DOI: 10.4274/anatoljmed.2026.73645
Anatol J Gen Med Res 2026;36(2):161-70

CLINICAL RESEARCH / KLİNİK ARAŞTIRMA

ANATOLIAN JOURNAL of GENERAL MEDICAL RESEARCH

Performance of Large Language Models in Differentiating High-grade Gliomas and Solitary Brain Metastases

Yüksek Dereceli Gliomlar ile Soliter Beyin Metastazlarının Ayırt Edilmesinde Büyük Dil Modellerinin Performansı

Raşit Eren Büyüktoka[1], Ali Salbas[2], Miran Nihal Otlu[2], Ali Murat Koc[2], Asli Dilara Buyuktoka[2], Aslı Kahraman[3]

[1]University of Health Sciences Türkiye, İzmir City Hospital, Department of Radiology, İzmir, Türkiye
[2]İzmir Katip Çelebi University, İzmir Atatürk Training Research Hospital, Department of Radiology, İzmir, Türkiye
[3]İzmir Katip Çelebi University, İzmir Atatürk Training Research Hospital, Department of Pathology, İzmir, Türkiye

Cite as: Büyüktoka RE, Salbas A, Otlu MN, Koc AM, Buyuktoka AD, Kahraman A. Performance of large language models in differentiating high-grade gliomas and solitary brain metastases. Anatol J Gen Med Res. 2026;36(2):161-70

Abstract

Objective: To evaluate the diagnostic capability of three general-purpose multimodal large language models (LLMs) in differentiating high-grade gliomas from solitary brain metastases (SBMs) on conventional magnetic resonance imaging (MRI) and to compare their performance with radiologists of different expertise.

Methods: In this single-center retrospective study, 109 patients with pathologically confirmed supratentorial high-grade gliomas (n=67) or SBMs (n=42) were included. Axial T2-weighted, pre-contrast T1-weighted, and post-contrast T1-weighted images (single slice per sequence) were interpreted by three LLMs (ChatGPT-4.5, Gemini 2.5 Pro, Claude 4 Opus) and by four radiologists under different reading conditions. Accuracy, precision, recall, and weighted F1-scores were calculated; differences were tested with Cochran's Q and McNemar tests.

Results: Among all readers, diagnostic accuracy ranged from 31.19% to 77.06% across different reading conditions. With all sequences combined, the neuroradiologist achieved the highest accuracy (77.06%); the best-performing LLM was Claude 4 Opus. Combined-sequence evaluation improved accuracy compared with the best single-sequence performance for the neuroradiologist (from 72.48% to 77.06%) and for Claude 4 Opus (from 31.19% on pre-contrast T1WI to 66.06%).

Conclusion: General-purpose LLMs show potential in differentiating high-grade gliomas from SBMs on conventional MRI, but they remain inferior to experienced radiologists. Performance improves when multiple sequences are provided, but the error rate remains high. Future advancements in model architectures and volumetric data integration may enhance their clinical utility.

Keywords: Large language models, glioma, brain neoplasms, magnetic resonance imaging, neuroimaging

Address for Correspondence/Yazışma Adresi: Raşit Eren Büyüktoka, MD, University of Health Sciences Türkiye, İzmir City Hospital, Department of Radiology, İzmir, Türkiye
E-mail: rasiterenbuyuktoka@hotmail.com
ORCID ID: orcid.org/0000-0002-6989-6613

Received/Geliş tarihi: 27.01.2026
Accepted/Kabul tarihi: 23.03.2026
Published date/Yayınlanma tarihi: 31.08.2026

Öz

Amaç: Bu çalışmanın amacı, konvansiyonel manyetik rezonans görüntüleme (MRG) kullanılarak yüksek dereceli gliomlar ile soliter beyin metastazlarının ayırt edilmesinde üç genel amaçlı çok modlu büyük dil modelinin (LLM) tanısal performansını değerlendirmek ve bu performansı farklı deneyim düzeylerine sahip radyologlarla karşılaştırmaktır.

Yöntem: Bu tek merkezli retrospektif çalışmaya, patolojik olarak doğrulanmış supratentoryal yüksek dereceli gliom (n=67) veya soliter beyin metastazı (n=42) tanısı bulunan toplam 109 hasta dahil edildi. Aksiyel T2 ağırlıklı, kontrast öncesi T1 ağırlıklı ve kontrast sonrası T1 ağırlıklı görüntüler (her sekans için tek kesit) üç LLM (ChatGPT-4.5, Gemini 2.5 Pro, Claude 4 Opus) ve farklı okuma koşullarında dört radyolog tarafından değerlendirildi. Doğruluk, kesinlik, duyarlılık ve ağırlıklı F1 skorları hesaplandı; gruplar arası farklar Cochran Q ve McNemar testleri ile analiz edildi.

Bulgular: Tüm değerlendiriciler arasında tanısal doğruluk, farklı okuma koşullarına bağlı olarak %31,19 ile %77,06 arasında değişti. Tüm sekanslar birlikte değerlendirildiğinde, en yüksek doğruluk oranına nöroradyolog ulaştı (%77,06); en iyi performans gösteren LLM ise Claude 4 Opus oldu. Birden fazla sekansın birlikte değerlendirilmesi, en iyi tek sekans performansına kıyasla hem nöroradyolog için (%72,48'den %77,06'ya) hem de Claude 4 Opus için (kontrast öncesi T1 ağırlıklı görüntülerde %31,19'dan %66,06'ya) doğruluğu artırdı.

