MM-NeuroOnco: A Multimodal Benchmark and Instruction Dataset for MRI-Based Brain Tumor Diagnosis

Feng Guo[1,2*] **Jiaxiang Liu**[1*] **Yang Li**[1] **Qianqian Shi**[3] **Mingkun Xu**[1†]
[1]Guangdong Institute of Intelligence Science and Technology, Hengqin, China
[2]School of Information Science and Technology, ShanghaiTech University, Shanghai, China
[3]Center for Brain-Inspired Computing Research (CBICR),
Department of Precision Instrument, Tsinghua University, Beijing, China
{guofeng,liujiaxiang,liyang,xumingkun}@gdiist.cn
qqshi@mail.tsinghua.edu.cn

Abstract

Brain tumor MRI diagnosis requires models not only to identify lesions but also to perform fine-grained medical reasoning based on imaging modality, tumor location, and key imaging findings. However, existing medical multimodal datasets often lack structured diagnostic semantics, and closed-ended evaluations may overestimate model ability due to option cues or elimination shortcuts. To address this issue, we introduce **MM-NeuroOnco**, a large-scale multimodal instruction dataset and benchmark for brain tumor MRI diagnostic reasoning, containing 24,726 MRI slices from 20 public data sources and approximately 200,000 multimodal instructions grounded in imaging evidence. We develop a multi-model collaborative pipeline for semantic completion and quality control, which supplements original tumor labels and segmentation masks with key diagnostic attributes, and further construct **MM-NeuroOnco-Bench** with a no-LLM-inference policy for ground-truth labels. The benchmark adds a "None of the above" option to closed-ended questions to reduce evaluation bias caused by option shortcuts. Experiments show that existing general-purpose and medical-specialist multimodal large models remain limited on this task, with the strongest baseline achieving only 41.88% accuracy on diagnosis-related questions. Fine-tuning on MM-NeuroOnco yields **NeuroOnco-GPT**, achieving a 27% absolute accuracy improvement and demonstrating the effectiveness of our dataset in enhancing fine-grained multimodal medical reasoning. Code and dataset documentation are available at https://github.com/gfnnnb/MM-NeuroOnco.

1 Introduction

Brain tumor MRI diagnosis requires models not only to identify lesions but also to make diagnostic judgments by integrating imaging modality, lesion location, and key imaging findings (Bray et al., 2024; Csutak et al., 2020; Price et al., 2024). Although existing brain tumor MRI benchmarks have advanced segmentation and pixel-level lesion modeling (Menze et al., 2014), segmentation masks mainly describe the spatial extent of lesions and do not directly capture the medical semantics required for diagnosis. A model that accurately delineates a tumor region may still fail to understand the relationship between MRI manifestations and tumor types (Setyawan et al., 2024). Therefore, brain tumor MRI understanding needs to move beyond lesion localization toward evidence-grounded multimodal diagnostic reasoning.

Existing medical multimodal reasoning datasets and benchmarks provide important foundations for this direction. For example, Step-CoT and MedThinkVQA mainly rely on radiology reports or teaching cases to organize diagnostic reasoning processes (Fan et al., 2026; Yao et al., 2026). CXReasonBench and MedLesionVQA construct medical reasoning and lesion-level question answering tasks using anatomical segmentation, medical rules, or expert annotations (Lee et al., 2025; Yu et al., 2026). However, public brain tumor MRI datasets usually lack paired radiology reports and fine-grained diagnostic descriptions, and more commonly provide only tumor categories, MRI modality information, or segmentation masks. Therefore, how to construct data for diagnostic reasoning training and evaluation under such sparse annotation settings remains insufficiently explored.

In addition, existing medical VQA evaluations often adopt closed-ended multiple-choice formats. Although this format enables automatic scoring, models may rely on option cues, language priors, or elimination strategies to select the correct answer without truly understanding the image content. Therefore, benchmarks for brain tumor MRI should also reduce shortcut biases in closed-ended

[*]These authors contributed equally to this work.
[†]Corresponding author.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2818–2845
November 6-10, 2026 ©2026 Association for Computational Linguistics

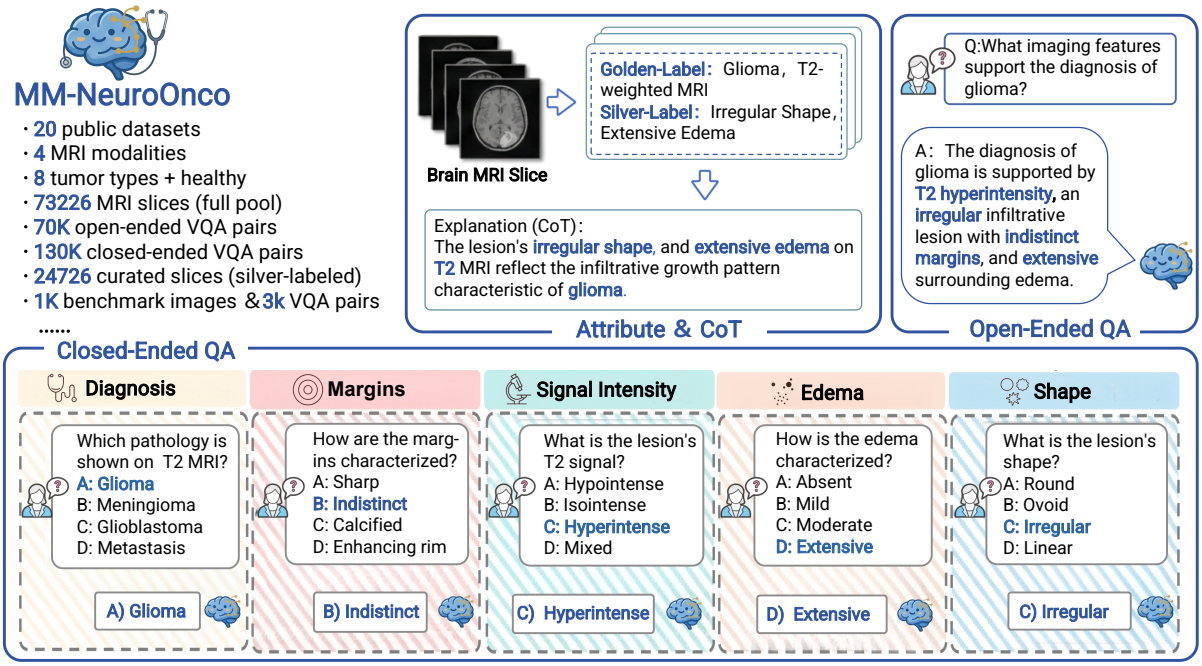


Figure 1: Overview of MM-NeuroOnco. The dataset is constructed from 20 public brain MRI sources and contains 24,726 curated MRI slices. Original labels are augmented with fine-grained diagnostic attributes and evidence-grounded CoT rationales. MM-NeuroOnco supports both open-ended diagnostic reasoning and closed-ended QA across multiple clinical categories.

question answering, so as to more faithfully reflect the model's diagnostic understanding ability.

To address these challenges, as illustrated in Figure 1, we introduce MM-NeuroOnco, a multimodal instruction dataset and benchmark for brain tumor MRI diagnostic reasoning. Unlike existing paradigms that rely on paired radiology reports, teaching cases, or expert-intensive annotations, our work starts from the types of supervision commonly available in public brain tumor MRI datasets, including original tumor labels, MRI modality information, and segmentation masks. We then construct diagnosis-related semantics through a multi-model collaborative semantic completion process, and further design a rejection-aware evaluation protocol to more reliably assess models' diagnostic reasoning ability.

The main contributions of this work are threefold:

- **A multimodal dataset for brain tumor MRI diagnostic reasoning.** We construct MM-NeuroOnco, consisting of 24,726 MRI slices and approximately 200,000 semantically enriched instructions, covering multiple MRI modalities, eight brain tumor subtypes, and healthy controls. We further build MM-NeuroOnco-Bench for evaluation.
- **A data construction pipeline from sparse annotations to diagnostic semantics.** We use multi-model collaborative semantic completion to supplement diagnosis-related attributes based on original labels, MRI modality information, and segmentation masks, thereby alleviating the lack of fine-grained medical semantics in public brain tumor MRI data and reducing the dependence on costly large-scale manual annotation.
- **Rejection-aware evaluation and specialist model validation.** We introduce a rejection-aware evaluation protocol to reduce option-shortcut bias in closed-ended questions. We also fine-tune NeuroOnco-GPT based on MM-NeuroOnco, validating the effectiveness of semantically enriched instruction data for improving brain tumor MRI diagnostic understanding.

2 Related Work

2.1 Brain Tumor Datasets and Benchmarks

Brain tumor MRI analysis has long been an important topic in medical imaging and has motivated public datasets such as BraTS (Brain Tumor Segmentation) (Menze et al., 2014; Bakas et al., 2017;

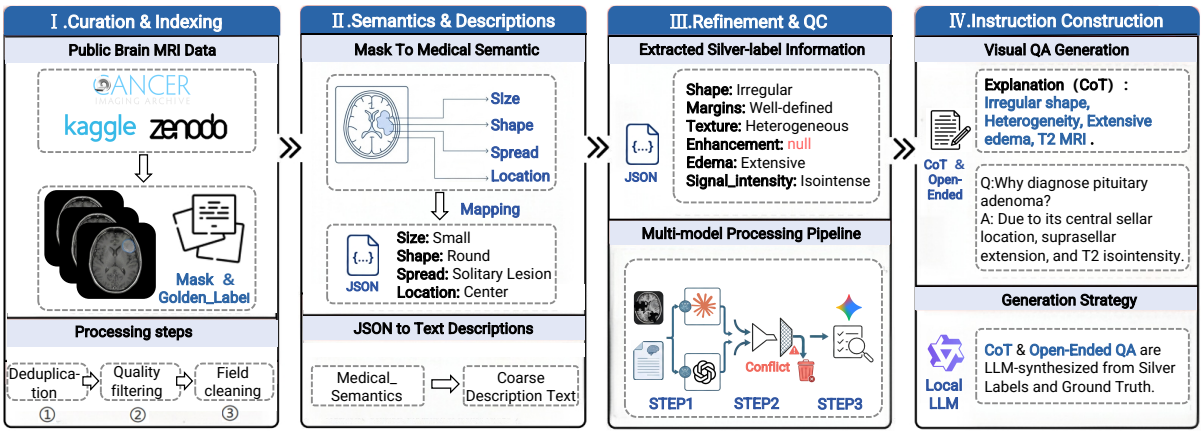


Figure 2: The MM-NeuroOnco dataset curation pipeline, which consists of four sequential steps transforming raw MRI data into semantically enriched multimodal instructions.

Baid et al., 2021; Adewole et al., 2023). BraTS provides multi-modal MRI scans, including T1, T2, FLAIR, and T1ce, with pixel-level annotations, and has become a standard benchmark for brain tumor segmentation. The Cancer Imaging Archive (TCIA) also collects brain tumor MRI datasets covering diverse tumor types, acquisition protocols, and clinical scenarios (Clark et al., 2013). Project Imaging-X reviews over 1,000 public medical imaging datasets and analyzes their modalities, task types, and anatomical coverage (Deng et al., 2026). These resources have greatly supported brain tumor research.

2.2 Evaluation Paradigms in Medical VQA

Medical VQA evaluation commonly includes two forms: closed-ended multiple-choice questions and open-ended generation. Closed-ended questions are convenient for scoring and model comparison, but when complex clinical judgments are compressed into a limited set of options, models may rely on option cues or elimination strategies, leading to overestimated diagnostic understanding (Chen et al., 2024; Yue et al., 2024). Open-ended generation provides a larger response space and is more suitable for evaluating reasoning coherence, but its scoring remains unstable. Traditional text metrics such as BLEU and ROUGE correlate poorly with clinical correctness (Lau et al., 2018; He et al., 2020), and generative models may produce medically plausible but factually incorrect content (Pal et al., 2023; Xia et al., 2024). Prior studies have explored structured constraints, auxiliary supervision, reasoning traces, and parameter-efficient adaptation in medical vision-language learning (Gai et al., 2024; Liu et al., 2024; Wang et al., 2025; Liu et al., 2023, 2026). Building on this line of work, we further focus on option-shortcut bias in closed-ended question answering and introduce a rejection-aware evaluation setting in MM-NeuroOnco-Bench, as detailed in Section 5.1.

3 MM-NeuroOnco Dataset Construction

The construction of MM-NeuroOnco consists of four stages: data curation and standardization, semantic attribute extraction, multi-model semantic completion and quality control, and instruction data construction. The overall workflow is illustrated in Figure 2.

