

BrainTRACE: Tracing Longitudinal, Multimodal, and Volumetric Evidence in Brain MRI Clinical Reasoning

Qizhen Lan^{1,*} Mengchen Fan^{2,*} Hang Zhang^{3,*} Jingwei Duan⁴
 Moule Lin⁵ Jialin Chen⁶ Baocheng Geng² Xiaoqian Jiang^{1,†}

¹University of Texas Health Science Center at Houston

²University of Alabama at Birmingham ³University of Pittsburgh

⁴The University of Texas MD Anderson Cancer Center

⁵University of Dublin, Trinity College ⁶Yale University

*Equal contribution. †Corresponding author: Xiaoqian Jiang.

Contact: qizhen.lan@uth.tmc.edu

Abstract

Brain MRI interpretation is a longitudinal clinical reasoning problem: radiologists compare serial studies, integrate information across MRI sequences, localize findings within volumetric anatomy, and translate this evidence into report-grounded assessments. Existing medical VQA and 3D imaging benchmarks capture important parts of this workflow, but often evaluate brain MRI through isolated images, static volumes, or ungrounded report-style answers, thereby obscuring failures in the evidence chain that support clinical validity. We introduce **BrainTRACE**, a report-grounded benchmark for evaluating whether vision-language models can trace the evidence structure required for longitudinal brain MRI interpretation. BrainTRACE contains 7,273 scored VQA instances derived from 1,778 longitudinal patients, 7,299 MRI studies, and approximately 29k co-registered 3D MRI sequence volumes. The benchmark is organized by five levels of clinical reasoning, from acquisition recognition to case-level synthesis, and by evidence demands covering longitudinal comparison, report-grounded references, multi-sequence integration, and volumetric spatial evidence. BrainTRACE supports rendered inputs compatible with standard VLM interfaces, a 3D-evidence condition, and a decomposed case-reasoning track that audits six steps in a longitudinal evidence chain. Evaluation of 20 VLM configurations shows that current systems can identify isolated visual cues but rarely compose them into grounded longitudinal interpretations. We release the benchmark specification, evaluation lists, scoring implementation, scoring rubrics, and audit-record format to support reproducible progress in brain MRI VLM evaluation. ¹

1 Introduction

Visual models trained end to end have achieved strong performance in object detection and 3D medical image segmentation [11, 25, 26]. General and medical VLMs extend image analysis to natural-language descriptions, question answering, and visual reasoning [4, 46, 49, 51, 56]. Disease interpretation on brain MRI, however, requires reasoning beyond an isolated image or sequence. It often depends on jointly evaluating lesion location, sequence-specific appearance, and longitudinal change, including where abnormalities are located, how they appear across different tissue contrasts,

¹Availability. The dataset release is available at <https://huggingface.co/datasets/lanqz7766/BrainTRACE>. Code and documentation are available at <https://github.com/lanqz7766/BrainTRACE>. Machine-readable RAI meta-data is shipped as a Croissant file alongside the dataset on HuggingFace.

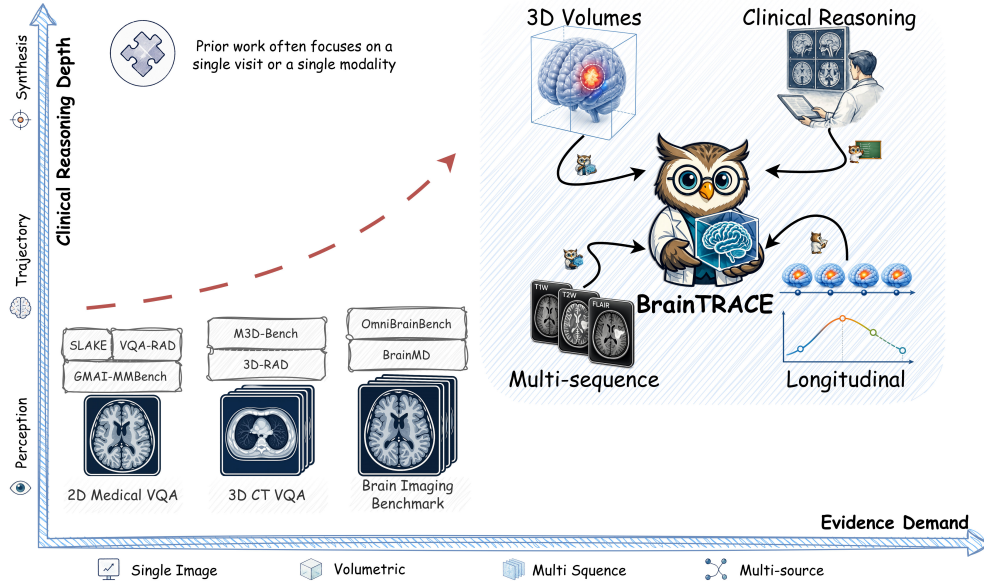


Figure 1: Positioning of BrainTRACE relative to representative medical VLM/VQA benchmarks. BrainTRACE makes longitudinal brain MRI follow-up the central evaluation unit, requiring models to integrate serial, multi-sequence, volumetric, and report-grounded clinical evidence.

and whether they are new, improved, stable, or progressive across serial studies. A trustworthy interpretation should follow an auditable evidence chain rather than merely describe the current disease state. This distinction is especially important for recent VLM studies, where fluent clinical prose can mask failures in true image-grounded reasoning. Studies of large reasoning models show that fluent outputs may reflect memorized patterns rather than input-grounded reasoning [47]. Thus, the key question is not whether a VLM can describe an isolated brain MRI, but whether it can reason from the evidence needed for clinically meaningful follow-up interpretation.

Existing benchmarks cover important parts of this problem, but they have challenges to evaluate the full follow-up reasoning process required in clinical settings. General medical VQA benchmarks make broad image question evaluation measurable [8, 27, 32]. Volumetric and temporal benchmarks extend evaluation toward 3D medical imaging, CT volume report modeling, or longitudinal CT reasoning [2, 13, 16], and brain imaging benchmarks address multi-stage clinical workflows [40], anomaly localization [6], 2D to 3D bridging [48], or cross-sectional brain MRI report generation [29]. However, real brain MRI interpretation requires joint reasoning over longitudinal change, multi-sequence findings, volumetric anatomy, and report-grounded clinical evidence (as shown in Figure 1). This leaves open a distinct evaluation question:

Evaluation Question: Can a VLM trace evidence through longitudinal brain MRI follow-up studies while leveraging multi-sequence MRI cues and volumetric spatial context to perform real clinical reasoning?

Answering this question requires a *report-grounded* benchmark that evaluates four linked demands, including *longitudinal* comparison across serial studies, *multi-sequence* MRI integration, *volumetric* localization and burden, and *clinical reasoning* beyond surface description.

To close this gap, we introduce **BrainTRACE**, a report-grounded brain MRI VQA benchmark containing 7,273 scored instances derived from 1,778 longitudinal patients, 7,299 MRI studies, and approximately 29k co-registered 3D MRI sequence volumes. BrainTRACE is designed around follow-up interpretation rather than generic image question answering. Each instance separates model-facing evidence from hidden clinical references so that models receive only the permitted visual evidence, question, and answer format, while radiology reports and segmentation-derived

Table 1: Positioning representative medical VLM/VQA benchmarks relative to BrainTRACE along four evidence attributes. *Longitudinal*: longitudinal or temporal visual evidence. *Multi-seq.*: item-level multi-sequence evidence. *Volumetric*: native-volume or explicit volumetric evidence. *Clinical*: reasoning depth beyond visual recognition. ● = full, ○ = partial, ◐ = absent.

Benchmark	Domain	Longitudinal	Multi-seq.	Volumetric	Clinical
VQA-RAD [27]	Mixed 2D radiology	◐	◐	◐	○
SLAKE [32]	Mixed 2D medical imaging	◐	◐	◐	○
Medical-Diff-VQA [19]	Paired chest X-rays	●	◐	◐	●
GMAI-MMBench [8]	Broad medical imaging	◐	○	○	●
M3D-Bench [2]	3D medical imaging	◐	◐	●	○
3D-RAD [13]	Longitudinal 3D CT	●	◐	●	●
BrainMD [48]	Brain MRI	◐	○	○	○
OmniBrainBench [40]	Broad brain imaging	◐	○	○	●
BrainTRACE	Longitudinal brain MRI	●	●	●	●

measurements define reference answers, required facts, and clinically material contradictions. This design allows the benchmark to evaluate whether a model’s answer is supported by available studies, not merely whether it is plausible.

BrainTRACE then localizes failures through two complementary annotations. The *clinical reasoning depth* axis follows the workflow of longitudinal brain MRI interpretation, from acquisition recognition and single study characterization to interval comparison, trajectory tracking, and case-level synthesis. The *evidence demand* axis records what support is required at each step: longitudinal comparison, multi-sequence integration, volumetric spatial evidence, or clinical reasoning. Together, this separation makes the benchmark more diagnostic so that errors can be interpreted as failures of the requested operation, failures to use the required evidence, or hallucinated facts unsupported by the record. Our contributions can be summarized as follows:

- We formalize brain MRI VLM evaluation as longitudinal evidence tracing, where clinical validity depends on comparing serial studies and grounding conclusions in the relevant imaging and clinical references.
- We instantiate this definition as BrainTRACE, a 7,273 instance benchmark with hidden report and measurement derived references, diagnostic two-axis annotation, three evaluation tracks, exact closed-form scoring, and factual audit scoring for open-ended and case-level answers.
- We evaluate 20 model configurations covering general, medical, frontier, and native volume VLMs, and identify a longitudinal evidence composition gap: the best broad compatible closed-form accuracy is 41.9%, but the strongest open-ended clinical pass is 19.8%, and the best case success rate is 1.4%.

