Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00225-5

Original Article

Analysis of Large Language Models for Magnetic Resonance Imaging Brain Tumor Segmentation Reporting

Vennila M [1] , Kalaiselvi T [1,]

1. *Department of Computer Science and Applications, The Gandhigram Rural Institute - Deemed to be University, Dindigul, IND*

Received: May 12, 2026 | Review began: June 15, 2026 | Review ended: August 04, 2026 | Published: August 14, 2026

Abstract

This paper builds an end-to-end model that automates the process of analyzing and reporting brain tumors based on multiparametric MRI scans, which is a step toward overcoming the bottleneck of translating specific imaging health records into specific clinical reports in neuroradiology. The pipeline takes post-treatment glioma cases from the Multimodal Brain Tumor Image Segmentation (BraTS) 2024 dataset, which has been divided into 80% for training and 20% for validation, and Swin UNETR is used for 3D segmentation of tumor subregions to improve the identification of the tumor, tumor core, and whole tumor, followed by the extraction of quantitative features of subregion volumes, geometric centroids, and radiomics descriptors. These structured queries prompt three large language models: GPT-5 for higher-level diagnostic reasoning, Claude 4 for safety-conscious, clinically aligned phrasing, and Llama 4 for concise, high-throughput triage reporting. The methodology entails using high-precision segmentation and valid multi-metric assessment, which consists of Dice similarity for visual consistency and BERTScore for semantic consistency, GREEN score for clinical safety, and RadGraph F1 for factual entity-relation overlap. The comparative analysis compares the models on the basis of diagnostic reasoning, liability-conscious documentation, and rapid triage use cases and indicates the task-specific advantages that can be used to direct clinical deployment. The framework bridges the image-to-text gap that has existed in neuroradiology, with multimodal integration providing production-scale artificial intelligence suitable for picture archiving and communication systems and demonstrating that model selection should be based on workflow priorities - depth of analysis versus generation speed. This method creates a clinically translational standard for neuro-oncology reporting, which can provide real-time support for treatment planning with quantitative confidence and informative results.

Categories: AI/ML-based decision support systems, Explainable AI, Natural Language Processing (NLP)

Keywords: generative ai, llm, mri report, gpt-5, claude 4, llama 4

Introduction

Artificial intelligence (AI) in clinical neuro-oncology is on the brink of a paradigm shift, with the emergence of cutting-edge Transformer-based models [1] from traditional Convolutional Neural Networks for automated medical image analysis. The initial volumetric models, such as V-Net, were only capable of medical segmentation due to their local convolutional operations, which were not suitable for processing complex 3D magnetic resonance imaging (MRI) data with long-range global dependencies [2]. These spatial constraints are substantially alleviated by Swin UNEt TRansformers (UNETR), which uses a hierarchical, multi-scale feature extraction Swin Transformer encoder to preserve

high spatial resolution through shifted-window self-attention [3]. Standardized benchmarking programs have made the evaluation of these sophisticated vision models easier, such as the Multimodal Brain Tumor Image Segmentation (BraTS) projects, which created expert-annotated ground truths for heterogeneous tumor regions [4,5].

At the same time, Large Language Models (LLMs) are rapidly evolving and offering a new paradigm of multimodal reasoning and clinical inference [6]. While these individual developments in visual detection and language processing have made strides, there is still a major gap in using these images to produce language or image-to-text. Deep learning systems can produce highly precise segmentation masks, but these are quantitative measures that need to be manually summarized into clinical narratives, which is time-consuming, cognitively demanding, and can lead to inter-observer variability [7]. Besides, medical text generation is just one of the applications where general-purpose LLMs show promise, which warrants stringent assessment with specialized semantic and factual metrics, including BERTScore, RadGraph, and the GREEN score [8-10].

The objective of this work is to fill the gap between volumetric visual identification and narrative clinical record-keeping by proposing an end-to-end framework. This system combines the Swin UNETR model with a powerful LLM to automate the conversion of HFS data into a structured and standardized diagnostic report. The core idea is that proprietary reasoning models will provide the deepest analysis of a patient's condition, architectures designed for safety will facilitate compliance with standardized reporting, and open-source models will provide efficient, privacy-preserving solutions for high-throughput clinical triage.

Literature review

The U-Net architecture introduced by Ronneberger et al. [11] provided the basis for medical image segmentation. This simple model proposed a symmetric contraction route for capturing local contextual information and an expansion route for accurate spatial localization, and introduced a commonly used technique, skip connections, to retain high-frequency features. Building on this architecture, Milletari et al. [2] extended the 2D U-Net concept to the volumetric space by introducing the V-Net, with Punn and Agarwal [12] later providing an extensive survey of the subsequent encoder-decoder designs this enabled for neuro-oncology applications. One of their greatest achievements was the formulation of the Dice loss function, a probabilistic objective function that explicitly addresses the extreme imbalance in class size that is usually found in brain scans, in which pathological tissue represents a small percentage of the volume. Following this approach, Sudre et al. [13] performed a series of tests with generalized versions of these overlap losses, a line of work that more recent studies have since built on, including the improvements of Ma et al. [14] and the self-configuring nnU-Net of Isensee et al. [15], which automatically configures these losses to achieve strong baseline performance on medical data.

Swin UNETR was proposed by Hatamizadeh et al. [3] for mapping global context in a viable way, albeit with its own limitations in the form of local convolutions. It uses a hierarchical Swin Transformer encoder that computes self-attention in shifted windows, therefore emphasizing long-range dependencies across large 3D voxel spaces. The shifting mechanism is key to enabling the model to represent spatial interactions between distant anatomical landmarks, as discussed in the extensive overview of medical transformers by Shamshad et al. [16], and is critical for delineating ambiguous tumor boundaries. El Badaoui et al. [17] also confirmed this in their research on dense semantic labeling of complex gliomas. The first multi-institutional, expert-annotated dataset for the evaluation of necrotic, edematous, and enhancing subregions was introduced through the Multimodal BraTS benchmark. Later, it was extended to include various scanner profiles and protocols to accommodate modern scanners.

Large-scale transformer networks have learned to reason with multiple hops in highly specialized clinical domains for tasks such as language modeling and clinical text generation. Frontier foundation models have been shown to exhibit high-level synthesis capabilities and can be used as effective supporting agents when dealing with structured medical data by converting it into descriptive clinical summaries that can be used for diagnosis. Thirunavukarasu et al. [18] highlighted the promise of such LLMs and the ethical questions that come with their use in a hospital's sensitive and tightly controlled environment.

How to cite this article:
M V, T K (August 14, 2026) Analysis of Large Language Models for Magnetic Resonance Imaging Brain Tumor Segmentation Reporting. Cureus J Comput Sci 3 : es44389-026-00225-5. DOI https://doi.org/10.7759/s44389-026-00225-5

To assess the clinical validity of the created narratives, however, more complex metrics are required than n-gram comparisons. Tools that use contextual embeddings to compute semantic similarity have been well tested for medical radiology reporting and clinical intent. To depict spatial relationships between clinical entities, a system called RadGraph was proposed. The GREEN score framework, which explicitly penalizes hallucinations and omissions, was created to ensure the safety of AI-generated text for patient records when assessing real-world safety. The need for structural and anatomical alignment is also emphasized by Yang et al. [19], who showed that learned knowledge bases can be integrated to drastically reduce hallucinations and increase factual correctness in multimodal radiology report generation. Lastly, Kickingereder et al. [20] demonstrated the critical need to match these automated narratives directly with objective, multicenter radiomic features.