3.1 Data Curation and Standardization

We aggregate over 100,000 brain MRI scans from public repositories, including Kaggle, Zenodo, and TCIA. Since these sources differ substantially in modality naming, file organization, and diagnostic labels, we design a Unified Metadata Schema to convert all samples into a standardized JSON index format, providing a consistent access interface for multi-source data.

Quality filtering removes unreadable, blank, or severely corrupted images and samples without valid MRI content, regardless of diagnostic category or sample difficulty. After deduplication and quality filtering, we establish a Full Master Index containing 73,226 valid samples, of which 19,086 are paired with pixel-level segmentation masks. This index supports subsequent semantic completion, task sampling, and benchmark construction. To reduce potential data leakage between the training set and the evaluation benchmark, for samples derived from 3D volumes, each case contributes at most one representative 2D slice. For 2D public

datasets without reliable case identifiers, we apply pixel-level deduplication and similarity-based hash checking after unified preprocessing to reduce duplicates or highly similar images.

3.2 Semantic Attribute Extraction

To convert pixel-level masks into structured information usable for language-based reasoning, we follow Vepa et al. (2025) and extract lesion morphology, localization, and spatial spread through deterministic geometric mapping.

First, we use the circularity metric C to describe lesion shape regularity:

$$C = \frac{4\pi A}{P^2}, \tag{1}$$

where A and P denote the lesion area and perimeter, respectively. A lower value of C indicates a more irregular lesion shape.

Second, we compute the lesion centroid (C_x, C_y) using spatial image moments to describe its location within the imaging plane:

$$C_x = \frac{M_{1,0}}{M_{0,0}}, \quad C_y = \frac{M_{0,1}}{M_{0,0}}, \tag{2}$$

where $M_{p,q}$ denotes the (p, q)-th order spatial moment of the binary lesion mask.

Finally, we use the dominant component ratio f_{core} to characterize the spatial distribution of the lesion:

$$f_{\text{core}} = \frac{A_{\max}}{\sum_i A_i}, \tag{3}$$

where A_i denotes the area of the i-th connected component, and $A_{\max}$ denotes the largest connected component. A lower value of f_{core} indicates a more dispersed lesion distribution.

These mask-derived attributes provide deterministic geometric labels for the benchmark and constraints for subsequent semantic extraction, whereas richer silver-standard attributes are extracted from images to support instruction construction. For missing metadata, unavailable fields are explicitly marked as "unknown" to avoid treating missing information as certain facts. The complete threshold settings and mapping rules are provided in the Appendix.

3.3 Multi-Model Refinement and Quality Control

Although segmentation masks are available for some samples, raw multi-source MRI datasets generally lack descriptive radiological attributes required for clinical reasoning, such as enhancement patterns, edema severity, and intratumoral texture. To supplement silver-standard semantics for instruction construction at scale, we design an automated multi-model collaborative pipeline, as illustrated in Stage III of Figure 2.

The framework follows a "dual-path extraction–consensus fusion–visual verification" mechanism, using cross-checking among heterogeneous models to reduce the risk of unreliable medical semantics generated by a single model.

Stage 1: Dual-Path Independent Extraction. We use two heterogeneous vision-language models (VLMs) to independently analyze each MRI slice and extract medical semantics into a unified structured format.

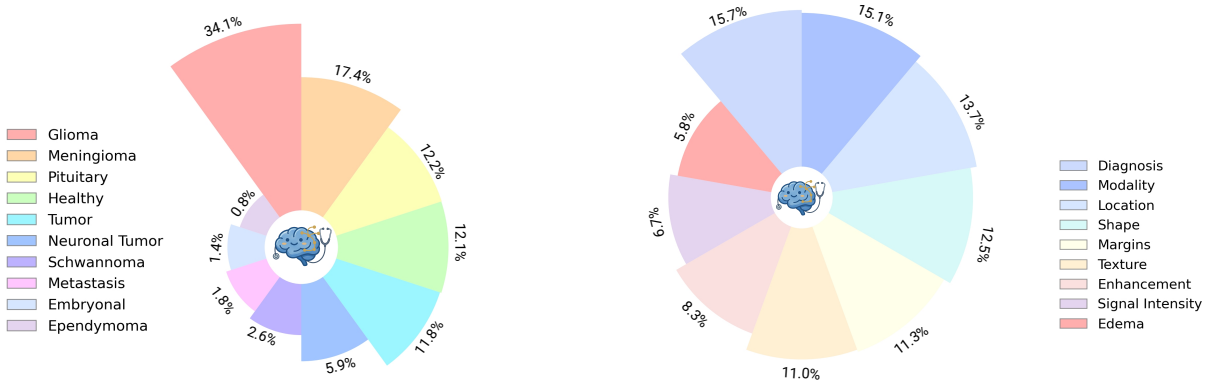
Stage 2: Consensus-Based Fusion. We adopt a field-level fusion strategy to merge the outputs of the two models. Fields with exact agreement are directly retained. For minor discrepancies in degree descriptions, we apply a downgrade strategy. For example, conflicts such as "mild" versus "moderate" are mapped to a broader label, "Present". For fields with substantial semantic conflicts, we set them to null to reduce unreliable attributes entering the final data.

Stage 3: Visual Final Review and Quality Control. To further reduce subtle errors, we introduce a third visual model to review the candidate labels. This stage follows a "subtraction-only" mechanism: the model is allowed only to reject untrustworthy fields or remove high-risk samples, and is not allowed to add new diagnostic information. This design further reduces unreliable content in automatically generated semantics.

Through this process, we partially alleviate the lack of diagnostic semantics in the original data and provide structured medical attributes for multimodal instruction construction. Detailed extraction prompts and explanations are provided in the Appendix.

3.4 Instruction Data Construction

Instruction generation is grounded in gold labels from the original datasets and silver-standard labels generated by the multi-model pipeline. We use a locally deployed Qwen3-80B-Instruct to synthesize these labels into closed-ended and open-ended VQA samples. Closed-ended questions include visually or pathologically similar hard distractors, while CoT answers are constructed by combining



(a) The category distribution of the MM-NeuroOnco dataset.

(b) The statistics of question types in the training set.

Figure 3: Statistical distributions of the MM-NeuroOnco dataset.

Table 1: Silver-standard annotation quality audit results.

Verifier	Fields	Error (%)	Omission (%)
Human1	900	7.00	17.22
Human2	225	5.78	16.00
Total	**1125**	**6.76**	**16.98**

silver-standard attributes with mask-derived geometric features. Detailed prompts, distractor construction rules, and CoT templates are provided in the Appendix.

4 MM-NeuroOnco Dataset Analysis

4.1 Instruction Dataset Statistics

We use the multi-model semantic completion pipeline to process over 20,000 samples and construct the MM-NeuroOnco instruction dataset. During sampling, we do not enforce a fully balanced category distribution; instead, we preserve the natural long-tailed distribution of multi-source public brain tumor datasets as much as possible. As shown in Figure 3(a), common categories such as glioma, meningioma, and pituitary account for a relatively large proportion, while several rare tumor subtypes contain fewer samples. This distribution is closer to the category imbalance in public brain tumor data resources and provides a basis for analyzing model performance in long-tailed diagnostic scenarios.

Through semantic completion, we supplement each sample with fine-grained medical attributes. On average, each sample is augmented with 2.4 diagnosis-related attributes, covering enhancement, edema, texture, morphology, and other information. The distribution of question types is shown in Figure 3(b).

4.2 Benchmark Composition

To establish an independent evaluation standard, we construct MM-NeuroOnco-Bench, which contains 1,000 images and over 3,000 QA samples, including more than 2,000 closed-ended questions and over 1,000 open-ended questions. Its modality distribution is shown in Figure 8, where the four MRI modalities are relatively balanced in both the instruction dataset and the benchmark. MM-NeuroOnco-Bench adopts a coverage-oriented stratified sampling strategy that considers both tumor subtype and MRI modality, and moderately supplements rare tumor categories to avoid domination by head classes. Detailed sampling rules and category distributions are provided in the Appendix.

MM-NeuroOnco-Bench adopts a "no-LLM-inference" policy for ground-truth labels. All evaluation labels are derived from human annotation or deterministic rule-based mappings from segmentation masks. LLMs are used only for question rephrasing and linguistic diversification, and do not participate in generating diagnostic fact labels.

4.3 Silver-Standard Label Quality Audit

To assess the quality of silver-standard attributes generated by the automated semantic completion pipeline, we manually audit structured medical attributes in the training set. Two human verifiers examine 1,125 structured medical attribute fields, covering diagnosis-related semantics such as enhancement, edema, texture, and signal intensity. As shown in Table 1, the overall error rate is 6.76%, indicating that the multi-model collaborative pipeline can control the introduction of incorrect attributes to a certain extent. More details on the audit setup and result analysis are provided in the Appendix.

Table 2: Overall performance comparison on the closed-ended benchmark. Best results are highlighted in **bold**, and second-best results are underlined. Colored deltas indicate performance changes introduced by Chain-of-Thought (CoT) supervision. Qwen3-VL-8B is evaluated in a zero-shot setting as an open-source baseline.

Model	Diagnosis	Size	Shape	Spread	Location	**Overall**
General-Purpose LVLMs						
GPT-5.1	37.0	32.3	21.2	58.0	38.1	37.2
Gemini-3-Flash	41.9	29.8	30.8	57.9	41.8	40.9
Claude-Sonnet-4.0	33.2	34.5	28.3	54.4	35.5	35.9
Qwen3-VL-8B	24.3	31.6	30.0	47.2	30.6	30.1
Medical Specialist LVLMs						
HuLuMed-32B	38.2	27.7	32.9	52.1	33.2	37.3
Lingshu-7B	39.2	30.0	32.9	45.0	37.5	37.6
Lingshu-32B	36.0	28.3	26.1	52.8	33.9	35.6
HuLuMed-7B	36.4	30.3	**40.7**	28.0	29.0	34.0
MedGemma-27B	33.1	29.0	32.9	33.2	27.7	31.8
MedGemma-1.5-4B	26.9	28.0	28.0	23.5	25.1	26.5
Ours (Domain-Adapted LVLMs)						
NeuroOnco-GPT [Ours]	40.6	42.0	27.7	34.5	**53.8**	40.0
NeuroOnco-GPT (CoT) [Ours]	**51.4** (+10.8)	**42.4** (+0.4)	40.4 (+12.7)	**70.4** (+35.9)	52.1 (−1.7)	**51.4** (+11.4)

Table 3: Rejection-mechanism ablation across 4-N, 5-N, and 5-R settings.

Model	**4-N**	**5-N**	**5-R**
HuLuMed-32B	59.6	56.9 (−2.7)	49.4 (−10.2)
Lingshu-32B	57.2	55.8 (−1.4)	46.9 (−10.3)
GPT-5.1	59.0	56.7 (−2.3)	49.6 (−9.4)
Average	**58.6**	**56.5** (−2.1)	**48.7** (−10.0)

5 MM-NeuroOnco Bench

5.1 Rejection-Aware Evaluation Strategy

We introduce a rejection-aware evaluation strategy for closed-ended questions in MM-NeuroOnco-Bench. Specifically, we add a fifth option, "None of the above," to each question. This setting means that the correct answer is not always provided by the regular options, thereby reducing the tendency of models to select a specific answer even when the candidates do not match the imaging evidence.

To distinguish the effect of the rejection mechanism from the effect of simply increasing the number of options, we compare three evaluation settings: the standard four-option setting, the five-option setting with a regular distractor, and the five-option setting with "None of the above." By comparing model performance under these settings, we can analyze whether models rely on option cues or elimination strategies. The ablation results are presented in Section 6.3.

5.2 Evaluation Metrics

We adopt different evaluation metrics for closed-ended and open-ended questions. For closed-ended questions, we use Accuracy as the evaluation metric. The model output is parsed into one of the candidate options, including either a regular option or "None of the above," and then matched against the ground-truth answer.

For open-ended questions, following previous works (Hao et al., 2026a,b), we construct a specialized evaluation prompt and use a locally deployed Qwen3-80B-Instruct to assist with scoring. During evaluation, the judge model considers the input question, the ground-truth answer, and the model output, and assigns a score from 0 to 10 according to a hierarchical scoring rubric. To improve the rigor of medical evaluation, the rubric includes constraints for critical errors.

For errors such as laterality confusion, unsupported medical details, or diagnostic conclusions that contradict the ground truth, the score is capped. The full evaluation prompt and scoring rubric are provided in the supplementary materials.

6 Experiments

6.1 Experimental Setup

To evaluate the capability boundaries of existing models on brain tumor MRI diagnostic understanding, we conduct experiments on 10 representative LVLMs and group them into two categories.