Sonuç: Genel amaçlı LLMs, konvansiyonel MRG üzerinde yüksek dereceli gliomlar ile soliter beyin metastazlarını ayırt etmede potansiyel göstermektedir; ancak deneyimli radyologların performansının gerisinde kalmaktadır. Birden fazla sekansın birlikte sunulması performansı artırsa da, hata oranları halen yüksektir. Model mimarilerindeki gelişmeler ve volumetrik verilerin entegrasyonu, gelecekte bu modellerin klinik kullanım değerini artırabilir.

Anahtar Kelimeler: Büyük dil modelleri, gliom, beyin neoplazileri, manyetik rezonans görüntüleme, nörogörüntüleme

Introduction

Gliomas and brain metastases are the most frequent neoplastic diseases in neuro-oncology and are associated with high morbidity and mortality rates. Gliomas account for approximately 75% of all primary brain tumors in adults. In contrast, brain metastases represent the most frequently encountered malignant brain neoplasms overall, occurring at a rate nearly ten times higher than that of primary malignant brain tumors[1,2].

Accurate differentiation between solitary brain metastases (SBMs) and high-grade gliomas is important because each entity necessitates distinct therapeutic approaches[3,4]. Because their clinical presentations often overlap, this distinction typically relies on imaging findings. However, the conventional magnetic resonance imaging (MRI) appearances of SBMs and high-grade gliomas can be similar, and accurate early classification is pivotal because of differences in treatment. From a morphologic standpoint, gliomas, especially high-grade gliomas, grow infiltratively, permeating adjacent white-matter tracts, whereas SBMs expand along a pushing border that displaces surrounding parenchyma[5-7]. Multiparametric MRI protocols, particularly those quantitative perfusion metrics such as relative cerebral blood volume (rCBV), improve the differentiation of high-grade gliomas from SBMs; however, access to these advanced techniques remains constrained in many clinical settings[8-10].

The use of artificial intelligence (AI) applications in radiology has increased rapidly over the past decade. However, AI applications, such as deep learning algorithms, require large, carefully curated datasets with diverse and unbiased case distributions; this dependency limits the development of generalizable, high-quality models[11,12]. In contrast, publicly available large language models (LLMs) and their vision-language derivatives have demonstrated capabilities in relatively simple radiological works[13-19].

Whether multimodal LLMs can harness their broad visual-textual priors to distinguish SBMs from high-grade gliomas on conventional MRI remains unexplored. Understanding their performance in complex tasks, such as differentiating SBMs from high-grade gliomas, is critical, as success could enable the integration of automated image interpretation with natural language explanations, reduce the time from diagnosis to treatment, and facilitate rapid, cross-institutional deployment. This study is the first to evaluate the performance of multimodal LLMs in differentiating high-grade gliomas from SBMs.

Accordingly, this study evaluates the diagnostic accuracy and reliability of three publicly available closed-source multimodal LLMs in differentiating SBMs from high-grade gliomas on conventional MRI and benchmarks their performance against radiologists with varying levels of expertise. The primary objective of this study is to evaluate whether state-of-the-art multimodal LLMs can reliably discriminate SBMs from high-grade gliomas using conventional MRI and whether their diagnostic performance surpasses that of radiologists across levels of expertise.

Materials and Methods

This study was conducted as a single-center retrospective analysis. All methods performed in this study were in accordance with national and local laws, institutional guidelines, and ethical standards. The study was approved by İzmir Katip Çelebi University Institutional Ethics Committee (approval no: 0395, date: 19.06.2025). Informed consent was waived by the committee due to the design of the study.

The study design is illustrated in Figure 1. We identified a cohort of patients from our institution's pathology database who had pathologically confirmed supratentorial high-grade gliomas or solitary supratentorial brain metastases. These patients were reviewed to identify those who had undergone cranial MRI within one week prior to their surgical intervention. The primary inclusion criterion for this study was the presence of a solitary, newly diagnosed supratentorial intracranial lesion.

All relevant MRI examinations were retrieved from the institution's picture archiving and communication system. For each included patient, three axial image sets were extracted: T2-weighted, pre-contrast T1-weighted, and post-contrast T1-weighted. Fluid-attenuated inversion recovery (FLAIR) and diffusion weighted imaging were not analyzed because of heterogeneous imaging planes and incomplete coverage.

From each sequence, the single axial slice that demonstrated the most central portion of the lesion and had the largest diameter was selected for analysis. To ensure standardization across the dataset, all selected images were resized to 1024x1024 pixels. Prior to analysis, all patient-identifying information was removed from the images to ensure complete anonymization. The finalized case list was then randomized to create a unique order for interpretation.

Three distinct LLMs (ChatGPT 4.5, Google Gemini 2.5 Pro, and Claude 4-Opus) were used for image analysis. A single prompt was developed and used for all evaluations performed by the LLMs to ensure consistency and comparability of the results. The LLMs analyzed the images under the same conditions as the human readers by evaluating each of the three sequences individually and all sequences combined.