Materials And Methods

Data preprocessing and volumetric normalization

The data include multi-parametric MRI sequences: T1-weighted imaging, T1-weighted contrast-enhanced imaging (T1ce), T2-weighted imaging, and fluid-attenuated inversion recovery (FLAIR). The experimental cohort consists of the post-treatment glioma challenge dataset from the Multimodal BraTS 2024 challenge, with N = 150 subjects completely segmented following a strict protocol. To be methodologically transparent and prevent spatial data leakage, the dataset was divided into a clear 80/20 split (120 training, 30 validation) between an independent training set (n = 120 subjects) and a completely held-out validation set (n = 30 subjects), corresponding to 80% and 20% of the total cohort, respectively. Random stratification was performed using a fixed seed (seed = 3194026842) to ensure reproducibility of the splits. Subjects were eligible if they had a confirmed pathological diagnosis of high-grade glioma (WHO grade III-IV), complete multiparametric magnetic resonance imaging, and validated expert-annotated ground truths for segmentation, including all four core sequences. Conversely, incomplete images, significant motion artifacts, and subject dropout were excluded from the study population. Figure *1* shows a representative subject with the four MRI modalities (T1, T1ce, T2, and FLAIR) and the corresponding ground-truth segmentation. For each tumor, the enhancing tumor (ET), peritumoral edema (ED), and tumor necrotic core (NCR) regions were manually annotated to obtain the ground-truth labels for each subregion, and the location of each subregion was uniformly arranged in a 240 × 240 × 155 voxel canonical grid for all cases.

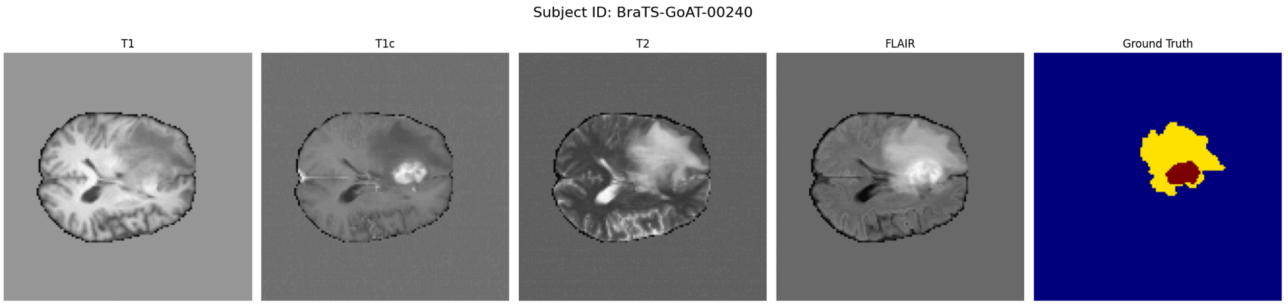


FIGURE 1: Example MRI modalities and ground-truth segmentation from the BraTS 2024 dataset

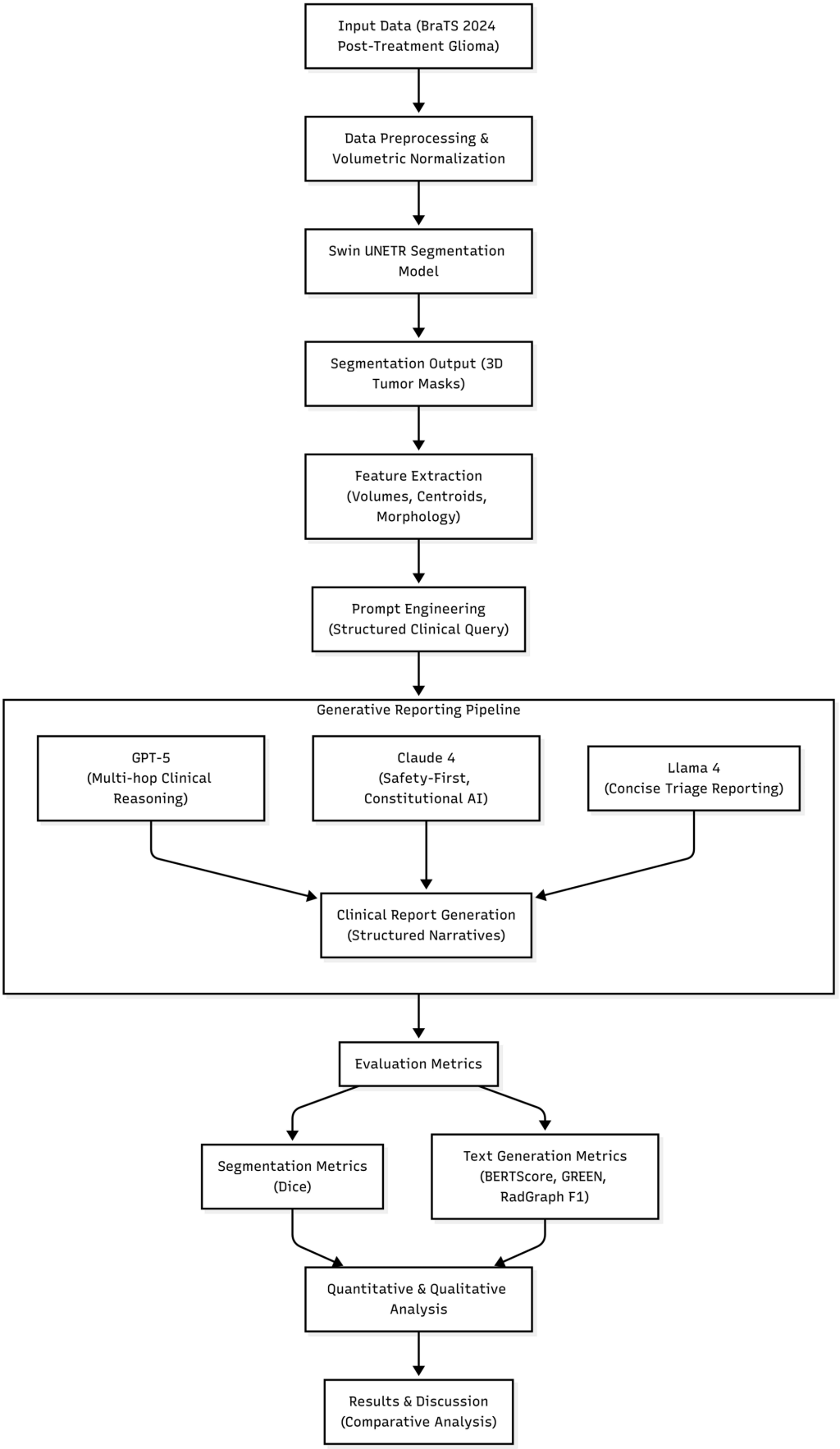
For a representative subject (BraTS-GoAT-00240), the four input modalities are shown: T1-weighted (T1), T1-weighted contrast-enhanced (T1c), T2-weighted (T2), and fluid-attenuated inversion recovery (FLAIR). The Ground-Truth panel shows the expert-annotated tumor subregions.