The first category is general-purpose multimodal large models. We select closed-source models such



(a) The word cloud for Closed-ended questions (b) The word cloud for Open-ended questions (c) The word cloud for Benchmark

Figure 4: Word cloud visualizations illustrating the linguistic distributions of (a) closed-ended questions, (b) open-ended questions, and (c) the evaluation benchmark.

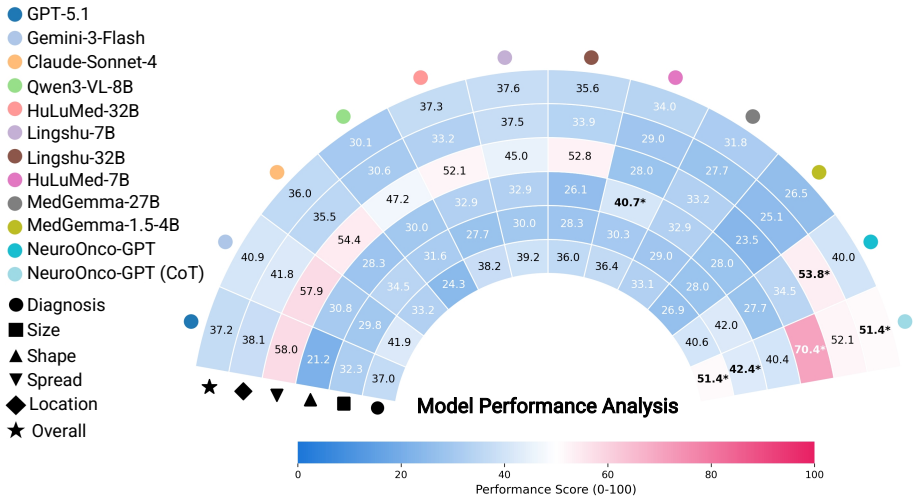


Figure 5: Heatmap comparing the performance of multiple LVLMs on closed-ended QA tasks across different dimensions.

Table 4: Performance comparison on common and rare tumor categories.

Model	Common	Rare	Gap
Lingshu-32B	29.4	19.4	10.0
GPT-5.1	49.6	21.2	28.3
Gemini-3-Flash	53.1	32.9	20.2
HuLuMed-32B	49.7	20.5	29.2
Claude-Sonnet-4.0	**57.9**	10.1	47.8
NeuroOnco-GPT	53.9	**49.2**	**4.6**

as GPT-5.1, Gemini-3-Flash, and Claude-Sonnet-4.0, which represent strong general multimodal reasoning capabilities.

The second category is medical specialist models. We include open-source vertical-domain models with parameter sizes ranging from 1.5B to 32B, such as HuLuMed, Lingshu, and MedGemma. These models have been continually pre-trained or fine-tuned on biomedical tasks and thus have certain medical-domain adaptation.

Implementation Details. Closed-source models are accessed through official APIs. Open-source models are evaluated on a server with 4 NVIDIA H100 80G GPUs. For closed-ended tasks, we use Accuracy as the metric. For open-ended questions, we adopt the LLM-as-a-Judge paradigm described in Section 5 and use Qwen3-80B-Instruct as the judge model.

Validation via SFT. To validate the effectiveness of the MM-NeuroOnco instruction data, we perform supervised fine-tuning (SFT) on Qwen3-VL-8B as the base model. Training is conducted with the LLaMA-Factory framework using LoRA. All hyperparameters are kept at their default settings, and the model is trained for one epoch. The resulting model is named NeuroOnco-GPT.

6.2 Experimental Results

Based on the overall performance on MM-NeuroOnco-Bench (Table 2), we compare general-

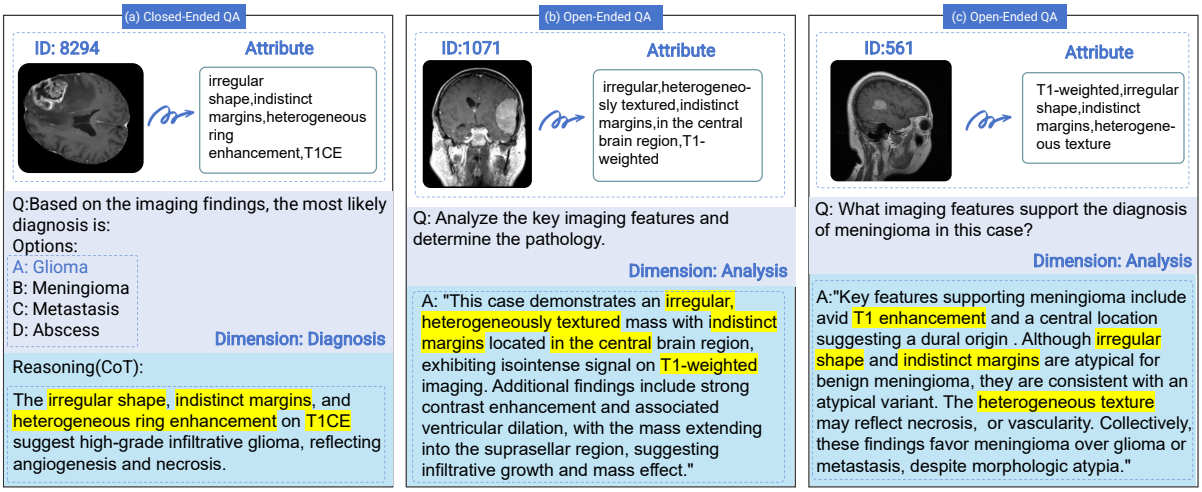


Figure 6: Three representative case studies for closed-ended and open-ended QA.

purpose multimodal large models, medical specialist models, and NeuroOnco-GPT fine-tuned on our dataset. The results show that MM-NeuroOnco-Bench remains challenging for existing LVLMs. Even the strongest commercial general-purpose model, Gemini-3-Flash, achieves only 40.9% overall accuracy. Medical specialist models also show no consistent advantage: HuLuMed-32B and Lingshu-7B achieve 37.3% and 37.6% accuracy, respectively, both lower than Gemini-3-Flash. This suggests that relying solely on general multimodal capability or existing biomedical pre-training remains insufficient for brain tumor MRI diagnostic reasoning.

In contrast, NeuroOnco-GPT, fine-tuned on MM-NeuroOnco, achieves a clear improvement on diagnosis-related questions, validating the effectiveness of semantically enriched instruction data. To further analyze performance under long-tailed categories, we compare models on common and rare tumor categories, as shown in Table 4. Most models perform substantially worse on rare categories, while NeuroOnco-GPT reduces the common–rare gap to 4.6%, suggesting that MM-NeuroOnco helps mitigate long-tail bias.

6.3 Ablation Study

We conduct two ablation studies to examine the impact of the rejection-aware evaluation setting and the effect of CoT supervision on model fine-tuning.

Impact of the Rejection Setting. To analyze model performance under different closed-ended evaluation settings, we compare three configurations: a standard four-option setting, a five-option setting with a regular distractor, and a five-option setting with "None of the above", as shown in Table 3. Adding a regular distractor leads to only a minor performance decrease (-2.13%), whereas introducing the rejection option results in a larger accuracy drop of 9.95%, corresponding to an additional 7.82% decrease relative to the regular five-option setting. This indicates that the rejection-aware setting does not merely increase the number of options, but also makes it more difficult for models to rely on option cues or elimination strategies, thereby providing a stricter evaluation of whether models can judge the correspondence between imaging evidence and candidate answers.

The Effect of CoT Supervision. We further evaluate the effect of CoT data on fine-tuning. Compared with instruction data without CoT, incorporating explicit CoT reasoning paths further improves model performance, with an overall gain of 11.4 percentage points. However, Location accuracy decreases from 53.8% to 52.1%, indicating that the gains from CoT supervision are not uniform across tasks. This result suggests that organizing imaging findings, lesion attributes, and diagnostic decisions into a structured reasoning process helps models learn more stable diagnostic evidence, rather than relying only on shallow image–label associations.

6.4 Case and Failure Mode Analysis

Figure 6 shows representative examples from our instruction dataset, covering closed-ended QA and open-ended diagnostic reasoning. The questions, options, and CoT rationales are generated from structured medical attributes, illustrating how imag-

ing evidence is linked to diagnostic answers. More examples are provided in the Appendix.

We further summarize recurring failure modes, including rejection failure, rare-category confusion, insufficient evidence grounding, and over-interpretation. These observations show that existing LVLMs struggle with evidence utilization, long-tail generalization, and rejection ability in fine-grained brain tumor MRI diagnostic reasoning.

7 Conclusion

In this work, we introduce MM-NeuroOnco, a multimodal instruction dataset and benchmark for brain tumor MRI diagnostic reasoning. MM-NeuroOnco enriches public MRI data with diagnosis-related attributes and evidence-grounded instruction samples through a multi-model semantic completion and quality control pipeline.

We further construct MM-NeuroOnco-Bench with no-LLM-inference ground-truth labels and a rejection-aware evaluation setting. Experiments show that existing general-purpose and medical-specialist LVLMs remain limited on this task, while NeuroOnco-GPT, fine-tuned on MM-NeuroOnco, improves diagnosis-related performance and achieves more balanced results on rare tumor categories. Overall, MM-NeuroOnco provides a useful resource for brain tumor MRI reasoning and evaluation.

Data and Code Availability. Code, dataset metadata, documentation, and construction scripts are available at: `https://github.com/gfnnnb/MM-NeuroOnco`. The organized image files are available through: `https://huggingface.co/datasets/gfnnnb/MM-NeuroOnco-Images`, subject to the repository's access conditions and the licenses of the original datasets. BraTS 2021 images are not redistributed and must be obtained separately from the original source.

AI Assistance Statement

In addition to the model-based components explicitly described in the dataset construction and evaluation pipeline, AI tools were used for language polishing, formatting assistance, and code debugging. All scientific claims, experimental design, data analysis, and final manuscript content were reviewed and verified by the authors.

Limitations

Although this work constructs a dataset and evaluation benchmark for brain tumor MRI diagnostic understanding, it still has several limitations. First, MM-NeuroOnco-Bench is mainly built from single 2D MRI slices and does not fully utilize the spatial context in complete 3D volumes. Second, part of the diagnostic semantics in MM-NeuroOnco comes from an automated multi-model collaborative pipeline. Although we adopt conservative fusion, visual verification, and manual audit for quality control, the silver-standard attributes may still be affected by image quality, errors in the original annotations, or model judgment bias. In addition, MM-NeuroOnco-Bench moderately supplements rare tumor categories to ensure sufficient coverage, so it is more focused on evaluating brain tumor MRI diagnostic understanding ability and does not represent a complete diagnostic workflow under real clinical distributions.

Ethical Considerations

All data used in this work are from publicly available brain tumor MRI datasets, following a data usage paradigm consistent with medical imaging benchmarks such as BraTS. These datasets were released by their original providers in de-identified form and do not contain personally identifiable information or protected health information. All samples used in this work can be traced back to their original sources, and we strictly follow the corresponding data usage agreements and licenses.

The dataset, benchmark, and model developed in this work are intended only for academic research. They aim to promote the development of multimodal diagnostic understanding methods for brain tumor MRI and should not be used for direct clinical diagnosis or as a substitute for professional medical decision-making.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 62506084 and 62606111, and the High-Level Talents Innovation Team Project of the Guangdong-Macao In-Depth Cooperation Zone in Hengqin (Grant No. 2630004018939).

References

Maruf Adewole, Jeffrey D Rudie, Anu Gbadamosi, Oluyemisi Toyobo, Confidence Raymond, Dong Zhang, Olubukola Omidiji, Rachel Akinola, Mohammad Abba Suwaid, Adaobi Emegoakor, et al. 2023. The brain tumor segmentation (BraTS) challenge 2023: Glioma segmentation in sub-saharan africa patient population (BraTS-Africa). *arXiv preprint arXiv:2305.19369*.

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. 2021. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. 2017. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13.

Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. 2024. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 74(3):229–263.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, et al. 2024. GMAI-MMBench: A comprehensive multimodal evaluation benchmark towards general medical AI. *Advances in Neural Information Processing Systems*, 37:94327–94427.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. 2013. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6):1045–1057.

Csaba Csutak, Paul-Andrei Ștefan, Lavinia Manuela Lenghel, Cezar Octavian Moroșanu, Roxana-Adelina Lupean, Larisa Șimonca, Carmen Mihaela Mihu, and Andrei Lebovici. 2020. Differentiating high-grade gliomas from brain metastases at magnetic resonance: the role of texture analysis of the peritumoral zone. *Brain Sciences*, 10(9):638.

Zhongying Deng, Cheng Tang, Ziyan Huang, Jiashi Lin, Ying Chen, Junzhi Ning, Chenglong Ma, Jiyao Liu, Wei Li, Yinghao Zhu, et al. 2026. Project Imaging-X: A survey of 1000+ open-access medical imaging datasets for foundation model development. *arXiv preprint arXiv:2603.27460*.