Four radiologists with varying levels of experience [a junior resident (M.N.O.) with 3 years of radiology experience, a senior resident (A.D.B.) with 5 years of radiology experience, a general radiologist (A.S.), and a subspecialist neuroradiologist (A.M.K.)] independently interpreted the images. Image interpretation was performed in four separate reading sessions: (1) T2-weighted images alone, (2) pre-contrast T1-weighted images alone, (3) post-contrast T1-weighted images alone, and (4) all three sequences combined.

To mitigate recall bias, a minimum washout period of two weeks was enforced between consecutive reading sessions for each radiologist. Furthermore, the order of the patient cases for each sequence was re-randomized before each reading session.

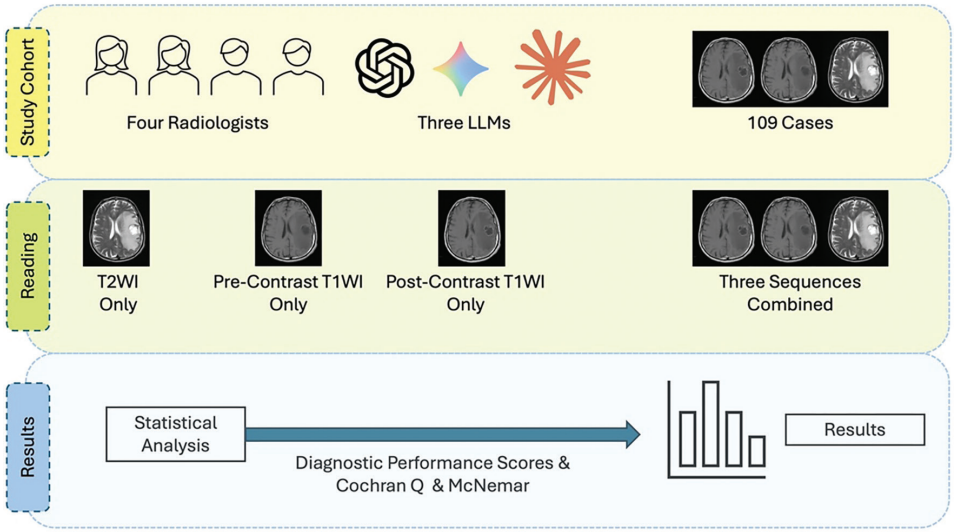

Figure 1. Overview of the study design

LLMs: Large language models

Statistical Analysis

All statistical analyses were performed using IBM SPSS Statistics for Windows, version 22.0 (IBM Corp., Armonk, NY, USA).

Descriptive variables included age, maximum lesion diameter, gender, and World Health Organization grade. The diagnostic performance of each interpreter was evaluated by calculating accuracy, precision, recall, and weighted F1-score. To ensure robust estimation, 95% confidence intervals for each of these metrics were generated using a 1.000-repetition bootstrapping procedure. The diagnostic accuracy of seven interpreters (three LLMs and four radiologists) was compared. Thus, failure to detect a lesion was scored as a diagnostic error, regardless of the ground-truth class. To assess the overall difference in diagnostic accuracy across all interpreters for each imaging condition, Cochran's Q test was employed. In cases where the Cochran's Q test yielded a statistically significant result, post-hoc pairwise comparisons were conducted using the McNemar test. A Bonferroni correction was applied to the significance level for these multiple comparisons to control for type I error.

A p-value of less than 0.05 was considered to indicate statistical significance for all analyses, except where analyses were adjusted for multiple comparisons.

Results

The final study cohort comprised 109 patients. The patient selection process is depicted in the study flowchart (Figure 1).

The median age of the cohort was 58 years (interquartile range, 49-65 years). When stratified by ground-truth diagnosis, 67 (61.47%) cases were categorized as high-grade glial tumors and 42 (38.53%) cases were categorized as SBMs. Baseline demographic and clinical characteristics of the final study cohort are summarized in Table 1.

Diagnostic performance scores for each of the seven interpreters (3 LLMs and 4 radiologists) were summarized in Table 2, and Figures 2 and 3.

During analysis of T2-weighted images, LLMs failed to identify some lesions. Specifically, ChatGPT-4.5 did not identify a lesion in 3 patients (2.75%), Gemini 2.5 Pro did not identify a lesion in 6 patients (5.50%), and Claude 4 Opus did not identify a lesion in 10 patients (9.17%). In contrast, all radiologists successfully identified lesions in every case.

Table 1. Baseline demographic and clinical characteristics of patients with high-grade glioma and solitary brain metastases

Diagnosis	n (%)	Gender	Age (mean ± SD)	Maximum diameter (mm) [median (Q1-Q3)]
High-grade glioma	67 (61.47%)	Male=43 Female=14	53.06±15.24	44.0 (33.5-57.5)
Metastasis	43 (38.53%)	Male=30 Female=13	61.02±10.90	32.5 (24.25-45.75)

SD: Standard deviation

Table 2. Diagnostic performance metrics for each reader across magnetic resonance imaging sequence types. For each reader-sequence pairing, accuracy, precision, recall, and F1 score are reported with their 95% confidence intervals (CIs)