A detailed preprocessing pipeline was developed to minimize the effects of domain shift and differences between scanners. The voxel spacing of all imaging volumes was kept isotropic at a 1.0 × 1.0 × 1.0 mm^3 voxel size, with trilinear interpolation used to preserve anatomical integrity during isotropic resampling. After that, Z-score normalization was performed with respect to intensity for each imaging modality. The intensities of the voxels were then normalized to:

$$I_{\text{norm}} = \frac{I - \mu_m}{\sigma_m} \quad (1)$$

where I_{norm} is the normalized intensity, I is the intensity for the voxel, and μ_m and σ_m are the mean and standard deviation of the non-zero voxels within the brain mask for each modality m. During the training stage, volumes were resized to a size of 3 × 128 × 128 × 128 using spatial cropping. Space-wise, the amount of data used for validation was kept the same to ensure strict operational consistency. During the training phase, data augmentation was used only to achieve structural and intensity invariance, where each part of the data was randomly flipped along each axis (probability = 0.5) and randomly rotated around each of the three axes (range = ±0.3 radians per axis, probability = 0.4). Gaussian noise was added (standard deviation = 0.1, probability = 0.2). The validation profiles were processed deterministically without augmentation to obtain an unbiased and completely unwarped validation set.

The overall methodology is summarized in Figure *2*. The workflow begins with the BraTS 2024 post-treatment glioma dataset, which is then preprocessed and normalized to a common volume. The Swin UNETR model performs 3D tumor segmentation, followed by feature extraction (volumes, centroids, and morphology). A structured clinical prompt is engineered and passed to three LLMs (GPT-5, Claude 4, and Llama 4) for clinical report generation. The generated reports are evaluated using segmentation metrics (Dice) and text-generation metrics (BERTScore, GREEN score, and RadGraph F1), followed by quantitative and qualitative analyses.



How to cite this article:
M V, T K (August 14, 2026) Analysis of Large Language Models for Magnetic Resonance Imaging Brain Tumor Segmentation Reporting. Cureus J Comput Sci 3 : es44389-026-00225-5. DOI https://doi.org/10.7759/s44389-026-00225-5

FIGURE 2: Methodology flowchart – end-to-end pipeline for brain tumor segmentation and reporting

Architecture: Swin UNETR

The segmentation framework is based on the Swin UNETR architecture that applies the Swin Transformer, a hierarchical encoder, to three-dimensional medical image segmentation. The first step in the input volume is to perform patch partitioning, partitioning into non-overlapping 3D patches of size 2 × 2 × 2 and projecting the inputs to a linear embedding dimension of 48. The architecture adopts Shifted Window-based Multi-Head Self-Attention to learn the long-range spatial dependency without incurring quadratic complexity in computation. The attention mechanism is mathematically represented as:

$$\mathbf{Attention}(Q, K, V) = \mathbf{Softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V \tag{2}$$

where Q is the query matrix, K is the key matrix, V is the value matrix, d is the scaling factor, and B is the relative position bias. The exact spatial resolution of the lost downsampling process is recovered by concatenating feature maps from the different resolutions of the encoder with the pathway of the decoder using residual skip connections.

Optimization objective and loss formulation

A compound loss function was used to optimize the network, minimizing both distribution- and region-based spatial errors. We used Dice loss, L_{Dice}, and cross-entropy loss to define a total loss function, L_{total}, as a weighted sum of these two loss functions:

$$L_{\text{total}} = \lambda_{\text{dice}} L_{\text{Dice}} + \lambda_{\text{ce}} L_{\text{CE}} \tag{3}$$

L_{Dice} is used to maximize the spatial overlap between the predicted probability map (P) and the ground-truth annotations (G). A smoothing factor, ϵ, is added to prevent division by zero and to make the expression numerically stable, giving the following formulation:

$$L_{\text{Dice}} = 1 - \frac{2\sum_i P_i G_i + \epsilon}{\sum_i P_i^2 + \sum_i G_i^2 + \epsilon} \tag{4}$$

At the same time, the Cross-Entropy Loss (L_{CE}) punishes for the wrong classification at each voxel in the probability distribution. The training procedure was performed using the AdamW optimizer with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-5} for network regularization.

Clinical feature extraction and prompt engineering

After the model is inferred, clinical features are extracted to connect the quantitative image analysis and natural language generation steps. A deterministic extractor function is used to transform the resulting segmentation mask into a full-featured vector:

$$V_{\text{feat}} = [V_{\text{ET}}, V_{\text{ED}}, V_{\text{NCR}}, C_{xyz}, \psi_{\text{solidity}}] \tag{5}$$

In the same vector, V represents the computed volumes of different tumor subregions (enhancing tumor (ET), peritumoral edema (ED), and necrotic core (NCR)). In addition, C_{xyz} represents the three-dimensional geometric centroid coordinates of the lesion, and ψ solidity is the shape morphology metric. This is transformed into a structured feature vector, which is then serialized to form a formatted prompt template for use as the clinical context window for the generative language models.

How to cite this article:
M V, T K (August 14, 2026) Analysis of Large Language Models for Magnetic Resonance Imaging Brain Tumor Segmentation Reporting. Cureus J Comput Sci 3 : es44389-026-00225-5. DOI https://doi.org/10.7759/s44389-026-00225-5

Generative reporting pipeline

Three different architectures were considered for the generative reporting pipeline to produce clinical reports from quantitative features extracted from the image, and these architectures were based on transformer networks and large language models. These encompassed a Mixture-of-Experts model, which aimed for multi-hop clinical reasoning (GPT-5); a model built on a safety-first principle that uses Constitutional AI to reduce hallucinatory outputs (Claude 4); and a dense decoder-only foundational model fine-tuned for structured instruction following (Llama 4). BERTScore was used to quantitatively measure the semantic quality of the text generated across the three LLM outputs. This measure computes the cosine similarity between the contextual embedding of the same token in the reports of one model and the corresponding token in the report of another model, allowing for the comparison of semantics between pairs of tokens.

$$R_{\mathrm{BERT}} = \frac{1}{|y|} \sum_{x_j \in y} \max_{\hat{x}_i \in \hat{y}} V(x_j)^T V(\hat{x}_i) \tag{6}$$

where $V(x_j)$ and $V(\hat{x}_i)$ are contextual embedding vectors of the tokens in the two model-generated reports being compared.

Evaluation metrics

To evaluate the model, a hybrid evaluation suite comprising both upstream segmentation accuracy and downstream clinical text generation accuracy was used. The main metric used for 3D volumetric segmentation was the Dice Similarity Coefficient (DSC), which measures the spatial overlap between the predicted segmentation (P) and the ground truth (G):

$$\mathrm{DSC} = \frac{2|P \cap G|}{|P| + |G|} \tag{7}$$

The metrics traditionally used for assessing generative clinical reporting - based on the n-gram approach - were found to be inadequate for this task. Instead, clinical validity was measured using a combination of three advanced metrics. BERTScore was first used to assess semantic fidelity. Second, clinical safety was evaluated using the Generalized Radiology Evaluation (GREEN) score, giving a high penalty for the presence of hallucinations or the lack of relevant biomarkers, and a score above 0.9 was regarded as clinically safe. Last, the RadGraph F1 score was used, which defines the overlap of specific anatomical entities and relations to generate a clinical narrative that is transformed into knowledge graphs to compute the overlap of specific anatomical entities and relations, giving a rigorous measurement of factual accuracy.

Statistical analysis

All experimental data were statistically analyzed for significance. The normality of the metric distributions (BERTScore, GREEN score, and RadGraph F1) was evaluated in 30 independent subjects from the test group using the Shapiro-Wilk test. The metric distribution for Llama-4 was not significantly different from a Gaussian distribution at $\alpha = 0.05$ ($W = 0.942$, $p = 0.0797$). In contrast, the metric distribution for GPT-5 ($W = 0.891$, $p = 0.0001$) and the metric distribution for Claude-4 ($W = 0.905$, $p = 0.0009$) were found to be non-Gaussian. To maintain methodological consistency across all three models, the non-parametric Wilcoxon signed-rank test was used for all pairwise model comparisons. For the evaluation, 30 paired observations were examined for each model pair, strictly at the per-subject level ($n = 30$). The significance level (α) used for all statistical tests was 0.05, with Bonferroni correction applied as appropriate, giving an adjusted significance level of $\alpha_{adj} = 0.0167$ for all pairwise comparisons.

