



OmniBrainBench: A Comprehensive Multimodal Benchmark for Brain Imaging Analysis Across Multi-stage Clinical Tasks

Supplementary Material

A. Dataset Details

A.1. Benchmark Construction

To ensure a comprehensive and multifaceted evaluation of model performance, we develop a set of distinct question templates, as illustrated in Fig. 13. These templates are meticulously designed to systematically outline the specific prompts associated with each diagnostic task. Specifically, our approach incorporates two complementary augmentation strategies: rule-wise question augmentation and GPT-wise question augmentation. The rule-based strategy employs structured templates with randomized answer choices, such as identifying hemorrhage types or imaging modalities, to ensure consistency and control over question formulation. In parallel, the GPT-based strategy leverages advanced language models guided by detailed system prompts to generate clinically relevant multiple-choice questions. These prompts require the generation of plausible distractors, randomization of the correct answer’s position, and strict adherence to a standardized output format, thereby enhancing clinical authenticity and variety. Furthermore, to support granular performance analysis across diagnostic subtasks, we introduced a clinical category tagging prompt mechanism. This allows each generated question to be classified into one of 15 predefined clinical categories.

Our pipeline involved a rigorous curation process by a three-tier phase, including the initial VQA generation by GPT-5, preliminary filtration by three junior physicians, and random sampling verification by the chief radiologist. The mapping of questions to clinical domains is conducted under the supervision of a board-certified radiologist with over 13 years of experience, with GPT-5 performing the classification task. This process ensures that the assignments meet the highest standards of clinical relevance. This structured framework facilitates the generation of a diverse and targeted question bank, encompassing a wide spectrum of clinical scenarios and varying levels of complexity. Consequently, it enables the robust and rigorous testing of MLLMs across the entire spectrum of brain imaging analysis capabilities, from foundational anatomical recognition to advanced clinical reasoning.

A.2. Details of Abbreviation

Brain imaging analysis relies on a diverse set of imaging modalities to visualize internal anatomy and function, which are interpreted through a hierarchy of clinical tasks

to support diagnosis and treatment. The details of the abbreviations are given in Table 1.

A wide array of modalities provides complementary information for clinical assessment. Cross-sectional imaging techniques like Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) form the cornerstone. MRI itself encompasses numerous specialized sequences, each highlighting different tissue properties. These include T2-weighted (T2W) and T1-weighted (T1W) imaging for anatomical detail, Fluid-attenuated Inversion Recovery (FLAIR) for suppressing cerebrospinal fluid, and Diffusion-Weighted Imaging (DWI) for detecting cellular density. Further sequences like Magnetic Resonance Angiography (MRA) visualize vasculature, while T1-weighted Contrast-Enhanced (T1CE) imaging assesses vascular permeability and inflammation. Functional MRI (fMRI) maps brain activity, and Susceptibility-Weighted Imaging (SWI) is sensitive to blood products. In nuclear medicine, Positron Emission Tomography (PET) and Single-Photon Emission Computed Tomography (SPECT) provide metabolic and functional data. Proton Density Weighted (PD) imaging offers another contrast mechanism in MRI, and Histopathology Imaging (HI) remains the gold standard for definitive diagnosis. Anatomical Diagrams (ADiag) are often used for reference and education.

The tasks performed using these modalities can be categorized from foundational to advanced. The foundation begins with Imaging Modality Identification (IMI) and Anatomical Structure Identification (ASI), which are prerequisites for higher-level reasoning. Anatomical Function Understanding (AFU) builds upon this structural knowledge. The core of radiological analysis involves Abnormal Screening (AS), Lesion Localization (LL), and Lesion Feature Description (LFD). These can be grouped under the broader umbrella of Lesion Identification and Localization (LIL). Advanced tasks integrate these findings for clinical decision-making. This includes Clinical Sign Prediction (CSP), Disease Diagnosis Reasoning (DDR), and understanding the Pathophysiological Mechanism Correlation (PMC). Together, ASI, AFU, and IMI form the basis of a comprehensive Anatomical and Imaging Assessment (AIA), while DDR and PMC are key components of Diagnostic Synthesis and Causal Reasoning (DSCR). Management-focused tasks include Preoperative Assessment (PA), Treatment Plan Selection (TPS), and Prognostic Factor Analysis (PFA), which contributes to Prognostic

Table 1. Details of The Abbreviation

Modality	
CT	Computed Tomography
MRI	magnetic resonance imaging
DWI	diffusion-weighted imaging
T2W	T2-weighted imaging
T1W	T1-weighted imaging
FLAIR	fluid-attenuated inversion recovery
MRA	magnetic resonance angiography
T1CE	T1-weighted contrast-enhanced
fMRI	functional magnetic resonance imaging
SWI	susceptibility-weighted imaging
PET	positron emission tomography
PD	proton density weighted imaging
HI	histopathology imaging
ADiag	anatomical diagram
SPECT	single-photon emission computed tomography
Task	
AFU	Anatomical Function Understanding
AS	Abnormal Screening
ASI	Anatomical Structure Identification
CSP	Clinical Sign Prediction
DDR	Disease Diagnosis Reasoning
DRP	Drug Response Prediction
IMI	Imaging Modality Identification
LFD	Lesion Feature Description
LL	Lesion Localization
PFA	Prognostic Factor Analysis
PMC	Pathophysiological Mechanism Correlation
POA	Postoperative Outcome Assessment
PA	Preoperative Assessment
RS	Risk Stratification
TPS	Treatment Plan Selection
AIA	Anatomical and Imaging Assessment
DSCR	Diagnostic Synthesis and Causal Reasoning
LIL	Lesion Identification and Localization
PJRF	Prognostic Judgment and Risk Forecasting
TCM	Therapeutic Cycle Management

Judgment and Risk Forecasting (PJRF). Risk Stratification (RS) is another critical prognostic task. Following intervention, Postoperative Outcome Assessment (POA) and Drug Response Prediction (DRP) are essential for monitoring, both falling under the scope of Therapeutic Cycle Management (TCM).

A.3. Diverse Disease Coverage

We have collaborated closely with board-certified radiologists to systematically categorize all diseases appearing in the dataset into two distinct groups: independent diseases and descriptive (non-independent) diseases.

On the one hand, independent diseases refer to well-defined, standalone clinical entities with specific histopathological, genetic, or etiological characteristics that allow them to be diagnosed as distinct nosological units. Examples from our dataset include Meningioma (the frequency count is 60), Glioblastoma (36), Pituitary Adenoma (24), Metastasis (19), Astrocytoma (12), Schwannoma (11), and rare but highly specific entities such as Dysplastic Cerebellar Gangliocytoma (1). These conditions typically present characteristic imaging features and are the final clinical diagnoses recorded in radiology reports.

On the other hand, descriptive diseases are descriptive pathological terms or imaging findings that do not constitute a final, standalone diagnosis but instead describe structural, developmental, or secondary abnormalities that may occur across a wide spectrum of underlying etiologies. Representative examples include Vermis and Midbrain Malformation, Agenesis of the Corpus Callosum, Cortical Dysplasia, Arachnoid Cyst, Chiari I Malformation, Leigh Syndrome, Periventricular Leukomalacia, Holoprosencephaly, Lissencephaly, and vascular anomalies such as Cavernous Malformation or Developmental Venous Anomaly. These findings are important radiological signs, but they usually require integration with clinical context and additional workup to reach a definitive diagnosis. A key motivation behind this fine-grained categorization and the construction of OmniBrainBench stems from a critical limitation observed in existing public brain imaging benchmarks created for Disease Diagnosis and Reasoning (DDR) tasks. Most prior datasets and challenges predominantly focus on a handful of broad, nonspecific categories such as “Tumor”, “Stroke”, “Edema”, “Hemorrhage”, or “Normal/Mild Atrophy.” While these labels are clinically frequent, they fail to reflect the true complexity and diversity that radiologists and neurologists encounter in daily practice, where hundreds of rare and highly specific diagnoses must be considered in the differential.