Reader	Sequence type	Accuracy (95% CIs)	Precision (95% CIs)	Recall (95% CIs)	F1 score (95% CIs)
ChatGPT 4.5	Combined	0.58 (0.49-0.66)	0.68 (0.57-0.77)	0.58 (0.49-0.66)	0.57 (0.47-0.66)
	Post-contrast T1WI	0.58 (0.49-0.67)	0.67 (0.57-0.77)	0.58 (0.49-0.67)	0.59 (0.49-0.68)
	Pre-contrast T1WI	0.52 (0.43-0.61)	0.63 (0.53-0.74)	0.52 (0.43-0.61)	0.56 (0.46-0.66)
	T2WI	0.61 (0.51-0.70)	0.59 (0.49-0.71)	0.61 (0.51-0.70)	0.58 (0.47-0.68)

Table 2. Continued

Reader	Sequence type	Accuracy (95% CIs)	Precision (95% CIs)	Recall (95% CIs)	F1 score (95% CIs)
Claude 4 Opus	Combined	0.66 (0.57-0.75)	0.65 (0.56-0.75)	0.66 (0.57-0.75)	0.65 (0.55-0.75)
	Post-contrast T1WI	0.61 (0.51-0.69)	0.61 (0.51-0.71)	0.61 (0.51-0.69)	0.60 (0.50-0.69)
	Pre-contrast T1WI	0.31 (0.23-0.39)	0.58 (0.46-0.70)	0.31 (0.23-0.39)	0.41 (0.30-0.49)
	T2WI	0.52 (0.42-0.61)	0.54 (0.42-0.66)	0.52 (0.42-0.61)	0.51 (0.41-0.61)
Gemini 2.5 Pro	Combined	0.58 (0.49-0.67)	0.58 (0.48-0.68)	0.58 (0.49-0.67)	0.58 (0.48-0.67)
	Post-contrast T1WI	0.58 (0.49-0.67)	0.59 (0.49-0.69)	0.58 (0.49-0.67)	0.58 (0.49-0.67)
	Pre-contrast T1WI	0.48 (0.39-0.57)	0.50 (0.40-0.61)	0.48 (0.39-0.57)	0.49 (0.39-0.58)
	T2WI	0.54 (0.45-0.63)	0.53 (0.43-0.63)	0.54 (0.45-0.63)	0.53 (0.43-0.62)
Neuroradiologist	Combined	0.77 (0.70-0.84)	0.79 (0.71-0.86)	0.77 (0.70-0.84)	0.77 (0.70-0.84)
	Post-contrast T1WI	0.72 (0.64-0.81)	0.74 (0.66-0.83)	0.72 (0.64-0.81)	0.73 (0.65-0.81)
	Pre-contrast T1WI	0.61 (0.51-0.81)	0.66 (0.56-0.75)	0.61 (0.51-0.71)	0.62 (0.53-0.71)
	T2WI	0.72 (0.63-0.80)	0.75 (0.67-0.83)	0.72 (0.63-0.80)	0.72 (0.64-0.80)
General radiologist	Combined	0.74 (0.66-0.82)	0.77 (0.70-0.85)	0.74 (0.66-0.82)	0.75 (0.67-0.82)
	Post-contrast T1WI	0.67 (0.58-0.75)	0.71 (0.63-0.80)	0.67 (0.58-0.75)	0.67 (0.59-0.76)
	Pre-contrast T1WI	0.58 (0.49-0.67)	0.61 (0.52-0.71)	0.58 (0.49-0.67)	0.58 (0.49-0.67)
	T2WI	0.71 (0.62-0.79)	0.74 (0.67-0.82)	0.71 (0.62-0.79)	0.71 (0.63-0.79)
Senior resident	Combined	0.72 (0.64-0.81)	0.74 (0.66-0.82)	0.72 (0.64-0.81)	0.73 (0.64-0.81)
	Post-contrast T1WI	0.67 (0.58-0.76)	0.71 (0.61-0.79)	0.67 (0.58-0.76)	0.67 (0.58-0.76)
	Pre-contrast T1WI	0.56 (0.47-0.65)	0.58 (0.49-0.68)	0.56 (0.47-0.65)	0.57 (0.48-0.65)
	T2WI	0.68 (0.60-0.76)	0.71 (0.63-0.80)	0.68 (0.60-0.76)	0.68 (0.60-0.77)

Table 2. Continued

Reader	Sequence type	Accuracy (95% CIs)	Precision (95% CIs)	Recall (95% CIs)	F1 score (95% CIs)
Junior resident	Combined	0.67 (0.57-0.75)	0.66 (0.56-0.75)	0.67 (0.57-0.75)	0.66 (0.55-0.75)
	Post-contrast T1WI	0.66 (0.57-0.74)	0.65 (0.55-0.74)	0.66 (0.57-0.74)	0.65 (0.55-0.73)
	Pre-contrast T1WI	0.64 (0.56-0.73)	0.65 (0.56-0.74)	0.64 (0.56-0.73)	0.64 (0.56-0.73)
	T2WI	0.74 (0.66-0.83)	0.75 (0.67-0.83)	0.74 (0.66-0.83)	0.74 (0.66-0.83)