Lin Fan, Yafei Ou, Zhipeng Deng, Pengyu Dai, Chongxian Hou, Jiale Yan, Yaqian Li, Kaiwen Long, Xun Gong, Masayuki Ikebe, et al. 2026. Step-CoT: Stepwise visual chain-of-thought for medical visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2049–2059.

Xiaotang Gai, Chenyi Zhou, Jiaxiang Liu, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. Medthink: Explaining medical visual question answering via multimodal decision-making rationale. *arXiv preprint arXiv:2404.12372*.

Jing Hao, Yuxuan Fan, Yanpeng Sun, Kaixin Guo, Lizhuo Lin, Jinrong Yang, Qi Yong H Ai, Lun M Wong, Hao Tang, and Kuo Feng Hung. 2026a. Towards better dental AI: A multimodal benchmark and instruction dataset for panoramic X-ray analysis. *Advances in Neural Information Processing Systems*, 38.

Jing Hao, Yuci Liang, Lizhuo Lin, Yuxuan Fan, Wenkai Zhou, Kaixin Guo, Zanting Ye, Yanpeng Sun, Xinyu Zhang, Yanqi Yang, et al. 2026b. OralGPT-Omni: A versatile dental multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 38509–38519.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):180251.

Hyungyung Lee, Geon Choi, Jung-Oh Lee, Hangyul Yoon, Hyuk Gi Hong, and Edward Choi. 2025. CXReasonBench: A benchmark for evaluating structured diagnostic reasoning in chest X-rays. *arXiv preprint arXiv:2505.18087*.

Jiaxiang Liu, Tianxiang Hu, Juwei Guan, Yujie Wu, Yusong Wang, Yao Mu, Zuozhu Liu, and Mingkun Xu. 2026. Biomedvr: Confusion-aware mixture-of-prompt experts for biomedical visual reprogramming. *arXiv preprint arXiv:2606.24740*.

Jiaxiang Liu, Tianxiang Hu, Yan Zhang, Yang Feng, Jin Hao, Junhui Lv, and Zuozhu Liu. 2023. Parameter-efficient transfer learning for medical visual question answering. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(4):2816–2826.

Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, and Zuozhu Liu. 2024. MedCoT: Medical chain of thought via hierarchical expert. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17371–17389.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. 2014. The multimodal brain

tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334.

Mackenzie Price, Christine Ballard, Julia Benedetti, Corey Neff, Gino Cioffi, Kristin A Waite, Carol Kruchko, Jill S Barnholtz-Sloan, and Quinn T Ostrom. 2024. Cbtrus statistical report: primary brain and other central nervous system tumors diagnosed in the united states in 2017–2021. *Neuro-oncology*, 26(Suppl 6):vi1.

Nurhuda Hendra Setyawan, Lina Choridah, Hanung Adi Nugroho, Rusdy Ghazali Malueka, and Ery Kus Dwianingsih. 2024. Beyond invasive biopsies: using vasari mri features to predict grade and molecular parameters in gliomas. *Cancer Imaging*, 24(1):3.

Arvind Murari Vepa, Yannan Yu, Jingru Gan, Anthony Cuturrufo, Weikai Li, Wei Wang, Fabien Scalzo, and Yizhou Sun. 2025. A multimodal LLM approach for visual question answering on multiparametric 3D brain MRI. *arXiv e-prints*.

Javier E Villanueva-Meyer, Marc C Mabray, and Soonmee Cha. 2017. Current clinical brain tumor imaging. *Neurosurgery*, 81(3):397–415.

Yuan Wang, Jiaxiang Liu, Shujian Gao, Bin Feng, Zhihang Tang, Xiaotang Gai, Jian Wu, and Zuozhu Liu. 2025. V2t-cot: From vision to text chain-of-thought for medical reasoning and diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 658–668. Springer.

Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. 2024. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *Advances in Neural Information Processing Systems*, 37:140334–140365.

Zonghai Yao, Benlu Wang, Yifan Zhang, Junda Wang, Iris Xia, Zhipeng Tang, Shuo Han, Zhichao Yang, Arman Cohan, and Hong Yu. 2026. Medical thinking with multiple images. In *International Conference on Learning Representations*.

Deli Yu, Shengzhi Wang, Xiaozhong Ji, Bo Cui, Jieqiong Cao, Huichao Wang, Boyuan Jiang, Xu Wang, Qian Xu, Yi Zhao, et al. 2026. MedLesion-VQA: A multimodal benchmark emulating clinical visual diagnosis for body surface health. In *The Fourteenth International Conference on Learning Representations*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Mu Zhou, Jacob Scott, Baishali Chaudhury, Lawrence Hall, Dmitry Goldgof, Kristen W Yeom, Michael Iv, Yangming Ou, Jayashree Kalpathy-Cramer, Sandy Napel, et al. 2018. Radiomics in brain tumor: image assessment, quantitative feature descriptors, and machine-learning approaches. *American Journal of Neuroradiology*, 39(2):208–216.

MM-NeuroOnco: A Multimodal Benchmark and Instruction Dataset for MRI-Based Brain Tumor Diagnosis

Supplementary Material

Contents

A Details on Training Data Construction **13**

A.1 Multi-Source Data Standardization 13

A.2 Medical Semantic Mapping 13

A.3 Coarse Description Generation 14

A.4 Dual-Model Sign Extraction 14

A.5 Train–Benchmark Split 16

B Process Monitoring and Hallucination Mitigation **17**

B.1 Average Information Rate (AIR) 17

B.2 Multi-Model Cross-Validation 17

B.3 Silver-Standard Label Quality Audit 18

C Task and Evaluation Design **19**

C.1 Automated QA Generation 19

C.2 Feature Selection and Clinical Relevance 19

C.3 Open-Ended Evaluation Protocol 19

C.4 Rejection Behavior Analysis 20

D Supplementary Analysis and Qualitative Examples **20**

D.1 Benchmark Sampling Details 20

D.2 Common–Rare Category Analysis 20

D.3 Failure Mode Analysis 21

D.4 Clarifications on Evaluation Scope and Reporting 21

D.5 Supplementary Figures 21

A Details on Training Data Construction

A.1 Multi-Source Data Aggregation and Standardization.

To build a more comprehensive brain tumor MRI data resource, we aggregate 20 distinct datasets, including multi-center challenge datasets such as BraTS and multiple public repositories. For source datasets containing multiple imaging modalities, only MRI images were retained for inclusion in MM-NeuroOnco; CT images were excluded. The specific sources and distributions of these datasets are provided in Table 5 and Figure 7.

These data sources exhibit substantial heterogeneity in storage formats and annotation styles, including 3D NIfTI volumes, 2D DICOM sequences, and non-standard annotation formats such as YOLO/XML. To process these heterogeneous data, we implement seven specialized data processing scripts for directory parsing, modality alignment, and annotation conversion. In particular, for YOLO annotations represented by normalized vector coordinates, we rasterize them into unified binary masks to support subsequent semantic attribute extraction and task construction.

After format standardization, we further conduct large-scale duplicate removal. Specifically, after unified preprocessing, we perform pixel-level comparison and regard MRI slices with identical pixel matrices as duplicates. This process removes approximately 29,000 redundant images, reducing the data size from over 100,000 raw slices to more than 70,000 unique MRI slices. All deduplication is performed globally on the aggregated data pool before dataset partitioning, in order to reduce the impact of cross-source duplicates on subsequent training and evaluation.

Through the above standardization and integrity filtering workflow, we convert the original multi-source, heterogeneous, and structurally inconsistent brain tumor MRI data into a unified structured index, providing the basis for downstream multimodal instruction construction and benchmark design.

A.2 Medical Semantic Mapping.

Since the aggregated datasets contain MRI scans from different and unregistered planes, such as axial, coronal, and sagittal views, conventional anatomical localization may be ambiguous. To reduce this uncertainty, we design a purely visual morphological pipeline that maps binary segmentation masks into structured semantic labels along four dimensions:

1. **Relative Tumor Size.** To quantify tumor burden, we calculate the relative area ratio of the tumor region within the entire image:

$$R_{\text{area}} = \frac{N_{\text{tumor}}}{H \times W}, \tag{4}$$

where N_{tumor} denotes the number of tumor pixels, and H and W denote the image height and width, respectively. Based on empirical thresholds, lesions are categorized as:

- **Small/Focal:** $R_{\text{area}} < 1\%$
- **Medium:** $1\% \leq R_{\text{area}} < 5\%$
- **Large/Extensive:** $R_{\text{area}} \geq 5\%$

2. **Morphological Characterization.** Following the IBSI guidelines, we use Circularity (C) and Elongation (E) to describe lesion morphology.

Circularity C measures the deviation of a lesion from a perfect circle:

$$C = \frac{4\pi A}{P^2}, \tag{5}$$

where A and P denote the lesion area and perimeter, respectively. Elongation E is defined as the ratio between the major and minor axes derived by PCA:

$$E = \frac{\lambda_{\text{major}}}{\lambda_{\text{minor}}}. \tag{6}$$

Classification rules:

- **Irregular/Spiculated:** $C < 0.5$
- **Round/Oval:** $C \geq 0.8$ and $E \leq 1.5$
- **Lobulated:** intermediate cases

3. **Spread and Multiplicity.** We use Connected Component Analysis (CCA) to identify the number of independent lesion components N_c and compute the dominant component ratio f_{core}:

$$f_{\text{core}} = \frac{A_{\max}}{\sum_i A_i}, \tag{7}$$

where A_i denotes the area of the i-th connected component, and $A_{\max}$ denotes the largest connected component. Classifications include:

- **Solitary:** $N_c = 1$
- **Dominant with Satellites:** $N_c > 1$ and $f_{\text{core}} \geq 0.7$

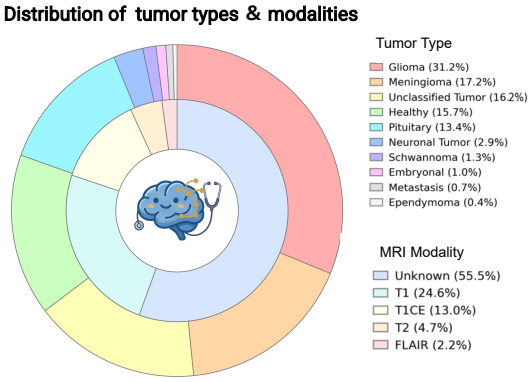


Figure 7: Distribution of tumor types and modalities. The outer ring shows the distribution of brain tumor types, while the inner ring displays the distribution of MRI modalities.

- **Scattered/Multifocal:** $N_c > 1$ and $f_{\text{core}} < 0.7$

4. **Grid-Based Visual Localization.** We compute the lesion centroid based on spatial image moments. Given a binary lesion mask $I(x, y)$, the (p, q)-th order spatial moment is defined as:

$$M_{pq} = \sum_x \sum_y x^p y^q I(x, y). \tag{8}$$

The lesion centroid (C_x, C_y) is then computed as:

$$C_x = \frac{M_{1,0}}{M_{0,0}}, \quad C_y = \frac{M_{0,1}}{M_{0,0}}. \tag{9}$$

We further discretize the centroid into a 3×3 spatial grid, generating descriptors such as "Upper-Left" or "Center." This visual coordinate system does not rely on anatomical priors, provides relatively stable location descriptions across different scan planes and unregistered images, and guides the VLM to attend to the pixel distribution of the lesion.

A.3 Automated Coarse Description Generation.

To convert the extracted structured metadata into textual formats that can be processed by Large Language Models (LLMs), we design a rule-based Natural Language Generation (NLG) engine. This engine serializes image metadata and semantic attributes into coherent natural language descriptions following a three-part hierarchical template:

$$D = [T_{\text{base}}, T_{\text{pathology}}, T_{\text{morphology}}]. \tag{10}$$

The overall generation workflow and template instantiation are illustrated in Figure 9.

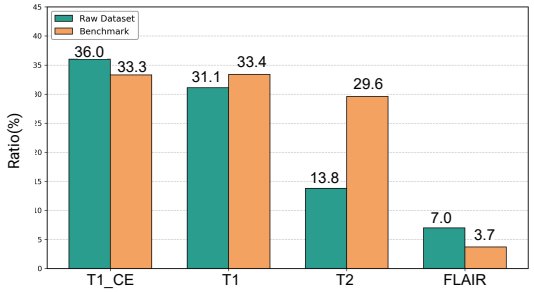


Figure 8: Modality distribution of the MM-NeuroOnco dataset and benchmark.The figure compares the proportional distributions of different MRI modalities in the dataset and the benchmark.