2 Related Work

2.1 Medical VQA Benchmarks

Medical VQA benchmarks have expanded from selected 2D radiology images to broader, volumetric, temporal, and reasoning-oriented evaluation. VQA-RAD [27] and SLAKE [32] established radiology VQA on selected 2D medical images, while GMAI-MMBench [8] broadened evaluation across departments, modalities, and task types. Medical-Diff-VQA [19] tests paired chest-X-ray difference reasoning, and CXReasonBench [28] evaluates structured diagnostic reasoning in chest X-rays. M3D-Bench [2] evaluates 3D medical image understanding, CT-RATE [16] pairs CT volumes with radiology text, SGMRI-VQA [36] introduces frame indexed spatial grounding for volumetric MRI, and 3D-RAD [13] adds longitudinal temporal diagnosis on CT scans. These benchmarks cover important axes of medical VLM evaluation, but they do not define brain MRI evaluation as report-grounded longitudinal evidence tracing. Brain MRI follow-up is sequence, volume, and time dependent: a clinically valid answer may require comparing serial studies, integrating sequence-specific findings, localizing volumetric burden, and preserving that evidence through clinical synthesis.

2.2 Brain Imaging Benchmarks and Resources

Brain imaging benchmarks address brain data more directly, but they still do not make longitudinal brain MRI follow-up the central unit of report-grounded VLM evaluation. OmniBrainBench [40] evaluates multimodal brain imaging comprehension clinical tasks, largely through 2D visual evidence. NOVA [6] stress tests out-of-distribution anomaly localization across rare brain MRI pathologies. BrainMD [48] bridges 2D VLMs and 3D brain MRI through slice selection, while AutoRG-Brain/RadGenome-Brain [29] grounds cross-sectional brain MRI report generation. Complementary resources provide important but different signals: BraTS [34] established a multi-contrast tumor segmentation, OASIS-3 [24] provides longitudinal neuroimaging cohorts, and longitudinal MRI methods such as Reuter et al. [44] motivates registration-aware serial comparison.

BrainTRACE differs by making longitudinal brain MRI follow-up the central unit of VLM evaluation. It jointly tests report-grounded comparison across longitudinal studies, multi-sequence MRI integration, volumetric localization, burden assessment, and clinical reasoning beyond surface description. Table 1 summarizes benchmark-level coverage of these evidence attributes. Broader datasets and non-benchmark resources are included in Appendix A.

3 BrainTRACE Design and Construction

3.1 From clinical records to evaluation instances

BrainTRACE treats each VQA item as a report-grounded clinical evidence test rather than a generic image question pair. Each item is derived from a *longitudinal brain MRI record* containing serial studies, available MRI sequences, radiology reports, and segmentation-derived measurements. A key design is to decouple *what the model is allowed to observe* from *what defines clinical correctness*. Concretely, each released instance is organized into three components:

- **Model-facing input:** the permitted visual evidence, the question, the required answer format, and allowed non-leaking metadata.
- **Hidden reference:** the report or measurement derived answer target, required facts, accepted variants, and clinical material contradictions used for scoring.
- **Diagnostic labels:** the clinical reasoning depth, evidence demand tag, and evaluation track used to localize failures.

For closed-form questions, the hidden reference is a normalized target answer. For open-ended and case-level questions, it further specifies the facts that must be present and the contradictions that would make an answer clinically unacceptable. Reports and measurements determine what is clinically supported, while models receive only the permitted visual evidence.

3.2 Tasks, evidence demands, and evaluation tracks

Each BrainTRACE item is annotated along three diagnostic dimensions: what clinical operation is being tested, what evidence is required, and under which input condition the item is evaluated. This separation is central to the benchmark: a failure may reflect the wrong clinical operation, missing evidence use, or an unsupported clinical synthesis.

Clinical reasoning depth. We first instantiate the clinical reasoning depth axis as a five-level task taxonomy. L1 covers acquisition and basic detection. L2 covers single study characterization, including burden, location, and appearance. L3 tests interval comparison and disease-specific effects. L4 tests trajectory, timing, and treatment response. L5 covers case-level synthesis. This taxonomy makes failures clinically localizable: an answer can fail because the model missed the abnormality, compared the wrong timepoints, misread burden evolution, or produced a synthesis unsupported by the available evidence. Appendix B.4 defines the clinical reasoning depth and example question families. Table 6 in Appendix gives the per-track denominators and task family inventory.

Evidence demand. The evidence demands tag records what support is required for a valid answer. Longitudinal questions require comparison across time points from the same patient. Multi-sequence

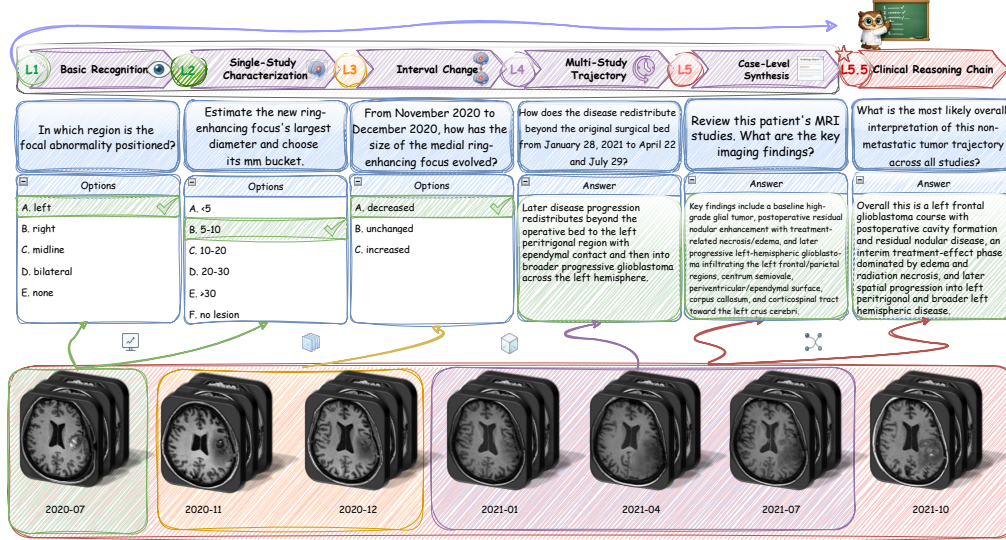


Figure 2: A running longitudinal case illustrates how BrainTRACE turns one patient record into localization, interval comparison, trajectory, and case synthesis tasks. The case reasoning track further decomposes the longitudinal interpretation into six report-grounded VQA steps.

questions require complementary MRI sequences within a study, such as FLAIR abnormality, enhancement, susceptibility, edema, or mass effect. Volumetric questions require spatial evidence beyond a single slice, either through multi-slice rendering or supporting native-volume input. Clinical reasoning questions require interpretation beyond surface recognition, such as treatment response, disease status, disease burden, or differential diagnosis. These tags are orthogonal to the operation levels: the same operation may require different forms of evidence, and the same evidence demand may appear across multiple levels. Table 6 in Appendix lists the demand tags for each task family.

Evaluation tracks. BrainTRACE instantiates these annotations through three evaluation tracks. The broad compatible track contains 5,653 instances and renders evidence into standard image inputs, allowing both open-weight and closed-source VLMs to be evaluated under the same interface. The 3D-evidence track contains 1,200 items and evaluates volumetric clinical targets using rendered multi-slice evidence or supported native-volume inputs. The case reasoning track contains 70 longitudinal cases and 420 decomposed VQA steps. It makes the longitudinal evidence chain explicit (as shown in Figure 2) by decomposing each case into six report-grounded VQA steps: anchor localization, early interval change, interval characterization, burden trajectory, latest status, and final synthesis. This decomposition turns an opaque case-level failure into a localized failure along the evidence path.

3.3 Report-grounded benchmark construction

BrainTRACE is derived from the publicly released MR-RATE corpus [12]. We retain patients with sufficient longitudinal context for evidence chain questions: each retained patient has at least three MRI studies with paired imaging, radiology reports, and segmentation-derived anatomical measurements. Each retained study is centered on a T1-weighted acquisition and may include co-registered T2-weighted, FLAIR, SWI, or other clinically relevant sequences. The benchmark targets longitudinal brain MRI follow-up studies from MR-RATE, rather than non-longitudinal referrals or the full neuroradiology spectrum. Appendix B.2 reports the upstream corpus scale and retained cohort composition.

The clinical references come from authentic MR-RATE radiology reports rather than synthetic report text. In the MR-RATE release, reports are anonymized and structured into clinical information, technique, findings, and impression fields through an iterative LLM-assisted construction procedure.

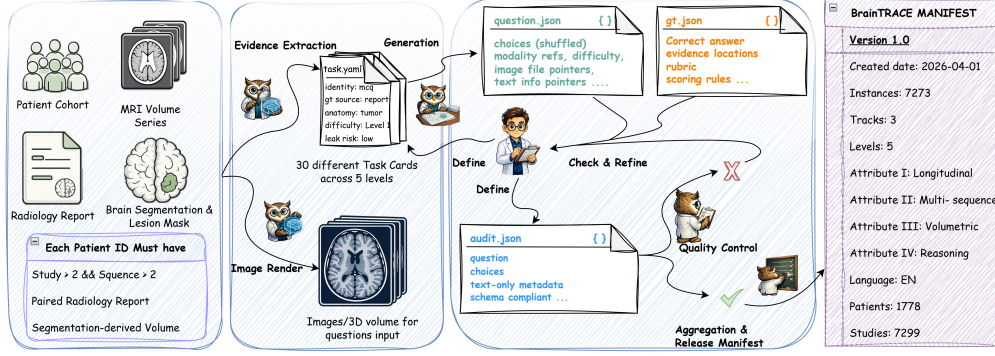


Figure 3: BrainTRACE constructs report-grounded VQA instances through a human-in-the-loop pipeline that separates hidden clinical references from model-facing visual evidence. Reports and segmentation-derived measurements define answer targets, required facts, contradictions, and scoring rubrics. Models receive only the permitted visual inputs.

BrainTRACE uses these structured reports as source clinical documents, then links report-supported findings to available imaging studies and segmentation-derived measurements.

Figure 3 summarizes the human-in-the-loop construction pipeline. The process first identifies report-supported evidence statements, including findings, anatomical locations, interval relations across studies, disease burden statements, disease status language, and quantitative measurements. Template eligibility rules then decide which statements can support a task: for example, an interval change question requires a finding described at comparable timepoints, whereas a measurement question requires a segmentation-derived value aligned to the relevant study. Eligible evidence is converted into a task card containing the question, answer format, hidden ground truth target or rubric, diagnostic labels, and the rendered visual input under the specified evidence condition. Human reviewers inspect the task cards and intermediate outputs throughout this process, checking report support, timepoint alignment, visual evidence availability, answer format validity, redaction, and scoring consistency. Appendix Section B.1 describes the model-facing and non-model-facing records used to reconstruct and audit these steps.

