npj Digital Medicine

https://doi.org/10.1038/s41746-026-02581-x

Article in Press

A robust vision language model for molecular status prediction and radiology report generation in adult-type diffuse gliomas

Received: 6 December 2025

Accepted: 16 March 2026

Cite this article as: Park, Y.W., 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 Digit. Med.* (2026). https://doi.org/10.1038/s41746-026-02581-x

Yae Won Park, Myeongkyun Kang, Huiseung Ryu, Kyunghwa Han, Yongsik Sim, Ji Eun Park, Jong Hee Chang, Se Hoon Kim, Seung-Koo Lee, Sang Hyun Park & Sung Soo Ahn

A Robust Vision Language Model for Molecular Status Prediction and Radiology Report Generation in Adult Type Diffuse Gliomas

Yae Won Park[1*], Myeongkyun Kang[2*], Huiseung Ryu[2], Kyunghwa Han[1], Yongsik Sim[3], Ji Eun Park[4,5], Jong Hee Chang[6], Se Hoon Kim[7], Seung-Koo Lee[1], Sang Hyun Park[8§], Sung Soo Ahn[1§]

[1]Department of Radiology and Research Institute of Radiological Science and Center for Clinical Imaging Data Science, Yonsei University College of Medicine, Seoul, Korea

[2]Department of Robotics and Mechatronics Engineering, Daegu Gyeongbuk Institute of Science and Technology, Daegu, Korea

[3]Department of Radiology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea

[4]Department of Radiology, Johns Hopkins University, Baltimore, Maryland, United states

[5]Department of Radiology and Research Institute of Radiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea

[6]Department of Neurosurgery, Yonsei University College of Medicine, Seoul, Korea

[7]Department of Pathology, Yonsei University College of Medicine, Seoul, Korea

[8]Department of Computer Science and Engineering, Pohang University of Science and Technology (POSTECH), Pohang, Korea

[*]**These authors contributed equally to the manuscript.**

Running title: A vision language model in adult type diffuse gliomas

§Co-corresponding authors

Sang Hyun Park, Associate Professor, PhD

Department of Computer Science and Engineering, Pohang University of Science and Technology (POSTECH), Pohang, Korea

77 Cheongam-ro, Nam-gu, Pohang-si, Gyeongsangbuk-do, 37673, Korea

E-mail: sanghyunpark@postech.ac.kr

Sung Soo Ahn, Associate Professor, MD, PhD

Department of Radiology and Research Institute of Radiological Science and Center for Clinical Imaging Data Science, Yonsei University College of Medicine

50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea

Tel: 82-2-2228-7400

Fax: 82-2-393-3035

E-mail: sungsoo@yuhs.ac

ABSTRACT

We aimed to establish a robust vision-language model ("Glio-LLaMA-Vision") for molecular status prediction and radiology report generation (RRG) in adult-type diffuse gliomas. Multiparametric MRI data and paired radiology reports from 1,001 patients with adult-type diffuse gliomas were included in the institutional training set. A vision-language model, Glio-LLaMA-Vision, was developed from LLaMA 3.1 pre-trained on 2.79 million biomedical image-text pairs from PubMed Central and further fine-tuned from the institutional training set. The performance was validated in 100 patients and 75 patients with paired MRI-radiology reports from an institutional validation set and another tertiary institution (AMC), and in 170 and 477 patients with MRI from TCGA and UCSF datasets, respectively. In terms of IDH mutation status prediction, Glio-LLaMA-Vision showed AUCs ranging from 0.85-0.95 in the internal validation and external datasets. In terms of RRG, the BLEU-1 and ROUGE-L scores were 0.50 and 0.49 in the internal validation, respectfully, and 0.32 and 0.36 on the AMC dataset, respectively. Overall, 37.8% of generated reports were considered superior or equal to the original reports, while overall 91.0% of generated reports were considered clinically acceptable by neuroradiologists. Glio-LLaMA-Vision demonstrates promising performance in molecular status prediction and RRG in adult-type diffuse gliomas, showing potential for clinical assistance.

INTRODUCTION

Adult-type diffuse gliomas are the most common malignant primary brain tumors.[1] According to the 2021 World Health Organization (WHO) classification, isocitrate dehydrogenase (IDH) mutation status is the most important molecular marker in diagnosis, treatment planning, and predicting prognosis.[2] Clinical studies have shown that patients with IDH-mutant, 1p/19q-codeleted oligodendroglioma (herein oligodendroglioma) and IDH-mutant astrocytoma show significantly favorable prognosis than IDH-wildtype glioblastoma given their distinct biological behavior.[3,4] However, genetic testing is costly and time-consuming and may not be available in resource-limited regions,[5] while tissue insufficiency from biopsy may result in incomplete diagnosis. Therefore, a complementary noninvasive method to predict molecular information is crucial for appropriate treatment planning. Furthermore, due to the worldwide growing volume of MRI exams and the specialized knowledge needed to interpret glioma patients' images,[6] radiologists face excessive workloads and time constraints.[7] A radiology-specific artificial intelligence (AI) model that predicts molecular status and generates high-quality radiology reports may benefit both radiologists and clinicians.

Vision-language models (VLMs) combine the perceptual strengths of vision models with the generative capabilities of large language models (LLMs).[8,9] VLMs have demonstrated potential to clinically assist radiologists in downstream tasks such as classification and radiology report generation (RRG). Despite these advances, the application of VLMs in medical imaging has largely been confined to 2D modalities, particularly in the chest X-ray domain.[10-12] Expanding the use of VLMs to a more complex imaging field requiring in-depth domain knowledge, in particular for adult-type diffuse gliomas, remains underexplored but potentially transformative.

LLaMA (Large Language Model Meta AI) is a significant advancement in AI research,

offering an efficient and capable family of models that excel across diverse natural language tasks.[13] However, in the specialized domain of radiology, general-purpose LLMs often lack the domain-specific knowledge required to accurately interpret complex medical data. Given these challenges, we hypothesized that developing a VLM based on LLaMA 3.1, pre-trained on a large dataset of biomedical image-text pairs from PubMed Central (PMC) and fine-tuned on paired MRI-report datasets, could be clinically useful for molecular status prediction and RRG in adult-type diffuse gliomas.

Therefore, this study aimed to establish a VLM ("Glio-LLaMA-Vision") for molecular status prediction and RRG in adult-type diffuse gliomas.

RESULTS

Patient Characteristics

This study included a total of 1,723 patients with adult-type diffuse gliomas: 1,001, 75, 170, and 477 patients from Severance, Asan Medical Center (AMC), TCGA, and UCSF datasets, respectively. The clinicopathological characteristics of each cohort are summarized in **Table 1**. While there was no statistically significant difference in sex distribution across the four cohorts ($P = 0.274$), significant difference was observed in age ($P = 0.024$), CNS WHO grade ($P < 0.001$), and molecular subtype ($P < 0.001$) among four cohorts, reflecting the heterogeneity inherent in real-world datasets.

Table 1. Characteristics of the included 1,723 adult-type diffuse glioma patients according to datasets.

Dataset	Severance (n = 1,001)	AMC (n = 75)	TCGA (n = 170)	UCSF (n = 477)	P^*
Age (years)	55.3 ± 14.4	56.5 ± 13.6	56.8 ± 15.2	53.0 ± 15.0	0.024
Sex (male)	578 (58.0)	38 (50.7)	90 (52.9)	285 (59.7)	0.274
CNS WHO grade					< 0.001
Grade 2	147 (14.7)	15 (20.0)	28 (16.5)	46 (9.6)	
Grade 3	107 (10.7)	9 (12.0)	30 (17.6)	29 (6.1)	
Grade 4	747 (74.6)	51 (68.0)	112 (65.9)	402 (84.3)	
Molecular subtype					< 0.001
Oligodendroglioma	134 (13.4)	18 (24.0)	22 (12.9)	13 (2.7)	
IDH-mutant astrocytoma	147 (14.7)	8 (10.7)	41 (24.1)	90 (18.9)	
IDH-wildtype glioblastoma	720 (71.9)	49 (65.3)	107 (62.9)	374 (78.4)	

*Differences between datasets were analyzed using a 1-way ANOVA for continuous variables and the chi-squared test for categorical variables.