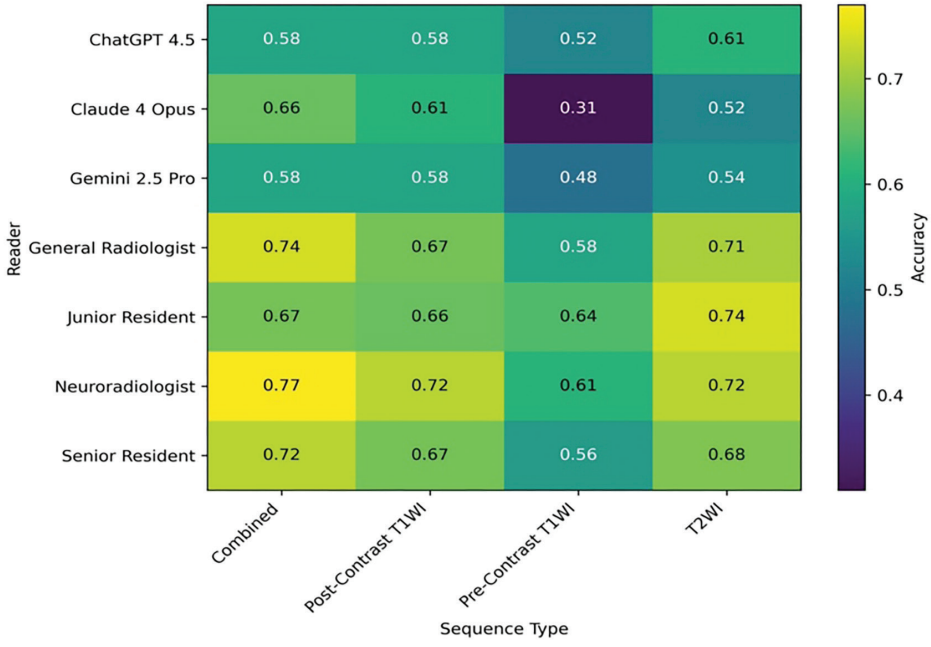

Figure 2. Accuracy heatmap by readers and sequence type. Rows represent the individual readers, including large language models and human radiologists. Columns represent reading types: combined (all sequences provided), post-contrast T1WI, pre-contrast T1WI, and T2WI. Color intensity reflects accuracy values, with lighter shades indicating lower accuracy and darker shades indicating higher accuracy. Numerical values in each cell denote the exact accuracy

Diagnostic accuracy rates for each of the seven readers on T2-weighted images ranged from 52.29% to 71.56%. A statistically significant difference in the correct classification rates among the interpreters was found (p=0.002). However, in subsequent pairwise comparisons, no statistically significant differences were identified between any pair of interpreters (a significance threshold of p<0.0024 was applied for multiple comparisons; all pairs of interpreters had p<0.0024).

LLMs exhibited a higher frequency of non-detected lesions when evaluating pre-contrast T1-weighted images. ChatGPT-4.5 failed to identify lesions in 21 patients (19.27%), Gemini 2.5 Pro failed to identify lesions in 9 patients (8.26%), and Claude 4 Opus failed to identify lesions in 50 patients (45.87%). The diagnostic accuracy of the readers ranged from 31.19% to 61.47%. A significant difference in performance among the interpreters was identified (p<0.001). Post-hoc pairwise comparisons revealed statistically significant differences between the performance of Claude 4 Opus and that of the neuroradiologist (p<0.001),

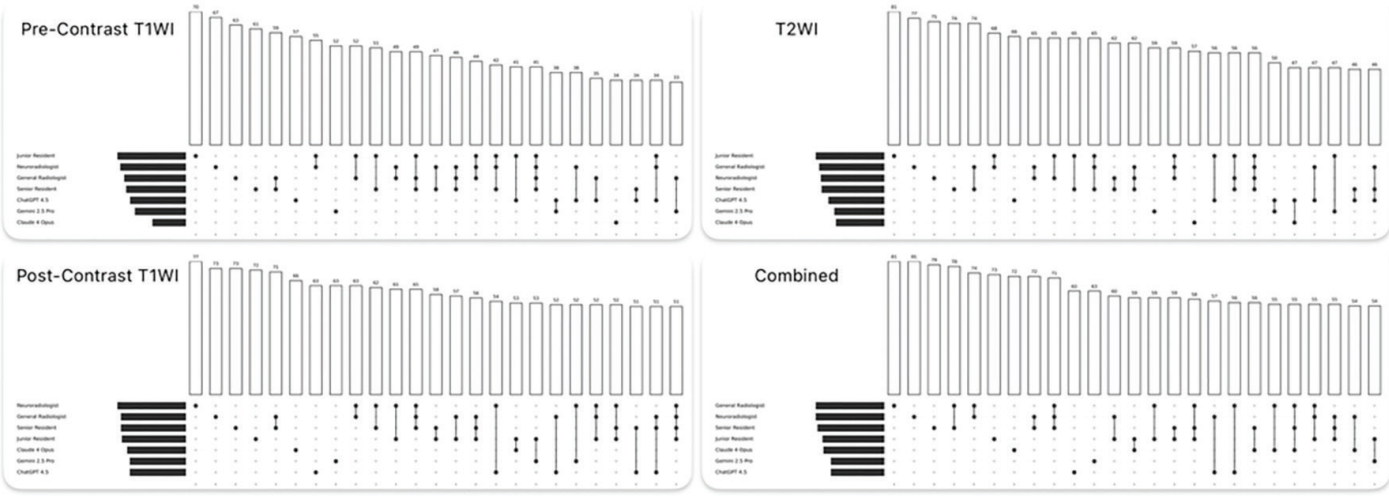

Figure 3. UpSet plot illustrating the overlap of correct classifications across readers and magnetic resonance imaging protocols. Each set corresponds to an individual reader (ChatGPT-4.5, Gemini 2.5 Pro, Claude 4 Opus, junior resident, senior resident, general radiologist, neuroradiologist), and protocol suffixes indicate the evaluated sequences (T2-weighted, pre-contrast T1WI, post-contrast T1WI, combined). Vertical bars represent the number of cases correctly classified in each specific intersection of readers, while filled dots below denote the readers included in that intersection. Horizontal bars display the total number of correct classifications per reader