How to cite this article:
M V, T K (August 14, 2026) Analysis of Large Language Models for Magnetic Resonance Imaging Brain Tumor Segmentation Reporting. Cureus J Comput Sci 3 : es44389-026-00225-5. DOI https://doi.org/10.7759/s44389-026-00225-5

Results And Discussion

Qualitative segmentation performance

The proposed framework was tested for volumetric segmentation tasks using the Swin UNETR architecture. The model's segmentation showed strong agreement with the ground-truth annotations in the validation set of the BraTS challenge, with a mean region Dice similarity coefficient of 0.7883. The tumor boundaries were segmented and visually compared with the gold-standard tumor outlines (Figure *3*). This boundary separation enabled quantitative imaging measurements, such as the total computed tumor volume of 28,718 mm^3. The volumetric measurements were then used in the prompt data payload of the downstream generative reporting stage.

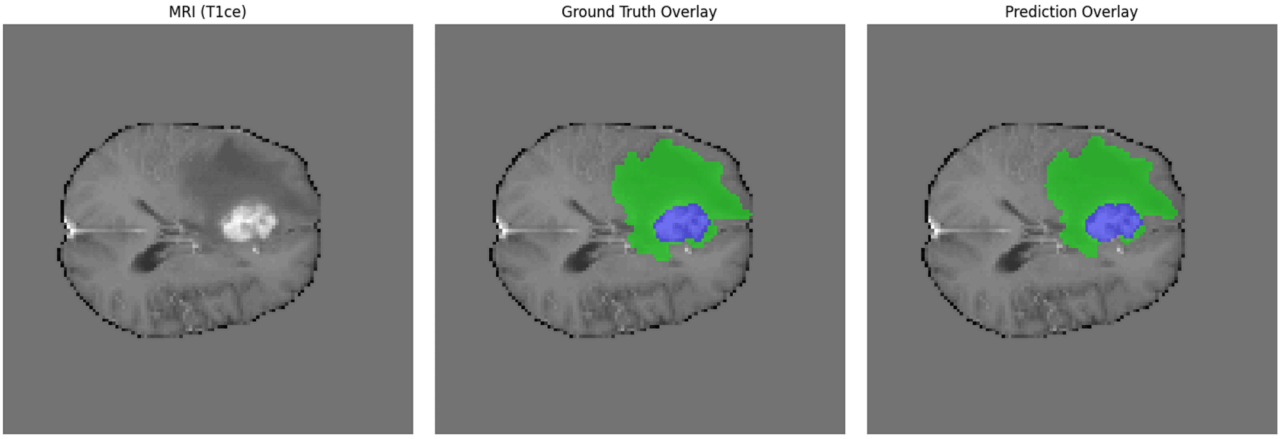


FIGURE 3: Segmentation results compared with ground truth

From left to right: Original T1ce MRI slice; ground-truth annotation overlay; and model prediction overlay on the T1ce image. Green and blue regions represent peritumoral edema and enhancing tumor, respectively. The overlay demonstrates high fidelity in subregion classification.

The per-subregion results above are broadly consistent with those reported by Sheikh et al. [21] for the MICCAI BraTS-25 challenge, who trained on post-operative glioma cases. It should be noted that they also report several evaluation conditions: on their own internal 20% validation split, their Dice scores are 0.903 (Whole Tumor [WT]), 0.761 (Tumor Core [TC]), and 0.764 (ET); however, on their externally validated Synapse platform and BraTS-GLI test-set evaluations, which are more similar in rigor to the present study's held-out evaluation, their lesion-wise scores are 0.775 (WT), 0.735 (TC), and 0.746 (ET). The differences from the present study's WT Dice score of 0.8390 may partly reflect differing treatment of resection cavities between the two datasets. The WT Dice score of 0.8390 in the present study is based on a test-set framing that is more directly comparable and is competitive with that of Sheikh et al. The agreement on ET and TC is also close between the reported ranges of both studies. This suggests that, rather than representing a limitation of the pipeline itself, the segmentation performance in these subregions is aligned with the level of difficulty of the post-operative/post-treatment glioma population as a whole. Among the relatively few published Swin UNETR studies with a similar scope, restricting training and evaluation to post-operative/post-treatment glioma, this agreement supports the assertion that segmentation performance in these three subregions reflects the inherent complexity of the post-operative/post-treatment glioma cohort rather than a limitation of the pipeline itself.

To put this in perspective, when testing Swin UNETR models on general, mostly pre-treatment glioma datasets, they achieve noticeably higher macro-average Dice scores; the difference is therefore more likely attributable to treatment status than to architectural differences, a pattern also consistent with the comparisons summarized in Table *1*. In their 2026 Semi-Swin UNETR study [22], Tian et al. showed that the Dice scores improved from 84.93% to 87.83% as the amount of labeled data increased from 10% to 80%. Jin et al. [23] achieved a mean Dice of 85.89% (±2.7%) on BraTS 2024

using Swin UNETR as a baseline and obtained 88.31% with their own model. The multi-scale enhancement strategy, along with UNETR, was used to create Alhwyji and Kurniawardhani's SwinME model, which achieved a mean validation Dice of 0.91 on BraTS 2023 [24]. In Maurya et al. [25], a combined Swin UNETR/U-Net achieved median Dice scores of 0.8811 (WT) and 0.8754 (ET). In their comparison set, Kim et al. [26] presented a multitask Swin UNETR model for joint segmentation and survival prediction, outperforming other models in terms of segmentation and survival prediction performance. Reduced dependence on manual preprocessing, multitask learning, ensembling with nnU-Net/Swin UNETR, and self-supervised pretraining for small metastatic lesions have been explored in other directions over the last year (Aktar et al. [27]; Satushe et al. [28]; Murmu et al. [29]; Zhang et al. [30]), suggesting a parallel development toward ensembling, multitasking, and reduced reliance on manual preprocessing.

An additional study from 2026 offers more context for direct benchmarking. To investigate whether self-supervised pretraining alone could close the post-training performance gap, Fougang et al. [31] built a Swin UNETR backbone and pretrained it using masked image modeling self-supervision before fine-tuning with the BraTS 2021 and BraTS-Africa 2024 datasets. They reported Dice scores of 0.7987 (ET), 0.8311 (TC), and 0.8518 (WT), which are comparable to, though modestly outperforming, the present single-model results. This suggests that self-supervised pretraining alone can meaningfully narrow the performance gap even without treatment-status-matched fine-tuning data, although the comparison is confounded by Fougang et al.'s cohort not being treatment-matched. It was also shown that post-processing refinements alone boosted the ranking score for the adult glioma task by 0.9% and by 14.9% for the sub-Saharan Africa cohort when applied to pretrained ensembles of existing nnU-Net/Swin UNETR models. This suggests that the remaining gap in the present single-model pipeline might be bridged through post-processing alone, without retraining.

These comparisons suggest that the per-subregion or macro-average Dice scores reported here are similar to those reported for other Swin UNETR models on similarly difficult post-treatment/post-operative cohorts, and are especially close for ET and TC. This indicates that automated preprocessing, cross-validation, ensembling, or multitask training could be evidence-based steps toward improving WT performance specifically. These 2025-2026 benchmarks are compared with the current results in Table *1*.