3.4 Quality Control

Ground truth control. BrainTRACE stores a non-model-facing hidden reference for each item, including the target answer, required supporting facts, clinically material contradictions, and scoring criteria. Closed-form items use a lightly normalized target answer. Open-ended items use a rubric of required facts, optional facts, and record contradicting statements. For case reasoning items, each decomposed VQA step has its own reference, so the evaluation can identify where the evidence path breaks. The case reasoning track receives item level review by an American Board of Radiology (ABR) certified radiation oncologist, including the question, ground truth, and audit rubric for each decomposed step.

Leakage control. BrainTRACE controls leakage by separating inputs from hidden clinical records. During evaluation, models receive only the rendered visual evidence, the question, the answer format, and allowed non-leaking metadata. They do not receive radiology reports, parsed report fields, segmentation masks, measurement tables, or hidden reference fields. Metadata is redacted when it could reveal the answer. Each instance passes quality control checks for redaction, report support, visual evidence availability, and scoring consistency. We also run text-only leakage probes to test whether the question wording alone is sufficient to infer the answer without visual evidence.

Table 2: Closed-form accuracy (%) is fragmented across clinical subclasses. Subclass columns pool broad-compatible and 3D-evidence items. Overall columns separate tracks.

Model	L1 recognition		L2 single-study			L3 interval			L4 trajectory			Overall	
	ACQ	DET	BUR	LOC	APP	QNT	EVT	EFF	TRJ	TMP	RSP	Broad	3D
<i>Open-weight general VLMs</i>													
Qwen2.5-VL-7B [4]	47.5	38.0	7.1	44.9	27.8	26.8	<u>53.7</u>	26.2	25.6	20.3	38.5	33.2	33.2
Qwen3-VL-4B [3]	47.2	41.3	8.7	44.6	31.4	<u>36.9</u>	52.6	28.0	13.9	16.8	27.3	32.7	35.4
Qwen3-VL-8B [3]	49.8	42.0	12.0	37.0	32.4	<u>33.7</u>	51.1	30.9	25.4	17.2	25.9	33.7	33.4
Qwen3-VL-30B [3]	64.5	<u>44.4</u>	<u>23.2</u>	33.5	30.4	37.2	52.6	28.6	25.4	15.8	10.5	35.1	34.8
Qwen3.5-27B [43]	81.5	38.9	13.9	28.0	30.2	23.8	35.8	43.4	24.5	14.0	18.9	37.3	12.6
InternVL3-8B [56]	63.0	40.7	6.0	48.2	38.6	34.8	49.1	29.7	24.9	15.2	25.2	35.8	35.3
InternVL3-14B [56]	56.8	40.2	7.7	42.9	37.4	31.0	53.1	27.7	29.3	18.6	16.1	35.1	34.0
InternVL3-78B [56]	60.5	40.2	10.7	43.2	35.0	25.9	53.6	30.3	24.5	19.4	35.7	35.5	33.3
Janus-Pro-7B [9]	1.0	34.2	1.1	3.2	0.6	3.8	44.8	0.3	7.5	15.1	4.2	10.9	17.6
<i>Open-weight medical VLMs</i>													
HuatuoGPT-Vision-7B [7]	76.2	38.7	7.1	44.3	31.6	25.6	47.9	25.9	26.8	18.6	34.3	36.3	29.1
Lingshu-7B [52]	78.8	37.8	7.9	<u>46.3</u>	27.6	30.4	54.2	29.7	26.2	19.5	30.1	36.5	35.4
MedGemma-4B-it [45]	42.2	35.5	2.7	20.2	28.2	7.5	47.8	25.4	15.2	14.5	0.7	24.9	18.9
LLaVA-Med-v1.5 [30]	0.0	2.9	0.0	0.2	0.4	0.0	0.7	0.0	0.0	0.0	0.0	0.6	0.0
<i>Closed-source frontier VLMs</i>													
GPT-5.4 [37]	<u>87.5</u>	42.2	15.3	37.9	40.6	35.1	50.2	<u>49.6</u>	<u>32.0</u>	<u>23.1</u>	<u>44.1</u>	<u>40.4</u>	42.2
GPT-5 [37]	88.8	40.5	10.7	40.5	<u>39.0</u>	31.5	51.9	58.3	33.9	21.7	49.7	41.9	<u>38.3</u>
GPT-5-mini [37]	84.2	42.4	25.2	30.7	34.4	28.0	50.4	49.3	27.7	19.0	42.7	38.2	37.5
Claude Opus 4.6 [1]	79.2	14.2	8.5	30.9	22.8	12.0	9.2	33.5	20.6	20.4	18.9	29.8	1.2
Gemini 2.5 Pro [14]	85.2	49.3	19.5	21.7	29.8	29.6	48.3	34.4	26.6	24.0	14.7	36.5	32.0
<i>Native-volume medical baselines</i>													
M3D-LaMed [2]	–	36.7	18.5	40.4	30.4	24.0	51.7	14.3	18.4	17.1	26.1	–	29.9
RadFM [49]	33.8	28.2	8.2	9.9	16.4	16.0	28.3	5.5	12.3	16.5	2.1	18.1	15.0
<i>n</i>	400	550	366	657	500	575	601	344	482	558	143	4,053	1,123

ACQ=acquisition, DET=detection/localization, BUR=burden, LOC=location, APP=appearance, QNT=quantitative change, EVT=lesion event, EFF=structural effect, TRJ=trajectory, TMP=timing, RSP=response. Best values are bold. Second-best values are underlined. Only the Overall columns use background shading.

4 Experiments

4.1 Experimental Setup

We evaluate 20 model configurations spanning open-weight general VLMs, open-weight medical VLMs, closed-source frontier VLMs, and available native-volume medical baselines (Table 2). All models are evaluated zero-shot using only the permitted model-facing visual evidence, question, answer format, and allowed non-leaking metadata. Reports, segmentation masks, parsed findings, measurements, and hidden reference fields are never shown. Each model receives one attempt per item. Malformed outputs, refusals, and abstentions remain in the denominator. For 3D evidence items, general VLMs use released rendered inputs, while native-volume medical baselines use their supported volume input interfaces. Thus, 3D evidence results compare supported interfaces on shared clinical targets rather than isolating native-volume representation as a controlled variable.

4.2 Evaluation Metrics

Closed-form accuracy. Closed-form questions are scored by exact match after light answer normalization and uniform removal of visible inline reasoning blocks. We report subclass accuracy by clinical reasoning depth and overall accuracy separately for the broad compatible and 3D evidence tracks.

Open-ended and case-level clinical pass. Open-ended answers and decomposed case-level steps are scored by a fixed factual clinical audit. The audit judge receives the question, the model’s final answer, and the item-level rubric, including the reference answer, required clinical facts, critical slots,

Table 3: Open-ended acceptability and case-reasoning remain low even for frontier models; entries are Pass %.

Model	Open subclasses				Case synthesis			Overall		Case reasoning	
	L1-ABN	L2-APP	L3-EFF	L4-TRJ	IMP	CMP	DIF	Broad	3D	Step	Case
👨 ABR diplomate oncologist	100	90	80	100	100	100	100	92.9	94.3	91.7	80
<i>Open-weight general VLMs</i>											
🌀 Qwen2.5-VL-7B [4]	12.0	8.5	7.6	1.4	2.3	2.8	2.3	4.9	3.9	9.3	1.4
🌀 Qwen3-VL-4B [3]	15.0	15.5	17.5	5.2	9.4	10.8	9.4	11.8	5.2	8.3	0.0
🌀 Qwen3-VL-8B [3]	11.5	18.0	<u>21.3</u>	6.6	11.7	24.9	2.8	13.8	9.1	3.8	0.0
🌀 Qwen3-VL-30B [3]	14.0	9.5	10.9	6.2	<u>12.7</u>	32.9	15.0	<u>14.4</u>	11.7	<u>11.7</u>	0.0
🌀 Qwen3.5-27B [43]	17.5	7.5	10.0	0.9	7.5	9.4	3.3	8.1	3.9	2.9	0.0
🌀 InternVL3-8B [56]	17.0	10.5	3.5	3.0	3.8	6.0	5.0	6.6	–	8.3	0.0
🌀 InternVL3-14B [56]	9.0	14.0	11.0	1.0	8.3	9.5	7.5	8.6	–	11.2	0.0
🌀 InternVL3-78B [56]	11.0	18.5	7.5	2.5	4.8	6.0	5.5	7.6	–	0.0	0.0
🌀 Janus-Pro-7B [9]	3.5	7.5	0.9	0.0	0.0	0.0	2.3	1.8	0.0	0.5	0.0
<i>Open-weight medical VLMs</i>											
🏥 HuatuoGPT-Vision-7B [7]	<u>23.5</u>	15.5	4.7	<u>7.6</u>	1.9	7.5	<u>14.1</u>	9.9	0.0	0.2	0.0
🏥 Lingshu-7B [52]	20.5	10.5	8.5	1.9	3.1	1.4	5.6	6.8	3.9	6.9	0.0
🏥 MedGemma-4B-it [45]	15.5	17.0	4.3	0.0	0.0	0.5	6.6	5.4	3.9	7.6	0.0
🏥 LLaVA-Med-v1.5 [30]	4.0	<u>19.0</u>	0.9	0.5	0.5	2.8	1.9	3.8	1.3	0.2	0.0
<i>Closed-source frontier VLMs</i>											
🌀 GPT-5.4 [37]	24.5	30.0	26.1	17.1	14.6	17.8	13.6	19.8	15.6	<u>24.0</u>	0.0
🌀 GPT-5 [37]	15.0	10.0	19.5	0.5	12.1	12.6	2.5	10.5	11.7	21.7	1.4
🌀 GPT-5-mini [37]	16.0	6.0	12.5	3.5	11.6	22.1	1.5	10.6	10.4	24.5	0.0
🌟 Claude Opus 4.6 [1]	22.0	6.0	12.5	1.0	9.8	15.1	1.0	9.6	6.5	23.1	0.0
🌟 Gemini 2.5 Pro [14]	20.5	15.0	16.6	2.4	10.1	<u>28.6</u>	10.3	14.2	<u>13.0</u>	10.7	0.0
<i>Native-volume medical baselines</i>											
🏥 M3D-LaMed [2]	–	–	0.0	0.0	0.0	0.0	0.0	–	0.0	4.8	0.0
🏥 RadFM [49]	6.0	13.5	4.3	1.4	1.4	4.7	1.4	4.3	2.6	0.0	0.0
<i>n</i>	200	200	211	211	426	213	213	1,597	77	420	70

APP=appearance description, EFF=interval effect, TRJ=trajectory summary, IMP=impression, CMP=comparative summary, DIF=differential diagnosis. Step denotes Step Pass over decomposed VQA steps; Case denotes Case Success, requiring all six steps in a case to pass. Broad denotes the broad-compatible track. The physician row is a separate 10-case image-grounded MCQ sanity check, not a directly comparable model baseline.

accepted paraphrases, and auto fail conditions. It does not receive the source report, segmentation tables, evidence quotes, or hidden reference record.