the general radiologist (p<0.001), and the senior resident (p<0.001). Claude 4 Opus was inferior in diagnostic accuracy to all radiologists. A significant difference was also observed between ChatGPT-4.5 and Claude 4 Opus (p=0.001). No other pairwise comparisons yielded statistically significant results.

For post-contrast T1-weighted images, ChatGPT-4.5, Gemini 2.5 Pro, and Claude 4 Opus failed to identify a lesion in 6 (5.50%), 2 (1.83%), and 3 (2.75%) cases, respectively. Diagnostic performance, expressed as accuracy rates, ranged from 57.80% to 72.48%. A statistically significant difference was found among the interpreters (p=0.037). Nevertheless, subsequent post-hoc pairwise analysis revealed no significant differences among the interpreters (all p>0.0024, Bonferroni threshold for significance). In the analysis using the complete set of MRI sequences, ChatGPT-4.5 failed to identify a lesion in a single case (0.92%), whereas Gemini 2.5 Pro and Claude 4 Opus identified lesions in all cases. Diagnostic accuracy ranged from 57.80% to 77.06%. A statistically significant difference in performance was observed among the readers (p<0.001). Pairwise comparisons demonstrated that the neuroradiologist's performance was significantly superior to that of ChatGPT-4.5 (p<0.001) and Gemini 2.5 Pro (p=0.002). No other pairwise comparisons were statistically significant.

Discussion

In this study, we benchmarked the diagnostic performance of three multimodal LLMs against radiologists of varying expertise in the critical task of differentiating SBMs from high-grade gliomas. Our principal finding is that, while LLMs demonstrate some capability in this classification task, their performance is surpassed by expert human interpreters; a subspecialist neuroradiologist achieved the highest performance, with an accuracy score of 0.77. The best-performing LLM is Claude 4 Opus, which, when provided with all images combined, achieved an accuracy score of 0.66. Except for ChatGPT-4.5, all readers individually perform at their best across all sequences presented during the reading session. ChatGPT-4.5 performs best in a reading session when only T2WI is provided, achieving an accuracy score of 0.61. Another pivotal and concerning finding is the frequent failure of LLMs to identify lesions, particularly on pre-contrast T1WI. Claude 4 Opus missed lesions in 50 cases (45.87%).

The differentiation of high-grade gliomas from SBMs on conventional MRI presents a significant diagnostic challenge due to their overlapping morphological features. While advanced techniques such as perfusion imaging with rCBV can improve diagnostic accuracy, their application is not universal across all clinical settings[1-6]. With the advent of AI,

numerous studies have explored computational methods to distinguish brain metastases from glial tumors, particularly glioblastomas[20-24]. Many of these investigations have relied on radiomics, an approach that necessitates segmentation by expert radiologists for feature extraction. Other studies employing deep learning models, such as the work by Park et al.[24], have demonstrated promising results, achieving an area under the curve (AUC) of 0.83.

To our knowledge, there is no similar study assessing general-purpose LLMs' capabilities for such a complex binary classification task in neuro-oncology. Several attempts have been made to evaluate the performance of LLMs in challenging neuroradiology cases. For instance, Horiuchi et al.[25] compared LLMs with radiologists on complex cases and found that GPT-4 and GPT-4V achieved final diagnostic accuracies of only 22% and 16%, respectively. These results fell short of the performance of both radiology residents, who had accuracy rates between 28% and 31%, and board-certified radiologists, whose accuracy ranged from 38% to 47%[25]. Another investigation evaluating ChatGPT-4o, Grok, and Gemini for brain MRI interpretation found that for pathology prediction, accuracies of different pathologies such as gliomas, metastasis and cavernomas etc. ranged between 13.8% to 57.7%[26]. A separate analysis of ChatGPT-4o's diagnostic ability in diagnosing brain tumors highlighted a disparity between feature recognition and actual diagnosis. While it demonstrated high accuracy in identifying features such as perilesional edema (81%) and contrast enhancement (82.2%), its accuracy for determining the single most likely diagnosis was only 29.5%. This was significantly lower than the 65.9% to 70.5% accuracy achieved by radiologists in the same task[13]. These studies collectively suggest that while LLMs show promise in identifying specific imaging features, they are not yet able to match the diagnostic accuracy of human experts in complex neuroradiological tasks. Similarly, our results suggest that without domain-specific tuning or volumetric context, general-purpose multimodal LLMs do not yet surpass experienced radiologists on this task and may introduce a distinct failure mode.