Study	Year	Architecture	Dataset/Cohort	Dice (ET/TC/WT/Macro)
Proposed Work	2026	Swin UNETR (single model)	BraTS 2024, post-treatment glioma, single split	0.7640/0.7620/0.8390/0.7883
Sheikh et al. [21] (internal validation split)	2026	Swin UNETR (single model)	BraTS-25, post-operative glioma	0.764/0.761/0.903/0.797*
Sheikh et al. [21] (BraTS-GLI test set, lesion-wise)	2026	Swin UNETR (single model)	BraTS-25, post-operative glioma	0.746/0.735/0.775/—
Tian et al. [22]	2026	Swin UNETR (baseline)/Semi-SwinUNeTR	BraTS 2019	—/—/—/0.8783 (80% labeled data)
Jin et al. [23]	2026	Swin UNETR (baseline)	BraTS 2024, general cohort, 5-fold cross-validation	—/—/—/0.8589
Jin et al. [23]	2026	SwinCLNet (proposed)	BraTS 2024	—/—/—/0.8831
Alhwyji and Kurniawardhani [24]	2026	SwinME (Swin V2 + UNETR)	BraTS 2023, validation	—/—/—/0.91
Maurya et al. [25]	2026	Swin UNETR + U-Net	BraTS, longitudinal response assessment	0.8754/—/0.8811/—
Kim et al. [26]	2025	Swin UNETR (multi-task)	BraTS	—/—/—/Survival prediction
Aktar et al. [27]	2026	Swin UNETR + multitask learning	BraTS 2025 Challenge	0.68/0.76/0.73/0.72
Satushe et al. [28]	2026	nnU-Net + Swin UNETR (ensemble)	BraTS-GoAT	0.90/0.87/0.84/—
Murmu et al. [29]	2026	Swin UNETR + Detectron2	Brain tumor imaging	—/—/—/Hybrid model
Zhang et al. [30]	2025	Swin UNETR (self-supervised)	Small brain tumor MRI	—/—/—/Self-supervised pretraining
Fougang et al. [31]	2026	Swin UNETR (self-supervised pretraining)	BraTS 2021 + BraTS-Africa 2024	0.7987/0.8311/0.8518/—

TABLE 1: Comparison of segmentation performance with recently published (2025–2026) studies

Macro-value is the mean of the four reported subregions (ET, TC, WT, and RC) for the post-operative model. RC is only applicable to post-operative cohorts and is not reported separately in this study. Jin et al. [23] SwinCLNet (proposed) macro-average of 0.8831 corresponds to the 88.31% average Dice score reported in the source.

Note: In some studies, the reported macro-averages are not a simple mean of the ET/TC/WT values shown. This reflects either additional subregions in the source authors' own macro-average or different experimental settings; readers are encouraged to consult the primary sources for detailed breakdowns.

ET, Enhancing Tumor; RC, Resection Cavity; TC, Tumor Core; WT, Whole Tumor

Comparative analysis of generative reporting

Table *2* situates these results against recently published comparisons. Results from the automated metrics show a trade-off between clinical depth and computational efficiency (Figure *4*). GPT-5 achieved a semantic similarity (BERTScore) of 0.9133, a GREEN score of 0.92, and a RadGraph F1 score of 0.85. Claude 4 scored 0.89 on the GREEN score and 0.8773 on BERTScore. Much of the difference between the models may be due to the use of qualifying language (e.g., "consistent with") rather than diagnostic statements. The average number of words per report also varies significantly, with Llama 4 generating the shortest reports (24 words) and GPT-5 generating the longest (40 words). The lower scores for Llama 4 on text evaluation metrics that rely on structure (GREEN score of 0.78 and RadGraph F1 of 0.72) could be partially attributed to its shorter outputs.

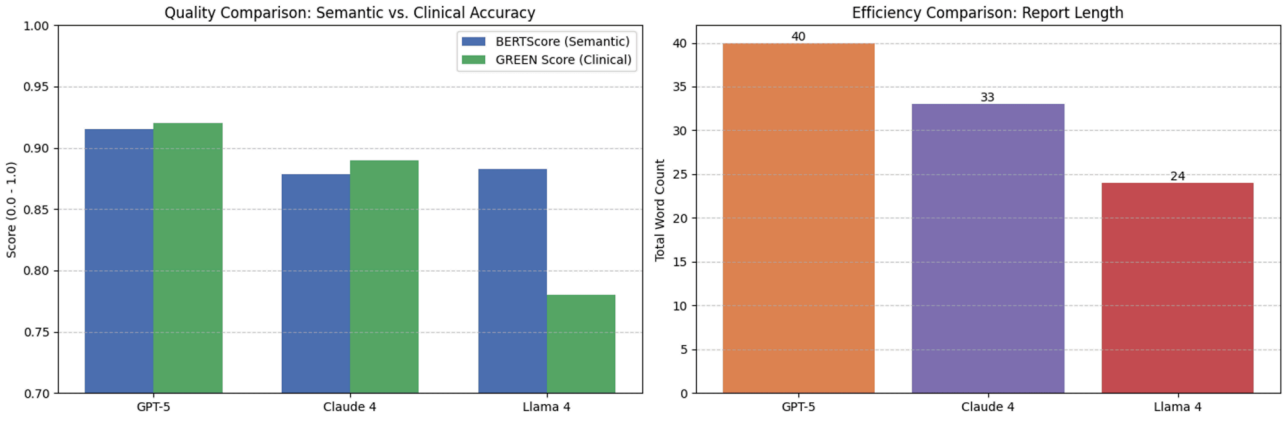


FIGURE 4: (Left) Quality Comparison: GPT-5 outperformed other models in semantic (BERTScore) and clinical (GREEN score) accuracy. (Right) Efficiency Comparison: Llama 4 produced the most succinct reports compared to the other proprietary models

Claude 4 was compared with ChatGPT-4o and Gemini 2.5 Pro on gliomas (the same disease area as this study) for brain MRI sequence classification by Salbas and Buyuktoka [32], with the results showing that Claude 4 Opus performed less well than the other two models. Similarly, Licu et al. [33] compared the generation of cardiac MRI protocols and found that Claude Opus 4 was not the best option. In contrast, Claude 4 achieved the highest accuracy in Takita et al.'s [34] comparison of head CT report structuring. Hahnfeldt et al. [35] evaluated Sonnet 3.5 and GPT-4o for the fully autonomous generation of reports from glioblastoma MRI scans and reported that Sonnet 3.5 achieved higher scores for anatomical region identification (100% compared with GPT-4o's 89.47%) and pathology detection (76.32% compared with 55.26%). Unlike the present work, these studies used different models and task designs: Hahnfeldt et al. [35] evaluated autonomous image-to-report generation using visual LLMs, whereas the current pipeline provides all three LLMs with the

same pre-extracted quantitative features rather than raw images, thereby avoiding the need for visual anatomical recognition. This difference in task design could account for the conflicting results between studies and should be taken into account when extrapolating either result.

In a separate study, Park et al. [36] fine-tuned LLaMA 3.1 on 2.79 million biomedical image-text pairs to predict molecular status and generate reports from images using data from 1,001 glioma patients, achieving 91.3% acceptable report generation according to clinician review. This is directly relevant to Llama 4's current role in the pipeline, as a model fine-tuned specifically on glioma imaging data, rather than a general-purpose triage model, could plausibly have narrowed the gap observed here (GREEN score of 0.78, RadGraph F1 of 0.72).