Data are presented in either number with percentage in parenthesis or mean ± standard deviation.

AMC = Asan Medical Center; TCGA = The Cancer Genome Atlas; UCSF = University of San Francisco; CNS = Central nervous system; WHO = World Health Organization; IDH = isocitrate dehydrogenase

Classification Performance of Glio-LLaMA-Vision

Table 2 shows the performance of IDH mutation status prediction on internal validation, AMC, TCGA, and UCSF datasets. Glio-LLaMA-Vision achieved an area under curve (AUC) of 0.90 (95%

confidence interval [CI]: 0.82–0.95) on the internal validation, 0.85 (95% CI: 0.72–0.94) on the AMC, 0.87 (95% CI: 0.81–0.92) on the TCGA, and 0.95 (95% CI: 0.92–0.96) on the UCSF datasets, demonstrating consistently high performance across both internal and external validation datasets.

Table 2. Performance of molecular status prediction of Glio-LLaMA-Vision on the internal validation, AMC, TCGA, and UCSF datasets.

Dataset	AUC (95% CI)	Accuracy (%)	Sensitivity (%)	Specificity (%)
Internal validation	0.90 (0.82–0.95)	83.0	86.1	75.0
AMC	0.85 (0.72–0.94)	84.0	81.3	88.9
TCGA	0.87 (0.81–0.92)	79.4	89.7	61.9
UCSF	0.95 (0.92–0.96)	75.5	69.3	98.1

AMC = Asan Medical Center; TCGA = The Cancer Genome Atlas; UCSF = University of San Francisco; AUC = area under the curve; CI = confidence interval

RRG Performance of Glio-LLaMA-Vision

Table 3 shows the quantitative performance of RRG on internal validation and AMC datasets. The BLEU-1 and ROUGE-L scores were 0.50 and 0.49 in the internal validation, and 0.32 and 0.36 on the AMC dataset, respectively, indicating moderate lexical quality and adequate content coverage of the RRG compared with the original reports.

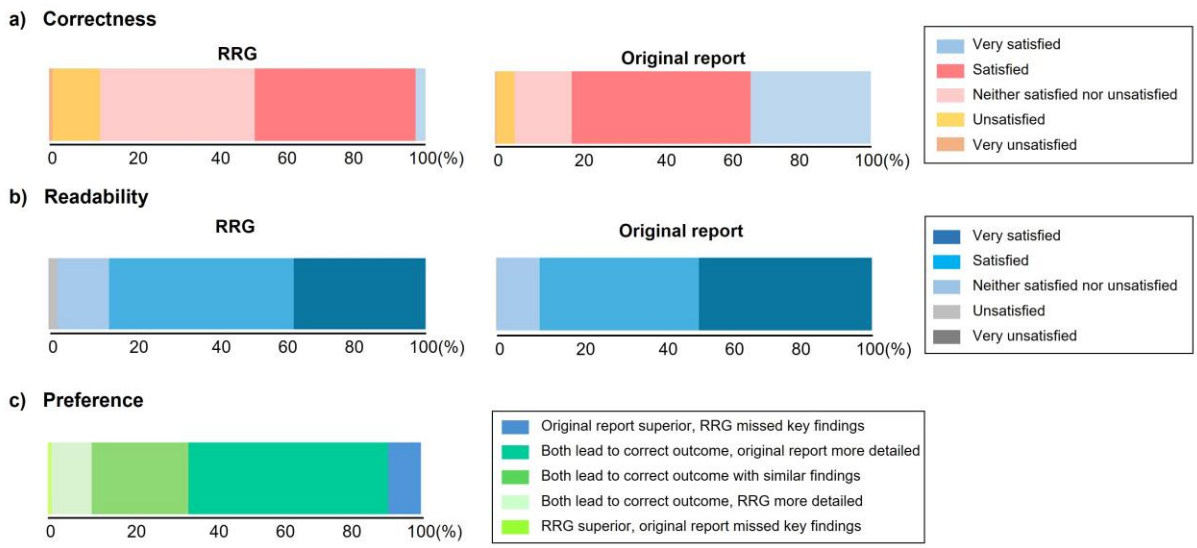
Table 4 presents the qualitative performance of RRG in terms of correctness and

readability on both the internal validation and AMC datasets. For correctness, all three readers showed significantly lower ratings for RRGs compared with original reports on internal validation and AMC datasets (all *P*s < 0.05). The reader-averaged ratings for correctness were also lower for RRGs compared with original reports on internal validation and AMC datasets (3.30 vs. 3.93, *P* < 0.001; 3.39 vs. 4.23, *P* < 0.001). For readability, readers generally gave significantly lower ratings to RRGs compared with the original reports on internal validation and AMC datasets (all *P*s < 0.05), with the only exception being reader 3 on the AMC dataset which gave significantly higher ratings to RRGs (*P* = 0.003). The reader-averaged readability ratings were also significantly lower for RRGs compared with original reports on internal validation and AMC datasets (4.16 vs. 4.34, *P* < 0.001; 4.15 vs. 4.35, *P* < 0.001).

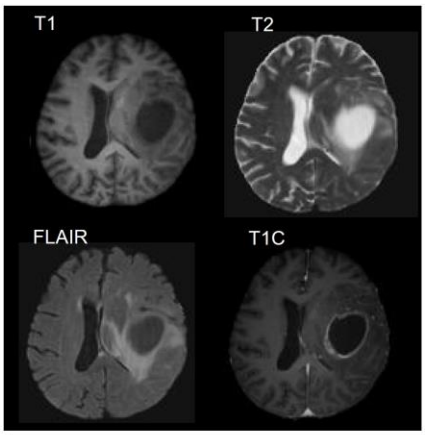
Table 5 summarizes the preference results between the RRG and original report. Readers consistently preferred the original reports over the RRGs, with mean preference scores significantly above the neutral value of 3 across all readers and in the reader-averaged analyses for both the internal validation (all *P*s < 0.001) and AMC datasets (all *P*s < 0.001). Overall, 37.8% of RRGs were considered superior or equal to the original report (52.6%, 16.0%, and 44.7% by Readers 1, 2, and 3, respectively), while overall 91.0% of RRGs were considered clinically acceptable (90.3%, 87.4%, and 95.4% by readers 1, 2, and 3, respectively) compared to original reports. **Figure 1** shows the distributions of overall correctness and readability for RRGs and original reports, as well as the overall preference scores comparing RRGs with original reports. **Figure 2** shows representative cases of molecular status prediction and RRG on the AMC dataset.

A total of nine cases (5.1%; five from Severance and four from AMC) demonstrated hallucinations in the RRG. Among these, two (22.2%) mentioned a separate tumor lesion outside the tumor involvement area, two (22.2%) arachnoid cysts, two (22.2%) minimal subarachnoid

hemorrhages, one (11.1%) an aneurysm, one (11.1%) a calcified granuloma, and one (11.1%) a dural tail. (representative cases in **Figure 3**).



a)



Molecular status prediction: "IDH-wildtype"

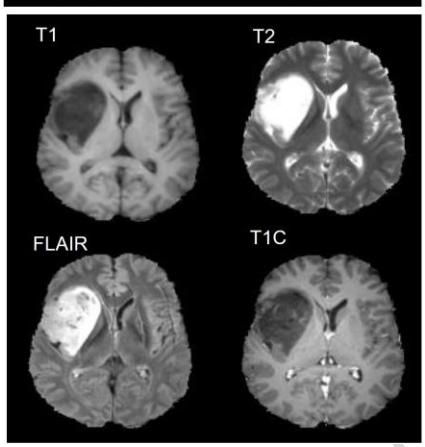
Original Report

A patient presented with a 7.4x 6.1x 5.6cm irregular rim enhancing mass with necrotic portion and heterogenous enhancing portion in left temporal lobe. The mass showed perilesional edema extending to left basal ganglia. Midline shifting to the right side was seen.