First, the generator constructs the imaging context T_{base} based on modality fields. When the modality is known, the system specifies the exact MRI sequence, such as *"a T1-weighted brain MRI scan"*. When modality information is missing, it uses a generic description and explicitly marks the modality as *"unknown"*, avoiding the generation of uncertain modality information.

The system then appends a pathology statement $T_{\text{pathology}}$. For healthy samples, it outputs *"no visible pathological findings"* and terminates generation. For tumor samples, the system generates diagnostic descriptions according to the granularity of available labels. When the tumor category is explicit, it produces a specific expression such as *"signs of Glioma"*; when the label is coarse or incomplete, it falls back to a more general description such as *"an abnormal mass"*.

After the baseline diagnostic description, the system checks whether pixel-level annotations are available to construct morphological details $T_{\text{morphology}}$. For samples with segmentation masks, the system fills syntactic slots with the size, location, shape, and spread attributes extracted in Appendix A.2, generating fine-grained descriptions such as *"The mass is large and located in the upper-left region."* For samples without mask annotations, the system explicitly states that morphological details are unavailable. This knowledge-boundary definition avoids generating unsupported medical descriptions for unlabeled regions and improves the consistency and credibility of the generated text under varying data quality conditions.

A.4 Dual-Model Visual Sign Extraction Framework.

To extract fine-grained radiological signs from 2D MRI slices, we construct an automated mul-

Table 5: Original data sources incorporated in MM-NeuroOnco.

ID	Dataset	Samples	Key Characteristics	Link
01	Brain Tumor Classification (MRI)	3,264	4-class MRI tumor classification: glioma, meningioma, pituitary, and no tumor.	URL
02	Br35H: Brain Tumor Detection 2020	1,500	Binary tumor classification; balanced dataset.	URL
03	Brain Tumor MRI Dataset	7,023	4-class tumor classification; MRI modality unspecified.	URL
04	Brain Cancer MRI Dataset	6,056	Glioma, meningioma, and pituitary tumor cases.	URL
05	BRISC Brain Tumor Segmentation	6,000	T1-weighted MRI; 4-class labels with segmentation masks.	URL
06	Jun Cheng Brain Tumor Dataset	3,064	T1-weighted contrast-enhanced MRI; three tumor subtypes.	URL
07	Figshare Brain Tumor Dataset	3,000	T1-weighted contrast-enhanced MRI; benchmark subset.	URL
08	Brain Tumor MRI Images (44 Classes)	4,478	Multimodal MRI; 44 fine-grained tumor classes.	URL
09	Brain Tumor MRI Images (17 Classes)	4,449	Multimodal MRI; 17 fine-grained tumor classes.	URL
10	Brain Tumor Classification 2D	6,813	Multimodal MRI with pixel-level segmentation masks.	URL
11	Brain Tumor Classification (14 Classes)	4,455	14-class brain tumor subtype classification.	URL
12	Brain Tumor Classification (15 Classes)	4,288	15-class brain tumor subtype classification.	URL
13	Brain Tumor Multimodal Image Dataset	9,618	MRI subset covering common brain tumor types.	URL
14	Brain Tumor MRI Dataset for DL	9,257	4-class MRI tumor classification for deep learning.	URL
15	Brain Tumor Segmentation Dataset	3,064	T1-weighted contrast-enhanced MRI with binary masks.	URL
16	RSNA–MICCAI BraTS 2021	8,330	Multimodal MRI with expert segmentation labels.	URL
17	Brain Tumors 256×256 Dataset	3,096	4-class tumor classification; standardized resolution.	URL
18	Brain Tumor Dataset: Images and Masks	6,129	Brain tumor images with segmentation masks.	URL
19	Medical Image Dataset: Detection	3,903	4-class medical image detection with YOLO boxes.	URL
20	Brain Tumor MRI Classification 2025	5,000	Binary MRI tumor classification; recent dataset.	URL

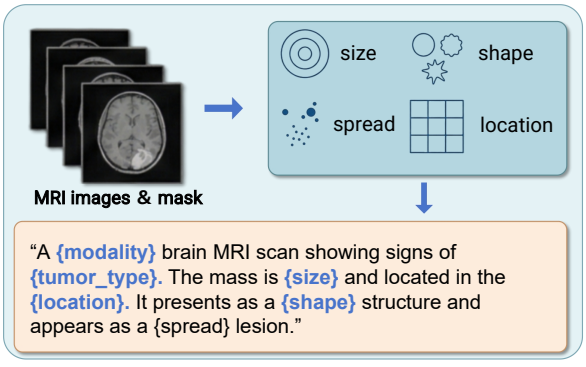


Figure 9: Coarse-grained text description template converted from structured JSON fields into natural language.

timodal analysis pipeline that integrates two heterogeneous multimodal large models via API services: GPT-5.1 and Claude-Sonnet-4.0. Considering their differences in model architecture and instruction-following behavior, we apply targeted System Prompt adaptations to improve reasoning consistency and output stability.

Despite these model-specific prompt adaptations, the overall framework follows a unified radiological sign extraction and evidence verification protocol. Specifically, both models are instantiated as conservative neuroradiologists following evidence-based principles, with *"Omission over Fabrication"* as the core principle to reduce hallucinations and ensure that outputs are grounded as much as possible in verifiable visual evidence.

To maintain reasoning controllability, we adopt a Default Null Policy: all diagnostic fields are initialized as `null` and can only be updated after passing

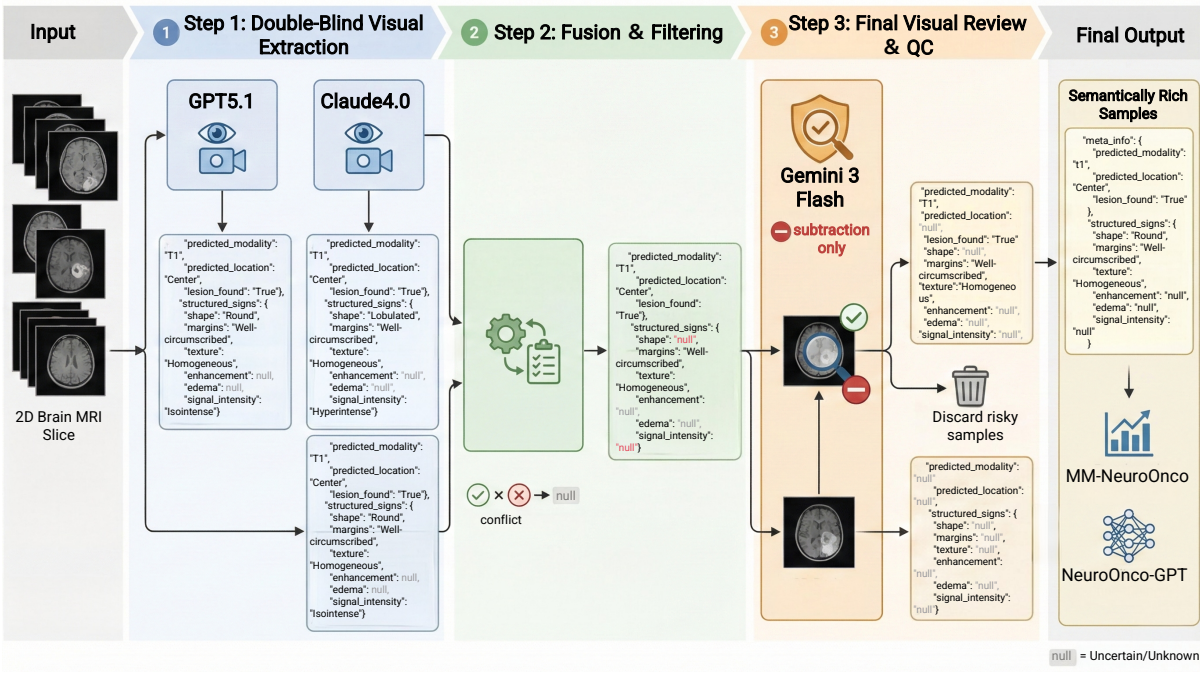


Figure 10: Multi-model Assisted Automated Medical Information Extraction and Quality Control Process.

a structured **"Location–Appearance–Certainty"** validation. That is, the model must identify where the lesion is located, what visual characteristics it presents, and why these observations support the corresponding conclusion.

In addition, we introduce a Pixel Authority Principle as an implementation constraint, which requires visual signals in the image to take precedence over textual priors during attribute extraction. To reduce grayscale ambiguity and modality-specific uncertainty, we further incorporate domain constraints, including a normal-appearing white matter reference, where contralateral normal-appearing white matter is used as the baseline for signal comparison, and MRI physics consistency rules, such as cerebrospinal fluid appearing hypointense on T1-weighted images. These constraints help reduce errors caused by modality confusion, unsupported textual priors, and inconsistent signal interpretation.

API Reproducibility. Table 6 lists the API model identifiers and first invocation timestamps used in our pipeline. Retaining the generated silver-standard labels allows downstream data construction and evaluation to proceed without re-calling these APIs.

A.5 Train–Benchmark Split Protocol.

To improve evaluation rigor and reduce potential data leakage, we establish structured isolation between the instruction dataset and MM-NeuroOnco-Bench. First, before dataset partitioning, we perform global pixel-level deduplication on the aggregated data pool. Specifically, after unified preprocessing, MRI slices with identical pixel matrices are treated as duplicates and removed, avoiding exact duplicate images appearing across different splits.

For volumetric datasets, such as brain tumor segmentation benchmarks, we select only one representative 2D slice from each 3D volume, so that each volumetric case contributes at most one sample. This design provides a degree of volume-level isolation and prevents different slices from the same 3D case from appearing simultaneously in the training and evaluation sets, thereby reducing the impact of cross-slice correlations.

After deduplication and single-slice sampling, we perform independent partitioning on the deduplicated data pool. We then apply stratified sampling based on tumor categories and imaging modalities to maintain a relatively balanced distribution across splits. For datasets originally provided as independent 2D slices without reliable patient- or case-level identifiers, strict patient-level grouping cannot be enforced. In such cases, we combine pixel-level deduplication, similarity-based hash checking, and stratified partitioning to ensure slice-level disjointness and reduce the risk of leakage caused by highly similar images.

Model	API Model ID	First Invocation (UTC)
GPT-5.1	`gpt-5.1`	2026-01-09 03:26:15
Claude Sonnet 4.0	`claude-sonnet-4-0`	2026-01-08 13:31:14
Gemini 3 Flash	`gemini-3-flash-preview`	2026-01-09 03:28:51

Table 6: API model identifiers and first invocation timestamps for the semantic completion pipeline.

Overall, this protocol reduces the risk of data leakage caused by exact duplicate images, highly similar slices, and cross-slice correlations from the same volumetric case, while providing more stable benchmarking conditions.

B Process Monitoring and Hallucination Mitigation

B.1 Process Monitoring via Average Information Rate (AIR).

To dynamically monitor information coverage in the semantic extraction pipeline and quantify the degree of conservatism under the "Default Null Policy," we introduce the **Average Information Rate (AIR)** as a statistical control metric.

AIR measures the proportion of non-null core radiological attributes extracted for each valid tumor sample. For a sample x_i, its information extraction rate $R(x_i)$ is defined as:

$$R(x_i) = \frac{1}{|S|} \sum_{s \in S} \mathbb{I}(s \neq \text{null}), \tag{11}$$

where S denotes the set of six predefined core radiological signs, including Shape, Margin, Texture, Enhancement, Edema, and Signal Intensity, with $|S| = 6$. $\mathbb{I}(\cdot)$ is the indicator function. The dataset-level AIR is computed as the mean of $R(x_i)$ over all positive samples, i.e., samples with `Lesion Found = True`.

Importantly, AIR is not intended to maximize information density. Instead, it serves as a calibration signal to balance attribute coverage and factual reliability. An overly low AIR may indicate excessive conservatism and the loss of valid diagnostic information, whereas an unusually high AIR may indicate over-generation or violations of MRI modality constraints. For example, reporting "enhancement" in a non-contrast T1 scan may introduce unreliable semantics.

To ensure that increases in AIR reflect valid semantic extraction rather than hallucination amplification, each prompt or policy adjustment is followed by manual audits on randomly sampled subsets. Only configurations that improve information coverage without degrading factual correctness are retained. Through this iterative calibration process, AIR helps maintain a balance between information richness and authenticity, while supporting the design principle of *"Omission over Fabrication."*

B.2 Hallucination Mitigation via Multi-Model Cross-Validation.

To reduce hallucination risks during medical semantic extraction, we incorporate multiple structural and procedural safeguards into the pipeline.