Wang et al. [37] conducted a reader study to assess the performance of LLMs in generating brain MRI diagnostic impressions, with DeepSeek-R1 demonstrating high performance in this task. BrainFound [38] is a multitask, diffusion-based generative model that integrates image-text contrastive learning for seven neuro-oncology tasks. Nakaura et al. [39] showed that LLMs can serve as assistive tools in the diagnosis of primary brain tumors based on structured and reportable MRI scans. Musthafa et al. [40] combined segmentation-integrated classifiers with LLMs to generate structured and human-readable brain tumor reports. Le Guellec et al. [41] assessed the safety and effectiveness of privacy-protective models (LLaMA 3.3, Athene V2, and Mistral Small) for producing lay summaries of brain MRI reports. Previous studies on AI-based information extraction from free-text radiology reports include Le Guellec et al.'s [42] study on open-source LLM information extraction from free-text radiology reports and Rudnicka et al.'s [43] AI-based segmentation and content-generation algorithms.

Kim et al. [44] assessed four LLMs - GPT-4o, o3-mini, DeepSeek-R1, and Qwen2.5-72B - for the task of producing structured brain MRI protocols using an error-index metric, with lower scores representing higher performance. The results demonstrated that model rankings are highly task-specific, with o3-mini achieving the best performance (base error index: 2.65 ± 1.61; enhanced: 1.94 ± 1.25), followed by GPT-4o (base: 3.11 ± 1.83; enhanced: 2.23 ± 1.48), DeepSeek-R1 (base: 3.42 ± 1.84; enhanced: 2.37 ± 1.53), and Qwen2.5-72B (base: 5.95 ± 2.61; enhanced: 2.75 ± 1.63). This result is consistent with the overall observation that no single LLM performs best on every clinical task, meaning that the choice of model needs to be tailored to the task at hand.

As a whole, this broader literature supports the overall conclusion presented in this study: LLM performance on medical tasks is highly task- and fine-tuning-dependent, and there is no single model that is consistently superior for image interpretation, report structuring, or safety-conscious narrative generation. These comparisons are also summarized in Table *2* alongside the present results.

Study	Year	Task/Basis	GPT Family	Claude Family	Other Model
Proposed Work	2026	GREEN score, brain MRI reports	GPT-5: 0.92	Claude 4: 0.89	Llama 4: 0.78
Proposed Work	2026	BERTScore, brain MRI reports	GPT-5: 0.9133	Claude 4: 0.8773	Llama 4: 0.8832
Salbas and Buyuktoka [32]	2025	Brain MRI sequence classification accuracy	ChatGPT-4o: 97.7%	Claude 4 Opus: 73.1%	Gemini 2.5 Pro: 93.1%
Licu et al. [33]	2026	CMR protocol guideline concordance	—	—	Gemini 2.5 Pro: 71.5% (highest)
Takita et al. [34]	2025	Head CT report structuring accuracy	Lower than Claude ($p<0.0001$)	Highest accuracy	Gemini: lowest
Hahnfeldt et al. [35]	2026	Glioma report generation, anatomical ID accuracy	GPT-4o: 89.47%	Sonnet 3.5: 100%	—
Hahnfeldt et al. [35]	2026	Glioma report generation, pathology detection accuracy	GPT-4o: 55.26%	Sonnet 3.5: 76.32%	—
Park et al. [36]	2025	Glioma report generation (fine-tuned VLM)	—	—	LLaMA 3.1; VLM: 91.3% acceptable
Wang et al. [37]	2026	Diagnostic impression generation	—	—	DeepSeek-R1: effective
Zhang et al. [38]	2026	Report generation (7 tasks)	—	—	Brainfound (multi-modal)
Nakaura et al. [39]	2025	Tumor differentiation	—	—	LLMs as assistive tools
Musthafa et al. [40]	2025	Diagnosis report generation	—	—	LLMs + classifier

Le Guellec et al. [41]	2026	Lay summaries, brain MRI reports	—	—	LLaMA 3.3, Athene V2, Mistral Small
Le Guellec et al. [42]	2024	Information extraction, radiology	—	—	Open-source LLM
Rudnicka et al. [43]	2024	Medical image segmentation and content generation	—	—	Overview
Kim et al. [44]	2026	Brain MRI protocol generation, error index (lower=better)	GPT-4o: 3.11 ± 1.83	—	o3-mini: 2.65 ± 1.61 (best); DeepSeek-R1: 3.42 ± 1.84; Qwen2.5-72B: 5.95 ± 2.61

TABLE 2: A comparison of LLM Report-Generation and Radiology-Task Evaluation with recently published studies

CMR, Cardiovascular Magnetic Resonance; CT, Computed Tomography; MRI, Magnetic Resonance Imaging; VLM, Vision-Language Model

Model output comparison and structural verification

A comparative tracking case study was performed among the three language architectures using a representative, anonymized target volume from the public dataset cohort, with a total computed tumor volume (CTV) of 28,718 mm^3 at coordinates (49, 41, 62). The text generated by the GPT-5 model included all the quantitative elements from the data payload, including explicitly parsing "large enhancing component (3,036 mm^3)" to provide a localized narrative. The Claude 4 model used conventional radiologic terminology and qualifiers, e.g., "irregular borders are seen consistent with glioblastoma." The Llama 4 model produced a linear text configuration that included a typical string to identify the lesion and a baseline recommendation to "correlate clinically." This short format is well suited for rapid data extraction or indexing in high-throughput settings where detailed reasoning is not a requirement of the architecture. These metrics are summarized in Table *3*.

S. No	Metric/Feature	GPT-5 (Proprietary)	Claude 4 (Safety)	Llama 4 (Open Source)
1	Segmentation Model	Swin UNETR	Swin UNETR	Swin UNETR
2	Dice Score (Visual Accuracy)	0.7883	0.7883	0.7883
3	BERTScore (Text Similarity)	0.9133	0.8773	0.8832
4	GREEN score (Clinical Quality)	0.92	0.89	0.78
5	RadGraph F1 (Factuality)	0.85	0.81	0.72
6	Report Style	Detailed & Analytic	Standard Radiological	Concise & Direct
7	Clinical Focus	Diagnosis & Reasoning	Descriptive Safety	Key Findings Extraction
8	Word Count	40	33	24

TABLE 3: Comparative performance metrics of generative reporting models

Statistical performance analysis

A Wilcoxon signed-rank test was used to compare the differences between the three LLMs (GPT-5, Claude 4, and Llama 4) to determine whether there was a statistically significant difference. The non-parametric test was chosen because the evaluation metrics used to assess the models were not normally distributed for GPT-5 ($p = 0.0001$) and Claude 4 ($p = 0.0009$). Although the distribution of the evaluation metrics for the third model, Llama 4, was not significantly different from a normal distribution ($p = 0.0797$), the test was used for methodological consistency across the three models. In total, 30 independent subject evaluations from the validation cohort ($n = 30$) were used for analysis. Thirty paired observations were used for each model pair in the Wilcoxon signed-rank test on the combined clinical text metrics (BERTScore, GREEN score, and RadGraph F1). The three metrics were pooled at the per-subject level before testing; that is, the BERTScore, GREEN score, and RadGraph F1 values for each of the 30 validation subjects were combined into a single paired distribution for each model-pair comparison, rather than each metric being tested independently. This pooling increases statistical power but reduces metric-specific resolution; testing each metric separately is recommended as a sensitivity analysis in future work. To test the statistical differences between all pairs, the Bonferroni-corrected significance level ($\alpha_{adj} = 0.0167$) was used. The comparisons between GPT-5 vs. Claude 4 ($p = 0.008362$), GPT-5 vs. Llama 4 ($p = 0.000183$), and Claude 4 vs. Llama 4 ($p = 0.000419$) were all statistically significant ($p < \alpha_{adj} = 0.0167$), with the latter two achieving $p < 0.001$. These findings indicate that the observed differences in performance between models are unlikely to be due to chance; that is, they reflect differences in the clinical text metrics measured for report generation. Because these metrics assess text-level semantic, safety, and factuality proxies rather than validated clinical outcomes,

this interpretation should be limited to the measured metrics rather than generalized to overall clinical report generation ability. As demonstrated in Table *4*, the performance gap varies significantly between open-source and proprietary models, particularly in the comparisons involving Llama 4 ($p < 0.001$), which is greater than for either of the other model comparisons.