As evidenced in Table 2 and the accompanying frequency distribution, OmniBrainBench contains 218 unique, radiologist-verified diagnosis labels—an order of magnitude greater than the typical 8–20 classes found in existing benchmarks. Every single label in the dataset has been individually reviewed and validated by at least one senior neuroradiologist to ensure diagnostic accuracy and clinical meaningfulness. This rigorous annotation process guarantees that OmniBrainBench not only dramatically exceeds prior benchmarks in breadth and depth of disease coverage but also provides a clinically authentic testing ground for evaluating the true diagnostic and reasoning capabilities of modern medical vision-language models. By forcing models to distinguish between subtle yet critical entities, such as differentiating a low-grade Dysplastic Cerebellar Gangliocytoma from a Medulloblastoma,

or recognizing the characteristic imaging pattern of Leigh Syndrome versus hypoxic-ischemic injury, we establish a significantly more challenging and clinically representative benchmark for the DDR task.

A.4. OmniBrainBench-Extended

OmniBrainBench-Extend is a patient-level extension of OmniBrainBench, supporting patient-level evaluation via subject IDs. This subset groups multiple question types for the same patient (and potentially multiple visits), covering the five clinical phases of the OmniBrainBench workflow: (1) Anatomical & Imaging Assessment → (2) Lesion Identification & Localization → (3) Diagnostic Synthesis → (4) Prognostic Judgment & Risk Forecasting → (5) Therapeutic Cycle Management. We curate an ADNI longitudinal subset, where abnormality localization is derived from white-matter probability maps and tasks from its metadata. While ADNI is a restricted dataset, we release annotations & metadata for this extension to allow compliant ADNI users to reconstruct it, inviting global neuroscientists to co-develop a full-workflow AI. We will publicly release it at [link](#).

A.5. Ethics Statement

Ethical use of brain imaging data. OmniBrainBench is developed with a strong commitment to ethical practices in handling brain imaging data. All data included in the benchmark are sourced from open-access repositories and published articles. The dataset has been rigorously anonymized and contains no patient-identifiable information, ensuring full compliance with applicable privacy and research ethics guidelines. This collection and benchmarking process does not constitute human subjects research.

Potential societal impacts. Both positive and negative effects on brain imaging study and clinical practice are possible with AI models. Advanced MLLMs have the potential to improve scientific research discovery and speed up diagnostic procedures, but they also run the risk of reinforcing biases in training data, which could result in uneven performance across various demographic groups or neurological conditions. OmniBrainBench uses structured metadata to support the analysis of model biases and fairness in order to help reduce these risks. In order to proactively address these and other new ethical issues, we are dedicated to continuing to engage with the research community.

Data licensing and usage. OmniBrainBench is distributed under the Creative Commons Attribution-ShareAlike 4.0 International license (CC BY-SA 4.0). This licensing framework is chosen to promote transparency, collaboration, and the responsible open-sourcing of resources within the research community. It enables both academic and commercial applications of the benchmark while ensuring that subsequent adaptations and distributions adhere to the same open and ethical principles.

B. Evaluation Details

B.1. Additional experimental verification

Fine-tuning. We fine-tune Qwen3-VL-8B on the training set of OmniBrainVQA (1 epoch, 1e-5 lr). Accuracy on OmniBrainBench surged from 49% (zero-shot) to 76%, validating it as a high-quality instruction-tuning resource for domain alignment.

LLM-as-a-judge. Using a GPT-5-based judge protocol to score clinical relevance (e.g., Gemini-2.5-Pro with 25.86) showed alignment with the standard metric ROUGE1 (Gemini-2.5-Pro with 20.71), confirming that low scores reflect benchmark difficulty rather than metric failure and the judge can effectively identify semantic equivalence.

B.2. Case Study

In this section, we conduct a comprehensive case study analysis of multiple MLLMs in our OmniBrainBench under various scenarios. The evaluation is structured into two primary tracks: closed-ended VQA and open-ended VQA, allowing for a nuanced assessment of model capabilities across different task formats.

Correct Samples. From Figs. 1 to 12, our closed- and open-ended evaluations reveal that state-of-the-art models demonstrate a high degree of proficiency in both accurately interpreting brain imaging data and generating clinically actionable insights. The prevalence of open-ended VQA instances further underscores the models' strong performance in generating detailed, free-form explanations, which is critical for comprehensive diagnostic support. These capabilities indicate a promising role for such models in assisting real-world brain imaging analysis.

Error Case Analysis. A fine-grained analysis of these errors reveals some predominant failure modes:

- **Perception Error.** It occurs when the model fails to correctly identify or localize fundamental visual elements within the brain scan, where a MLLM might misidentify a specific brain structure, overlook a small lesion, or perceive the boundaries of an anomaly. For example, in Fig. 5, Gemini-2.5-Pro [2] likely suffered from a subtle perception error. The MLLM may have correctly perceived the bilateral thalamic hyperintensities but failed to correctly perceive or localize the specific vascular territory involved on the imaging slices. In addition, in Fig. 11, Lingshu-32B [6] demonstrates a critical perception error by failing to observe the fundamental imaging finding of an absent apparent diffusion coefficient signal, which is essential for confirming the true nature of the DWI hyperintensity and making a correct diagnosis.
- **Understanding Error.** The model accurately perceives the visual features but fails to grasp their clinical significance or context, confusing one type of lesion for another or failing to link a radiographic finding to a poten-

tial pathology. For example, in Fig. 1, GPT-5 [4] demonstrates a clear understanding error by failing to grasp the critical epidemiological context that cerebellar metastases are the most common cerebellar tumor, while hemangioblastomas are rare in this age group. Moreover, in Fig. 7, Deepseek-V3.1 [3] demonstrates a fundamental understanding error by failing to grasp the basic clinical significance of increased glucose metabolism on a PET scan, constructing its explanation around a pathophysiologically implausible and factually incorrect premise.

- **Reasoning Error.** The model may correctly perceive and understand individual elements but then make an incorrect clinical deduction. The cases often fall, where the reasoning process of the MLLM becomes opaque or logically inconsistent, resulting in nonsensical or unjustified conclusions. For example, in Fig. 3, Claude-4.5-Sonnet [1] demonstrates a reasoning error by constructing a detailed, post-hoc justification for a decision-making that is epidemiologically improbable, while completely failing to consider the most likely cause of adult-onset focal dystonia. Moreover, in Fig. 9, Qwen3-VL-30B [5] demonstrates a classic reasoning error by substituting general textbook knowledge for a specific brain imaging analysis of the provided options, leading to a conclusion that is logically disconnected from the most probable and contextually appropriate answer.
- **CoT-induced Error.** CoT prompting paradoxically degrades performance (Qwen3-VL-8B: 49%→46%, Gemini-2.5-Pro: 67%→57%). Lengthy reasoning exacerbates hallucinations on fine-grained features, exposing a critical disconnect between medical visual perception and logical reasoning, confirming tasks demand precise visual grounding.

These findings emphasize that while the leading models are highly capable, their deployment in sensitive medical contexts requires careful validation and further refinement to mitigate these specific error types and ensure consistent, interpretable, and reliable results.

References

- [1] Anthropic. Claude sonnet 4.5, 2025. 4
- [2] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blisstein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. [arXiv:2507.06261](https://arxiv.org/abs/2507.06261), 2025. 3
- [3] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. [Nature](https://www.nature.com/articles/s41586-025-05016-2), 2025. 4
- [4] OpenAI. Gpt-5, 2025. 4
- [5] Qwen Team. Qwen3-vl: Sharper vision, deeper thought, broader action, 2025. 4
- [6] Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. [arXiv:2506.07044](https://arxiv.org/abs/2506.07044), 2025. 3

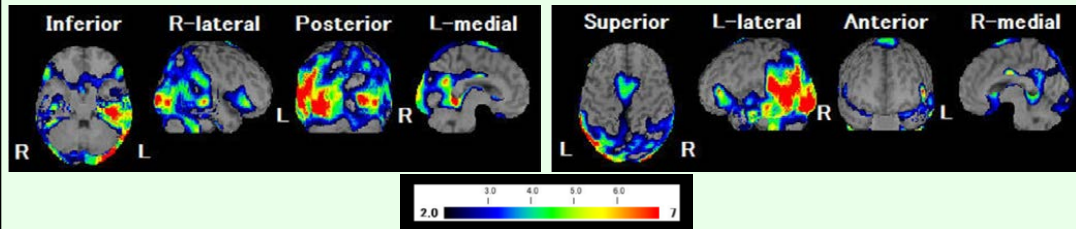
Table 2. Diverse disease coverage on our OmniBrainBench.