The audit measures factual clinical acceptability rather than prose quality. Let F_i , S_i , E_i , and U_i denote required fact coverage, critical slot coverage, absence of critical error, and absence of a clinically material unsupported claim. The item-level pass indicator is $P_i = F_i S_i E_i U_i$. For the case reasoning track, Step Pass averages P_i over decomposed steps, while Case Success requires all six steps in the same longitudinal case to pass: $\text{CaseSuccess}_c = \prod_{s=1}^6 P_{c,s}$. Appendix C.1 details the scoring fields, derived metrics, edge case policy, and released audit records. Appendix C.5 and Appendix C.6 report agreement and judge family sensitivity analyses.

4.3 Experimental Results

Closed-form accuracy is fragmented across clinical skills. Table 2 reports closed-form accuracy across clinical subclasses. The strongest frontier models cluster near 40% broad compatible accuracy: GPT-5 obtains the best broad score at 41.9%, and GPT-5.4 obtains the best 3D evidence score at 42.2%. However, the aggregate score hides large variation across acquisition recognition, burden, localization, interval change, trajectory, timing, and response assessment. Closed-form accuracy indicates partial success on isolated evidence tests, not reliable use of disease burden, temporal change, or treatment response evidence.

Open-ended clinical acceptability remains much lower than recognition accuracy. Table 3 shows that factual clinical pass is substantially harder than closed-form recognition. GPT-5.4 is strongest on this axis, but passes only 19.8% of broad open-ended questions and 15.6% of 3D

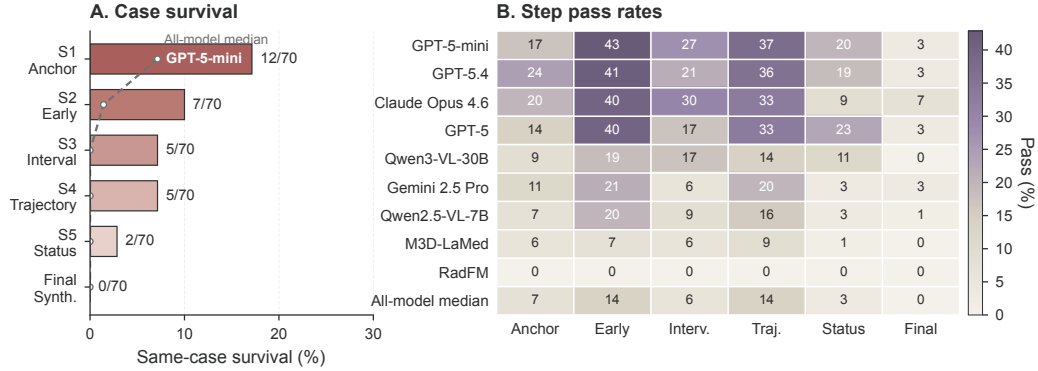


Figure 4: Case-level reasoning fails by attrition. Local step successes rarely survive through all six decomposed VQA steps in the same longitudinal case.

evidence open-ended questions. Failures are clinically material: models often mention a relevant abnormality but omit disease status, reverse interval direction, assign evidence to the wrong time point, or add unsupported anatomy. As a clinical answerability sanity check, one board-certified professional confirms that the selected cases are clinically answerable under expert reading, but it is not a directly comparable model baseline. Appendix C.2 reports the full open-ended audit leaderboard and secondary text metrics, while Appendix F gives paired success/failure examples that make the factual audit failure modes inspectable.

Case-level reasoning fails by evidence attrition. The case-reasoning results in Table 3 and Figure 4 exposes a stronger compositional failure. The best aggregate Step Pass is 24.5%, but this overstates the case-level ability because different steps often pass in different cases. Only two model rows complete any case, both at 1.4% Case Success. Models may identify an anchor abnormality or describe one interval, but these local successes rarely survive through the latest status interpretation and final synthesis. Local perception and interval reasoning are necessary but not sufficient for report-grounded longitudinal interpretation. Appendix C.3 provides the released per-step and per-case audit records, and Appendix C.4 further decomposes the attrition pattern in Figure 4.

Evidence-channel ablations confirm dependence on retained evidence. Table 4 reports matched diagnostic ablations for removing longitudinal evidence by using only the latest timepoint drops accuracy by 37.5 points for Qwen3.5-72B and 23.3 points for GPT-5.4. Restricting inputs to a single T2 sequence also reduces performance, and removing images entirely causes large drops for both models. The drops show that performance is not explained only by answer priors or question wording. It depends on retained longitudinal, multi-sequence, and visual evidence. Appendix E details the matched evidence channel ablations, and Appendix D reports a complementary 3D interface, rank sensitivity, and chance calibrated difficulty diagnostics.

Table 4: Evidence-channel ablations reduce accuracy on matched diagnostic items. All ablation rows use $n = 120$ with a constructed 50.0% full-evidence baseline, so drops are comparable across conditions, but conditional on the balanced sample.

Condition	Evidence retained			Accuracy (%)		Drop (pp)	
	Long.	Multi-seq.	Image	Qwen	GPT	Qwen	GPT
Full evidence	✓	✓	✓	50.0	50.0	—	—
Latest timepoint only	—	✓	✓	12.5	26.7	-37.5	-23.3
T2 sequence only	✓	—	✓	29.2	38.3	-20.8	-11.7
No image	—	—	—	15.0	9.2	-35.0	-40.8

5 Conclusion

We introduced **BrainTRACE**, a report-grounded benchmark for longitudinal brain MRI VLM evaluation. BrainTRACE asks whether models can trace evidence across serial studies, MRI sequences, volumetric anatomy, and report-derived clinical references, rather than answer isolated image questions. The main finding is not simply that current VLMs score low, but that their successes do not compose: local recognition rarely survives through longitudinal, report-grounded clinical synthesis. This gap persists under current 3D evidence and native-volume interfaces, suggesting that volumetric access alone is insufficient without stronger multi-sequence, temporal, and evidence-grounded reasoning. We release the benchmark specification, evaluation lists, scoring implementation, factual audit protocol, and audit record format to support reproducible evaluation of future brain MRI VLMs. Limitations and responsible-use considerations are summarized in Appendix G.

References

- [1] Anthropic. Claude opus 4.6. <https://www.anthropic.com/claude/opus>, 2026. Accessed: 2026-04.
- [2] Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.
- [3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.
- [4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL <https://arxiv.org/abs/2502.13923>.
- [5] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15016–15027, 2023.
- [6] Cosmin I Bercea, Jun Li, Philipp Raffler, Evamaria O Riedel, Lena Schmitzer, Angela Kurz, Felix Bitzer, Paula Roßmüller, Julian Canisius, Mirjam L Beyrle, et al. Nova: A benchmark for anomaly localization and clinical reasoning in brain mri. *arXiv preprint arXiv:2505.14064*, 2025.
- [7] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, et al. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 7346–7370, 2024.
- [8] Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyang Huang, Yanzhou Su, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37:94327–94427, 2024.
- [9] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.
- [10] Katarina Trojachanec Dineva, Stefan Andonov, Ilinka Ivanoska, Ivan Kitanovski, Sasho Gramatkov, Tamara Kostova, Monika Simjanoska Misheva, and Kostadin Mishev. Neurovlm-bench: Evaluation of vision-enabled large language models for clinical reasoning in neurological disorders. *arXiv preprint arXiv:2603.24846*, 2026.
- [11] Mengchen Fan, Baocheng Geng, Xi Xiao, Tianyang Wang, Siyuan Mei, Pulin Che, Xiaoqian Jiang, and Qizhen Lan. Detail consistent stage-wise distillation for efficient 3D MRI segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI*, 2026. URL <https://arxiv.org/abs/2605.26382>.
- [12] Forithmus, NVIDIA Medical, University of Zurich, and Istanbul Medipol University. MR-RATE: A large-scale brain and spine MRI dataset with paired radiology reports. HuggingFace datasets repository, March 2026. URL <https://huggingface.co/datasets/Forithmus/MR-RATE>. Accessed 2026-04; companion paper in preparation. Released under CC BY-NC-SA 4.0.
- [13] Xiaotang Gai, Jiayang Liu, Yichen Li, Zijie Meng, Jian Wu, and Zuozhu Liu. 3d-rad: A comprehensive 3d radiology med-vqa dataset with multi-temporal analysis and diverse diagnostic tasks. *arXiv preprint arXiv:2506.11147*, 2025.