Model Comparison	Statistical Test Used	Sample Size (n)	Calculated p-value	Significance Assessment
GPT-5 vs. Claude 4	Wilcoxon Signed-Rank	30	0.008362	Significant ($p < 0.01$)
GPT-5 vs. Llama 4	Wilcoxon Signed-Rank	30	0.000183	Significant ($p < 0.001$)
Claude 4 vs. Llama 4	Wilcoxon Signed-Rank	30	0.000419	Significant ($p < 0.001$)

TABLE 4: Critical clinical text significance matrix – pairwise statistical comparison of LLM performance

LLM, Large Language Model

Conclusions

The study created and tested an end-to-end multimodal system to automatically analyze brain tumors, addressing a gap in the current neuroradiology workflow by moving from imaging features to textual reports. This proposed pipeline uses the volumetric segmentation capabilities of the Swin UNETR architecture and a powerful LLM to precisely delineate the tumor region and transform the intricate quantitative findings into a clinically actionable narrative report. Among the most salient lessons from our comparative analysis is that there is no one-size-fits-all approach to selecting an artificial intelligence architecture and that the choice needs to be appropriate for the clinical use case. GPT-5 achieved the highest values on the semantic and clinical safety metrics evaluated (BERTScore and GREEN score) in complex case reviews; Claude 4 achieved comparatively strong GREEN scores in generating standardized, documentation-style text; and Llama 4 produced the most concise reports, a property that may be well suited to rapid clinical triage workflows pending further validation. A key constraint of this study is that the evaluation of the generated radiology reports is wholly dependent on automated text metrics, without manual validation and input from practicing radiologists. Further, the validation scheme is limited because it relies on the data configuration from a single center and is assessed only on a local partition of unseen subjects. To overcome these limitations, future research should include human-in-the-loop qualitative grading by medical experts and further validation across independent, external, multicenter clinical datasets. Lastly, in the near future, we aim to enhance our processing workflows by directly integrating and processing the new structural layouts provided by the modern BraTS 2024 dataset, thereby improving the data's compatibility with current neuroradiology imaging practice.

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: Vennila M, Kalaiselvi T

Acquisition, analysis, or interpretation of data: Vennila M, Kalaiselvi T

Drafting of the manuscript: Vennila M, Kalaiselvi T

Critical review of the manuscript for important intellectual content: Vennila M, Kalaiselvi T

Supervision: Kalaiselvi T

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.

Data Availability Statements

The data (and/or code) supporting this study are openly available in BraTS 2024 Challenge on Synapse at https://doi.org/10.7303/syn59059776, reference syn53708249.

Acknowledgements

The authors would like to acknowledge that this paper has been submitted to the International Conference on Artificial Intelligence and Intelligent Computing Systems (IC-AIICS 2026), organized by The Gandhigram Rural Institute, for review and consideration. The authors also express their sincere gratitude to all contributors and institutions that supported this research work.

References

1. Khan RF, Lee BD, Lee MS: Transformers in medical image segmentation: a narrative review. Quantitative Imaging in Medicine and Surgery. 2023, 13:8747-8767. 10.21037/qims-23-542
2. Milletari F, Navab N, Ahmadi SA: V-Net: fully convolutional neural networks for volumetric medical image segmentation. 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA. 2016, 565-571. 10.1109/3DV.2016.79
3. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D: Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Crimi A, Bakas S (ed): Springer, Cham; 2022. 12962:272-284. 10.1007/978-3-031-08999-2_22
4. Menze BH, Jakab A, Bauer S, et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Transactions on Medical Imaging. 2015, 34:1993-2024. 10.1109/tmi.2014.2377694
5. Bakas S, Akbari H, Sotiras A, et al.: Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific Data. 2017, 4:170117. 10.1038/sdata.2017.117
6. Clusmann J, Kolbinger FR, Muti HS, et al.: The future landscape of large language models in medicine. Communications Medicine. 2023, 3:141. 10.1038/s43856-023-00370-1
7. Milad D, Antaki F, Milad J, et al.: Assessing the medical reasoning skills of GPT-4 in complex ophthalmology cases. British Journal of Ophthalmology. 2024, 108:1398-1405. 10.1136/bjo-2023-325053
8. Yu F, Endo M, Krishnan R, et al.: Evaluating progress in automatic chest X-ray radiology report generation. Patterns. 2023, 4:100802. 10.1016/j.patter.2023.100802

9. Jain S, Agrawal A, Saporta A, et al.: RadGraph: Extracting clinical entities and relations from radiology reports. PhysioNet. 2021, RRID:SCR_007345. 10.13026/hm87-5p47

10. Ostmeier S, Xu J, Chen Z, et al.: GREEN: generative radiology report evaluation and error notation. Findings of the Association for Computational Linguistics: EMNLP 2024. 2024, 374-390. 10.18653/v1/2024.findings-emnlp.21

11. Ronneberger O, Fischer P, Brox T: U-Net: convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. Navab N, Hornegger J, Wells W, Frangi A (ed): Springer, Cham; 2015. 9351:234-241. 10.1007/978-3-319-24574-4_28

12. Punn NS, Agarwal S: Modality specific U-Net variants for biomedical image segmentation: a survey. Artificial Intelligence Review. 2022, 55:5845-5889. 10.1007/s10462-022-10152-1

13. Sudre CH, Li W, Vercauteren T, Ourselin S, Cardoso MJ: Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Cardoso MJ, Arbel T, Carneiro G (ed): Springer, Cham; 2017. 10553:240-248. 10.1007/978-3-319-67558-9_28

14. Ma X, Song H, Jia X, Wang Z: An improved V-Net lung nodule segmentation model based on pixel threshold separation and attention mechanism. Scientific Reports. 2024, 14:5412. 10.1038/s41598-024-55178-3

15. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods. 2021, 18:203-211. 10.1038/s41592-020-01008-z

16. Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, Fu H: Transformers in medical imaging: A survey. Medical Image Analysis. 2023, 88:102802. 10.1016/j.media.2023.102802

17. El Badaoui R, Bonmati Coll E, Psarrou A, Asaturyan HA, Villarini B: Enhanced CATBraTS for brain tumour semantic segmentation. Journal of Imaging. 2025, 11:8. 10.3390/jimaging11010008

18. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW: Large language models in medicine. Nature Medicine. 2023, 29:1930-1940. 10.1038/s41591-023-02448-8

19. Yang S, Wu X, Ge S, Zheng Z, Zhou SK, Xiao L: Radiology report generation with a learned knowledge base and multi-modal alignment. Medical Image Analysis. 2023, 86:102798. 10.1016/j.media.2023.102798