Diverse Disease Coverage		
Tumor	Toxic Or Metabolic Encephalopathies	Metastatic Brain Lesions
Stroke	Granulomatosis With Polyangiitis	Pericallosal Lipoma
Aneurysm	Leptomeningeal Spread	Moyamoya Disease
Meningioma	Vasculopathy w. Cerebral Leukoencephalopathy	Fragile X-associated Tremor/ataxia Syndrome
Glioma	Familial Cerebral Cavemous Malformation	Cavemous Malformation
Glioblastoma	Cerebral Venous Air Embolism	Maxillary Sinusitis
Pituitary Adenoma	Hypothalamic Hamartoma	Chordoma
Metastasis	Dolichoectasia	Sinus Pericranii
Adenoma	Vexas Syndrome	Pontine Tuberculoma
Astrocytoma	Arachnoid Cyst	Lewy Body Dementia
Schwannoma	Cns Toxoplasmosis	Alpha-methylacyl-coa racemase deficiency
Arteriovenous Malformation	Hypomelanosis Of Ito	Pineoblastoma
Venous Sinus Thrombosis	Methylmalonic Acidaemia	Epidural Hematoma
Neoplasm	Cerebral Fat Embolism	Pilocytic Astrocytoma
Multiple Sclerosis	Intracranial Atherosclerosis	Oropouche Virus Encephalitis
Temporal Lobe Epilepsy	Olfactory Bulb Meningioma	Mature Teratoma
Cavemous Angioma	Uremic Encephalopathy	Neurosyphilis
Parkinson's Disease	Adamantinomatous Craniopharyngioma	Wilson Disease
Lymphoma	Small Vessel Ischemic Disease	Cavemous Cerebral Malformation
Fahr's Syndrome	Lipoma Of The Corpus Callosum	Sphenoid Mucocele
Tuberous Sclerosis Complex	Mitochondrial Echs1 Deficiency	Inflammatory Demyelinating Lesions
Epilepsy	Myeloid Sarcoma	Anterior Cerebral Artery Aneurysm
Vestibular Schwannoma	Vertebral Artery Dissection	Vermis Agenesis
Subdural Hematoma	Demyelinating Disease	Cerebellar tumor
Malignant Neoplasm	Frontal Sinusitis	Infratemporal Fossa Tumor
Wilson's Disease	Venolymphatic Malformation	Actinomyces Osteomyelitis
Vascular Malformation	Meningoencephalitis	Clivus chordoma
Hemangioblastoma	Lissencephaly	Small Vessel Disease
Focal Cortical Dysplasia	Reversible Cerebral Vasoconstriction Syndrome	Central Pontine Myelinolysis
Epidermoid	Late-infantile Metachromatic Leukodystrophy	Angiosarcoma
Acoustic Neuroma	HSV Encephalitis	Carotid Artery Dissection
Joubert Syndrome	Venous Malformation	Baló's concentric sclerosis
Alzheimer's Disease	Frontotemporal dementia	Progressive Supranuclear Palsy
Hypertrophic Olivary Degeneration	Mild Encephalitis/encephalopathy	Megalencephaly
Frontotemporal Dementia	Haemangioblastoma	Anterior Cerebral Artery Stroke
Hypopituitarism	Disorder Of Glycosylation Type-1a	Jugular Bulb Thrombosis
Pontocerebellar Hypoplasia	Midbrain Tectum Glioma	Pachygyria
Craniosynostosis	H3K27M-mutant glioma	Poretti-boltshauer Syndrome
Hemimegalencephaly	Dyke-Davidoff-Masson Syndrome	Human Immunodeficiency Virus Dementia
Wernicke's Encephalopathy	Agenesis of the Septum Pellucidum	Lipoma
Choroid Plexus Papilloma	Benign Or Low-grade Neoplasm	Lgi1 Autoimmune Encephalitis
Brainstem glioma	Atherosclerosis	Myelin Oligodendrocyte Glycoprotein Antibody
Central Neurocytoma	Vermis hypoplasia	Primary Angiitis Of The Central Nervous System
Central Nervous System Germinoma	Posterior Cortical Atrophy	Methanol Toxicity
Transient Ischemic Attack	Encephalitis	Anterior Choroidal & Thalamoperforate Arteries Syndrome
Rhinocerebral Mucormycosis	Juvenile Angiofibroma	Orbital Cellulitis
Chiari I Malformation	Fibromuscular Dysplasia	Corpus Callosum Agenesis
Dysplastic Cerebellar Gangliocytoma	Dysplastic Gangliocytoma	Reversible Posterior Leukoencephalopathy Syndrome
Paranasal Sinus Tumor	Congenital CMV Infection	Idiopathic Hyperatrophic Pachymeningitis
Enterovirus A71 Rhombencephalomyelitis	Rhino-orbital-cerebral Mucormycosis	Vestibulocochlear Nerve Schwannoma
Anti-Ig1 Receptor Encephalitis	Carotid-cavemous Fistula	Fabry disease
Textiloma	Toxoplasmosis	Acute Complete Occlusion Of Internal Carotid Artery
Limbic Encephalitis	Chronic Subdural Hematoma	Polymicrogyria
Gliosarcoma	Plasmacytoma	Cryptococcosis
Marchiafava-bignami Disease	Acute Subdural Hematoma	Medulloblastoma
Progressive External Ophthalmoplegia	Rhabdomyosarcoma	Chiari malformation type III
Herpes Simplex Encephalitis	Posterior Reversible Encephalopathy Syndrome	Acute Disseminated Encephalomyelitis
Hypoxic Ischemic Encephalopathy	Nocardia Asteroides Infection	Congenital Fusion Of The Radius And Ulna
Meningomyelocele	Acute Necrotizing Encephalitis Of Childhood	Cerebral Hydatid Disease
Rabies	Tuberculomas	X-linked Adrenoleukodystrophy
Krabbes Disease	Hypoglycaemic Encephalopathy	Intraventricular Migration Of Intra-ocular Silicone Oil
Leigh's Disease	HHV-6 Encephalitis	Autoimmune Subacute Encephalitis
Ependymoma	Basilar Artery Thrombosis	Ethmoid Sinusitis
Ischemic Stroke	Meningioangiomas	Beta-propeller Protein-associated Neurodegeneration
Vertebral artery dissection	Neurodegeneration w. Brain Iron Accumulation	Sphenoid Sinus Mucocele
Arteriovenous Fistula	Prolactinoma	Vein of Galen malformation
Cerebrotendinous Xanthomatosis	Thrombosis Of The Dural Sinuses	Agenesis Of The Corpus Callosum
Chronic Cerebrovascular Disease	Transverse Sinus Thrombosis	Granulomatous Amebic Encephalitis
Bacterial Meningitis	Maxillary Sinus Tumor	Huntington's Disease
Linear Scleroderma	Diabetic Striopathy	Low-grade Fibromyxoid Sarcoma
High-grade Neoplasm	Craniopharyngioma	Apert Syndrome
Amyotrophic Lateral Sclerosis	Vitamin B1 (thiamine) Deficiency	Pericallosal Artery Aneurysm
Colpocephaly	Cerebral Air Embolism	

< Closed-ended VQA >

Q: A 72-year-old patient presents with progressive memory loss and cognitive decline and SPECT imaging. SPECT imaging reveals bilateral hypoperfusion in temporo-parietal lobes and posterior cingulate gyrus, more prominent on the left. Which condition is most consistent with these findings?

- (A) **Alzheimer's disease** (B) Acute bacterial meningitis (C) Glioblastoma multiforme
(D) Recent ischemic stroke (E) Herpes simplex encephalitis



GPT-5 response:

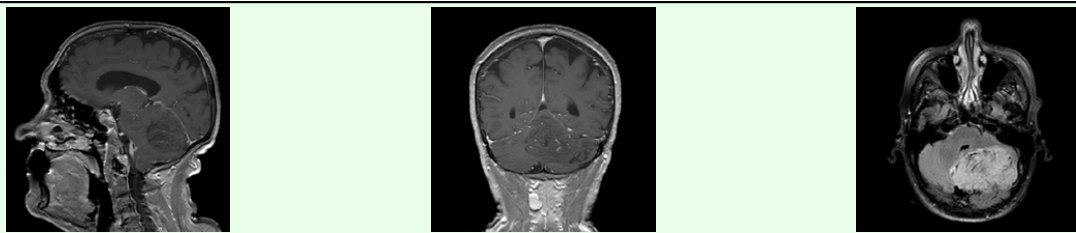
The SPECT scan shows bilateral hypoperfusion in the temporo-parietal lobes and posterior cingulate gyrus, more prominent on the left. This is most consistent with ***(A) Alzheimer's disease****.