RRG

A patient has a tumor located in the temporal region, specifically in the left side, which is approximately 8.0 cm in size and has a necrotic portion. The tumor is causing a midline shift to the right. It is compressing the left lateral ventricle and has an irregular solid portion. The tumor is also causing a mass effect.

b)



Molecular status prediction: "IDH-mutant"

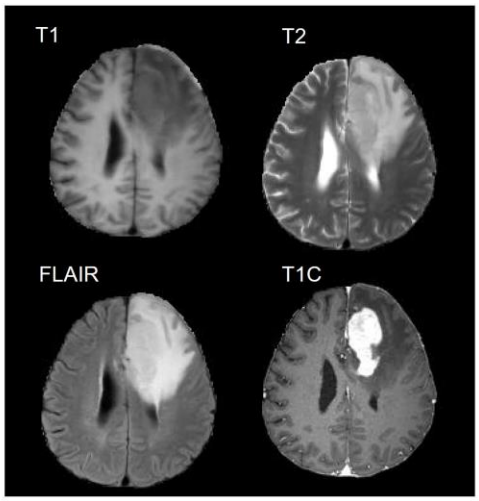
Original Report

A patient has a non-enhancing tumor located in the insula, measuring a maximum of 5.6cm, identified as a glioma situated in the right insula, which has displaced the basal ganglia and internal capsule.

RRG

A patient has a mass located in the insula of the right side of the brain. The mass is T2 hyperintense. It is not enhanced. The mass is a diffuse glioma located in the insula.

a)



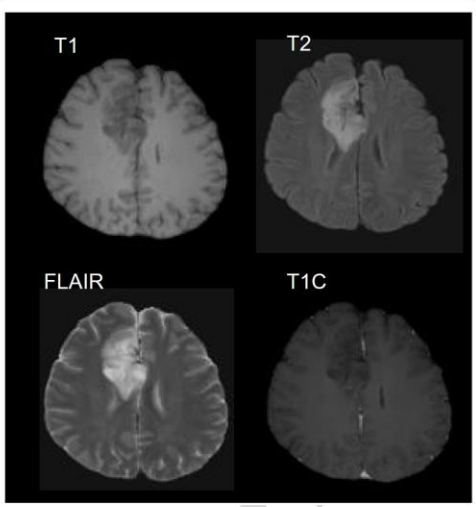
Original Report

A patient presented with a homogeneously enhancing mass located at the left frontal lobe of the brain, which is accompanied by perilesional edema, and there is no internal hemorrhage within the mass.

RRG

A patient has a 5.5-cm enhancing mass located at the left frontal lobe. The mass is accompanied by marked peritumoral edema. There is also a 1.5-cm enhancing nodule located at the right frontal lobe.

b)



Original Report

A patient presented with a 6.4 cm lesion located in the right parasagittal frontal lobe, extending to the corpus callosum. The lesion appeared as a non-enhancing area of high signal intensity on T2-weighted and FLAIR sequences. Notably, a focal hemorrhagic component was identified within the lesion.

RRG

A patient has a tumor located in the frontal area, which is approximately 6.3-cm in size. The tumor is expansile and nonenhancing. It involves the corpus callosum and has a right frontal location. Additionally, the patient has a probable arachnoid cyst.

Table 3. Quantitative performance of RRG on the internal validation and AMC datasets. The BLEU-1 and ROUGE-L were used as evaluation metrics.

Dataset	BLEU-1	ROUGE-L
Internal validation	0.50	0.49
AMC	0.32	0.36

RRG = radiology report generation; AMC = Asan Medical Center; BLEU-1; Bilingual Evaluation Understudy-1; ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence

Table 4. Qualitative performance of RRG on the internal validation and AMC datasets in terms of correctness and readability.

	Reader 1			Reader 2			Reader 3			Reader-averaged		
	RRG	Original	P^*	RRG	Original	P^*	RRG	Original	P^*	RRG	Original	$P^§$
Internal validation												
Correctness	3.32 ± 0.72	3.63 ± 0.80	0.009	3.36 ± 0.88	4.36 ± 0.93	< 0.001	3.21 ± 0.81	3.81 ± 0.72	< 0.001	3.30 ± 0.81	3.93 ± 0.87	< 0.001
Readability	3.78 ± 0.91	4.01 ± 0.80	0.053	4.40 ± 0.71	4.91 ± 0.29	< 0.001	4.29 ± 0.71	4.09 ± 0.29	0.008	4.16 ± 0.83	4.34 ± 0.70	< 0.001
AMC												
Correctness	3.43 ± 0.70	4.03 ± 0.59	< 0.001	3.39 ± 0.79	4.87 ± 0.45	< 0.001	3.35 ± 0.63	3.80 ± 0.70	< 0.001	3.39 ± 0.71	4.23 ± 0.74	< 0.001
Readability	3.73 ± 0.66	3.99 ± 0.69	0.020	4.40 ± 0.57	4.70 ± 0.29	< 0.001	4.31 ± 0.61	4.05 ± 0.43	0.003	4.15 ± 0.68	4.35 ± 0.66	< 0.001

Data are indicated as mean with standard deviation.

*For the reader-wise comparison, the Wilcoxon signed-rank test was used to assess differences between the RRGs and original reports.

§For the reader-averaged analysis, a cumulative linear mixed model was employed.

RRG = radiology report generation; AMC = Asan Medical Center

Table 5. Qualitative performance of RRG on the internal validation and AMC datasets in terms of preference.

	Reader 1	P^*	Reader 2	P^*	Reader 3	P^*	Reader-averaged	$P^§$
Internal validation	3.35 ± 0.88	< 0.001	3.96 ± 0.60	< 0.001	3.33 ± 1.00	< 0.001	3.55 ± 0.89	< 0.001
AMC	3.35 ± 0.79	< 0.001	3.97 ± 0.43	< 0.001	3.39 ± 0.85	< 0.001	3.64 ± 0.76	< 0.001

Data are indicated as mean with standard deviation.

*For the reader-wise comparison, the Wilcoxon signed-rank test was conducted across all subjects.

§For the reader-averaged analysis, a cumulative linear mixed model was employed.

AMC = Asan Medical Center; RRG = radiology report generation

Ablation Results

Table 6 shows results from the ablation experiments of Glio-LLaMA-Vision for molecular status prediction on the internal validation dataset, AMC set, TCGA dataset, and UCSF dataset. The AUCs were lower in the ablated models (the models without the classifier or without the RRG) compared to the model with both the classifier and RRG ("Glio-LLaMA-Vision") on the internal validation set, TCGA set, and UCSF dataset, respectively. On the AMC set, the model without classifier showed lower performance than Glio-LLaMA-Vision, whereas the model without RRG achieved a higher AUC than Glio-LLaMA-Vision. The ablation experiments generally support the idea that employing both objectives (molecular subtype prediction with a classifier and RRG with an LLM) is essential for achieving a high performance on molecular status prediction.

Table 6. Results from the ablation studies of Glio-LLaMA-Vision for molecular status prediction on the internal validation set, AMC set, TCGA set, and UCSF set. The performances the molecular status prediction without the classifier, the molecular status prediction without RRG, and the molecular status prediction with both the classifier and RRG ("Glio-LLaMA-Vision") are shown.

Test set	AUC (95% CI)	Accuracy (%)	Sensitivity (%)	Specificity (%)
Molecular status prediction without the classifier				
Internal	0.83 (0.73-0.91)	85.0	87.5	78.6
AMC	0.80 (0.70-0.89)	80.0	79.2	81.5
TCGA	0.72 (0.65-0.78)	76.5	89.7	54.0
UCSF	0.80 (0.76-0.84)	75.1	71.4	88.4
Molecular status prediction without RRG				
Internal	0.87 (0.79-0.94)	76.0	73.6	82.1
AMC	0.88 (0.78-0.96)	72.0	60.4	92.6
TCGA	0.84 (0.78-0.90)	74.1	72.9	76.2
UCSF	0.93 (0.90-0.96)	56.0	44.1	99.0
Molecular status prediction with both the classifier and RRG ("Glio-LLaMA-Vision")				
Internal	0.90 (0.82–0.95)	83.0	86.1	75.0
AMC	0.85 (0.72-0.94)	84.0	81.3	88.9
TCGA	0.87 (0.81–0.92)	79.4	89.7	61.9
UCSF	0.95 (0.92–0.96)	75.5	69.3	98.1

TCGA = The Cancer Genome Atlas; UCSF = University of San Francisco; RRG = radiology report generation; AUC = area under the curve; CI = confidence interval

Impact of Radiology Report Preprocessing

Table 7 shows the results of our method and without LLM-based preprocessing. LLM-based preprocessing did not degrade performance for molecular status prediction but instead resulted in a slight improvement. The degraded performance of the model without LLM-based preprocessing is likely due to discrepancies in writing style between the predicted reports and the ground truth, or to degradation resulting from training on inconsistent reports. Therefore, we applied LLM-based preprocessing to mitigate inconsistencies in the reports. **Table 8** shows the qualitative results of LLM-based preprocessing. Qualitatively, LLM-based preprocessing preserves the semantic content of radiology reports while improving consistency in writing style.