- [14] Google DeepMind. Gemini 2.5 Pro model card. Google DeepMind model card, June 2025. URL <https://deepmind.google/models/model-cards/>. Updated 2025-06-27; accessed 2026-04-27.
- [15] Feng Guo, Jiayang Liu, Yang Li, Qianqian Shi, and Mingkun Xu. Mm-neuroonco: A multi-modal benchmark and instruction dataset for mri-based brain tumor diagnosis. *arXiv preprint arXiv:2602.22955*, 2026.
- [16] Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihhan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*, pages 1–19, 2026.
- [17] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.
- [18] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- [19] Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M Summers, and Yingying Zhu. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4156–4165, 2023.
- [20] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimed-vqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183, 2024.
- [21] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghighi, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.
- [22] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.
- [23] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.
- [24] Pamela J LaMontagne, Tammie LS Benzinger, John C Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G Vlassenko, et al. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *medrxiv*, pages 2019–12, 2019.
- [25] Qizhen Lan and Qing Tian. ACAM-KD: Adaptive and cooperative attention masking for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3957–3966, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Lan_ACAM-KD_Adaptive_and_Cooperative_Attention_Masking_for_Knowledge_Distillation_ICCV_2025_paper.html.
- [26] Qizhen Lan, Yu-Chun Hsu, Nida Saddaf Khan, and Xiaoqian Jiang. ReCo-KD: Region- and context-aware knowledge distillation for efficient 3D medical image segmentation, 2026. URL <https://arxiv.org/abs/2601.08301>.
- [27] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1): 180251, 2018.

- [28] Hyungyung Lee, Geon Choi, Jung-Oh Lee, Hangyul Yoon, Hyuk Gi Hong, and Edward Choi. Cxreasonbench: A benchmark for evaluating structured diagnostic reasoning in chest x-rays, 2025. URL <https://arxiv.org/abs/2505.18087>.
- [29] Jiayu Lei, Xiaoman Zhang, Chaoyi Wu, Lisong Dai, Ya Zhang, Yanyong Zhang, Yanfeng Wang, Weidi Xie, and Yuehua Li. Autorg-brain: Grounded report generation for brain mri. *arXiv preprint arXiv:2407.16684*, 2024.
- [30] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564, 2023.
- [31] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- [32] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE, 2021.
- [33] Răzvan V Marinescu, Neil P Oxtoby, Alexandra L Young, Esther E Bron, Arthur W Toga, Michael W Weiner, Frederik Barkhof, Nick C Fox, Polina Golland, Stefan Klein, et al. Tadpole challenge: Accurate alzheimer’s disease prediction through crowdsourced forecasting of future data. In *International workshop on predictive intelligence in medicine*, pages 1–10. Springer, 2019.
- [34] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.
- [35] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine learning for health (ML4H)*, pages 353–367. PMLR, 2023.
- [36] Lama Moukheiber, Caleb M Yeung, Haotian Xue, Alec Helbling, Zelin Zhao, and Yongxin Chen. Beyond a single frame: Multi-frame spatially grounded reasoning across volumetric mri. *arXiv preprint arXiv:2604.15808*, 2026.
- [37] OpenAI. OpenAI GPT-5 Family model documentation. OpenAI API documentation, 2026. URL <https://platform.openai.com/docs/models>. Accessed 2026-04.
- [38] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, 2022.
- [39] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002. doi: 10.3115/1073083.1073135.
- [40] Zhihao Peng, Cheng Wang, Shengyuan Liu, Zhiying Liang, Zanting Ye, Min Jie Ju, Peter Y. M. Woo, and Yixuan Yuan. OmniBrainBench: A comprehensive multimodal benchmark for brain imaging analysis across multi-stage clinical tasks. *arXiv preprint arXiv:2511.00846*, 2025. doi: 10.48550/arXiv.2511.00846.
- [41] Fernando Pérez-García, Rachel Sparks, and Sébastien Ourselin. Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer methods and programs in biomedicine*, 208:106236, 2021.
- [42] Maja Popović. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, 2017.

- [43] Qwen Team. Qwen3.5: Towards native multimodal agents. Qwen blog, February 2026. URL <https://qwen.ai/blog?id=qwen3.5>. Official model citation from the Qwen3.5 model card.
- [44] Martin Reuter, Nicholas J. Schmansky, H. Diana Rosas, and Bruce Fischl. Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, 61(4):1402–1418, 2012. doi: 10.1016/j.neuroimage.2012.02.084.
- [45] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.
- [46] Pan Wang, Yihao Hu, Xiujiu Liu, Jingchu Yang, Hang Wang, and Zhihao Wen. AtlasVA: Self-evolving visual skill memory for teacher-free VLM agents, 2026. URL <https://arxiv.org/abs/2605.17933>.
- [47] Yuhui Wang, Changjiang Li, Guangke Chen, Jiacheng Liang, and Ting Wang. Reasoning or retrieval? a study of answer attribution on large reasoning models. *arXiv preprint arXiv:2509.24156*, 2025.
- [48] Yuli Wang, Jian Peng, Yuwei Dai, Craig Jones, Haris Sair, Jinglai Shen, Nicolas Loizou, Jing Wu, Wen-Chi Hsu, Maliha Imami, et al. Enhancing vision-language models for medical imaging: bridging the 3d gap with innovative slice selection. *Advances in Neural Information Processing Systems*, 37:99947–99964, 2024.
- [49] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866, 2025.
- [50] Joy T Wu, Nkechinyere N Agu, Ismini Lourentzou, Arjun Sharma, Joseph A Paguio, Jasper S Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*, 2021.
- [51] Xi Xiao, Chen Liu, Chih-Ting Liao, Yunbei Zhang, Qizhen Lan, Yuxiang Wei, Lin Zhao, Janet Wang, Jianyang Gu, Muchao Ye, Tianyang Wang, and Hao Xu. Staying VIGILant: Mitigating visual laziness via counterfactual visual alignment in MLLMs. In *European Conference on Computer Vision*. Springer, 2026. URL <https://arxiv.org/abs/2606.26387>.
- [52] 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 preprint arXiv:2506.07044*, 2025.
- [53] Zhijian Yang, Noel DSouza, Istvan Megyeri, Xiaojian Xu, Amin Honarmandi Shandiz, Farzin Haddadpour, Krisztian Koos, Laszlo Rusko, Emanuele Valeriano, Bharadwaj Swaminathan, et al. Decipher-mr: a vision-language foundation model for 3d mri representations. *npj Digital Medicine*, 2026.
- [54] Hang Zhang, Ruheng Wang, Yuelu Ji, Mingu Kwak, Xizhi Wu, Chenyu Li, Li Zhang, Wenqi Shi, Yifan Peng, and Yanshan Wang. Scaling medical reasoning verification via tool-integrated reinforcement learning. *arXiv preprint arXiv:2601.20221*, 2026.
- [55] Xiaoman Zhang, Hong-Yu Zhou, Xiaoli Yang, Oishi Banerjee, Julián N Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. Rexrank: A public leaderboard for ai-powered radiology report generation. *arXiv preprint arXiv:2411.15122*, 2024.
- [56] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternV13: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Benchmark positioning and scope

This appendix chapter provides the evidence-attribute positioning behind Section 2. The purpose is not to rank prior resources, many of which address important medical VQA, temporal imaging, 3D imaging, or brain-MRI tasks. Instead, the comparison asks a narrow benchmark-design question: whether a resource jointly evaluates report-grounded longitudinal brain-MRI reasoning across longitudinal, multi-sequence, volumetric, and clinically interpretive evidence.

Table 5 extends the main-text positioning table (Table 1) to additional benchmarks, datasets, cohorts, and imaging resources considered during benchmark design. The columns are evidence attributes rather than quality judgments: **L** denotes longitudinal or temporal evidence, **M** item-level multi-sequence evidence, **V** volumetric evidence, and **C** clinical reasoning beyond surface recognition. Scale is reported in each source’s native unit, such as images, studies, volumes, questions, or evaluated models; it should not be read as a normalized dataset-size comparison.

Table 5: Existing resources cover important parts of medical VQA, temporal imaging, 3D imaging, and brain-MRI evaluation, but do not simultaneously instantiate the four evidence attributes targeted by BrainTRACE. Scale is reported in each source’s native unit for provenance and is not a normalized dataset-size comparison.  = full,  = partial,  = absent.

Benchmark / resource	Year	Scale	L	M	V	C	Primary task
<i>2D medical VQA / report benchmarks</i>							
VQA-RAD [27]	2018	315 imgs / 3.5k QA					2D VQA
PathVQA [17]	2020	5k imgs / 33k QA					Pathology VQA
SLAKE [32]	2021	642 imgs / 14k QA					Bilingual VQA
MIMIC-CXR [23]	2019	65k pts / 227k studies					CXR report gen
CheXpert [21]	2019	65k pts / 224k imgs					CXR class.
ReXrank / ReXPref-Prior [55]	2024	10k studies					CXR report leaderboard
Medical-Diff-VQA [19]	2023	164k pairs / 700k QA					CXR pair VQA
MS-CXR-T [5]	2023	1.3k pairs + 361 sents					Temporal CXR class.
Chest ImaGenome [50]	2021	242k scene graphs					CXR progression
OmniMedVQA [20]	2024	118k imgs / 128k QA					Multi-domain VQA
Med-Flamingo eval [35]	2023	~1k USMLE items					Few-shot VQA
<i>Text-only MCQA (radiology-adjacent)</i>							
MedQA [22]	2020	12.7k Qs					USMLE MCQA
MedMCQA [38]	2022	194k Qs					MCQA (Indian PG)
MMLU (clinical) [18]	2021	~1.8k Qs					MCQA
<i>3D and multi-sequence medical imaging benchmarks</i>							
CT-RATE [16]	2024	21k pts / 25.7k vols					CT report gen
M3D-Bench [2]	2024	120k pairs / 662k instr					3D CT VQA
RadFM / RadBench [49]	2023–25	13M 2D + 615k 3D scans					Web 2D+3D VQA
3D-RAD [13]	2025	136k train / 170k QA					Long. 3D CT VQA
Decipher-MR [53]	2025	22k studies / 200k series					3D MRI Dx/retrieval
<i>Brain-MRI specialised</i>							
RadGenome-Brain / AutoRG-Brain [29]	2024	1,007 cases / 3,408 pairs					Grounded brain report gen
BrainMD / BrainBench [48]	2024	2,453 scans					Slice-selected VQA
MM-NeuroOnco [15]	2026	24.7k slices / 200k instr					2D tumor instr.-tune
OmniBrainBench [40]	2025	9,527 VQA / 31.7k imgs					Tri-plane 2D clinical VQA
NeuroVLM-Bench [10]	2026	20 MLLMs evaluated					2D clinical reasoning
BraTS (2012–24) [34]	2012–24	~4,500 cases (2024)					Tumor segmentation
IXI [41]	~2006	~600 subjects					Analysis / synthesis
<i>Longitudinal biomarker challenges</i>							
OASIS-3 [24]	2019	1,378 pts / 2,842 sess					AD progression
ADNI / TADPOLE [33]	2017–19	1,677 pts					Biomarker forecast
BrainTRACE (ours)	2026	1,778 pts / 7,299 studies					5-level / 4-property suite

B Dataset construction and task coverage

This appendix chapter documents how the retained MR-RATE follow-up cohort is converted into auditable evaluation instances. The emphasis is on provenance, release boundaries, leakage control,

and task coverage: which records are model-facing, which records define hidden report-grounded references, and which clinical reasoning depth and evidence attributes each task family is intended to test.