Other options are less likely: *****Acute bacterial meningitis****: Inflammation, not hypoperfusion. *****Glioblastoma****: Localized hypoperfusion, not widespread. *****Recent ischemic stroke****: More localized hypoperfusion. *****Herpes simplex encephalitis****: Temporal lobe involvement, not this pattern. **Thus, the answer is ***(A) Alzheimer's disease****.**

Correct Sample

Q: A 77-year old woman presented to our institution with persistent headache, gait instability, dysarthria and dysphagia. A head MRI was scheduled. What is the most likely diagnosis for this participant?

- (A) **Dysplastic gangliocytoma of the cerebellum** (B) Cerebellar hemangioblastoma
(C) Low-grade astrocytoma (D) Metastatic lesion to the cerebellum (E) Medulloblastoma



GPT-5 response:

Given the clinical presentation of headache, gait instability, dysarthria, and dysphagia, along with the age of the patient, cerebellar hemangioblastoma (Option B) is the most likely diagnosis. These tumors are associated with the cerebellum and are frequently observed in individuals around this age. The MRI would typically show a well-defined, enhancing mass with cystic features, which is consistent with the patient's presentation. Additionally, hemangioblastomas often present with symptoms like those seen in this case due to their mass effect on nearby structures and vascularity. **In conclusion, (B) Cerebellar hemangioblastoma is the most likely diagnosis for this patient.**

Error Sample

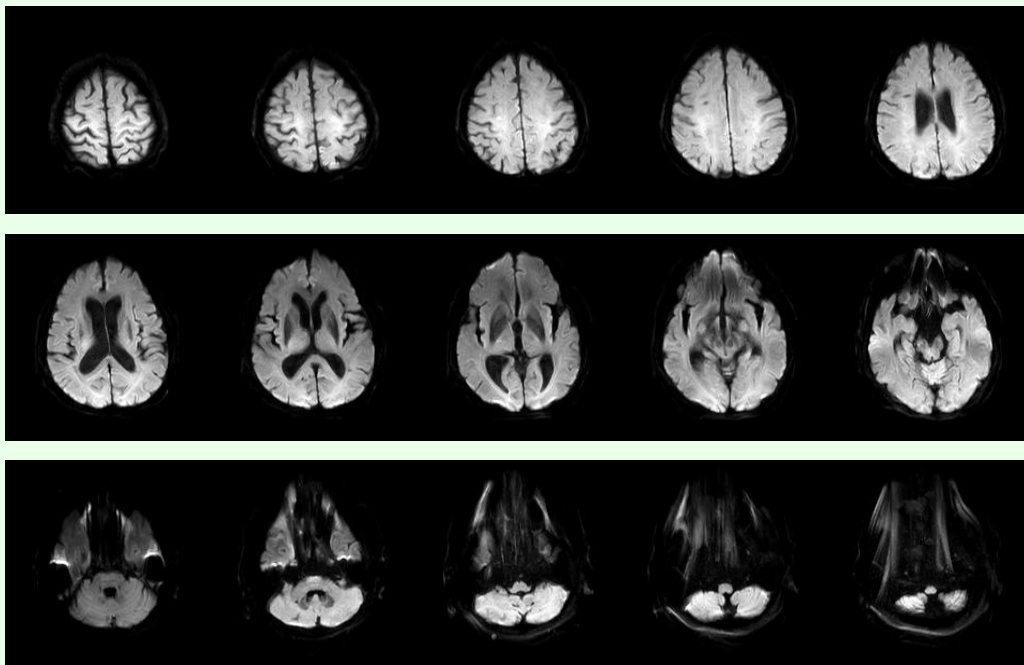
Figure 1. Correct/Error samples in GPT-5 closed-ended VQA.

< Open-ended VQA >

Q: A 68-year-old patient presents with cognitive decline and mild gait instability. Review the 18 available DWI images and describe any significant findings that could explain these symptoms, including potential acute/subacute pathology, chronic changes, and incidental findings.

A:

GT: **Findings: Multiple scattered patchy isointense T1 and hyperintense T2 signals are observed in the pons, bilateral thalami, basal ganglia, cerebral white matter, and cerebellar hemispheres, appearing as hyperintense and hypointense on FLAIR and as isointense to hypointense on DWI. Additionally, a punctate slightly hyperintense DWI signal is noted in the right cerebral peduncle. The cerebral sulci and cisterns are enlarged, and the ventricles are symmetrical. Symmetrical patchy hyperintense T1 and T2 signals are seen in the periventricular white matter, appearing as hyperintense on FLAIR and isointense on DWI. The midline structures are not displaced. Mucosal thickening is present in the bilateral ethmoid sinuses. **Conclusion:** - Hyperintense DWI signal in the right cerebral peduncle, suggesting possible subacute cerebral infarction—clinical correlation and follow-up recommended. - Multiple ischemic degenerative foci and small encephalomalacia in the brain. - Age-related cerebral changes. - Bilateral ethmoid sinusitis.**



GPT-5 response:

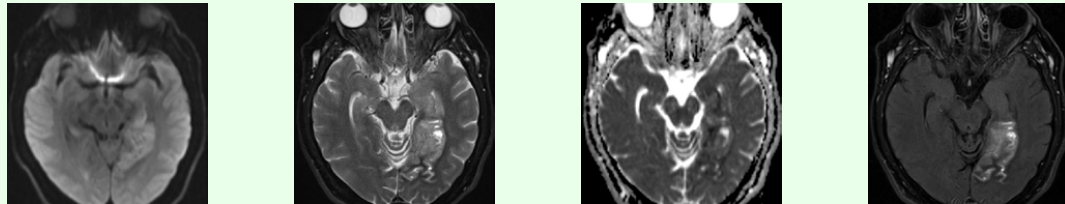
- Acute/subacute pathology on DWI: No focal diffusion-restriction identified. No evidence of acute ischemic infarct, mass effect, or hemorrhage on these images. - Chronic changes: Prominent sulci at the high convexities consistent with generalized cortical atrophy (age-related volume loss), which can correlate with cognitive decline. Pattern does not suggest normal-pressure hydrocephalus (no tight/high-convexity crowding). - Incidental findings: None evident on the available DWI slices. Overall, the DWI series does not show an acute cause for the symptoms. Consider correlation with FLAIR/T2 for chronic small-vessel disease or other neurodegenerative changes, and clinical workup for non-acute causes of gait instability and cognitive decline.

Figure 2. Correct/Error samples in GPT-5 open-ended VQA.

< Closed-ended VQA >

Q: 59-year-old male presenting with first occurrence of generalized tonic-clonic seizure. Patient had a history of mantle cell lymphoma 2-years prior and was on maintenance therapy. What is the most likely diagnosis for this participant?

- (A) Glioblastoma (B) Primary CNS lymphoma **(C) Secondary CNS lymphoma**
(D) Cerebritis/encephalitis (E) Subacute cerebral infarction



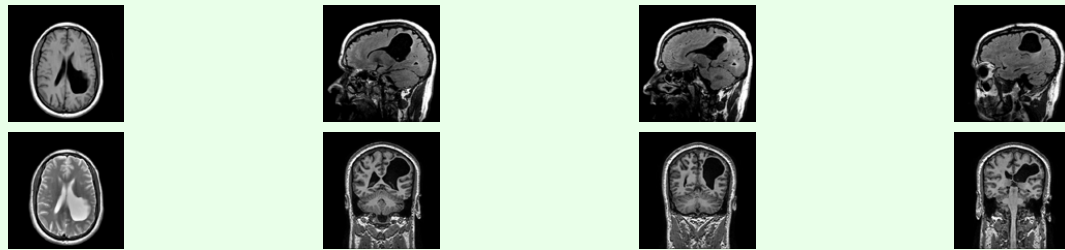
Claude-4.5-Sonnet response:

The answer is (C) Secondary CNS lymphoma. The combination of: 1- Known systemic mantle cell lymphoma; 2- Maintenance therapy status; 3- Bilateral deep gray matter involvement; 4- New neurological symptoms (seizure) strongly supports secondary CNS involvement by the patient's known lymphoma. CNS relapse or progression should be considered in any patient with a history of systemic lymphoma presenting with new neurological symptoms. This patient would require CSF analysis, consider brain biopsy if feasible, and revision of systemic therapy to include CNS-penetrating agents.