Table 9 shows the results of our method without the IDH mutation status. Excluding the IDH mutation status resulted in slight decreases in accuracy, sensitivity, and specificity for molecular status prediction, indicating that IDH mutation status still provides valuable information for molecular status prediction. Additionally, a decrease in RRG performance was observed compared with the model that included IDH mutation status. The results show that pathologically confirmed results on IDH mutation status may increase the model performance.

Table 7. Results of the ablation studies on LLM-based preprocessing for molecular status prediction and the performance of RRG on the internal validation set.

LLM-Preprocessing	AUC (95% CI)	Accuracy (%)	Sensitivity (%)	Specificity (%)	BLEU-1	ROUGE-L
With	0.90 (0.82-0.95)	83.0	86.1	75.0	0.50	0.49
Without	0.90 (0.84-0.95)	82.0	86.1	71.4	0.35	0.39

LLM = Large Language Model; AUC = area under the curve; CI = confidence interval; BLEU-1; Bilingual Evaluation Understudy-1; ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence

Table 8. Qualitative comparison of LLM-based preprocessing.

Filtered Original Radiology Report	LLM-preprocessed Radiology Report
Infiltrative T2 lesion at the right frontal lobe and right side corpus callosum genu. - focal area of enhancement with necrosis.	A patient has an infiltrative T2 lesion located at the right frontal lobe and the right side of the corpus callosum genu, with a focal area of enhancement and necrosis.
About 5cm size mass at Rt frontal lobe. Focal cystic change. Irregular enhancement is noted. Suspicious calcification.	A patient has a tumor located in the frontal lobe, specifically at the right side, where a mass approximately 5cm in size is present. The mass exhibits focal cystic change. Irregular enhancement is noted in the mass. Suspicious calcification is also observed.

LLM = Large Language Model

Table 9. Results of ablation studies excluding IDH mutation status from the radiology report for molecular status prediction and the performance of RRG on the internal validation set.

Provision of IDH mutation status	AUC (95% CI)	Accuracy (%)	Sensitivity (%)	Specificity (%)	BLEU-1	ROUGE-L
With	0.90 (0.82-0.95)	83.0	86.1	75.0	0.50	0.49
Without	0.90 (0.84-0.96)	81.0	84.7	71.4	0.40	0.35

IDH = isocitrate dehydrogenase; RRG = radiology report generation; AUC = area under the curve; CI = confidence interval; BLEU-1; Bilingual Evaluation Understudy-1; ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence

Impact of Segmentation Model Performance

Table 10 shows the results for the eroded and dilated segmentations. Although eroded and dilated segmentations were used, performance showed high consistency. This robustness arises because our method uses segmentation only to select slices containing tumors; therefore, our model is not sensitive to segmentation performance.

Table 10. Results of the ablation studies on the original, eroded, and dilated segmentations for molecular status prediction and the performance of RRG on the internal validation set.

Segmentation	AUC (95% CI)	Accuracy (%)	Sensitivity (%)	Specificity (%)	BLEU-1	ROUGE-L
Original	0.90 (0.82-0.95)	83.0	86.1	75.0	0.50	0.49
Erosion	0.89 (0.82-0.95)	84.0	87.5	75.0	0.50	0.50
Dilation	0.90 (0.83-0.96)	84.0	86.1	78.6	0.50	0.49

RRG = radiology report generation; AUC = area under the curve; CI = confidence interval; BLEU-1; Bilingual Evaluation Understudy-1; ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence

DISCUSSION

Our proposed Glio-LLaMA-Vision based on LLaMA 3.1 demonstrated promising performance in molecular status prediction and RRG in adult-type diffuse gliomas in a large multi-institutional dataset. The AUCs of molecular status prediction ranged from 0.85 to 0.95 on internal and external validations, showing robust performance in the heterogeneous datasets. The BLEU-1 and ROUGE-L scores ranged from 0.32 to 0.50, indicating adequate RRG performance in terms of quantitative metrics. Furthermore, to assess the potential real-world utility of our RRG, a comprehensive expert evaluation was performed alongside to quantitative evaluation, and overall, 91.0% of RRGs were considered clinically acceptable and would have resulted in correct clinical management of the patient. Notably, our current study offers a practical paradigm of adapting general domain LLMs to applications in a specific medical domain, potentially serving as an assistant to radiologists as well as clinicians. Considering the global trend of shortage of radiologists, especially in the niche field of adult-type diffuse gliomas that requires in-depth domain knowledge for accurate image interpretation, our model shows the potential to be implemented as a useful clinical tool.

Previous AI models for diffuse glioma imaging have mainly relied on convolutional neural networks (CNNs).[14-16] Studies on IDH mutation status prediction have been conducted either applying CNNs on representative tumor slices from orthogonal planes or 2D tumor images,[14,16] or by applying CNNs to full 3D images,[15] yielding AUCs ranging from 0.86 to 0.96 in external validation. However, CNNs are typically single-modal and task-specific, being primarily developed for the narrow purpose of molecular classification of gliomas. More recently, a multimodal transformer was proposed to predict the molecular status of adult-type diffuse gliomas by integrating imaging and clinical data. It outperformed CNN-based baselines, demonstrating

both the representational power of vision transformers and the benefit of clinical data integration through an LLM-based encoder.[17] Nonetheless, such approaches remain limited in their explainability and in their ability to support diverse downstream tasks, thereby providing little practical assistance in the radiological workflow. In real-world clinical practice, radiologists use natural language to communicate the imaging findings in the form of written radiology reports, while the CNN outputs are context-restrictive. Building on this, our study leverages LLMs for both molecular status prediction and report generation, aiming to provide practical support in the radiological workflow.

Recently, VLMs have been fine-tuned on large-scale medical image-text datasets curated from PMC,[18] demonstrating strong performance on diverse medical tasks. While various VLMs (including LLaVA-Med or Med-PaLM Multimodal) have reported their primary success in X-ray and single-slice CT report generation, these tasks were limited in 2D medical images with different modalities.[11,12,19,20] Recently, several VLMs have been developed for brain MRI analysis,[21,22] but none have been specifically trained to address crucial tasks in adult-type diffuse gliomas requiring domain knowledge. Furthermore, existing VLMs for brain MRIs have not addressed RRG, a task most closely aligned with routine radiologist practice and patient care. Thus, currently developed VLMs have little efficacy in the clinical workflow of adult-type diffuse gliomas.

To address these limitations, we propose Glio-LLaMA-Vision, which is a specialized VLM for adult-type diffuse gliomas. The Glio-LLaMA-Vision employs a 3D representation by averaging features across multiple axial slices, which enables more effective MRI analysis. Our framework jointly performs molecular status prediction and RRG by integrating an LLM with a vision encoder. To capture domain specific knowledge of adult-type diffuse gliomas, it leverages pre-training on 2.79 million PMC image-text pairs and is further refined through fine-tuning on an

institutional dataset. Also, as demonstrated by our ablation study, both the classifier and the RRG components play a critical role in improving prediction performance. This finding indicates that prediction alone, whether by a classifier or by report generation, is insufficient, whereas the generalized representation learned via RRG substantially improves the classifier's performance. Our results show promising performance in both molecular status prediction and RRG.