Another important finding in our study is the failure of LLMs to detect lesions. Claude 4 Opus missed lesions in 50 cases (45.87%). Performance patterns reflected the information content of sequences with respect to this differential. Post-contrast T1WI usually has ring-enhancement and nodular margins typical of metastases and an enhancing tumor core in high-grade gliomas, while T2WI depicts infiltrative T2 hyperintensity and peritumoral edema patterns[6]. LLMs' higher non-detection rates occurred mostly on pre-contrast T1WI, which might be due to the nature of the tumor feature on pre-contrast T1-weighted images.

In the current study, it has been observed that the diagnostic performance of both human readers and LLMs, including ChatGPT-4.5 and a junior radiology resident, improves when a combined set of multiparametric MRI images is provided, as opposed to single sequences. This underscores the importance of a holistic assessment. Specifically, the neuroradiologist's accuracy increased from 72% with single sequences to 77% with combined images, and Claude 4 Opus's accuracy rose from 31% to 66%.

Similar to our findings, prior studies have demonstrated that combining multiple imaging sequences or views can substantially enhance classification performance in various radiological tasks. However, to our knowledge, no prior study has examined whether providing additional MRI sequences enhances the diagnostic performance of general-purpose LLMs. The only similar work to date has involved domain-specific, fine-tuned, or task-specific models. Bai et al.[27] demonstrated with their 3D multimodal large language model that incorporating full volumetric context across spatial perspectives markedly improved performance in visual question answering, retrieval, and segmentation tasks compared with 2D analysis. Furthermore, Kong et al.[28] reported that integrating T1-CE, T2, and T2-fluid-FLAIR sequences in a 3D ResNet-18 framework increased accuracy for glioblastoma vs. solitary brain metastasis discrimination to 87.2%, outperforming any single sequence. Likewise, in chest radiography, Hashir et al.[29] observed that adding lateral views to posteroanterior chest X-rays improved the AUC for various diagnostic labels, with gains comparable to doubling the training set size when using posteroanterior images alone. Collectively, these results suggest that the integration of multi-sequence and multi-view data provides richer contextual and spatial cues that benefit both deep learning architectures and emerging multimodal LLMs, although the magnitude of improvement may vary across models and levels of reader expertise.

Study Limitations

This study has several limitations that should be acknowledged. First, it was conducted at a single center and had a retrospective design, which may limit the generalizability of the findings to other institutions with different patient populations, MRI protocols, or scanner hardware. Second, only three conventional MRI sequences

(axial T2-weighted, axial pre-contrast T1-weighted, and axial post-contrast T1-weighted) and a single representative slice per sequence were provided for analysis. This approach omits volumetric information, advanced sequences, such as perfusion imaging, and multiplanar reconstructions that may carry diagnostic cues. Third, only closed-source, general-purpose, multimodal LLMs were evaluated; the performance of open-source, domain-specific, fine-tuned medical models or of prompt engineering techniques such as few-shot learning was not assessed, and results may differ in such settings. Finally, although lesion detection failures were documented, the closed-source nature of these models meant we could not perform a detailed qualitative error analysis of the underlying failure modes in LLM decision-making, limiting insights into how these models process and prioritize imaging features. Despite these limitations, this study is the first to compare LLMs and radiologists in differentiating high-grade gliomas from SBMs, providing a novel contribution to the literature.

Future research should enhance general-purpose multimodal LLMs by incorporating detection-first pipelines that localize lesions and extract features before classification. While it is not currently possible to provide full 3D volumetric inputs to the general-purpose LLMs used in this study, advances in model architectures may eventually allow multi-slice or volumetric analysis, preserving critical spatial information lost in single-slice evaluation. Future directions might include prompt optimization and fine-tuning of open-source models such as large language-and-vision assistant[30].

Conclusion

In this head-to-head comparison, general-purpose multimodal LLMs demonstrated a measurable, albeit inferior, ability to differentiate high-grade gliomas from SBMs on conventional MRI compared with radiologists, particularly those with subspecialty neuroradiology expertise. While performance improved when multiple MRI sequences were provided, LLMs nonetheless failed because lesions were frequently not detected. These findings underscore that in their current form, LLMs cannot replace expert radiological interpretation for this challenging neuro-oncologic task. However, their potential to augment clinical workflows remains, particularly if future developments enable volumetric image processing, domain-specific fine-tuning, and integration of lesion detection pipelines. Such advancements could narrow the performance gap and support broader, cost-effective deployment of LLM-assisted diagnostic tools in neuroimaging.

Ethics

Ethics Committee Approval: The study was approved by İzmir Katip Çelebi University Institutional Ethics Committee (approval no: 0395, date: 19.06.2025).

Informed Consent: Informed consent was waived by the committee due to the design of the study.

