinical information (Thompson score) in terms of sensitivity, specificity, positive and negative predictive values (PPV, NPV), odds ratio (OR) and Fleiss Kappa coefficient results

	Sensitivity*	Specificity*	PPV*	NPV*	OR	Fleiss Kappa
Image-Based Evaluation	56.7 (46–67)	90.0 (84–96)	85.0 (76–94)	67.5 (59–76)	11.77	0.825
With Thompson Score	72.2 (62–81)	91.1 (84–97)	89.0 (81–96)	76.6 (68–85)	26.65	0.920

***:** *% (CI %)*

to resident physicians. In our study, a significant difference was observed between the responses provided by ChatGPT-4o when diagnosing based solely on imaging and when additional clinical information was provided. Among patients with HIE, in some cases with high Thompson scores indicating greater clinical severity, the model reported the presence of diffusion restriction after receiving clinical information. Conversely, in two patients with low Thompson scores, the model indicated no diffusion restriction following the provision of clinical data. Our findings are consistent with studies showing that clinical context can affect LLM interpretations. For example, Huppertz et al.(11) reported better accuracy for ChatGPT-4 when clinical information was provided, and Horiuchi et al. (10). found that ChatGPT-4 performed better with text-based clinical data than with images alone. Our study adds to this evidence by showing measurable changes after including the Thompson score in neonatal DWI for HIE. Although radiologists themselves may sometimes be influenced by clinical information during image interpretation, the detection of diffusion restriction in our study represents a relatively straightforward diagnostic task. The statistically significant difference observed in this limited patient sample raises important concerns regarding the clinical integration of current LLM versions in diagnostic workflows.

An important factor in the diagnostic process is repeatability. In a study by Frangi et al. (22), which evaluated emergency department triage, a 21% variation was observed across 30 repeated runs for each combination, indicating low repeatability. In a study by Khatri et al. (23) investigating medication-related information queries, the authors found that when the same prompts were tested on different days, ChatGPT-3.5 achieved an accuracy ranging from 10% to 30%, while ChatGPT-4 achieved an accuracy ranging from 40% to 60%, demonstrating limited reproducibility. Similarly, in a study by Krishna et al. (24) using a radiology board-style examination, both ChatGPT-3.5 and ChatGPT-4 demonstrated moderate intrarater agreement. By contrast, Lacaita et al.(16) reported a high level of intraobserver agreement in their study evaluating ChatGPT-4o's performance in interpreting chest and abdominal X-ray images. In our study, a high level of intraobserver agreement was observed both in evaluations based solely on imaging and in those where clinical information was provided. The varying results regarding reproducibility reported in the literature may be attributable to factors such as differences in imaging modalities, the complexity of the diagnostic tasks, the extent of the LLMs' prior exposure to the investigated imaging types, and differences between model versions. From the perspective of integrating LLMs into medical decision support systems, reproducibility must be systematically improved to ensure reliability for both patients and healthcare professionals.

The primary limitations of our study include its single-center and retrospective design. Although each image was evaluated at five different time points to increase the total number of assessments, the relatively small patient population may limit the generalizability of the findings. Another limitation is the repeated-measures design. Since each patient's images were evaluated five times, the McNemar test included non-independent observations, which may have led to an overestimation of statistical power. The study focused exclusively on ChatGPT-4o, a widely used and popular LLM, and did not compare its performance with other LLMs or artificial intelligence applications specifically developed for medical imaging. The results, therefore reflect the performance of this particular version, and it should be noted that future updates or emerging developments may influence these outcomes. Furthermore, although the most demonstrative slices were selected for evaluation, the analysis was performed based on single representative images rather than complete imaging series, which may not fully reflect routine radiological practice.