The quality of the generated reports was examined through quantitative and qualitative evaluations. BLEU-1 and ROUGE-L are quantitative metrics to assess the fidelity and informativeness of the generated report.[23,24] BLEU-1 scores ranged from 0.32 to 0.50, and ROUGE-L scores ranged from 0.36 to 0.49 in internal and external validations, indicating moderate translation quality and adequate content coverage. Importantly, to bridge the gap between automated metrics and real-world clinical utility, qualitative evaluation of the RRG was performed by neuroradiologists ("human-in-the-loop"), ensuring that the generated reports were both linguistically accurate and clinically meaningful in radiologic interpretation.[23,24] Although the correctness and readability were higher in the original reports, 91.0% of RRGs were considered clinically acceptable, demonstrating clinical promise. However, only 37.8% of generated reports were rated as superior or comparable to the original reports, indicating that further improvements are required before RRG in VLMs can be considered clinically valuable. Furthermore, we speculate that explainability or uncertainty estimation may be useful to support safer decision-making in VLM models and should be implemented in the future.[17,25,26]

Our study has several limitations. First, there was a relatively small number of patients with paired MRI-radiology reports. As open-source datasets such as TCGA and UCSF did not provide radiology reports, a future study with a larger number of patients with paired MRI-radiology reports on external datasets is warranted for further validation. Second, our approach

was based solely on standard clinical MRI sequences, omitting advanced modalities like diffusion-weighted or dynamic susceptibility contrast imaging. Our intention was to prioritize widely accessible scans, thereby enhancing the model's generalizability and to enabling evaluation on independent cohorts. Since most external validation datasets did not include these advanced sequences, their integration into our study was not feasible. Third, as radiology reports exhibit strong dependencies between earlier words (e.g., radiology findings) and later words (e.g., impressions), the influence of MRI features diminishes with increasing sequence length, while reliance on preceding words grows, potentially leading to hallucinations. Therefore, although the hallucination from RRGs were small in proportion (5.1%), caution is required for real-world clinical application of our developed model.[24]

In conclusion, Glio-LLaMA-Vision shows promising performance in molecular status prediction and RRG for adult-type diffuse gliomas, highlighting its potential for clinical assistance.

METHODS

Study Design and Ethical Approval

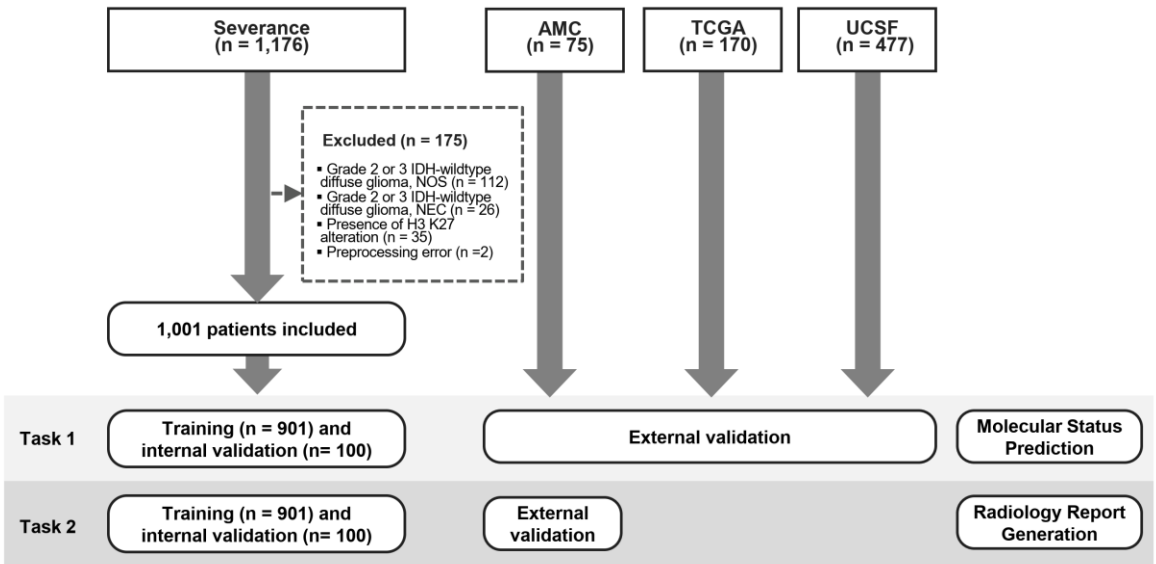
The requirement for patient consent was waived owing to the retrospective study design. This study was approved by the Institutional Review Board of Severance Hospital (No. 4-2024-0040). The study was conducted in accordance with the Declaration of Helsinki.

Patient Population

Between June 2005 and December 2022, 1,176 patients with diffuse gliomas were included in the institutional development set. The inclusion criteria were as follows: 1) histopathologically confirmed diffuse gliomas, 2) known IDH mutation and 1p/19q codeletion status, 3) aged ≥ 18

years, and 4) presence of corresponding MRI reports. The exclusion criteria were as follows: 1) histological grade 2 or 3 IDH-wildtype diffuse gliomas not tested for genetic parameters (*TERT* promoter mutation, *EGFR* gene amplification, or chromosome +7/-10), thereby diagnosed as IDH-wildtype diffuse glioma, not otherwise specified (n = 112),[27] 2) histological grade 2 or 3 IDH-wildtype diffuse gliomas which were negative for all three genetic parameters (*TERT* promoter mutation, *EGFR* gene amplification, and chromosome +7/-10), thereby diagnosed as IDH-wildtype diffuse glioma, not elsewhere classified (NEC) (n = 26)[27], 3) presence of H3 K27M alteration, leading to a diagnosis of diffuse midline glioma, H3 K27-altered (n = 35), and 4) preprocessing error (n = 2). A total of 1,001 patients were included in the institutional dataset.

For internal and external validation, identical criteria were applied to 100 and 75 patients with paired MRI-report datasets from the institutional validation set and AMC, respectively. In addition, 170 and 477 adult-type diffuse glioma patients with MRI data were included from the TCGA (http://cancergenome.nih.gov) and the UCSF datasets, respectively.[28,29] **Figure 4** shows the patient flowchart.



Molecular Classification

All patients were diagnosed according to the 2021 WHO classification.[2] IDH1/2 mutation and 1p/19q codeletion status were assessed. The presence of H3 K27M mutation was evaluated in tumors with midline location.

Targeted next-generation sequencing (NGS) was performed. In the institutional and TCGA external validation sets, 20 and 18 patients, respectively, with histological grade 2 or 3 IDH-wildtype gliomas with either a *TERT* promoter mutation, *EGFR* gene amplification, or chromosome +7/-10, were classified as IDH-wildtype glioblastoma according to their molecular profiles.[2]

MRI Protocol

Preoperative MRI was performed using a 3.0-T MRI scanner (Achieva or Ingenia, Philips Medical Systems) with an eight-channel sensitivity-encoding head coil in the institutional set. A 3T MRI unit (Achieva or Ingenia; Philips Healthcare) and an eight-channel sensitivity encoding head coil

were used for brain MRI. The protocol included T1-weighted turbo spin-echo images with inversion recovery (repetition time [TR], 2000 ms; echo time [TE], 10 ms; inversion time [TI], 1000 ms; field of view [FOV], 240 mm; section thickness, 5 mm; matrix, 256 × 256), T2-weighted turbo spin-echo (TR, 3000 ms; TE, 80 ms; FOV, 240 mm; section thickness, 5 mm; matrix, 256 × 256), and T2-weighted fluid-attenuated inversion recovery (FLAIR) (TR, 10,000 ms; TE, 125 ms; TI, 2500 ms; FOV, 240 mm; section thickness, 5 mm; matrix, 256 × 256) images. Postcontrast 3D T1-weighted turbo field echo images (TR, 9.8 ms; TE, 4.6 ms; FOV, 240 mm; section thickness, 1 mm; matrix, 224 × 224) were acquired 6 minutes after the injection of gadolinium-based contrast (0.1 mL/kg of gadobutrol, Gadovist; Bayer Schering Pharma).