B.1 Construction and audit records

The construction record has two purposes: it defines what an evaluated model may see, and it preserves the evidence needed to audit the reference after scoring. For item i , we write the model-facing record as

$$X_i = \{q_i, a_i, m_i, v_i\},$$

where q_i is the rendered question, a_i is the answer format and any visible options, m_i is the permitted metadata, and v_i is the rendered visual evidence. The non-model-facing reference record is

$$R_i = \{y_i, r_i, e_i, c_i\},$$

where y_i is the target answer or reference response, r_i is the rubric, e_i is the report- or measurement-supported evidence, and c_i records exclusion, redaction, and scoring checks. Only X_i is shown during model evaluation. R_i is retained for reconstruction, leakage review, factual audit, and reproducibility.

1. Clinical eligibility and task specification. Each task family first fixes the clinical reasoning depth to be tested, the required evidence attributes, the answer format, and the allowed metadata. Eligibility rules then determine whether a retained patient and study sequence can support that question. For example, an interval-change item requires comparable findings across timepoints, whereas a volumetric item requires a segmentation-derived measurement linked to the relevant study.

2. Reference formation. Eligible report statements and segmentation-derived measurements are converted into a hidden reference. Closed-form items store a normalized target answer. Open-ended items store required facts, critical rubric slots, accepted paraphrases, and clinically material contradictions. Case-reasoning items store the same information separately for each of the six decomposed steps, so a failure can be assigned to a specific part of the longitudinal reasoning sequence rather than only to the final case answer.

3. Model-facing instance construction. The released question record contains the text seen by the model, the answer format, any answer options, the permitted demographic or timepoint metadata, and the visual-evidence specification. It does not include the source report, evidence quotations, segmentation masks, measurement tables, or the hidden reference. Broad-compatible models receive rendered views; 3D-evidence items additionally preserve the volume pathway needed by native-volume baselines.

4. Leakage and evidence-availability review. Before release, each item is checked for redaction, report support, visual-evidence availability, and consistency between the visible question and the hidden reference. Closed-form items also receive a text-only leakage probe using a stripped version of the question and metadata. Items that fail rendering or reference checks are listed in the exclusion records rather than removed without trace.

5. Scoring and reproducibility. During evaluation, model outputs are joined back to the hidden reference only at scoring time. Closed-form answers are scored by deterministic normalization and exact match. Open-ended and case-level answers are graded against the factual-audit rubric described in Appendix C.1. The companion repository contains the corresponding scoring specification, rendering instructions, evaluation lists, exclusion records, and scoring implementation: <https://github.com/lanqz7766/BrainTRACE>. These materials define the released denominator and allow an external user to reconstruct the benchmark from the same MR-RATE access conditions.

B.2 Upstream MR-RATE source and retained follow-up cohort

BrainTRACE is constructed from MR-RATE [12], a public MRI corpus with paired clinical radiology reports. The upstream corpus contains 83,425 patients, 98,334 MRI studies, and 705,254 MRI series or metadata rows. BrainTRACE retains the longitudinal subset needed for follow-up reasoning: 1,778 patients, 7,299 MRI studies, and 65,974 series-level metadata rows, with at least three MRI studies per

retained patient. This source design makes provenance auditable and keeps the references grounded in authentic, anonymized radiology reports rather than generated report text. It does not imply population representativeness for all neuroradiology indications. BrainTRACE does not redistribute raw imaging or model-facing reports; users obtain upstream MR-RATE access under its terms, and evaluated models receive only the visual evidence permitted by each task.

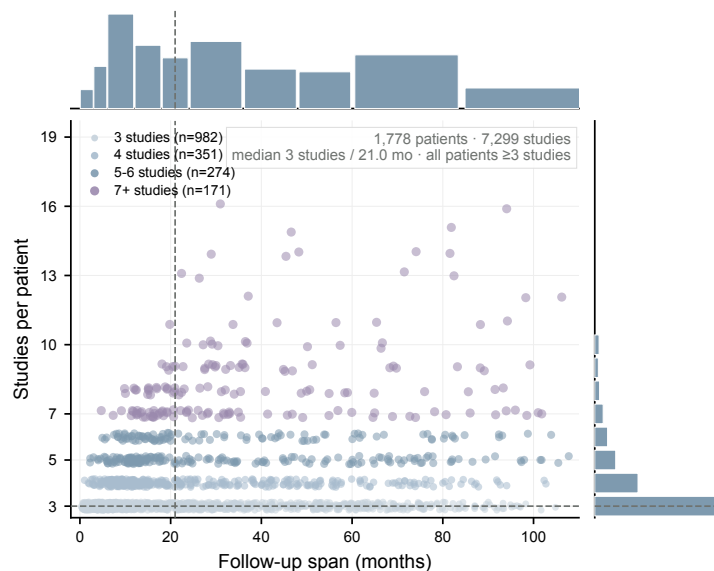


Figure 5: The retained MR-RATE subset provides the longitudinal follow-up structure required by BrainTRACE. Every retained patient has at least three MRI studies, allowing report-grounded questions to be tied to serial imaging rather than only single-visit findings.

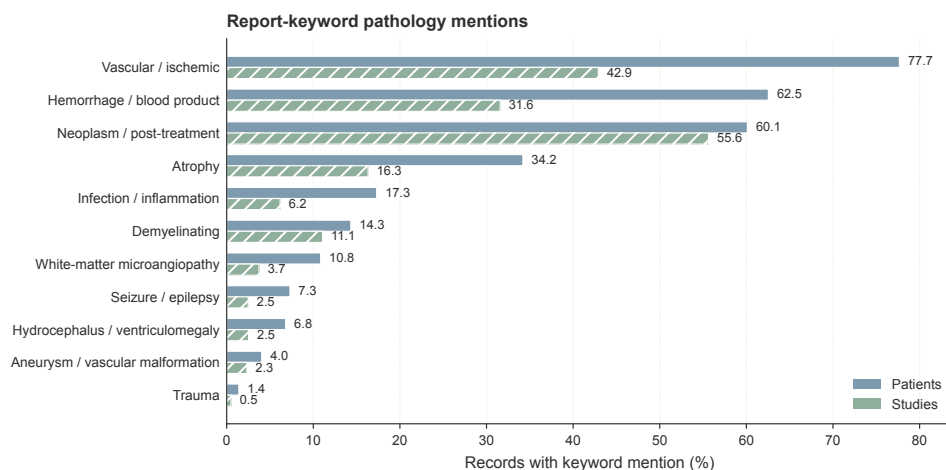


Figure 6: Report-keyword audits indicate that the retained cohort is not a single-pathology collection. Groups are non-mutually exclusive report-text keyword audits used for cohort characterization, not official diagnostic prevalence labels.

Cohort diversity. The retained follow-up cohort is not restricted to a single demographic, sequence protocol, or scanner condition. Figure 7 summarizes this evidence: patients span pediatric through older-adult age groups, nearly half have follow-up beyond two years, common MRI sequences are widely available, and study-level acquisition metadata includes multiple vendors and both 1.5T and 3.0T scans. Figure 6 complements this view by auditing report keyword mentions across broad

pathology groups, showing that the retained cohort is not dominated by a single report category. These summaries support the breadth of the retained cohort for benchmark construction; they are not intended as prevalence or population representativeness claims.

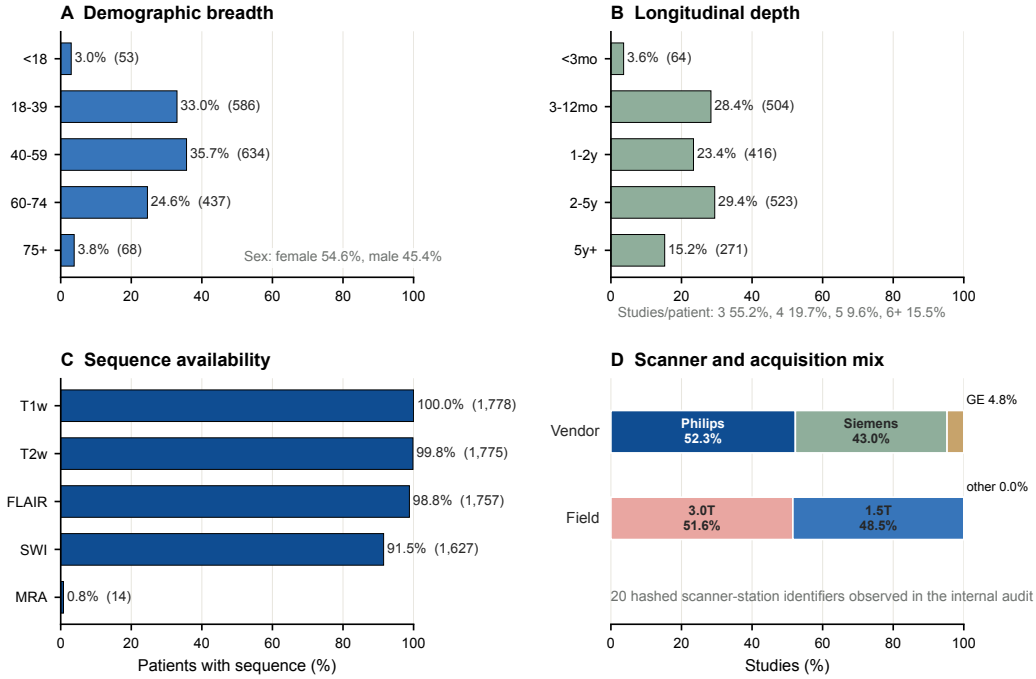


Figure 7: The retained cohort is not restricted to a single demographic or acquisition slice. Patient-level demographics and follow-up depth show broad age coverage, balanced sex distribution, and substantial multi-year follow-up. Sequence availability is high for T1w, T2w, FLAIR, and SWI, while study-level acquisition metadata spans multiple scanner vendors and both 1.5T and 3.0T field strengths. Percentages are patient-level except for vendor and field strength, which are study-level; the figure describes retained-cohort breadth rather than population prevalence.