Correct Sample

Q: 38-year-old male with Right hand and foot dystonia. What is the most likely diagnosis for this participant?

- (A) Porencephaly** (B) Schizencephaly (C) Cerebral palsy
(D) Neuroglial cysts (E) Wilson's disease



Claude-4.5-Sonnet response:

Based on the neuroimaging provided and the clinical presentation, schizencephaly is the most likely diagnosis for this 38-year-old male presenting with right hand and foot dystonia. The schizencephalic cleft disrupts normal motor pathways from the left motor cortex, resulting in the dystonic movements of the right hand and foot. The severity and distribution of symptoms depend on the size and location of the cleft. This patient's relatively limited symptoms suggest compensation by surrounding neural tissue throughout development, with dystonia emerging as the predominant manifestation.

Answer: (B) Schizencephaly

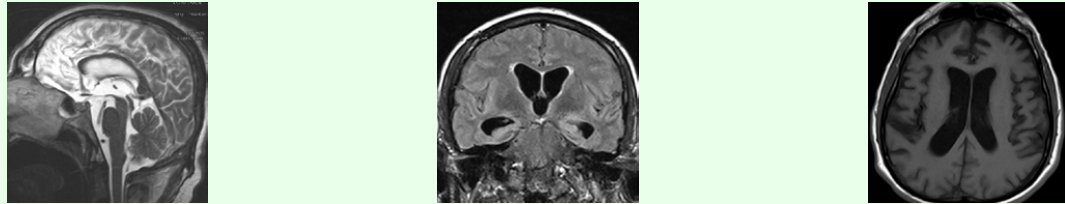
Error Sample

Figure 3. Correct/Error samples in Claude-4.5-Sonnet closed-ended VQA.

< Closed-ended VQA >

Q: 32-year-old gentleman with progressive cognitive decline and memory impairment for 2 years. What is the most likely diagnosis for this participant?

- (A) Psychosis (B) Frontotemporal dementia (C) Acute meningoencephalitis
(D) Temporal lobe atrophy and meningoencephalitis secondary to neurosyphilis
(E) Progressive dementia



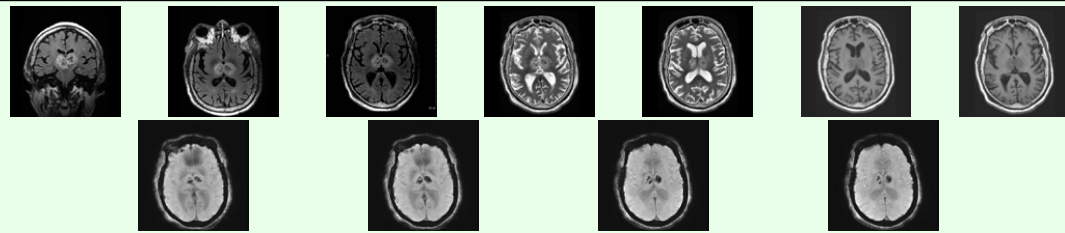
Gemini-2.5-Pro response:

The most likely diagnosis for this participant is (D) Temporal lobe atrophy and meningoencephalitis secondary to neurosyphilis. This conclusion is reached by carefully integrating the patient's clinical history with the striking findings on the provided brain MRI scans. The patient is a 32-year-old gentleman, which is an unusually young age for the onset of significant dementia. The presentation of progressive cognitive decline and memory impairment over a two-year period indicates a chronic, neurodegenerative, or chronic inflammatory process rather than an acute event.

Correct Sample

Q: 86-year-old female patient, found down, scalp haematoma, concern for seizure vs stroke. What is the most likely diagnosis for this participant?

- (A) Top of the basilar artery syndrome (B) Bilateral internal cerebral vein thrombosis
(C) Cerebral venous sinus thrombosis with cortical vein infarction
(D) Bilateral thalamic infarction (E) Acute subdural hematoma with transtentorial herniation



Gemini-2.5-Pro response:

The most likely diagnosis for this participant is (B) Bilateral internal cerebral vein thrombosis.

This conclusion is based on a careful analysis of the patient's clinical presentation and the highly specific findings on the provided multi-sequence brain MRI. MRI images reveals a distinct and pathognomonic pattern of injury centered on the deep gray matter structures of the brain. The clinical presentation of being "found down" is consistent with a sudden, severe neurological event causing altered consciousness. MRI provides classic evidence of bilateral, symmetric venous infarction of the thalami, a hallmark pattern of deep cerebral venous thrombosis involving the internal cerebral veins.

Error Sample

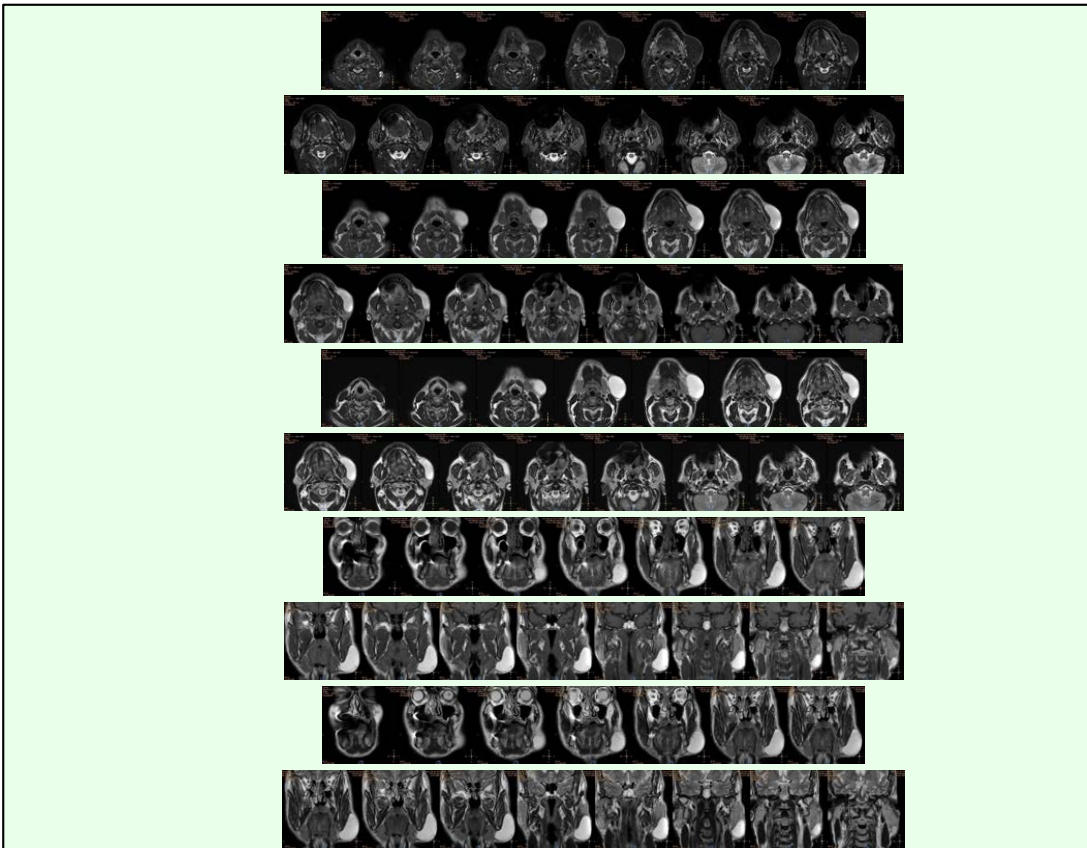
Figure 5. Correct/Error samples in Gemini-2.5-Pro closed-ended VQA.

< Open-ended VQA >

Q: A patient presents with a left submandibular mass. Based on the MRI sequences provided (T1, T2, STIR, and coronal images), what are the characteristic imaging features of this lesion and what would be the most likely differential diagnosis?