Footnotes

Authorship Contributions

Concept: R.E.B., Design: R.E.B., A.S., A.D.B., Data Collection or Processing: A.S., M.N.O., A.M.K., A.D.B., A.K., Analysis or Interpretation: R.E.B., A.S., Literature Search: R.E.B., Writing: R.E.B.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

References

1. Lamba N, Wen PY, Aizer AA. Epidemiology of brain metastases and leptomeningeal disease. Neuro Oncol. 2021;23:1447-56.
2. Lapointe S, Perry A, Butowski NA. Primary brain tumours in adults. Lancet. 2018;392:432-46.
3. Weller M, van den Bent M, Preusser M, et al. EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood. Nat Rev Clin Oncol. 2021;18:170-86.
4. Vogelbaum MA, Brown PD, Messersmith H, et al. Treatment for brain metastases: ASCO-SNO-ASTRO Guideline. J Clin Oncol. 2022;40:492-516.
5. Suh CH, Kim HS, Jung SC, Kim SJ. Diffusion-weighted imaging and diffusion tensor imaging for differentiating high-grade glioma from solitary brain metastasis: a systematic review and meta-analysis. AJNR Am J Neuroradiol. 2018;39:1208-14.
6. Jung BY, Lee EJ, Bae JM, Choi YJ, Lee EK, Kim DB. Differentiation between glioblastoma and solitary metastasis: morphologic assessment by conventional brain MR imaging and diffusion-weighted imaging. Investig Magn Reson Imaging. 2021;25:23-34.
7. Blanchet L, Krooshof PW, Postma GJ, et al. Discrimination between metastasis and glioblastoma multiforme based on morphometric analysis of MR images. AJNR Am J Neuroradiol. 2011;32:67-73.
8. Suh CH, Kim HS, Jung SC, Choi CG, Kim SJ. Perfusion MRI as a diagnostic biomarker for differentiating glioma from brain metastasis: a systematic review and meta-analysis. Eur Radiol. 2018;28:3819-31.
9. Mouthuy N, Cosnard G, Abarca-Quinones J, Michoux N. Multiparametric magnetic resonance imaging to differentiate high-grade gliomas and brain metastases. J Neuroradiol. 2012;39:301-7.

10. Abreu VS, Tarrio J, Silva J, et al. Multiparametric analysis from dynamic susceptibility contrast-enhanced perfusion MRI to evaluate malignant brain tumors. J Neuroimaging. 2024;34:257-66.
11. Tripathi S, Gabriel K, Dheer S, et al. Understanding biases and disparities in radiology AI datasets: a review. J Am Coll Radiol. 2023;20:836-41.
12. Büyüktoka RE, Adibelli ZH, Sürücü M, et al. Magnetic resonance imaging-based artificial intelligence model predicts neoadjuvant therapy response in triple-negative breast cancer. Diagn Interv Radiol. 2025.
13. 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;32:3608-17.
14. Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. Radiology. 2024;310:e232756.
15. Akinci D'Antonoli T, Stanzione A, Bluethgen C, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. Diagn Interv Radiol. 2024;30:80-90.
16. Elek A, Ekizalioğlu DD, Güler E. Evaluating microsoft bing with ChatGPT-4 for the assessment of abdominal computed tomography and magnetic resonance images. Diagn Interv Radiol. 2025;31:196-205.
17. Çamur E, Cesur T, Güneş YC, et al. Evaluating large language models for imaging modality selection: potential to reduce unnecessary contrast agent use and radiation exposure. Clin Imaging. 2025;125:110573.
18. Singh R, Hamouda M, Chamberlin JH, et al. ChatGPT vs. Gemini: comparative accuracy and efficiency in lung-RADS score assignment from radiology reports. Clin Imaging. 2025;121:110455.
19. Xia X, Wu W, Tan Q, Gou Q. Interpretable machine learning models for differentiating glioblastoma from solitary brain metastasis using radiomics. Acad Radiol. 2025;32:5388-400.
20. Büyüktoka RE, Salbas A. Multimodal large language models for pediatric bone-age assessment: a comparative accuracy analysis. Acad Radiol. 2025;32:6905-12.
21. Qian Z, Li Y, Wang Y, et al. Differentiation of glioblastoma from solitary brain metastases using radiomic machine-learning classifiers. Cancer Lett. 2019;451:128-35.
22. Artzi M, Bressler I, Ben Bashat D. Differentiation between glioblastoma, brain metastasis and subtypes using radiomics analysis. J Magn Reson Imaging. 2019;50:519-28.
23. Li Y, Liu Y, Liang Y, et al. Radiomics can differentiate high-grade glioma from brain metastasis: a systematic review and meta-analysis. Eur Radiol. 2022;32:8039-51.
24. Park YW, Eom S, Kim S, et al. Differentiation of glioblastoma from solitary brain metastasis using deep ensembles: empirical estimation of uncertainty for clinical reliability. Comput Methods Programs Biomed. 2024;254:108288.
25. 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;34:779-87.
26. Sozer A, Sahin MC, Sozer B, et al. Do LLMs have 'the eye' for MRI? Evaluating GPT-4o, Grok, and Gemini on brain MRI performance: First Evaluation of Grok in medical imaging and a comparative analysis. Diagnostics (Basel). 2025;15:1320.
27. Bai F, Du Y, Huang T, Meng MQH, Zhao B. M3D: advancing 3D medical image analysis with multi-modal large language models. arXiv. 2024.
28. Kong C, Yan D, Liu K, Yin Y, Ma C. Multiple deep learning models based on MRI images in discriminating glioblastoma from solitary brain metastases: a multicentre study. BMC Med Imaging. 2025;25:171.
29. Hashir M, Bertrand H, Cohen JP. Quantifying the value of lateral views in deep learning for chest X-rays. arXiv . 2020.
30. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. arXiv. 2023.