B.3 Benchmark task coverage

The retained source cohort is sampled into evaluated VQA instances that span clinical reasoning depth, evidence conditions, and evidence attributes. Figure 8 gives the visual overview; Table 6 records the evaluated denominators and the intended evidence demands for each task family.

B.4 Clinical-level taxonomy

BrainTRACE organizes questions by the clinical reasoning depth required to answer them, rather than treating all items as interchangeable visual recognition tasks. The levels therefore define what kind of evidence operation is being tested; they are not merely ordered difficulty bins.

- **L1: Image and acquisition recognition.** These questions test whether a model can identify acquisition properties and basic visual findings, including sequence, plane, laterality, focal abnormality, and concise image description.
- **L2: Anatomical localization and single-study finding recognition.** These questions require lesion counting, anatomical localization, size or volume categorization, signal and enhancement characterization, morphology assessment, and asymmetry recognition within a single study.
- **L3: Disease-specific interpretation and interval change.** These questions compare findings across studies, including size change, new lesion appearance, resolution, enhancement evolution, mass-effect change, interval summaries, and volume change.

- **L4: Longitudinal trajectory or quantitative change assessment.** These questions require reasoning over trajectories, including progression pattern, first onset, peak timepoint, narrative trajectory, response category, and serial volume change.
- **L5: Case-level and integrated clinical judgment.** These questions require synthesis across visual evidence and clinical context, including impression generation, disease-status assessment, comparative summaries, and differential diagnosis with supporting evidence.

Per-track denominators and evidence-property annotations are reported in Table 6.

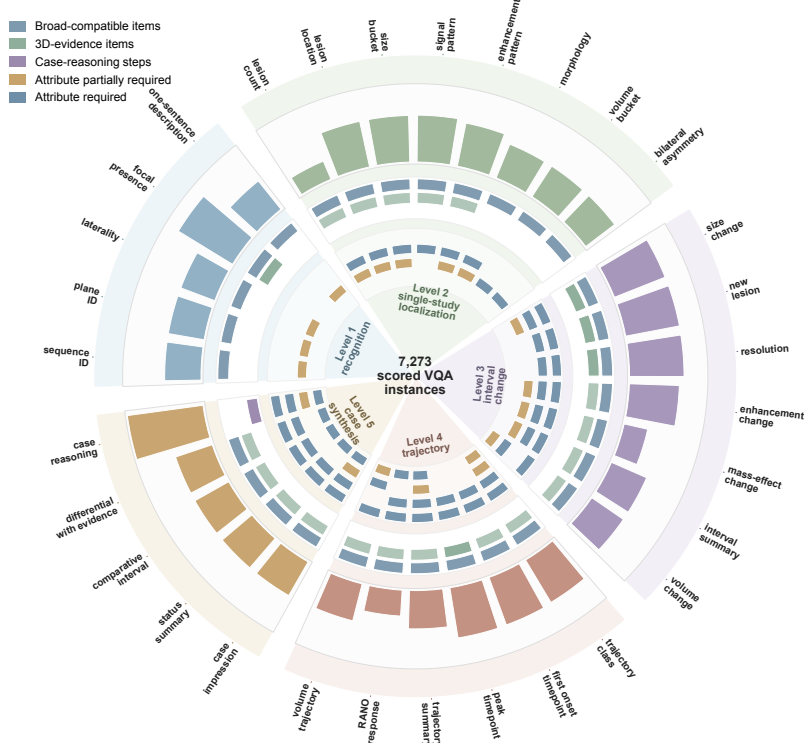


Figure 8: Scored VQA instances span five clinical-operation levels, three evidence conditions, and the evidence attributes used to audit longitudinal, multi-sequence, volumetric, and clinically interpretive reasoning.

B.5 Per-task counts and evidence-property coverage

Table 6 combines task-family sampling provenance with the evidence attributes used throughout the paper. The count columns define evaluated denominators for the broad-compatible, 3D-evidence, and case-reasoning tracks. The rightmost columns state the intended evidence demands, which makes the table a task design matrix rather than a second prose taxonomy. We split volumetric evidence into V_p for pseudo-3D multi-plane or multi-slice composites and V_n for native-volume segmentation-derived values.

C Scoring, audit, and reproducibility protocol

This appendix chapter defines the factual-audit rule used for open-ended answers and decomposed case-reasoning steps. The goal is to measure clinical acceptability under the released report-grounded rubric, not stylistic similarity to a reference sentence. The chapter also states how Pass %, Step Pass, and Case Success are computed and reports agreement and judge-family sensitivity audits of the scoring instrument.

Table 6: Task-family design matrix defines evaluated denominators and intended evidence demands. Broad, 3D, and Case give evaluated VQA-instance counts. **L** = longitudinal, **M** = multi-sequence, **V_p** = pseudo-3D volumetric, **V_n** = native-volume volumetric, and **C** = clinical reasoning beyond surface perception.  = required,  = partial or compatible, and  = not required;  marks a central purpose.

Task family	Ans.	Broad	3D	Case	Total	Notes	L	M	V _p	V _n	C
L1.1 pulse_sequence_identification	MCQ	200	0	0	200						
L1.2 plane_id	MCQ	200	0	0	200						
L1.3 laterality_of_focal_finding	MCQ	200	0	0	200						
L1.4 focal_abnormality_presence	binary	200	150	0	350						
L1.6 one_sentence_image_description	open	200	0	0	200						
L2.1 discrete_abnormality_count	ord.	85	24	0	109	pool-limited					
L2.2 location_bucket_dominant	MCQ	200	57	0	257	canonical n					
L2.3 size_bucket_dominant	ord.	200	57	0	257	canonical n					
L2.4 signal_pattern	MCQ	200	58	0	258						
L2.5 enhancement_pattern	MCQ	188	54	0	242	dedup					
L2.6 morphology_and_boundary	open	200	0	0	200						
L2.7 structure_volume_bucket	ord.	200	0	0	200						
L2.8 bilateral_volumetric_asymmetry	MCQ	200	0	0	200						
L3.1 size_change_at_site	ord.	200	100	0	300						
L3.2 new_focal_abnormality	binary	200	101	0	301						
L3.3 resolution_of_prior	MCQ	200	100	0	300						
L3.4 enhancement_change_at_site	MCQ	200	75	0	275						
L3.5 mass_effect_change_at_site	MCQ	88	45	0	133	pool-limited					
L3.6 open_interval_change_at_site	open	200	11	0	211						
L3.7 structure_volume_change	ord.	200	11	0	211	canonical n					
L4.1 trajectory_class_closed	MCQ	195	77	0	272						
L4.2 new_lesion_timing	ord.	200	78	0	278						
L4.3 peak_extent_timepoint	ord.	200	80	0	280						
L4.4 open_trajectory_summary	open	200	11	0	211						
L4.5 response_proxy_RANO_lite	MCQ	97	46	0	143	pool-limited					
L4.6 structure_volume_trajectory	ord.	200	10	0	210	canonical n					
L5.1 full_impression_3to5_sentences	open	200	14	0	214	canonical n					
L5.2 brief_imaging_status_summary	open	200	14	0	214	canonical n					
L5.3 comparative_interval_summary	open	200	14	0	214	canonical n					
L5.4 differential_with_visual_evidence	open	200	13	0	213						
Decomposed case reasoning	case	0	0	420	420	case-level					
Total		5,653	1,200	420	7,273						

C.1 Factual-audit scoring for open-ended and case-level answers

Open-ended and case-level answers are scored by factual clinical audit rather than lexical similarity. A response passes only if it covers the required clinical facts, satisfies the critical rubric slots, avoids clinically material contradictions, and does not introduce unsupported claims that would alter the interpretation. This distinction is necessary for radiology-style answers: fluent prose can still fail if it changes laterality, reverses interval direction, assigns a finding to the wrong anatomy, cites the wrong timepoint, or asserts an unsupported disease-status conclusion.

For each open-ended item or decomposed case step i , the evaluator receives the question q_i , the model answer y_i , the reference answer r_i , accepted paraphrases when available, required clinical facts, critical rubric slots, and pre-specified failure criteria. The evaluator does not receive the source report, segmentation tables, evidence quotations, or construction record. All evaluated outputs use the same judge model version, instruction text, response schema, and deterministic decoding setting, and the complete judge record is archived for each item.

C.1.1 Scoring fields and pass rule

For each item or decomposed case step, the judge record contains the following fields. These fields define the scoring variables used in the reported metrics:

- **required_fact_results**: checks whether named findings, locations, laterality, time-points, trajectory direction, and current status required by the reference are explicitly present or clinically equivalent.
- **slot_results**: evaluates each rubric slot against the model answer. Critical slots enter the pass rule; noncritical slots are retained for coverage and quality analyses.
- **critical_errors**: records hard contradictions such as wrong anatomy, wrong laterality, wrong interval direction, wrong timepoint, wrong modality-specific reading, wrong diagnosis, or clinically material severity errors.
- **unsupported_claims**: records claims not supported by the reference or rubric, such as a new lesion, disease progression, treatment response, or diagnosis absent from the clinical reference.
- **clinically_acceptable**: gives the final rubric-based factual verdict used for Pass %.
- **quality_score**: records graded answer quality on the same 1–5 clinical scale used in the human validation subset.