A:

GT: Left submandibular region oval shaped well defined mass. It exhibits high signal in T1, T2 and suppressed in STIR images. No bony or soft tissue infiltration.



Gemini-2.5-Pro response:

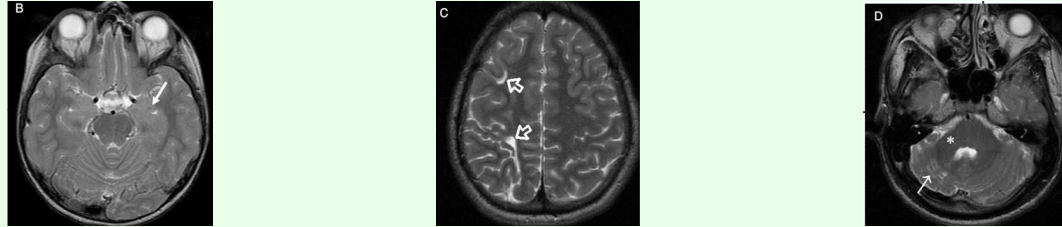
Based on the provided MRI sequences, here is a concise summary of the imaging features and differential diagnosis: **Imaging Features:** The images demonstrate a well-circumscribed, encapsulated mass in the left submandibular space. The key imaging characteristics are: **T1-weighted images (Image 2 & 4):** The lesion is uniformly hypointense (dark). **T2-weighted and STIR images (Image 1, 3, & 5):** The lesion is markedly and uniformly hyperintense (very bright), a feature often described as the "lightbulb sign." These features are characteristic of a benign, non-infiltrative lesion with a very high fluid or myxoid content. **Differential Diagnosis:** Given the location and classic imaging features, the most likely differential diagnosis includes: 1. **Schwannoma:** This is a top consideration due to the well-defined margins and classic "lightbulb" T2 hyperintensity, which is highly characteristic of this benign nerve sheath tumor. 2. **Pleomorphic Adenoma:** As the most common benign tumor of the submandibular gland, this is a strong possibility. While typically T2 hyperintense, they may not always be as uniformly bright as seen here. 3. **Branchial Cleft Cyst (Second):** This is a common congenital neck cyst that can present in this location and typically appears as a well-defined, T1-dark, T2-bright lesion.

Figure 6. Correct/Error samples in Gemini-2.5-Pro open-ended VQA.

< Closed-ended VQA >

Q: A T2-weighted MRI of the brain reveals a hyperintense area in the cerebellum. Which of the following is the most likely clinical significance of this finding?

- (A) Normal age-related atrophy **(B) Cerebellar stroke, tumor, or demyelinating lesion**
(C) Benign developmental variant (D) Artifact from patient motion
(E) Chronic ischemic changes without acute significance



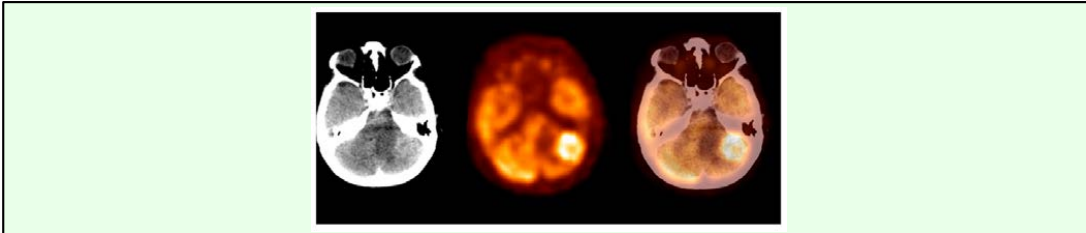
Deepseek-V3.1 response:

The most appropriate answer is (B) Cerebellar stroke, tumor, or demyelinating lesion. A hyperintense area on T2-weighted MRI indicates an increase in water content in that region of the brain. While small hyperintense signals can appear as benign, incidental findings, a focal, well-defined hyperintense area in the cerebellum often suggests a pathologic process rather than a normal variant. MRI shows a focal hyperintensity in the cerebellum, the clinically significant interpretation is that this indicates a cerebellar stroke, tumor, or demyelinating lesion. This finding warrants further clinical correlation, possibly additional imaging, or laboratory testing to determine the exact pathology.

Correct Sample

Q: A patient undergoes a PET scan that reveals bright spots in specific brain regions indicating increased glucose metabolism. What is the most likely clinical significance of these findings?

- (A) Benign calcifications (B) Acute ischemic stroke
(C) Metastatic brain lesions (D) Normal age-related changes
(E) Cerebral edema from trauma



Deepseek-V3.1 response:

The hyperintense spots on PET correspond to regions of increased glucose consumption, typical of acute ischemic injury before irreversible damage sets in. These areas may appear as bright spots on PET scans before clinical deficits become apparent or before structural changes are visible on other imaging modalities like MRI. In summary: Early phase of acute ischemic stroke can efficiently display focal hypermetabolism on PET scans, reflecting increased neuronal activity or metabolic demand during acute ischemic injury before irreversible damage causes decline. Therefore, the most appropriate answer is: (B) Acute ischemic stroke.

Error Sample

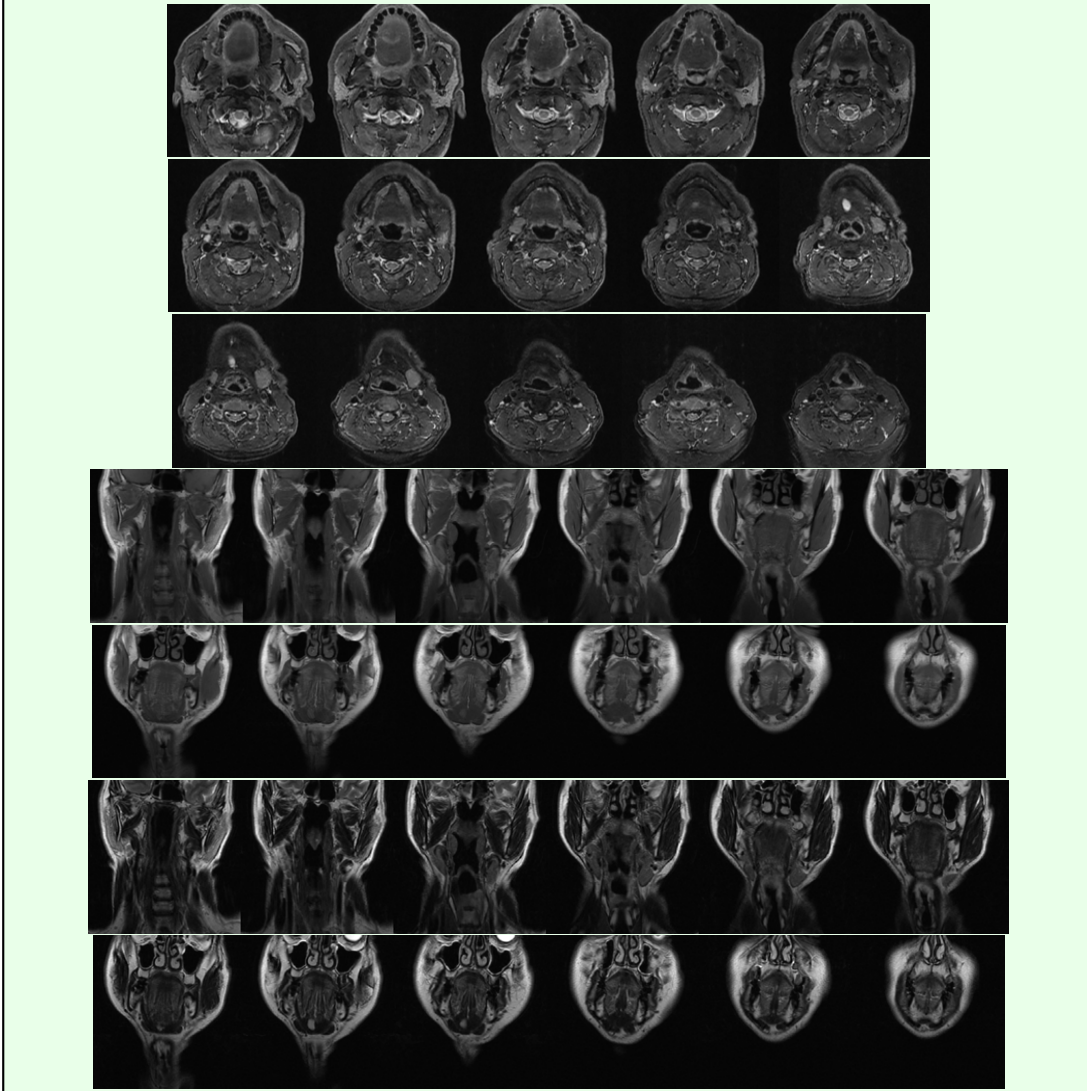
Figure 7. Correct/Error samples in Deepseek-V3.1 closed-ended VQA.