First, we use two heterogeneous multimodal large models to independently extract medical semantic attributes from the same visual input. Instead of relying on a single model output, we retain only attributes that are supported across models. This redundancy-based design reduces the risk of over-generation or idiosyncratic errors from a single model.

Second, we introduce a third heterogeneous model as an external verifier to perform a final consistency check over the extracted semantic fields. The verifier follows a strictly constrained policy: it is allowed only to remove or nullify uncertain or visually unsupported attributes, and is explicitly prohibited from introducing new information. This asymmetric verification mechanism ensures that the final stage can reduce suspicious attributes without amplifying or adding potential hallucinated content.

Third, before semantic extraction, all medical attribute fields are initialized as `null`. Fields are populated only when relatively explicit visual evidence is identified. This default-empty strategy encourages conservative behavior and aligns with the clinical reasoning principle of avoiding speculation when evidence is insufficient.

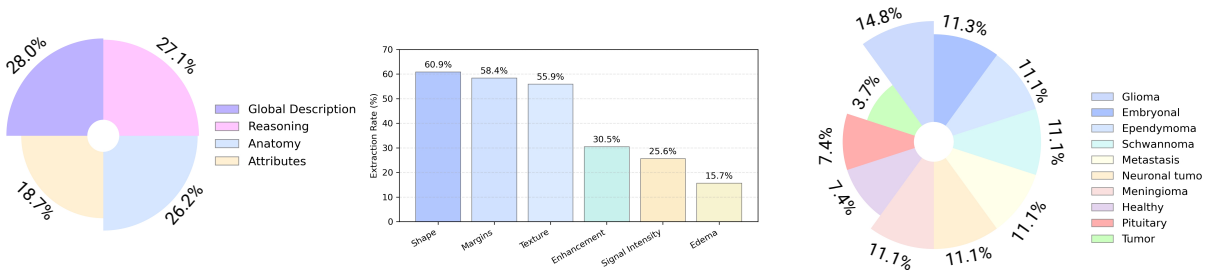
Finally, to evaluate the reliability of the overall pipeline, we conduct random manual audits at two key stages: (1) after the initial dual-model extraction and (2) after the final verification stage. The audits focus on whether the extracted attributes are visually supported and clinically plausible. The audit results suggest that the multi-stage design helps reduce unsupported attribute generation while pre-

Table 7: Overview of the MM-NeuroOnco data components and their sizes.

Component	Description	Size
Image Pool	Curated 2D MRI slices from 20 public datasets, processed with unified indexing, deduplication, and normalization.	73,226
Attributes	Gold labels such as tumor type, together with silver-standard radiological attributes such as edema and enhancement.	24,726
VQA Pairs	130k+ closed-ended QA pairs with challenging distractors and CoT rationales, plus 70k+ open-ended QA pairs.	200k+
Benchmark	Evaluation subset with 1,000 images, including 2k+ closed-ended and 1k+ open-ended questions.	3k+

Table 8: Overall performance comparison on the open-ended benchmark. Best results are highlighted in **bold**, and second-best results are underlined.

Model	Detail	Location	Reasoning	Overall
General-Purpose LVLMs				
GPT-5.1	**71.15**	63.99	**86.17**	**72.72**
Gemini 3 Flash	65.71	49.22	82.02	65.67
Qwen3-VL-8B	58.94	59.59	44.28	56.14
Medical-Specialized LVLMs				
HuLuMed-32B	60.80	46.58	78.23	61.44
Lingshu-7B	57.42	55.10	80.29	61.53
Lingshu-32B	54.06	47.86	81.15	58.24
HuLuMed-7B	48.27	28.97	72.14	49.14
MedGemma-27B	55.79	17.08	62.76	49.44
MedGemma-1.5-4B	50.89	20.78	71.15	48.92
Ours (Domain-Adapted LVLMs)				
NeuroOnco-GPT	52.51	**84.77**	68.40	62.14



(a) Distribution of Open-Ended Question Types (b) Extraction Rate of Medical Attribute Information (c) Distribution of Tumor Categories in the Benchmark

Figure 11: Data Distribution and Information Extraction Rate.

serving diagnostically meaningful medical information.

B.3 Supplementary Notes on Silver-Standard Label Quality Audit.

To assess the quality of silver-standard attributes generated by the automated pipeline, we manually audit structured medical attributes in the training set. Consistent with the main-text table, two human verifiers examine 1,125 structured attribute fields, covering diagnosis-related semantics such as enhancement, edema, texture, and signal intensity. The two verifiers were a medical student and a researcher experienced in medical imaging and brain tumor MRI, rather than licensed neuroradiologists. Human auditing complements automated quality control by identifying errors and omissions that model agreement alone may overlook. The audit results show an overall error rate of 6.76%, indicating that the multi-model collaborative pipeline can control incorrect attribute introduction to some extent.

In addition to the error rate, we also record information omissions during the audit. The overall omission rate is 16.98%, which mainly reflects the conservative semantic completion strategy adopted

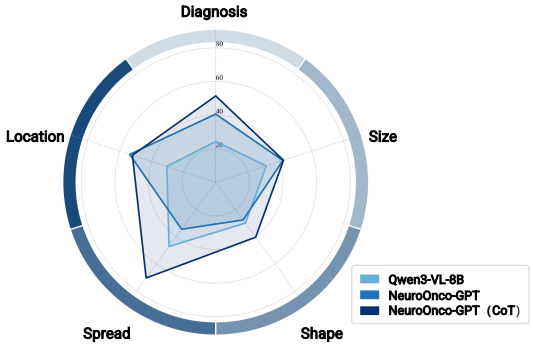


Figure 12: Radar chart comparing the performance of the base model and different fine-tuning strategies on MM-NeuroOnco-Bench.

in this work: when visual evidence is insufficient or models exhibit clear conflicts, the system tends to keep the corresponding field as `null` rather than forcing the generation of uncertain attributes. Therefore, the omission rate is not equivalent to factual error, but instead reflects the design principle of *"Omission over Fabrication."* Overall, the audit results suggest that the silver-standard attributes maintain a relatively low error rate while following a conservative medical semantic extraction strategy.

C Task and Evaluation Design

C.1 Automated QA Generation.

We adopt distinct generation paradigms for the instruction dataset and the evaluation benchmark to ensure functional separation between data synthesis and model assessment.

For the instruction dataset, we use Qwen3-80B-Instruct to synthesize linguistically diverse instruction-style QA data. To reduce template overfitting and improve expressive variability, the model is required to introduce natural phrasing variations during question construction. The QA format is adapted according to diagnostic category: binary classification questions are generated for healthy samples, while multiple-choice questions with distractors are constructed for confirmed tumor cases. This design exposes the model to a broader range of clinically relevant reasoning patterns during training.

In contrast, the evaluation benchmark is constructed with GPT-5.1 under an independent control protocol, ensuring separation from the instruction data generation process. To reduce textual information leakage, we exclude descriptive cues related to lesion size, morphology, or other diagnostic hints from question stems, encouraging models to rely more on visual evidence. In addition, distractors are not randomly sampled; instead, radiologically similar pathologies are selected as adversarial options, such as using Lymphoma as a distractor for Glioma. This adversarial distractor design enables a more fine-grained assessment of visual discrimination ability while reducing shortcut reasoning. No benchmark question is directly reused or paraphrased from the instruction dataset.

C.2 Rationale for Feature Selection and Clinical Relevance.

Brain tumor imaging diagnosis usually relies on the integration of multiple complementary imaging attributes rather than isolated cues. Prior studies have shown that morphology, margins, texture, enhancement patterns, signal intensity, and edema are closely related to tumor classification and grading (Zhou et al., 2018). Together, these attributes reflect tumor geometry, tissue interaction, biological activity, and the surrounding microenvironmental response.

Specifically, shape and margins describe the geometric configuration and boundary definition of lesions. Irregular contours or indistinct margins may indicate infiltrative growth or higher malignancy. Texture reflects intratumoral heterogeneity, while signal intensity encodes modality-specific contrast patterns in T1- and T2-weighted MRI sequences and provides diagnostic reference. Enhancement is usually associated with blood–brain barrier disruption and tumor vascular activity, while edema may reflect peritumoral reaction, infiltration, or mass effect, and is often related to tumor aggressiveness (Villanueva-Meyer et al., 2017).

Importantly, our feature selection is aligned with the evaluation protocol. Since the benchmark is mainly based on single-slice 2D MRI inputs, we intentionally exclude attributes that require three-dimensional or temporal context for reliable assessment, such as volumetric measurements, cross-slice growth patterns, or longitudinal progression (Zhou et al., 2018). The selected features are relatively observable within a single slice and diagnostically meaningful, ensuring consistency between the input format and evaluation criteria while preserving clinical relevance.

C.3 Open-Ended Evaluation Protocol.

Given the limitations of traditional n-gram metrics, such as BLEU and ROUGE, in capturing clin-

Model	Regular Acc.	False Reject.	Correct Reject.	Failed Reject.
Lingshu-32B	42.0	4.1	8.5	91.5
GPT-5.1	41.9	10.3	17.4	82.6
Gemini-3-Flash	48.8	1.0	7.3	92.7
HuLuMed-32B	45.1	2.7	4.0	96.0
Claude-Sonnet-4.0	42.4	3.1	8.7	91.3
NeuroOnco-GPT (CoT)	**56.1**	9.2	**31.3**	**68.7**

Table 9: Rejection behavior across five benchmark tasks. All values are percentages.

ically grounded reasoning quality, we follow the LLM-as-a-Judge evaluation paradigm adopted in recent medical multimodal benchmarks (Hao et al., 2026a) to assess open-ended responses.

Specifically, we use Qwen3-80B-Instruct as the automated evaluation model. The evaluation prompt, shown in Figure 18, follows the structured scoring scheme described in prior work and assesses responses along clinically relevant dimensions, including factual consistency, spatial correctness, safety compliance, and reasoning coherence. Critical errors, such as laterality conflicts or unsupported pathological claims, are penalized according to predefined scoring rules. Meanwhile, reasonable semantic variations in radiological descriptions are tolerated to avoid over-penalization caused by minor lexical differences.

This setup provides a structured and clinically aligned evaluation of open-ended diagnostic reasoning and is consistent with existing LLM-based assessment practices.

C.4 Rejection Behavior Analysis.

We analyze rejection behavior across five benchmark tasks. Table 9 reports regular-answer accuracy, false rejection, correct rejection, and failed rejection (false acceptance), where E denotes "None of the above." False rejection corresponds to GT$\neq$E and Pred=E; correct and failed rejection correspond to GT=E with Pred=E and Pred$\neq$E, respectively.

The baseline models exhibit low correct-rejection rates (below 18%). NeuroOnco-GPT (CoT) achieves the highest regular-answer accuracy and correct-rejection rate among these models, but still incorrectly accepts a non-E answer in 68.7% of cases where E is the ground truth. Thus, rejection remains challenging despite improved task performance.

D Supplementary Analysis and Qualitative Examples

D.1 Benchmark Sampling Details.

MM-NeuroOnco-Bench adopts a coverage-oriented stratified sampling strategy rather than simple random sampling according to the original data proportions. Specifically, we independently sample from the deduplicated data pool and jointly consider tumor subtype and MRI modality, so that the benchmark covers different diagnostic categories and imaging manifestations.

Public brain tumor MRI datasets usually exhibit a clear long-tailed distribution. If the benchmark were sampled strictly according to the original proportions, it would be easily dominated by head classes such as glioma and meningioma, thereby weakening the evaluation of rare tumor categories. Therefore, we moderately supplement rare tumor categories so that MM-NeuroOnco-Bench can more sufficiently evaluate models' diagnostic understanding across different tumor subtypes.

It should be noted that this sampling strategy mainly serves coverage-oriented capability analysis. Its goal is to construct a broader diagnostic understanding benchmark, rather than to fully reproduce real clinical distributions. Therefore, the training set focuses more on preserving the natural long-tailed distribution of public data, while the benchmark emphasizes coverage across different categories and modalities.

D.2 Common–Rare Category Analysis.

To further analyze model performance under long-tailed categories, we group tumor subtypes according to their sample proportions in the training set. Glioma, meningioma, and pituitary, which have relatively more samples, are treated as common categories, while the remaining tumor subtypes are treated as rare categories. This grouping is used to evaluate whether models exhibit substantial performance bias on low-frequency tumor types.

Experimental results show that most models achieve clearly lower accuracy on rare categories than on common categories, indicating that current LVLMs remain unstable in diagnosing low-

frequency brain tumor subtypes. In contrast, NeuroOnco-GPT, fine-tuned on MM-NeuroOnco, reduces the performance gap between common and rare categories to 4.6%, showing more balanced category generalization. This suggests that semantically enriched instruction data can help mitigate long-tail performance bias to some extent.