CONCLUSION

Our study demonstrates that ChatGPT-4o achieved notably high specificity in detecting diffusion restriction on brain DWI scans for the diagnosis of HIE. A key finding of this study is the significant improvement in diagnostic performance following the inclusion of clinical information, which raises important concerns regarding potential bias and prejudice. Despite these concerns, the increasing accessibility and widespread adoption of LLMs suggest that they may hold substantial promise as diagnostic support tools in radiology. However, future studies involving larger and more diverse patient populations, as well as the inclusion of additional influencing factors, are necessary to further evaluate the reliability, objectivity, and clinical utility of LLMs in medical imaging.

Ethics committee approval

This study was conducted in accordance with the Helsinki Declaration Principles. The study was approved by Izmir Tınaztepe University (December 5, 2024, reference number: MUKCE2024/79).

Contribution of the authors

Conception: **OC** - Design: **OC**; **IB** - Supervision: **IB** -Materials: **OC**; **IB** - Data Collection and/or Processing: **OC**; **IB** - Analysis and/or Interpretation: **OC**; **IB** - Literature: **OC**; **IB** - Review: **OC**; **IB** - Writing:

OC- Critical Review: **IB**. All authors reviewed the results and approved the final version of the article.

Source of funding
The authors declare the study received no funding.

Conflict of interest
The authors declare that there is no conflict of interest.

Acknowledgements
We would like to thank Mustafa Tayfun ALTINOK for his support in the evaluation of the imaging data.