< Open-ended VQA >

Q: Based on the MRI images (875_Axial_STIR_67.jpg, 875_Coronal_T1_347.jpg, 875_Coronal_T2_483.jpg), what is the most likely diagnosis for the fluid-containing structure in the right sublingual space, and what anatomical considerations are important for its management?

A: _____

GT: Cystic lesion in the right sublingual space.



Deepseek-V3.1 response:

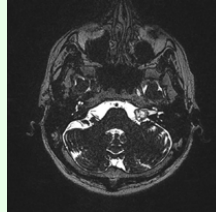
Based on the provided MRI images (875_Axial_STIR_67.jpg, 875_Coronal_T1_347.jpg, 875_Coronal_T2_483.jpg), the most likely diagnosis is a **ranula**, specifically a **plunging ranula** given its extension into the sublingual space. The key anatomical consideration for management is the relationship of the lesion to the **mylohyoid muscle**. A simple ranula is superior to the muscle, while a plunging ranula (as suggested by the sublingual location) has dissected posteriorly around the free edge of the mylohyoid muscle. Surgical management (e.g., excision of the ranula along with the sublingual gland) must address this pathway to prevent recurrence.

Figure 8. Correct/Error samples in Deepseek-V3.1 open-ended VQA.

< Closed-ended VQA >

QA 52-year-old patient presents with unilateral hearing loss and an MRI reveals a lesion in the pontocerebellar angle. Which of the following is the most likely diagnosis based on common tumors in this location?

- (A) Glioblastoma multiforme **(B) Vestibular schwannoma** (C) Pituitary adenoma
(D) Cerebral metastasis (E) Oligodendroglioma



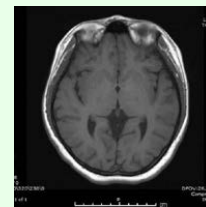
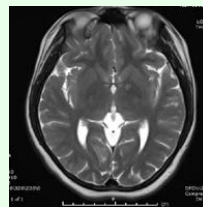
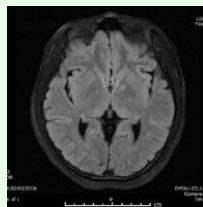
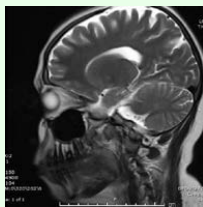
Qwen3-VL-30B response:

This conclusion is based on the classic clinical presentation and the characteristic location of the lesion revealed by the MRI. The patient is a 52-old with unilateral hearing loss. This symptom, particularly when it is progressive and affects one ear only, is highly suggestive of a disorder involving the auditory nerve. The most common cause of unilateral sensorineural hearing loss in adults is a tumor arising the vestibular portion of this nerve. The combination of unilateral hearing loss and a mass in the pontocerebellar angle is the classic presentation of a vestibular schwannoma.

Therefore, the most likely diagnosis (B) Vestibular schwannoma.

Correct Sample

Q: A 15-year-old female patient is being followed up for neurofibromatosis type 1. Cranial MR examination was taken for control purposes. What is the most likely diagnosis for this participant? (A) Plexiform neurofibroma of the trigeminal nerve (B) Optic nerve glioma
(C) **Cranial focal area of signal intensity in NF type 1** (D) Sphenoid wing dysplasia
(E) Cerebral aqueductal stenosis with hydrocephalus



Qwen3-VL-30B response:

The most likely diagnosis for this participant is (B) Optic nerve glioma. This conclusion is based on a comprehensive analysis of the provided cranial MRI images and the clinical context of neurofibromatosis type 1 (NF1). Neurofibromatosis type 1 is a genetic disorder characterized by the development of benign tumors along nerves. It is one of the most common inherited neurological conditions, affecting approximately 1 in 3,000 individuals. The hallmark features include café-au-lait spots, neurofibromas, and Lisch nodules. A aspect of NF1 management is the surveillance for central nervous system complications, particularly optic pathway gliomas.

Error Sample

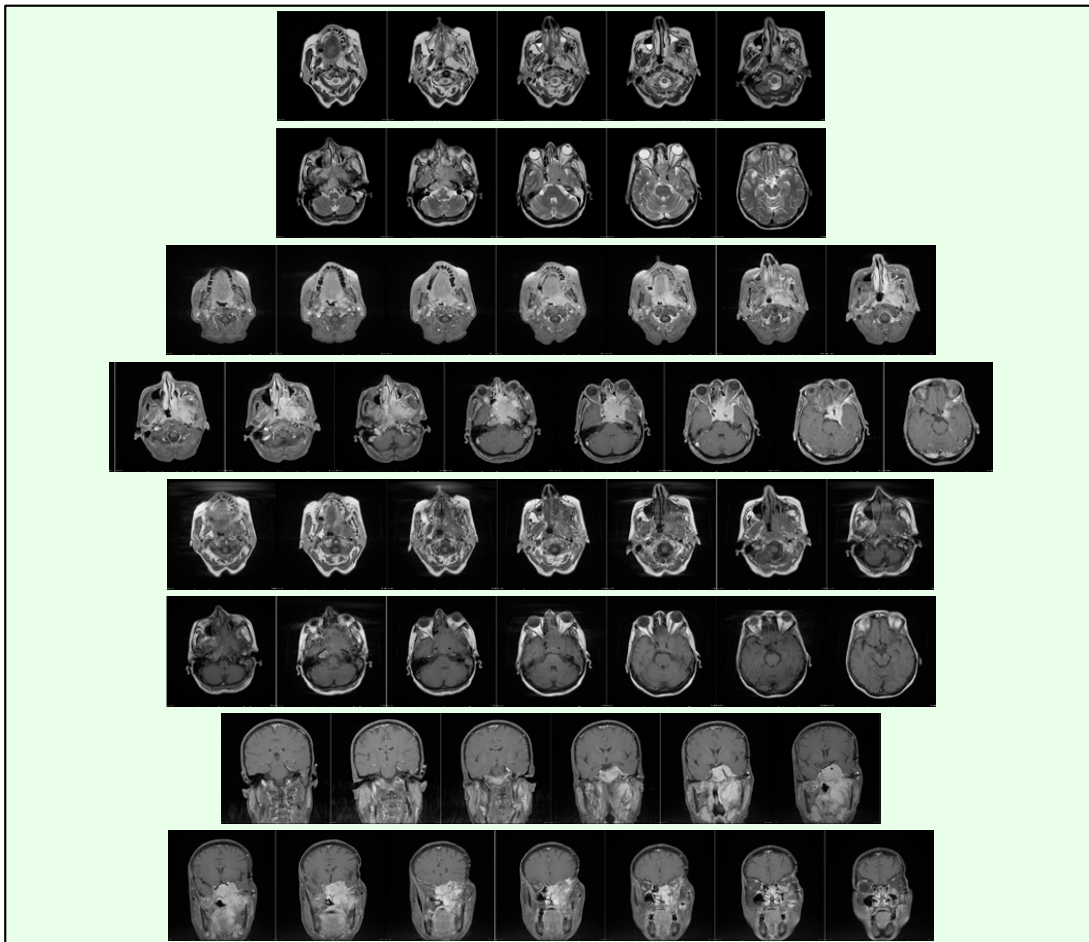
Figure 9. Correct/Error samples in Qwen3-VL-30B closed-ended VQA.

< Open-ended VQA >

Q: Based on the provided axial and coronal MRI images (FLAIR, T1, and post-contrast sequences with fat saturation), what are the key anatomical structures involved by this large skull base tumor, and what are the most concerning features regarding potential complications or surgical planning?