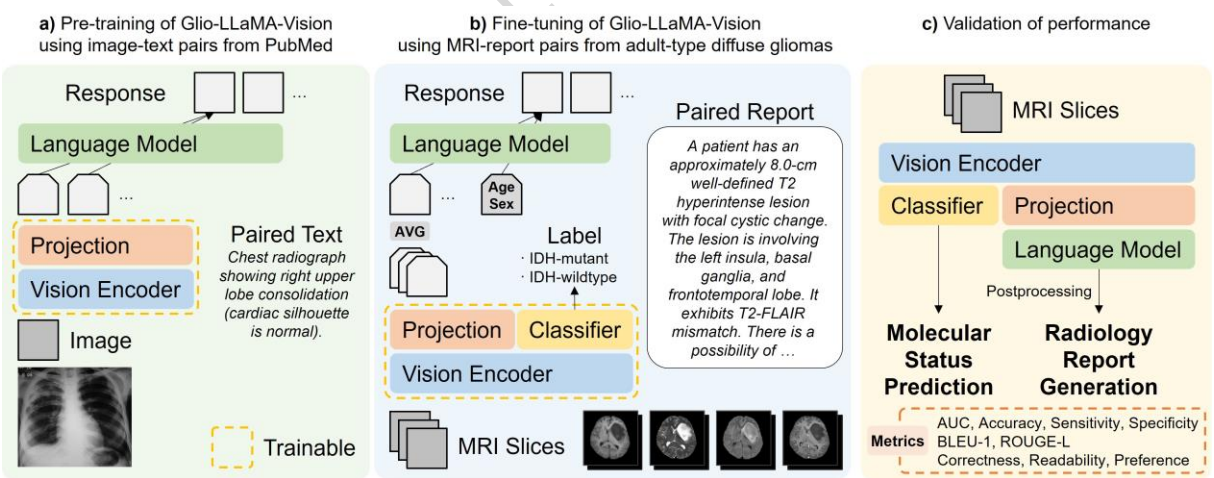
Image Processing

Image preprocessing involved isovoxel resampling to 1 mm^3, N4 bias field correction, and co-registration of T1, T2, and FLAIR images to T1C images, utilizing Advanced Normalization Tools (ANTs). Skull stripping was performed using HD-BET.[30] Signal intensity was z-score normalized. Tumor regions were segmented using a fully automated segmentation tool (HD-GLIO)[31] to select MRI slices containing tumors. Because the segmentation model is applied only to selected tumor slices, overall performance is not sensitive to segmentation accuracy. All axial slices of T1, T2, FLAIR, and T1C containing the tumor were selected and used for model development and testing.

Radiology Report

MRI radiology reports in the dataset were curated prior to analysis using the instruction-tuned LLaMA 3.1 with 70 billion parameters.[13] As an initial step, irrelevant information was removed, including references to names of MRI sequences, contrast usage, reporting radiologists, and

timestamps. Redundant statements (e.g., *"There are no other significant abnormalities."*) were excluded, and the impression section was excluded for later replacement with the IDH-mutant diagnostic ground truth. Once only MRI findings remained, the text prompt used for preprocessing was as follows: *"Rewrite the following report into a single paragraph that begins with 'A patient', using all the information provided, without adding or omitting facts, excluding subjective conjunctional words, and ensure no sentence becomes too long."* This prompt was designed to standardize the varied formats of the original reports into a standardized paragraph format. A statement indicating the pathologically confirmed IDH mutation status was added to the end of the revised report in place of the impression, such as *"The IDH mutation status indicates IDH-wildtype glioblastoma."*. Neuroradiologists verified that there was no discrepancy between the original report and preprocessed report.



Glio-LLaMA-Vision

The developed Glio-LLaMA-Vision framework consisted of a vision encoder, a projection layer, a classifier, and an LLM. **Figure 5** shows an overview of the Glio-LLaMA-Vision and study

pipeline. The training objective is to jointly learn two tasks: prediction of molecular status and RRG. We fine-tuned the vision encoder, projection layer, and classifier, while the LLM parameters were frozen during training. The vision encoder follows a Vision Transformer base architecture parameterized by BiomedCLIP,[32] using an input size of 224 and a patch size of 16. The projection layer and the classifier each consist of two multilayer perceptron layers. A total of 901 MRIs from the institutional dataset and their corresponding pre-processed radiology reports were utilized for training.

For molecular status prediction, MRI scans were input into the vision encoder, and the resulting features were taken from the classification (CLS) token, serving as a compact representation of the entire image. The 3D representation was constructed by averaging the global 2D features extracted from MRIs of the T1, T2, FLAIR, and T1C sequences. Subsequently, averaged features were fed into a classifier for label prediction. For training, cross-entropy loss was used as an objective function, using MRIs and corresponding information on molecular status.

For RRG, MRIs were sequentially input into the vision encoder, followed by a projection layer to align the feature dimensions between the vision encoder and the LLM. Note that the output features of the projection layer are not aligned with the LLM's word embeddings; therefore, pre-training is required before fine-tuning for downstream tasks (see the section **Pre-training of Glio-LLaMA-Vision**). A total of 2.79 million biomedical image-text pairs, collected from PMC, were employed in auto-regressive training to predict the paired text for feature alignment pre-training. A 3D representation was obtained by averaging the global 2D features extracted from MRIs of the T1, T2, FLAIR, and T1C sequences. In particular, the average global 2D features of MRIs can leverage generation capabilities pre-trained on large-scale datasets, enabling effective training of the vision encoder and a projection layer even with only 901 MRIs. The averaged features were

fed into the instruction-tuned LLM, LLaMA 3.1 with 8 billion parameters,[13] generating a text response for the given MRIs. For training, an auto-regressive training objective was employed using paired MRIs and radiology reports, along with age and sex information. Notably, employing RRG with an LLM was crucial for achieving high molecular status prediction performance.

For internal and external validation, MRIs were input into Glio-LLaMA-Vision to predict the molecular status and generate radiology reports. As in training, MRI features were averaged to construct a 3D representation and then fed into the subsequent layer for patient-level prediction. For postprocessing, we enhanced the quality of the generated radiology reports by eliminating repetitive and incomplete sentences. However, the proposed method did not enforce any explicit output format during inference; instead, the generated results adhered to the free-form writing style of the fine-tuned dataset.

Pre-training of Glio-LLaMA-Vision

Since the output features of the projection layer are not aligned with the LLM's word embeddings, pre-training was required before fine-tuning for downstream tasks. An image (e.g., an axial MR slice) was input into the vision encoder, which is a Vision Transformer parameterized by BiomedCLIP, to extract the global 2D features. Subsequently, the extracted global 2D features were fed into the projection layer consisting of two multi-layer perceptron layers and outputs 50 soft tokens, aligning the feature dimensions between the vision encoder and the LLM. The 50 soft tokens were then fed into the instruction-tuned LLM, LLaMA 3.1 with 8 billion parameters, which generates a text response for the given image. A total of 2.79 million biomedical image-text pairs collected from PMC were used for feature alignment pre-training. An autoregressive training objective was employed to enable the LLM to predict the next token in the text based on the paired

image. We optimized the parameters of the vision encoder and projection layers for two epochs while keeping the parameters of the LLM frozen.

Ablation Studies

To verify the impact of our two training objectives, molecular subtype prediction using a classifier and RRG using an LLM, we conducted ablation experiments by individually excluding each objective. In the ablation study without the classifier, the generated radiology report was re-input into the LLM with the following prompt: *"Derive the isocitrate dehydrogenase (IDH) mutation status. A. oligodendroglioma (IDH-mutant and 1p/19q codeleted) or IDH-mutant astrocytoma. B. IDH-wildtype glioblastoma. Based on the description, respond with only the letter of the correct option from the given choices, starting with 'The correct answer is [A or B]'."* to derive a molecular subtype label. In the ablation study without the RRG, we fine-tuned a vision encoder, parameterized by BiomedCLIP, along with a classifier.