D.3 Additional Case and Failure Mode Analysis.

To better understand model behavior on MM-NeuroOnco-Bench, we summarize representative prediction errors and identify several major failure modes, including rejection failure, rare-category confusion, insufficient evidence grounding, and over-interpretation.

Rejection failure refers to cases where the model still tends to select a specific regular category when the correct answer is not among the regular candidate options, instead of choosing "None of the above." This suggests that some models are still affected by the forced-choice paradigm and have difficulty judging whether candidate answers are consistent with imaging evidence.

Rare-category confusion refers to cases where models more easily misclassify low-frequency tumor subtypes as common categories or visually similar categories. This phenomenon is related to the long-tailed data distribution and also indicates that existing models remain limited in distinguishing rare brain tumor subtypes.

Insufficient evidence grounding refers to responses that fail to accurately cite or use key imaging evidence, such as lesion location, enhancement pattern, edema extent, or signal characteristics. In some cases, models may give correct or near-correct conclusions, but their reasoning remains generic and lacks explicit correspondence with visual evidence.

Over-interpretation refers to cases where models generate medical details that are not clearly supported by the image, such as speculating about enhancement patterns, tumor invasion extent, or histological features that are not visible. This type of error is more common in open-ended responses and reflects the remaining hallucination risk in medical generation tasks.

These failure modes indicate that existing LVLMs still face challenges in evidence utilization, long-tail generalization, and uncertainty handling in fine-grained brain tumor MRI diagnostic reasoning.

D.4 Clarifications on Evaluation Scope and Reporting.

Open-ended evaluation scores are normalized to a 0–100 scale for reporting in Table 8. The closed-ended evaluation in Table 2 excludes modality-identification QA pairs, as we do not consider MRI modality recognition to be a task of understanding tumor attributes. In contrast, the rejection-mechanism ablation in Table 3 includes modality-identification QA pairs. Our dataset also includes samples labeled simply as "Tumor", without further tumor-subtype annotations. These samples are used for binary tumor-versus-normal classification. Throughout the main text, NeuroOnco-GPT refers to the CoT-trained model by default, including the results reported in Table 4. When comparing training with and without CoT supervision in the ablation study, we explicitly distinguish the two variants as NeuroOnco-GPT (CoT) and NeuroOnco-GPT, respectively.

D.5 Supplementary Figures.

This section provides additional implementation details and qualitative examples, including the attribute extraction algorithm, prompt templates, and representative samples from the instruction dataset and benchmark. These materials complement the main results by illustrating the data construction workflow, prompt control strategy, and benchmark difficulty.

Algorithm 1: Knowledge-Guided 2D MRI Brain Tumor Attribute Extraction

Input : 2D MRI image I, clinical context text C_{txt}
Output : Meta-info M, extracted signs/attributes S

/ Step 1: Initialization */*
$M \leftarrow \{\text{predicted_modality} : \text{null},\ \text{lesion_found} : \text{null},\ \text{loc} : \text{null}\}$;
$S \leftarrow \{\text{signal} : \text{null},\ \text{enhance} : \text{null},\ \text{edema} : \text{null},\ \text{shape} : \text{null},\ \text{margins} : \text{null}\}$;
/ Step 2: Modality Determination & Validation */*
M.predicted_modality $\leftarrow$ `InferModality`(I, C_{txt});
; *// Uses text first, falls back to intensity heuristics; vetoes invalid T1CE*
/ Step 3: Lesion Detection ("Pixel Authority") */*
if Healthy $\in C_{txt}$ **then**
    M.lesion_found $\leftarrow$ False;
    return $\{M, S\}$;
$R \leftarrow$ `DetectLesion`(I);
; *// Scan + spatial coherence filtering*
if $R = \emptyset$ **then**
    M.lesion_found $\leftarrow$ False;
    return $\{M, S\}$;
M.lesion_found $\leftarrow$ True;
M.loc $\leftarrow$ `Localize`(R);
/ Step 4: Establish Reference (NAWM) */*
$P_{\text{ref}} \leftarrow$ `SelectRef`(I);
; *// Contralateral normal-appearing white matter*
/ Step 5: Attribute Extraction */*
$S \leftarrow$ `ExtractAttrs`$(R, P_{\text{ref}}, M.\text{predicted_modality})$;
; *// Signal/shape/margins + modality-specific logic*
/ Step 6: Consistency Check */*
$S \leftarrow$ `ResolveContr`$(S, M.\text{predicted_modality})$;
; *// Revert contradictory fields to* null
return $\{M, S\}$;

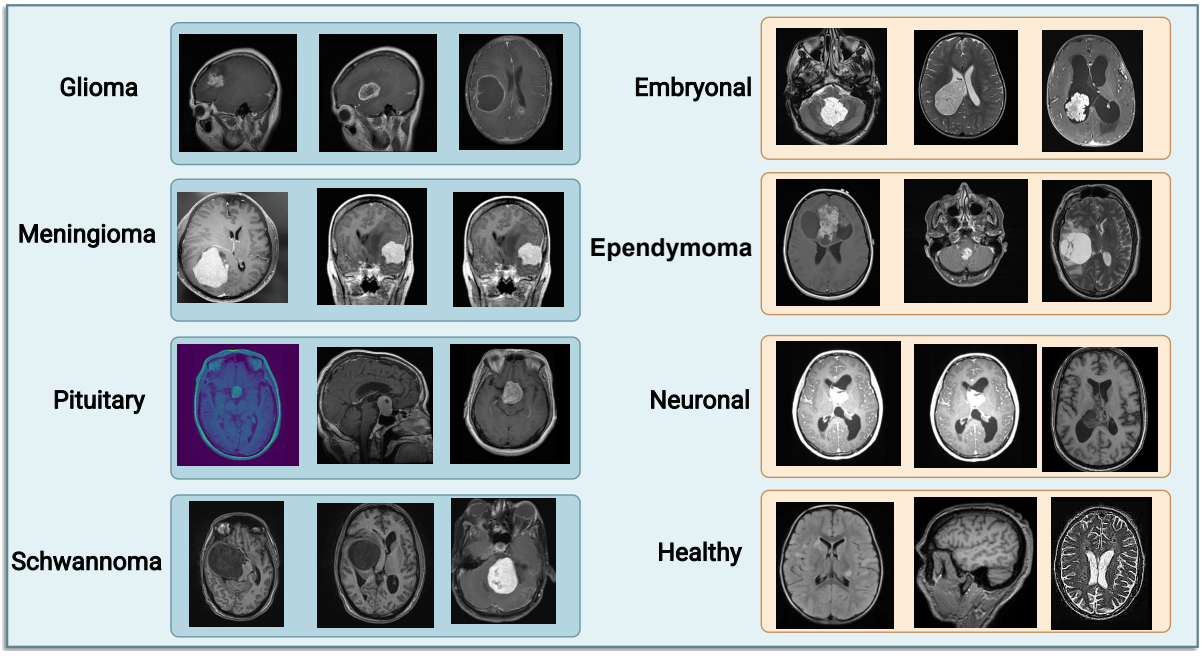


Figure 13: Sample Images of Meningioma, Glioma, Pituitary Tumors, and Others.

```
SYSTEM_PROMPT = r"""
### 1. ROLE & MISSION
You are a Neuroradiologist specialized in **2D MRI brain tumor diagnosis**.
**CORE PRINCIPLES:**
* **Initialize to null:** Default all fields to `null`.
* **Evidence-Based:** Only modify `null` if it satisfies the **"3W Rule"**.
* **Omission is professional; fabrication is a medical error.**

### 2. INPUT DATA
- **IMAGE:** A single 2D MRI slice (Primary Evidence).
- **TEXT:** Clinical context containing known facts or missing info (Guidance Hint).

### 3. BEHAVIORAL CONSTRAINTS & VETOS
1. **Pixel Authority:** **Image pixels are the ultimate diagnostic authority**.
2. **Mandatory Detection for Non-Healthy Samples:** If the TEXT indicates a tumor, you MUST set
`lesion_found`="True". Do NOT reject the image or set `lesion_found`="False" due to low contrast,
blurry pixels, or indistinct margins. **If specific features are unreadable, keep those specific fields
as `null`, but do NOT veto the entire sample.**
3. **Anatomy Filter:** Exclude normal structures (Vessels, Sinuses, Choroid Plexus, symmetric
anatomy).
4. **Unknown Lock:** If modality is "Unknown", force `signal_intensity`=null (also keep
`edema`/`enhancement` null as usual).

### 4. MEDICAL PRIOR & MODALITY ANCHORS
- **T1:** CSF is Dark; White Matter (WM) is brighter than Gray Matter. (Lock: Edema/Enhancement
= null).
- **T2:** CSF is Bright/White.
- **FLAIR:** CSF is Dark; Edema/Inflammation is Bright.
- **T1_Contrast (T1CE):** Vessels/Sinuses MUST be Bright/White.
    * **Veto:** If vessels are dark, treat as T1; force `enhancement`=null.
    * **Lock:** On confirmed T1CE, force `signal_intensity`=null (use `enhancement` field instead).

### 5. ENHANCED MODALITY RECOGNITION FOR T1
- **In T1 Modality:** If a lesion is detected, **enhancement** and **edema** fields should only
remain null if no evidence of enhancement or swelling is observed in the image. If there are
suspicious bright regions (possible enhancement), fill `enhancement` as **Present**.
- **Edema in T1:** If there is a visually discernible expansion or blurred boundary (possible edema),
fill `edema` as **Present**.
```

Figure 14: Prompt for the Medical Information Extraction Process

6. UNIFIED LABELING DICTIONARY (Mandatory Anchor: Healthy WM)
All brightness judgments MUST use contralateral normal-appearing deep White Matter (NAWM) in the same slice as the ZERO reference (avoid cortex/GM ribbon and CSF).
- **Signal Intensity:**
  - **Hyperintense:** Visually **brighter/whiter** than healthy WM.
  - **Hypointense:** Visually **darker/grayer** than healthy WM.
  - **Isointense:** Signal is **nearly identical** to healthy WM.
  - **Mixed:** Both Hyper and Hypo signals are present, each occupying >30%.
- **Texture:** Homogeneous | Heterogeneous.
- **Enhancement (T1CE only):** Solid | Ring | Patchy.
- **Edema (T2/FLAIR only):** Mild | Extensive.
- **Margins:** Well-circumscribed | Ill-defined.
- **Shape:** Round | Oval | Irregular | Lobulated.

7. DYNAMIC EXECUTION WORKFLOW
STEP 1: Initialization - Set all fields to `null`.
STEP 2: Healthy Short-Circuit - Set `lesion_found`="False" ONLY if text says "Healthy" OR the image is 100% normal with NO localized signal shift. **If TEXT says there is a tumor, you MUST skip the Veto and proceed to STEP 3.**
STEP 3: Modality Determination - **Adopt text hint if provided.** Otherwise, identify via Section 4.
STEP 4: Localization - **Focus on text-hinted location.** Any localized asymmetry or slight grey-scale shift vs. contralateral NAWM satisfies this.
STEP 5: Fine-Grained Extraction - Extract signs via Section 5. **If a specific sign is truly invisible due to quality, keep it `null` but do NOT revert `lesion_found` to False.**
STEP 6: Self-Reflection Audit - Verify physics. If the logic chain for signs is weak, revert signs to `null`, but **if the clinical text confirmed a tumor, `lesion_found` MUST remain "True".**

8. OUTPUT (STRICT JSON ONLY)
```
{
  "meta_info": {
    "predicted_modality": "T1 | T2 | FLAIR | T1_Contrast | Unknown",
    "predicted_location": "Upper-Left | Upper-Center | Upper-Right | Center-Left | Center | Center-Right | Lower-Left | Lower-Center | Lower-Right | null",
    "lesion_found": "True | False"
  },
  "structured_signs": {
    "shape": "Option | null",
    "margins": "Option | null",
    "texture": "Option | null",
    "enhancement": "Option | null",
    "edema": "Option | null",
    "signal_intensity": "Option | null"
  }
}

"""
```