A:

GT: Large destructive skull base meningioma (biopsy proven) demonstrated. The tumor extends into the ethmoid and sphenoid sinuses, left maxillary sinus, nasal cavity, pituitary fossa, left petrous apex and clivus. There is involvement of the left orbital apex with mild proptosis of the left globe and extension into the pterygopalatine fossa. Mass effect on the left mesial temporal lobe and inferior frontal lobe. The mass extends into the prepontine cistern. The tumor encases the intracranial left ICA and extends into the middle cranial fossa . There is extension of the tumor into the suprahyoid aspect of the neck.



Qwen3-VL-30B response:

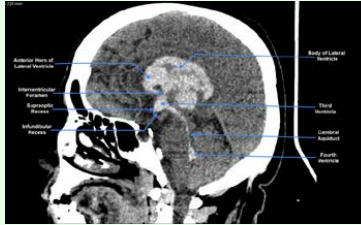
The key anatomical structures involved are the right petrous apex, the middle and posterior cranial fossae, the cerebellopontine angle, and the internal auditory canal. The most concerning features are the extensive bone destruction, the involvement of the internal auditory canal (suggesting potential hearing and vestibular nerve compromise), the compression of the brainstem and cerebellum, and the potential for vascular encasement (e.g., of the internal carotid artery or vertebrobasilar system), all of which significantly complicate surgical resection and increase the risk of neurological deficits.

Figure 10. Correct/Error samples in Qwen3-VL-30B open-ended VQA.

< Closed-ended VQA >

Q: A 65-year-old patient presents with acute neurological deterioration. A non-contrast CT scan reveals a large hematoma in the right basal ganglia with significant midline shift. Which of the following is the most likely underlying etiology for this finding?

- (A) **Chronic hypertension** (B) Bacterial meningitis (C) Ischemic stroke
(D) Brain tumor metastasis (E) Multiple sclerosis exacerbation



Lingshu-32B response:

The most likely underlying etiology for the large hematoma in the right basal ganglia with significant midline shift is (A) Chronic hypertension. Chronic hypertension is a well-known risk factor for intracerebral hemorrhage, particularly in the basal ganglia region. The high blood pressure can cause weakening of the small vessels, making them prone to rupture and leading to hemorrhage. The absence of other signs such as mass effect in the brain, which might be seen in brain tumor metastasis, or the inflammatory signs that would be present in bacterial meningitis or multiple sclerosis exacerbation, further supports chronic hypertension as the primary cause.

Correct Sample

Q: A patient's MRI shows both T2-weighted imaging (T2WI) and diffusion-weighted imaging (DWI) hyperintensities in the basal ganglia. These findings are most consistent with which of the following pathological processes?

- (A) **Vasogenic edema** (B) Cytotoxic edema (C) Demyelination (D) Cerebral amyloid angiopathy (E) Leptomeningeal enhancement



Lingshu-32B response:

The patient's MRI shows hyperintensities in the basal ganglia on both T2-weighted imaging (T2WI) and diffusion-weighted imaging (DWI). Additionally, red arrows highlight areas of interest corresponding to these hyperintensities across different sequences. Final Conclusion: Hyperintensities Within Basal Ganglion Regions Simultaneously Displayed Across BOTH T2 And Diffusion Weighted Sequences Strongest Consistent With Cytoplasmic Swelling Result From Cellular Injury Induced By Toxic Metabolites Or Hypoxia Ischemia. Therefore Most Likely Diagnosis Correspondingly Belongs TO "Cytotoxic edema". Correct answer: (B) Cytotoxic edema.

Error Sample

Figure 11. Correct/Error samples in Lingshu-32B closed-ended VQA.

< Open-ended VQA >

Q: A patient presents with a palpable left scalp mass. Based on the imaging findings, what is the most likely vascular diagnosis and what secondary intracranial finding suggests hemodynamic compromise?

A: _____

GT: Subcutaneous mass on the left side of the head showing multiple serpiginous flow voids within the lesion which enhance with contrast, in keeping with dilated vessels. The lesion shows feeding branches from the ipsilateral external carotid artery and numerous dilated draining extracranial veins. Cerebellar parenchyma appears to be shrunken probably due to reduced perfusion, known as the steal phenomenon.



Lingshu-32B response:

The most likely vascular diagnosis is a left transverse/sigmoid sinus dural arteriovenous fistula (AVF) with venous aneurysm. The secondary intracranial finding suggesting hemodynamic compromise is the presence of venous infarcts.

Figure 12. Correct/Error samples in Lingshu-32B open-ended VQA.

(a) Rule-wise Question Augmentation

QA Generation Prompt Example I	QA Generation Prompt Example II
<p>Question: " What is the type of the `intracranial hemorrhage?", Answer Choices:</p> <p>(A) [option1] (B) [option2] (C) [option3] (D) [option4] (E) [option5]</p> <p>Note that options are randomly selected from "intraparenchymal", "subdural", "subarachnoid", "intraventricular", and "epidural".</p>	<p>Question: " Which type of imaging modality was utilized to obtain this image?", Answer Choices:</p> <p>(A) [option1] (B) [option2] (C) [option3] (D) [option4] (E) [option5]</p> <p>Note that options are randomly selected from "FLAIR", "T1W", "fMRI", "DWI", and "PET".</p>

(b) GPT-wise Question Augmentation

QA Generation Prompt Example III
<p><System Prompt> You are an expert medical educator. Your task is to create a multiple-choice question based on the given question and answer. Make the question clinically relevant and ensure all options are plausible for medical professionals.</p> <p>{Requirements}: 1. Create a clear, clinical question that can be answered using the provided information; 2. Generate five answer choices (A, B, C, D, E); 3. One choice should be correct based on the answer information; 4. Four choices should be plausible but incorrect; 5. Randomize the position of the correct answer; 6. Format the output exactly as specified.</p> <p>{Output format}: Question: [Your generated question]. Answer Choices: (A) [option1] (B) [option2] (C) [option3] (D) [option4] (E) [option5]. Correct Answer: [A/B/C/D/E].</p>

(c) Clinical Category Tagging Prompt

Clinical Category Tagging Prompt
<p><System Prompt> You are a senior professor and expert in neurology, neurosurgery, and neuroradiology with decades of clinical and research experience. Your task is to classify medical questions related to brain imaging and neurological conditions into the most appropriate diagnostic/clinical category.</p> <p>Standard categories include: 1. Anatomical Structure Identification: Identifying normal brain anatomical structures in images 2. Imaging Modality Identification: Identifying and distinguishing between different imaging modalities 3. Anatomical Function Understanding: Understanding neuroanatomical functions and brain regions 4. Abnormal Screening: Detecting abnormalities in brain imaging or neurological tests 5. Lesion Feature Description: Describing imaging characteristics, morphology, or appearance of brain lesions 6. Lesion Localization: Identifying precise anatomical location of brain lesions or abnormalities 7. Disease Diagnosis Reasoning: Differential diagnosis and disease identification from imaging/clinical data 8. Pathophysiological Mechanism Correlation: Understanding disease mechanisms and pathophysiology 9. Risk Stratification: Clinical risk assessment, patient risk stratification, or risk factor evaluation 10. Prognostic Factor Analysis: Analyzing factors affecting patient prognosis, survival, or treatment outcomes 11. Clinical Sign Prediction: Predicting neurological symptoms or clinical manifestations 12. Drug Response Prediction: Medication efficacy prediction, drug selection, or treatment response assessment 13. Preoperative Assessment: Evaluating patient condition before neurosurgical procedures 14. Treatment Plan Selection: Choosing appropriate therapeutic interventions or management strategies 15. Postoperative Outcome Assessment: Evaluating surgical results and post-treatment changes"</p> <p>Classification Guidelines: - Apply your expert clinical judgment as a neuroradiology specialist - Consider the primary clinical intent behind each question - When multiple categories could apply, choose the most clinically relevant one - If none of the standard categories fit perfectly, suggest a new category</p> <p>Reply format: Category Number: [1-15 for existing categories, or 'NEW' for new category] Category Name: [Use existing name for 1-15, or suggest new descriptive name] Explanation: [Brief clinical rationale for classification]</p>

Figure 13. The prompt engineering is the process of designing and refining prompts (instructions) to guide generative AI models toward producing specific, high-quality outputs.