REFERENCES

1. Chalak LF, Rollins N, Morriss MC, Brion LP, Heyne R, Sánchez PJ. Perinatal Acidosis and Hypoxic-Ischemic Encephalopathy in Preterm Infants of 33 to 35 Weeks' Gestation. J Pediatr. 2012;160(3):388-94. doi:10.1016/j.jpeds.2011.09.001
2. Graham EM, Ruis KA, Hartman AL, Northington FJ, Fox HE. A systematic review of the role of intrapartum hypoxia-ischemia in the causation of neonatal encephalopathy. Am J Obstet Gynecol. 2008;199(6):587-95. doi:10.1016/j.ajog.2008.06.094
3. Chao CP, Zaleski CG, Patton AC. Neonatal Hypoxic-Ischemic Encephalopathy: Multimodality Imaging Findings. RadioGraphics. 2006;26(suppl_1):S159-72. doi:10.1148/rg.26si065504
4. Liauw L, van Wezel-Meijler G, Veen S, van Buchem MA, van der Grond J. Do Apparent Diffusion Coefficient Measurements Predict Outcome in Children with Neonatal Hypoxic-Ischemic Encephalopathy? AJNR Am J Neuroradiol. 2009;30(2):264-70. doi:10.3174/ajnr.A1318
5. Busch F, Hoffmann L, dos Santos DP, Makowski MR, Saba L, Prucker P, et al. Large language models for structured reporting in radiology: past, present, and future. Eur Radiol. 2024;35(5):2589-602. doi:10.1007/s00330-024-11107-6
6. Russe MF, Fink A, Ngo H, Tran H, Bamberg F, Reisert M, et al. Performance of ChatGPT, human radiologists, and context-aware ChatGPT in identifying AO codes from radiology reports. Sci Rep. 2023;13(1):14215. doi:10.1038/s41598-023-41512-8
7. Temperley HC, O'Sullivan NJ, Mac Curtain BM, Corr A, Meaney JF, Kelly ME, et al. Current applications and future potential of ChatGPT in radiology: A systematic review. J Med Imaging Radiat Oncol. 2024;68(3):257-64. doi:10.1111/1754-9485.13621
8. Keshavarz P, Bagherieh S, Nabipoorashrafi SA, Chalian H, Rahsepar AA, Kim GHJ, et al. ChatGPT in radiology: A systematic review of performance, pitfalls, and future perspectives. Diagn Interv Imaging. 2024;105(7-8):251-65. doi:10.1016/j.diii.2024.04.003
9. Büyüktoka RE, Surucu M, Erekli Derinkaya PB, Adibelli ZH, Salbas A, Koc AM, et al. Applying large language model for automated quality scoring of radiology requisitions using a standardized criteria. Eur Radiol. 2025 Aug 20. doi: 10.1007/s00330-025-11933-2. Epub ahead of print.
10. Horiuchi D, Tatekawa H, Oura T, Shimono T, Walston SL, Takita H, et al. ChatGPT's diagnostic performance based on textual vs. visual information compared to radiologists' diagnostic performance in musculoskeletal radiology. Eur Radiol. 2025;35(1):506-16. doi: 10.1007/s00330-024-10902-5
11. Huppertz MS, Siepmann R, Topp D, Nikoubashman O, Yüksel C, Kuhl CK, et al. Revolution or risk?-Assessing the potential and challenges of GPT-4V in radiologic image interpretation. Eur Radiol. 2025;35(3):1111-21. doi: 10.1007/s00330-024-11115-6
12. 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(1):e231040. doi: 10.1148/radiol.231040.
13. Kuzan BN, Meşe İ, Yaşar S, Kuzan TY. A retrospective evaluation of the potential of ChatGPT in the accurate diagnosis of acute stroke. Diagnostic and Interventional Radiology. Diagn Interv Radiol. 2025;31(3):187-95. doi:10.4274/dir.2024.242892
14. Zhang D, Ma Z, Gong R, Lian L, Li Y, He Z, et al. Using Natural Language Processing (GPT-4) for Computed Tomography Image Analysis of Cerebral Hemorrhages in Radiology: Retrospective Analysis. J Med Internet Res. 2024;26:e58741. doi:10.2196/58741
15. Ozenbas C, Engin D, Altinok T, Akcay E, Aktas U, Tabanli A. ChatGPT-4o's Performance in Brain Tumor Diagnosis and MRI Findings: A Comparative Analysis with Radiologists. Acad Radiol. 2025;32(6):3608-17. doi:10.1016/j.acra.2025.01.033
16. Lacaita PG, Galijasevic M, Swoboda M, Gruber L, Scharll Y, Barbieri F, et al. The Accuracy of ChatGPT-4o in Interpreting Chest and Abdominal X-Ray Images. J Pers Med. 2025;15(5):194. doi:10.3390/jpm15050194
17. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024;30(9):2613-2622. doi: 10.1038/s41591-024-03097-1.
18. Park YJ, Pillai A, Deng J, Guo E, Gupta M, Paget M, Naugler C. Assessing the research landscape and clinical utility of large language models: a scoping review. BMC Med Inform Decis Mak. 2024;24(1):72. doi: 10.1186/s12911-024-02459-6
19. Yu E, Chu X, Zhang W, Meng X, Yang Y, Ji X, Wu C. Large Language Models in Medicine: Applications, Challenges, and Future Directions. Int J Med Sci. 2025 ;22(11):2792-801. doi: 10.7150/ijms.111780
20. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health. 2024;6(1):e12-e22. doi:10.1016/S2589-7500(23)00225-X
21. Schmidt HG, Rotgans JI, Mamede S. Bias Sensitivity in Diagnostic Decision-Making: Comparing ChatGPT with Residents. J Gen Intern Med. 2025;40(4):790-5. doi:10.1007/s11606-024-09177-9
22. Franc JM, Cheng L, Hart A, Hata R, Hertelendy A. Repeatability, reproducibility, and diagnostic accuracy of a commercial large language model (ChatGPT) to perform emergency department triage using the Canadian triage and acuity scale. CJEM. 2024;26(1):40-6. doi:10.1007/s43678-023-00616-w
23. Khatri S, Sengul A, Moon J, Jackevicius CA. Accuracy and reproducibility of ChatGPT responses to real-world drug information questions. JACCP Journal of the American College of Clinical Pharmacy. J Am Coll Clin Pharm. 2025; 8(6): 432-8. doi:10.1002/JAC5.70038
24. Krishna S, Bhambra N, Bleakney R, Bhayana R. Evaluation of Reliability, Repeatability, Robustness, and Confidence of GPT-3.5 and GPT-4 on a Radiology Board–style Examination. Radiology. 2024;311(2): e232715. https://doi.org/10.1148/radiol.232715