Ablation of Radiology Report Preprocessing

We conducted ablation experiments by fine-tuning the model without LLM-based preprocessing. The radiology reports used for fine-tuning followed the same preprocessing steps, except that the LLM-based preprocessing step was excluded. In addition, we conducted an ablation experiment by fine-tuning the model without the pathologically confirmed IDH mutation status, such as *"The IDH mutation status indicates IDH-wildtype glioblastoma."*

Ablation of Segmentation Model Performance

We conducted ablation experiments to assess the sensitivity of the fine-tuned model to HD-GLIO segmentation performance. Specifically, we constructed pseudo segmentations by applying erosion and dilation with a 3-pixel kernel to the original HD-GLIO segmentations. Then, we selected MRI slices containing tumors using the eroded and dilated segmentations and used these slices for evaluation.

Statistical Analysis

The demographic and clinical characteristics of the patients in the training, internal validation, and external validation sets were compared with the chi-square test for categorical variables and ANOVA for continuous variables.

To evaluate the performance for molecular status prediction, the AUC, sensitivity, specificity, and accuracy were calculated. The 95% confidence intervals CI of AUCs were calculated using bootstrapping with 1,000 iterations.

Performance on RRG was quantitatively evaluated using BLEU-1 and ROUGE-L, both of which are widely used objective evaluation metrics.[23] In particular, BLEU-1 measures precision of n-gram matches between the generated report and the original report, while ROUGE-L measures word order and sequence similarity.

Qualitative evaluation of RRG was performed in the institutional internal validation and AMC datasets with paired MRI-reports. For qualitative evaluation of RRG, three qualified neuroradiologists (with 5, 9, and 15 years of experience in neuroradiology, respectively) performed a 5-point Likert scale evaluation in three categories: “correctness”, “readability”, and “preference”

for generated radiology reports and original reports.[33,34] The origin of reports was masked, and the reports (either generated reports or original reports by radiologists) were shown in random order to avoid bias (**Table 11** explains parameters for the qualitative evaluation of RRG).

In the "preference" category, the Likert scale evaluates the clinical relevance of RRGs and original reports and their impact on patient management: point 1 (or point 5) indicates that the RRG (or original report) is superior to the other, and captures key clinically relevant findings that the other report misses, leading to correct patient management while the other does not; point 2 (or point 4) indicates that the RRG (or original report) captures more relevant findings than the other, but both would still result in correct patient management; point 3 indicates that both reports capture similar findings and would lead to identical patient management. Overall, inter-rater agreement was high across centers and evaluation categories, with Gwet's AC2 (Agreement Coefficient) values generally indicating substantial to near-perfect agreement and consistently high weighted percent agreement (see **Table 12**). All incorrect findings were noted during the blinded review phase. After unblinding the report sources, radiologists reviewed incorrect RRG findings through discussion (Y.S., Y.W.P., and S.S.A, with 5, 9, and 15 years of experience in neuroradiology, respectively), and factual inaccuracies, generated findings absent from the imaging data, were labeled as hallucinations.

For the reader-wise comparison of "correctness" and "readability", the Wilcoxon signed-rank test was used to assess differences between the RRGs and original reports. To compare the "preference", a Wilcoxon signed-rank test was conducted across all subjects. For the reader-averaged analysis, a cumulative linear mixed model (CLMM) was employed. In this model, the report type (original report vs. RRG) and reader were treated as fixed effects, while individual images were included as a random effect to account for repeated evaluations of the same image by

different methods. In terms of "preference", if RRG captured the key clinically relevant findings and resulted in correct patient management, it was considered "clinically acceptable".[33] A *P* value < 0.05 was interpreted as statistically significant. All statistical analyses were performed by an expert biostatistician (K.H., with 16 years of experience as a clinical biostatistician) using R software (R version 4.0.2).

Table 11. Explanation of parameters for the qualitative evaluation of RRG.

	Definition	Scale
Correctness	The factual accuracy of the report in relation to the imaging findings	Ranged from 5 to 1, where 5 represented "very satisfied," 4 stood for "satisfied," 3 meant "neither satisfied nor unsatisfied," 2 indicated "unsatisfied," and 1 corresponded to "very unsatisfied."
Readability	The clarity and accessibility of the language, including vocabulary, sentence structure, and overall presentation	Ranged from 5 to 1, where 5 represented "very satisfied," 4 stood for "satisfied," 3 meant "neither satisfied nor unsatisfied," 2 indicated "unsatisfied," and 1 corresponded to "very unsatisfied."
Preference	The comparative clinical utility of two radiology reports (A and B), based on their ability to capture key findings and guide appropriate patient management	1: Report A captures key clinically relevant findings that are not found in B. Report A would result in correct patient management and report B would not. 2: Report A captures more relevant findings, but both would result in the same correct patient managements. 3: Both reports capture similar findings in the image and would result in correct patient management. 4: Report B captures more relevant findings, but both would result in the same correct patient managements. 5: Report B captures key clinically relevant findings that are not found in A. Report B would result in correct patient management and report A would not.

RRG = radiology report generation

Table 12. Inter-rater agreement in qualitative evaluation of RRG and original reports in terms of AC2 and percent of agreement.

Center	Category	AC2 (95% CI)	Percent of agreement (%)
AMC	Correctness of RRG	0.88 (0.86-0.91)	94.6
	Readability of RRG	0.74 (0.68-0.80)	88.0
	Correctness of original report	0.61 (0.53-0.69)	83.0
	Readability of original report	0.30 (0.17-0.43)	68.7
	Preference	0.71 (0.63-0.79)	86.6
Severance	Correctness of RRG	0.85 (0.81-0.89)	93.7
	Readability of RRG	0.81 (0.78-0.85)	92.1
	Correctness of original report	0.77 (0.72-0.81)	90.4
	Readability of original report	0.66 (0.61-0.72)	84.9
	Preference	0.80 (0.75-0.85)	91.8
Overall	Correctness of RRG	0.86 (0.84-0.89)	94.1
	Readability of RRG	0.83 (0.80-0.85)	92.4
	Correctness of original report	0.77 (0.73-0.80)	90.1
	Readability of original report	0.66 (0.62-0.70)	84.7
	Preference	0.81 (0.78-0.85)	91.9

RRG= radiology report generation; CI = confidence interval; AMC = Asan Medical Center; AC2 = Agreement Coefficient 2

Data availability:

The datasets analyzed during the current study are not publicly available due to institutional and ethical restrictions but are available from the corresponding author on reasonable request. Access to the data will be provided after review and approval by all authors and in compliance with applicable institutional policies and ethical guidelines.

Code availability:

A valid OSI-approved open-source license, the MIT License, is applied to our code. The code repository includes a clear LICENSE file specifying that the code is released under the MIT License. Code is publicly available at (https://github.com/myeongkyunkang/Glio-LLaMA-Vision). Other information is available from the corresponding author upon request.

Acknowledgments:

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00516124, RS-2025-00515423, and RS-2025-00515536).

Competing interests:

The authors declare no competing financial or non-financial interests.

Author contributions:

Y.W.P., K.H., Y.S., J.E.P., J.H.C., S.H.K., S.L., and S.S.A. collected and curated the data. Y.W.P., M.K., S.H.P., and S.S.A. conceptualized the study and developed the methodology. M.K. and H.R. developed the software. Y.W.P., M.K., H.R., K.H., Y.S., J.E.P., J.H.C., S.H.K., S.K., S.H.P., and

S.S.A. performed the validation. Y.W.P., M.K., and H.R. performed the statistical analysis. S.H.P. and S.S.A. designed the study and supervised the research. Y.W.P. and M.K. wrote the original draft. All authors reviewed and approved the final manuscript.

References

1 Ostrom, Q. T. *et al.* CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2016-2020. *Neuro Oncol* **25**, iv1-iv99, doi:10.1093/neuonc/noad149 (2023).

2 Louis, D. N. *et al.* The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro Oncol* **23**, 1231-1251, doi:10.1093/neuonc/noab106 (2021).

3 Brat, D. J. *et al.* Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *N Engl J Med* **372**, 2481-2498, doi:10.1056/NEJMoa1402121 (2015).

4 Turkalp, Z., Karamchandani, J. & Das, S. IDH mutation in glioma: new insights and promises for the future. *JAMA Neurol* **71**, 1319-1325, doi:10.1001/jamaneurol.2014.1205 (2014).