20. Kickingereder P, Isensee F, Tursunova I, et al.: Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. The Lancet Oncology. 2019, 20:728-740. 10.1016/s1470-2045(19)30098-1

21. Sheikh MT, Maurya S, Singh A: Segmentation of pre- and post-operative glioma tumors using Swin UNETR and BraTS-25 challenge data. Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. MICCAI 2025. Bakas S, Dennis E, Astaraki M, et al. (ed): Springer, Cham; 2026. 16376:63-73. 10.1007/978-3-032-16365-3_6

22. Tian Y, Wang Z, Guo L: Semi-SwinUNeTR: towards 3D Swin Vision Transformer-based UNet for medical image segmentation with limited annotations. Bioengineering. 2026, 13:695. 10.3390/bioengineering13060695

23. Jin S, Noh Y, Moon H, Lee M, Noh W: SwinCLNet: a robust framework for brain tumor segmentation via shifted window attention and cross-scale fusion. Scientific Reports. 2026, 16:2105. 10.1038/s41598-025-31937-8

24. Alhwyji AMAA, Kurniawardhani A: SwinME: a Swin Transformer V2-based framework for multimodal brain tumor segmentation. 2025 10th International Conference on Information Technology and Digital Application (ICITDA), Yogyakarta, Indonesia. 2025, 1-8. 10.1109/ICITDA68167.2025.11332434

25. Maurya S, Kassaw EA, Sheikh MT, Mehndiratta A, Singh A: Automated brain tumor response assessment from longitudinal multiparametric MRI data using Swin UNETR and a radiomics based classifier. Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. MICCAI 2025. Bakas S, Dennis E, Astaraki M, et al. (ed): Springer, Cham; 2026. 16377:246-257. 10.1007/978-3-032-16370-7_22

26. Kim G, Xing F, Kong HJ, et al.: Overall survival prediction of brain tumor patients with multimodal MRI using Swin UNETR. 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI), Houston, TX, USA. 2025, 1-5. 10.1109/ISBI60581.2025.10981128

27. Aktar M, Nasser T, Souza R: A multitask learning approach for segmenting brain tumor sub-regions: towards better generalization. Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. MICCAI 2025. Bakas S, Dennis E, Astaraki M, et al. (ed): Springer, Cham; 2026. 16376:471-480. 10.1007/978-3-032-16365-3_42

How to cite this article:
M V, T K (August 14, 2026) Analysis of Large Language Models for Magnetic Resonance Imaging Brain Tumor Segmentation Reporting. Cureus J Comput Sci 3 : es44389-026-00225-5. DOI https://doi.org/10.7759/s44389-026-00225-5

28. Satushe V, Arora M, Vyas V, et al.: Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. MICCAI 2025. Bakas S, Dennis E, Astaraki M, et al. (ed): Springer, Cham; 2026. 16376:538-548. 10.1007/978-3-032-16365-3_48

29. Murmu A, Kumar N, Khan SB, Albalawi E, Almusharraf A, Albalawi A: SU-FTD2: transformer-driven brain tumor imaging framework using explainable AI for consumer applications. IEEE Transactions on Consumer Electronics. 2026, 72:4632-4640. 10.1109/TCE.2026.3652504

30. Zhang W-J, Chen W-T, Liu C-H, Chen S-W, Lai Y-H, You SD: Feasibility study of detecting and segmenting small brain tumors in a small MRI dataset with self-supervised learning. Diagnostics. 2025, 15:249. 10.3390/diagnostics15030249

31. Fougang LT, Wacira JM, Jlassi A, Zhang D, Iorumbur A, Raymond C: Segmentation, classification, and synthesis for brain tumors and traumatic brain injuries. Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. MICCAI 2025. Bakas S, Dennis E, Astaraki M, et al. (ed): Springer, Cham; 2026. 16376:349-359. 10.1007/978-3-032-16365-3_32

32. Salbas A, Buyuktoka RE: Performance of large language models in recognizing brain MRI sequences: a comparative analysis of ChatGPT-4o, Claude 4 Opus, and Gemini 2.5 Pro. Diagnostics. 2025, 15:1919. 10.3390/diagnostics15151919

33. Licu RA, Muscogiuri G, Casartelli D, et al.: Reliability of Gemini 2.5 Pro, ChatGPT 4.1, DeepSeek V3, and Claude Opus 4 in generating standardized CMR protocols. European Radiology Experimental. 2026, 10:7. 10.1186/s41747-025-00671-1

34. Takita H, Walston SL, Mitsuyama Y, Watanabe K, Ishimaru S, Ueda D: Comparative performance of large language models in structuring head CT radiology reports: multi-institutional validation study in Japan. Japanese Journal of Radiology. 2025, 43:1445-1455. 10.1007/s11604-025-01799-1

35. Hahnfeldt R, Laukamp KR, Schönfeld MH, et al.: From image to report: Fully AI-generated radiology reports using visual LLMs — A feasibility study on glioma monitoring. European Journal of Radiology Artificial Intelligence. 2026, 5:3050-5771. 10.1016/j.ejrai.2025.100054

36. Park YW, Kang M, Ryu H, et al.: A robust vision language model for molecular status prediction and radiology report generation in adult-type diffuse gliomas. npj Digital Medicine. 2026, 9:406. 10.1038/s41746-026-02581-x

37. Wang ML, Zhang RP, Wu WJ, et al.: Evaluation of large language models for diagnostic impression generation from brain MRI report findings: a multicenter benchmark and reader study. npj Digital Medicine. 2026, 9:187. 10.1038/s41746-026-02380-4

38. Zhang G, Gao Z, Duan C, et al.: A multi-modal foundation model for brain disease diagnosis and medical imaging. Patterns. 2026, 7:2666-3899. 10.1016/j.patter.2026.101538

39. Nakaura T, Uetani H, Yoshida N, et al.: Intra-axial primary brain tumor differentiation: comparing large language models on structured MRI reports vs. radiologists on images. European Radiology. 2026, 36:1594-1604. 10.1007/s00330-025-11924-3

40. Musthafa N, Masud MM, Memon QA, Statsenko Y: Generating human understandable diagnosis report for brain tumors from MRI scans. 2025 IEEE 19th International Conference on Application of Information and Communication Technologies (AICT), Al Ain, United Arab Emirates. 2025, 1-6. 10.1109/AICT67988.2025.11268586

41. Le Guellec B, Bentegeac R, Shorten L, et al.: Safety and efficacy of privacy-preserving models to create Lay summaries of brain MRI reports. Scientific Reports. 2026, 16:6316. 10.1038/s41598-026-36081-5

42. Le Guellec B, Lefèvre A, Geay C, et al.: Performance of an open-source large language model in extracting information from free-text radiology reports. Radiology: Artificial Intelligence. 2024, 6:e230364. 10.1148/ryai.230364

43. Rudnicka Z, Szczepanski J, Pregowska A: Artificial intelligence-based algorithms in medical image scan segmentation and intelligent visual content generation—a concise overview. Electronics. 2024, 13:746. 10.3390/electronics13040746

44. Kim SH, Schramm S, Schmitzer L, et al.: Evaluating large language model-generated brain MRI protocols: performance of GPT4o, o3-mini, DeepSeek-R1 and Qwen2.5-72B. European Radiology. 2026, 36:1644-1655. 10.1007/s00330-025-11989-0

How to cite this article:
M V, T K (August 14, 2026) Analysis of Large Language Models for Magnetic Resonance Imaging Brain Tumor Segmentation Reporting. Cureus J Comput Sci 3 : es44389-026-00225-5. DOI https://doi.org/10.7759/s44389-026-00225-5