Figure 15: Continuation of the Prompt for the Medical Information Extraction Process.

```
SYSTEM_PROMPT = r"""
You are an experienced Neuroradiologist Auditor.
**Task:** "Sanity Check" Step 2 Labels against the MRI slice.
**Mindset:** Step 2 labels are your baseline. Be a conservative auditor. Only correct CLEAR and
GROSS errors.

### 1. Audit Logic (Strict Decision Tree)

#### A. Location (Threshold: "Gross Violation Only")
* **Rule:** Location is binary (Right or Wrong). No middle ground.
* **KEEP:** If there is ANY overlap, proximity, or the lesion is small/faint at the site.
* **ERASE:** ONLY for Gross Geographical Errors (e.g., Label says "Left Hemisphere", Image shows
lesion on "Right").
* **Note:** Do NOT downgrade location.

#### B. Modality (Physics Check - CRITICAL)
* **Goal:** Verify if the "Predicted Modality" matches the image physics.
* **T1 vs T1CE (Contrast) Differentiation:**
  * **LOOK FOR:** 1. **Nasal Turbinates/Mucosa** (Best Indicator), 2. Dural Venous Sinuses (e.g.,
Sagittal Sinus), 3. Choroid Plexus.
  * **T1 (Non-Contrast):** These structures appear **DARK** or Isointense.
  * **T1CE (Contrast):** These structures MUST be **BRIGHT WHITE** (Hyperintense).
* **ACTIONS:**
  * **KEEP:** If physics match or image is ambiguous.
  * **DOWNGRADE:** If label is "T1_Contrast" BUT nasal turbinates/sinuses are clearly DARK ->
Change to "T1".
  * **ERASE:** If label is "T2" but CSF is DARK (physically impossible).

#### C. Signs (Visual Cleaning & De-specification)
* **Goal:** Validate specific visual features.
* **KEEP:** If the sign is visible OR suggestive. (Benefit of the doubt).
* **ERASE:** Only if the sign is **100% absent** (Hallucination).
* **DOWNGRADE (De-specification):** * **Definition:** The feature exists (True Positive), but the
Step 2 description is **too specific** or **wrongly detailed**.
  * **Logic:** "I see the abnormality, but I don't see *that specific* pattern."
  * **Examples:**
    * Label="Ring-Enhancing" -> Image shows blurry white blob -> Action: **DOWNGRADE** (to
generic "Enhancing").
    * Label="Spiculated Margins" -> Image shows fuzzy edges -> Action: **DOWNGRADE** (to
generic "Indistinct").
    * Label="Necrotic Center" -> Image is solid gray -> Action: **DOWNGRADE** (to generic
"Heterogeneous").

### 2. Output Format (FULL JSON REQUIRED)
{
  "final_decision": "ACCEPT",
  "field_actions": {
    "predicted_modality": "KEEP|ERASE|DOWNGRADE",  // DOWNGRADE only for T1CE -> T1
correction
    "predicted_location": "KEEP|ERASE",          // NO DOWNGRADE allowed
    "shape": "KEEP|ERASE|DOWNGRADE",
    "margins": "KEEP|ERASE|DOWNGRADE",
    "texture": "KEEP|ERASE|DOWNGRADE",
    "enhancement": "KEEP|ERASE|DOWNGRADE",
    "edema": "KEEP|ERASE|DOWNGRADE",
    "signal_intensity": "KEEP|ERASE"             // NO DOWNGRADE allowed (Signal is objective: Bright
or Dark)
  },
  "additional_findings": "Max 3 words (e.g., 'Artifact present', 'Wrong orientation'), null if none."
}
"""
```

Figure 16: Prompt for Visual Final Review and Quality Control.

```
SYSTEM_PROMPT = """
You are a Board-Certified Neuroradiologist. Your task is to transform MRI metadata into
professional, VARIED, and IMAGE-BASED MCQs.

### 1. Style & Diversity Guidance (CRITICAL):
* **DIVERSITY**: Do NOT use the exact same sentence template for every case. Vary your phrasing
naturally.
  - *Example Diagnosis*: "What is the primary diagnosis?" / "Identify the pathology shown." / "Based
on the findings, the most likely diagnosis is:"
  - *Example Features*: "How are the margins described?" / "Describe the observed internal texture."
/ "The enhancement pattern is best characterized as:"
* **NO CLINICAL HISTORY**: Strictly FORBIDDEN to invent patient age, gender, symptoms, or
medical history.
* **BRIEF & VISUAL**: Keep questions concise (under 20 words). Focus only on imaging features.
* **STRICT SOURCE**: Use ONLY provided metadata. IGNORE 'additional_clues'.

### 2. Case Logic (Benchmark Consistent):
* **CASE A: Healthy / Normal (lesion_found is False)**:
    * Generate a **Binary (2-Option) Question**.
    * Question: "Is there a pathological lesion present in this MRI?" (or similar phrasing).
    * Options: {"A": "Tumor / Abnormal", "B": "Healthy / Normal"}.
    * Answer: "B".
* **CASE B: Generic Tumor** (e.g., "Tumor", "Mass"):
    * Generate a **Binary (2-Option) Question**.
    * Options: {"A": "Tumor / Abnormal", "B": "Healthy / Normal"}.
    * Answer: "A".
* **CASE C: Specific Pathology** (e.g., "Meningioma", "Glioma"):
    * Generate a **4-Option MCQ**.
    * Distractors: Must be specific pathologies (e.g., Glioblastoma, Metastasis).

### 3. Feature Logic (Strict Exclusion Rule):
* **IF Healthy / Normal**:
    - **STOP HERE**. Only generate the Diagnosis and Modality MCQs.
* **IF Pathological (Tumor)**:
    - Generate a standard **4-Option MCQ** for EACH provided field: (modality, location, shape,
margins, texture, signal_intensity, enhancement, edema).

### Output JSON Format (Strict Structure):
{
  "questions": [
    {
      "question_type": "diagnosis",
      "question": "varied_brief_question_text",
      "options": {"A": "...", "B": "...", "C": "...", "D": "..."},
      "answer": "A"
    }
  ]
}
"""
```

Figure 17: Prompt for Automatic Generation of Questions and Options Based on LLM.

Medical Scoring Rubric (Revised for Ambiguity)

Step 1: Check for CRITICAL SAFETY ERRORS (The "Kill" Switch)
If ANY of the following exist, the Maximum Score is **2**.
1. **Direct Laterality Conflict:** Explicitly saying **"Left"** when GT says **"Right"** (or vice versa).
  * ◁ **EXCEPTION:** If GT is "Central/Midline", confusing it with "Left" or "Right" is **NOT** a critical error (treat as Step 3 ambiguity).
  * ◁ **EXCEPTION:** If GT is "Bilateral", saying "Left" or "Right" is incomplete but **NOT** a critical error.
2. **Pathology Hallucination:** Inventing critical features (hemorrhage/necrosis) explicitly denied by GT.
3. **Missed Diagnosis:** Failing to find the tumor or answering "I don't know".

Step 2: Check for MAJOR DIAGNOSTIC ERRORS
If Step 1 is passed, deduct **3-4 points**. (Max Score: 6).
1. **Wrong Anatomical Region (Distant):** "Frontal lobe" vs "Occipital lobe" (Far apart).
2. **Wrong Enhancement:** "Ring-enhancing" vs "Non-enhancing" (Clinical mismatch).
* Note: Overlapping regions (e.g., "Frontal" vs "Fronto-parietal") are **NOT** major errors.

Step 3: Check for MINOR / SPATIAL AMBIGUITY
If Steps 1 & 2 are passed, deduct **0-2 points**. (Score range: 8-10).
1. **Spatial Ambiguity (Acceptable):** - "Central" vs "Upper Left/Right" (Common variance in visual interpretation).
  - "Frontal" vs "Anterior cerebrum".
  - **Action:** If the general location is correct, **DO NOT DEDUCT POINTS** or deduct max 1 point.
2. **Imprecise Margins:** "Irregular" vs "Lobulated".
3. **Missing Secondary Features:** Missed mild edema or slight mass effect.

***Step 4: CLINICAL EQUIVALENCE (Perfect Score)**
Score 10 if:
- The clinical diagnosis is correct.
- Location is roughly consistent (e.g., "Left Frontal" matches "Upper Left").
- No hallucinations.

Evaluation Protocol:
1. **Strictly punish** Left vs Right contradictions (Score = 2).
2. **Be tolerant** of "Central" vs "Paramedian/Lateral" ambiguities (Score 8-10).
3. **Be tolerant** of "Lobe" boundaries (e.g., Frontal vs Fronto-Parietal).

Output strictly in JSON format:
{{
  "score": <int 0-10>,
  "reasoning": "<Explain reasoning. Explicitly state if spatial ambiguity was forgiven.>"
}}
"""

Figure 18: Prompt for LLM-assisted Open Question Scoring.

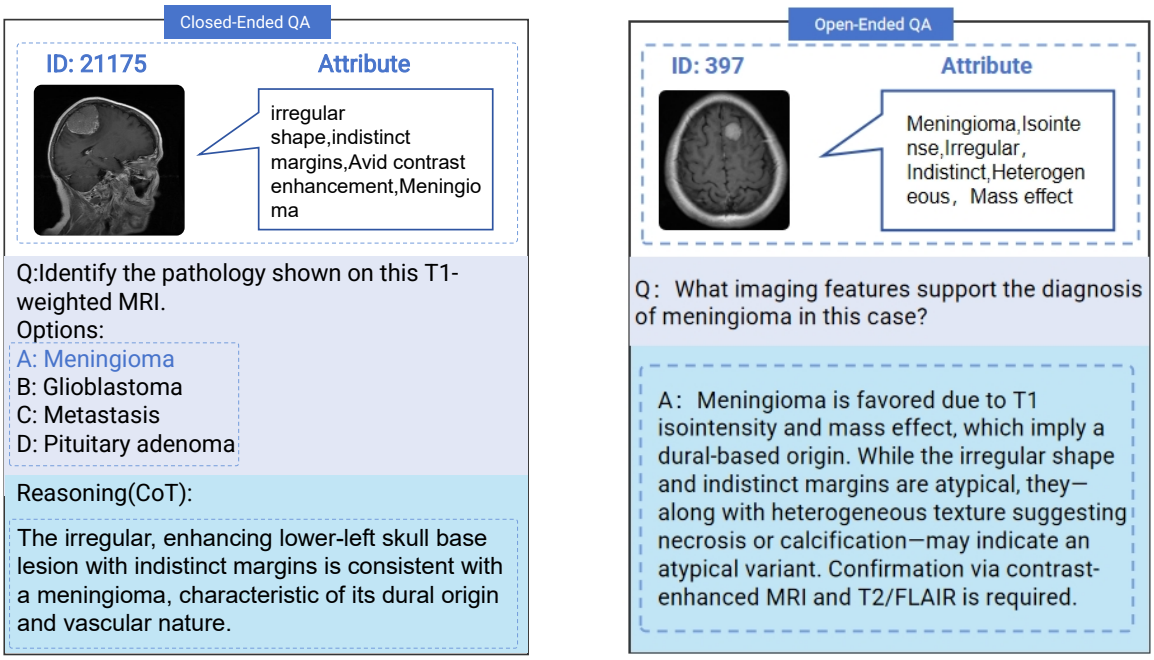


Figure 19: Sample Examples of Closed-ended and Open-ended Question Answering from the Instruction Dataset.

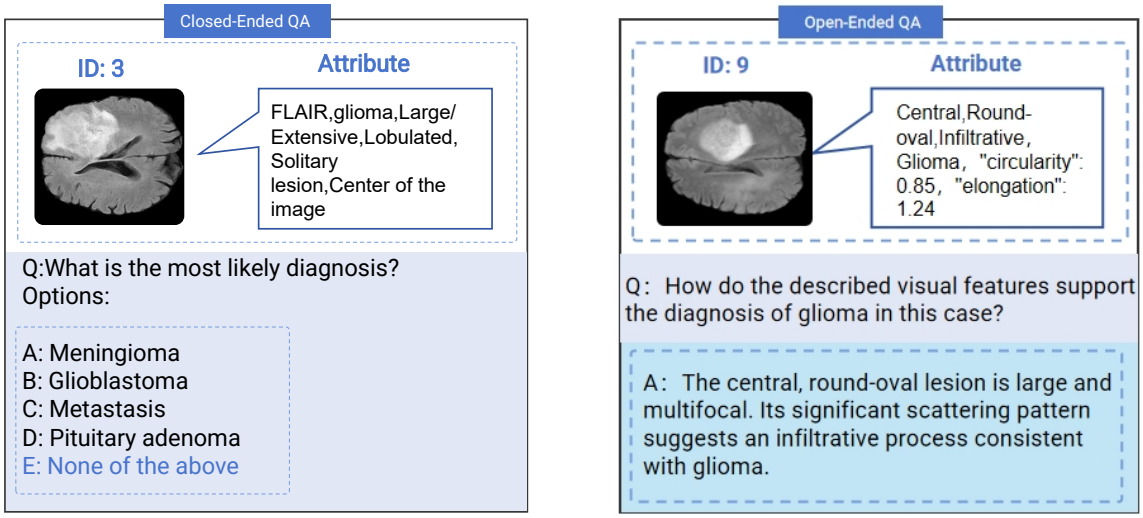


Figure 20: Sample Examples of Closed-ended and Open-ended Question Answering from the Benchmark.