5 Sarkar, C. *et al.* Resource availability for CNS tumor diagnostics in the Asian Oceanian region: A survey by the Asian Oceanian Society of Neuropathology committee for Adapting Diagnostic Approaches for Practical Taxonomy in Resource-Restrained Regions (AOSNP-ADAPTR). *Brain Pathol*, e13329, doi:10.1111/bpa.13329 (2025).

6 Park, Y. W. *et al.* The 2021 WHO Classification for Gliomas and Implications on Imaging Diagnosis: Part 1-Key Points of the Fifth Edition and Summary of Imaging Findings on Adult-Type Diffuse Gliomas. *J Magn Reson Imaging* **58**, 677-689, doi:10.1002/jmri.28743 (2023).

7 Smith-Bindman, R. *et al.* Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016. *Jama* **322**, 843-856, doi:10.1001/jama.2019.11456 (2019).

8 Kim, K. *et al.* Updated Primer on Generative Artificial Intelligence and Large Language Models in Medical Imaging for Medical Professionals. *Korean J Radiol* **25**, 224-242, doi:10.3348/kjr.2023.0818 (2024).

9 Nam, Y. *et al.* Multimodal Large Language Models in Medical Imaging: Current State and Future Directions. *Korean J Radiol* **26**, 900-923, doi:10.3348/kjr.2025.0599 (2025).

10 Huang, W. *et al.* Enhancing representation in radiography-reports foundation model: a granular alignment algorithm using masked contrastive learning. *Nat Commun* **15**, 7620, doi:10.1038/s41467-024-51749-0 (2024).

11 Tanno, R. *et al.* Collaboration between clinicians and vision-language models in radiology report generation. *Nat Med* **31**, 599-608, doi:10.1038/s41591-024-03302-1 (2025).

12 Chen, Z. *et al.* A Vision-Language foundation model to enhance efficiency of chest x-ray interpretation. *arXiv e-prints*, arXiv: 2401.12208 (2024).

13 Dubey, A. *et al.* The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

14 Cluceru, J. *et al.* Improving the noninvasive classification of glioma genetic subtype with deep learning and diffusion-weighted imaging. *Neuro Oncol* **24**, 639-652, doi:10.1093/neuonc/noab238 (2022).

15 van der Voort, S. R. *et al.* Combined molecular subtyping, grading, and segmentation of glioma using multi-task deep learning. *Neuro Oncol* **25**, 279-289, doi:10.1093/neuonc/noac166 (2023).

16 Choi, Y. S. *et al.* Fully automated hybrid approach to predict the IDH mutation status of gliomas via deep learning and radiomics. *Neuro Oncol* **23**, 304-313, doi:10.1093/neuonc/noaa177 (2021).

17 Byeon, Y. *et al.* Interpretable multimodal transformer for prediction of molecular subtypes

and grades in adult-type diffuse gliomas. *NPJ Digit Med* **8**, 140, doi:10.1038/s41746-025-01530-4 (2025).

18 Xiao, H. *et al.* A comprehensive survey of large language models and multimodal large language models in medicine. *Information Fusion* **117**, 102888 (2025).

19 Li, C. *et al.* Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems* **36**, 28541-28564 (2023).

20 Tu, T. *et al.* Towards generalist biomedical AI. *Nejm Ai* **1**, AIoa2300138 (2024).

21 Hoopes, A., Butoi, V. I., Guttag, J. V. & Dalca, A. V. Voxelprompt: A vision-language agent for grounded medical image analysis. *arXiv preprint arXiv:2410.08397* (2024).

22 Tak, D. *et al.* A foundation model for generalized brain MRI analysis. *medRxiv*, doi:10.1101/2024.12.02.24317992 (2024).

23 Park, S. H. & Kim, N. Challenges and Proposed Additional Considerations for Medical Device Approval of Large Language Models Beyond Conventional AI. *Radiology* **312**, e241703, doi:10.1148/radiol.241703 (2024).

24 Yi, P. H. *et al.* Best Practices for the Safe Use of Large Language Models and Other Generative AI in Radiology. *Radiology* **316**, e241516, doi:10.1148/radiol.241516 (2025).

25 Faghani, S. *et al.* Quantifying Uncertainty in Deep Learning of Radiologic Images. *Radiology* **308**, e222217, doi:10.1148/radiol.222217 (2023).

26 Park, Y. W. *et al.* Differentiation of glioblastoma from solitary brain metastasis using deep ensembles: Empirical estimation of uncertainty for clinical reliability. *Comput Methods Programs Biomed* **254**, 108288, doi:10.1016/j.cmpb.2024.108288 (2024).

27 Louis, D. N. *et al.* cIMPACT-NOW update 1: Not Otherwise Specified (NOS) and Not Elsewhere Classified (NEC). *Acta Neuropathol* **135**, 481-484, doi:10.1007/s00401-018-1808-0 (2018).

28 Gutman, D. A. *et al.* MR imaging predictors of molecular profile and survival: multi-institutional study of the TCGA glioblastoma data set. *Radiology* **267**, 560-569, doi:10.1148/radiol.13120118 (2013).

29 Calabrese, E. *et al.* The University of California San Francisco Preoperative Diffuse Glioma MRI Dataset. *Radiol Artif Intell* **4**, e220058, doi:10.1148/ryai.220058 (2022).

30 Isensee, F. *et al.* Automated brain extraction of multisequence MRI using artificial neural networks. *Hum Brain Mapp* **40**, 4952-4964, doi:10.1002/hbm.24750 (2019).

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

32 Zhang, S. *et al.* A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI* **2**, AIoa2400640 (2025).

33 Yang, L. *et al.* Advancing multimodal medical capabilities of Gemini. *arXiv preprint arXiv:2405.03162* (2024).

34 Hasani, A. M. *et al.* Evaluating the performance of Generative Pre-trained Transformer-4 (GPT-4) in standardizing radiology reports. *Eur Radiol* **34**, 3566-3574, doi:10.1007/s00330-023-10384-x (2024).

Figure Legends.

Figure 1. The distributions of a) overall correctness and b) readability for RRGs and original reports, as well as the c) overall preference scores comparing RRGs with original reports. Overall 37.8% of RRGs were considered superior or equal to the original report, while overall 91.0% of RRGs were considered clinically acceptable compared to original reports.

RRG = radiology report generation

Figure 2. Representative case of molecular status prediction and RRG on the AMC dataset. a) MRI, original report, and generated report of a 38-year-old male patient with IDH-wildtype glioblastoma. The model correctly predicted the IDH mutation status ("IDH-wildtype"). Compared with the original report, the generated report was generally correct and was evaluated as "clinically acceptable" by all three readers. In terms of readability, the RRG was labeled as "very satisfied" or "satisfied" by three readers.

b) MRI, original report, and generated report of a 65-year-old female patient with oligodendroglioma. The model correctly predicted the IDH mutation status ("IDH-mutant"). Compared with the original report, the generated report was generally correct and was evaluated as "clinically acceptable" by all three readers. In terms of readability, the RRG was labeled as "very satisfied" or "satisfied" by three readers.

RRG = radiology report generation; AMC = Asan Medical Center; IDH = isocitrate dehydrogenase

Figure 3. Representative cases of hallucination from the RRGs.

a) MRI, original report, and generated report of a 71-year-old male patient with IDH-wildtype glioblastoma. Compared with the original report, the generated report hallucinated that there was a separate contrast-enhancing lesion at the right frontal lobe (highlighted in red).

b) MRI, original report, and generated report of a 40-year-old female patient with oligodendroglioma. Compared with the original report, the generated report hallucinated that there was an arachnoid cyst (highlighted in red).

IDH = isocitrate dehydrogenase; RRG = radiology report generation

Figure 4. Patient flowchart.

AMC = Asan Medical Center, TCGA = The Cancer Genome Atlas, UCSF = University of San Francisco

Figure 5. Schematic of Glio-LLaMA-Vision. Training and validation framework of Glio-LLaMA-Vision. a) Pre-training of Glio-LLaMA-Vision using image-text pairs from PubMed Central. b) Fine-tuning of Glio-LLaMA-Vision using MRI-report pairs from adult-type diffuse gliomas. c) Performance validation.