Pass rule. Let $F_i = 1$ if all required facts for item i are present or clinically equivalent, $S_i = 1$ if all critical rubric slots are satisfied, $E_i = 1$ if no critical error is present, and $U_i = 1$ if no clinically material unsupported claim is present. The item-level pass indicator is

$$P_i = F_i \cdot S_i \cdot E_i \cdot U_i.$$

Equivalently, an answer fails if it omits a required fact, misses a critical rubric slot, contradicts the reference, or introduces a clinically material unsupported claim.

Reported metrics. For a set $\mathcal{I}$ of open-ended items, Pass % is

$$\text{Pass}(\mathcal{I}) = 100 \times \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} P_i.$$

For the case-reasoning track, each case c has six ordered steps $s = 1, \dots, 6$. With $P_{c,s}$ denoting the pass indicator for step s in case c , Step Pass and Case Success are

$$\text{StepPass} = 100 \times \frac{1}{6|\mathcal{C}|} \sum_{c \in \mathcal{C}} \sum_{s=1}^6 P_{c,s}, \quad \text{CaseSuccess} = 100 \times \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \prod_{s=1}^6 P_{c,s}.$$

Case Success is stricter than Step Pass because all six steps must pass within the same longitudinal case. We also report Mean Quality Score, Critical Error %, required-fact coverage, missing-required-fact counts, slot-match rates, unsupported-claim counts, and the full judge records. These quantities are diagnostic; the primary open-ended leaderboard is ranked by Pass %.

Let $Q_i \in \{1, \dots, 5\}$ denote the graded quality score, and let M_i and K_i denote the number of matched and required facts for item i . The diagnostic aggregates are

$$\text{MeanQS}(\mathcal{I}) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} Q_i, \quad \text{CriticalError}(\mathcal{I}) = 100 \times \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} (1 - E_i),$$

$$\text{RequiredFactCoverage}(\mathcal{I}) = 100 \times \frac{\sum_{i \in \mathcal{I}} M_i}{\sum_{i \in \mathcal{I}} K_i}.$$

These aggregates explain where failures arise, but they do not override the binary factual-audit pass rule.

The audit uses a conservative interpretation of clinical acceptability under the released rubric. Generic radiology language does not satisfy a named finding, laterality, trajectory, or disease-status requirement. Direction reversals, laterality errors, wrong-anatomy claims, wrong-diagnosis claims, and

unsupported new findings are failures even when the surrounding answer is fluent. Omissions and contradictions are recorded separately: an omission affects required-fact coverage, while a contradiction is counted as a critical error.

Human validation is reported in Appendix C.5. Two expert reviewers independently scored a case-reasoning subset under the same clinical quality and critical-error semantics, and the released records include per-step expert grades and judge fields so that external auditors can repeat or extend the agreement analysis. Appendix F provides annotated success and failure examples showing how the audit treats matched required facts, missing facts, wrong direction, wrong anatomy, and unsupported clinical claims.

C.2 Open-ended audit leaderboards

Table 7 gives the complete open-ended leaderboard behind the abbreviated main-text summary. It is retained in the appendix so the main paper can emphasize scoring design and headline patterns while still making every evaluated configuration auditable. The table reports the factual-audit metrics available for the completed broad-compatible open-ended runs, together with four traditional text-overlap or embedding-similarity diagnostics. These secondary diagnostics are included for comparability with prior VQA and report-generation evaluations, but the table remains ranked by factual audit because BLEU [39], ROUGE [31], chrF++ [42], and BERTScore [54] do not check required clinical slots, contradictions, laterality, timing, or unsupported disease-status claims.

Table 7: The full broad-compatible open-ended leaderboard is ranked by factual-audit Pass %, while traditional text metrics are reported only as secondary comparability diagnostics. QS is Mean Quality Score and Crit. is Critical Error %; BLEU4, ROUGE-L, chrF++, and BERTScore are mean scores over the same broadQA open-ended items in their native scales. Bold marks the best value in each metric column and underlining marks the second-best value; lower is better for Crit., and higher is better for all other metric columns.

Rank	Model	Pass	QS	Crit.	BLEU4	ROUGE-L	chrF++	BERTScore
1	GPT-5.4	19.8	2.55	22.2	0.065	0.198	<u>0.281</u>	0.043
2	Qwen3-VL-30B	<u>14.4</u>	2.27	36.6	0.039	0.193	<u>0.273</u>	0.149
3	Gemini 2.5 Pro	<u>14.2</u>	2.28	32.9	0.040	0.183	0.267	0.144
4	Qwen3-VL-8B	13.8	<u>2.29</u>	29.0	0.081	0.202	0.272	0.157
5	Qwen3-VL-4B	11.8	2.24	28.0	0.070	0.190	0.263	0.134
6	GPT-5-mini	10.6	2.07	40.3	0.056	0.176	0.264	0.102
7	GPT-5	10.5	2.12	33.8	<u>0.074</u>	0.179	0.263	0.128
8	HuatuoGPT-Vision-7B	9.9	2.13	25.6	<u>0.072</u>	<u>0.208</u>	0.289	0.157
9	Claude Opus 4.6	9.6	2.00	43.0	0.020	0.142	0.247	0.050
10	InternVL3-14B	8.6	2.11	32.2	0.044	0.206	0.276	<u>0.181</u>
11	Qwen3.5-27B	8.1	1.91	43.7	0.052	0.148	0.211	-1.116
12	InternVL3-78B	7.6	2.04	34.6	0.037	0.199	0.276	0.161
13	Lingshu-7B	6.8	2.03	25.7	0.042	0.211	0.273	0.185
14	InternVL3-8B	6.6	2.00	32.1	0.045	0.207	0.272	0.166
15	MedGemma-4B-it	5.4	1.90	33.2	0.026	0.176	0.239	0.141
16	Qwen2.5-VL-7B	4.9	1.96	27.4	0.037	0.204	0.251	0.164
17	RadFM	4.3	1.96	<u>21.4</u>	0.014	0.108	0.195	-0.107
18	LLaVA-Med-v1.5	3.8	2.06	8.6	0.010	0.070	0.105	-1.605
19	Janus-Pro-7B	1.8	1.86	24.4	0.007	0.082	0.180	-0.529

Pass %, QS, and Critical Error % are rubric-based factual-audit metrics: they score clinical acceptability under the released reference, graded answer quality, and disqualifying clinical errors. The lexical and embedding metrics are deliberately secondary. A high-overlap answer can still omit a required fact or add an unsupported diagnosis, while an acceptable answer can use wording that differs from the reference. BERTScore uses the rescaled RoBERTa-large F1 convention in the scored artifacts, so negative values are possible.

C.3 Case-level reasoning audit records

Table 3 reports the headline case-level columns in the main paper, so this appendix focuses on the released audit records rather than duplicating a case-only leaderboard. Figure 11 summarizes Step Pass uncertainty for the strongest case-level configurations. The release retains the per-step and per-case records needed to recompute Step Pass, Case Success, Mean QS, Critical Error, missing-required-fact counts, and required-fact coverage diagnostics.

C.4 Per-step attrition in case-reasoning

Figure 4 gives the per-step diagnostic: Panel A tracks cumulative case survival, while Panel B shows which decomposed reasoning steps pass before a case collapses. The main pattern is not a single weak model row but a consistent composition failure. Earlier steps can pass at roughly 30–43% for the strongest configurations, but final synthesis remains near zero. Claude Opus 4.6 has the highest final-synthesis pass rate among the displayed rows (7.1%), but still has 0/70 full-case survival; GPT-5-mini has the highest aggregate Step Pass (24.5%) but also reaches 0/70 full-case survival. Thus Step Pass is useful for localizing failures, whereas Case Success tests whether those local successes occur in the same longitudinal case.

The GPT-5-family rows remain closely spaced in aggregate Step Pass, as shown by the consolidated bootstrap intervals in Figure 11. Increasing GPT-5.4 reasoning effort redistributes successes across step types but does not improve aggregate Step Pass, reinforcing that the bottleneck is evidence propagation and final synthesis rather than a single isolated recognition step.

C.5 Two-reviewer human validation of the factual-audit judge

C.5.1 Protocol

This analysis evaluates agreement between the released factual-audit judge and expert review; it is not a clinical certification study. Two board-certified neuroradiologists (Reviewer 1 and Reviewer 2) independently graded the same 60-step subset, blind to each other and to the LLM-judge output. The subset comprises 10 longitudinal tumor cases $\times$ 6 reasoning steps, with model outputs split equally between GPT-5.4 (5 cases) and Qwen3-VL-30B (5 cases). Each step was rated on a 5-point clinical quality scale (1 = critical error; 5 = clinically excellent) and a binary critical-error flag. After an initial round, Reviewer 1 rescored the subset under finalized rubric semantics; we use the post-calibration grades. Total per-reviewer effort was approximately one hour.

C.5.2 Headline metrics

Quadratic-weighted Cohen’s κ_w on the 1–5 quality: 0.737 (Reviewer 1 vs. judge, 95% CI [0.56, 0.86]), 0.645 (Reviewer 2 vs. judge, [0.47, 0.78]), 0.821 (Reviewer 1 vs. Reviewer 2, [0.72, 0.89]). Intraclass correlation ICC(3,1) is 0.740 / 0.649 / 0.824 respectively. Binary critical-error Cohen’s κ : 0.475 (Reviewer 1 vs. judge, [0.29, 0.67]), 0.552 (Reviewer 2 vs. judge, [0.35, 0.74]), 0.767 (Reviewer 1 vs. Reviewer 2, [0.60, 0.90]). Judge specificity is 1.000 against both reviewer references; sensitivity is 0.484 and 0.536 respectively.

C.5.3 Per-step decomposition

Reviewer 1 vs. judge κ_w across S1–S6: 0.84, 0.82, 0.48, 0.77, 0.63, 0.42. Reviewer 2 vs. judge: 0.64, 0.76, 0.36, 0.56, 0.60, 0.50. Inter-human (Reviewer 1 vs. Reviewer 2): 0.86, 0.79, 0.94, 0.75, 0.65, 0.31. Final-synthesis agreement degrades for all three pairs, consistent with the greater ambiguity and evidence-composition burden of integrative reasoning.

C.5.4 Per-model decomposition

Reviewer 1 vs. judge κ_w is 0.699 on GPT-5.4 decomposed VQA steps ($n = 30$) and 0.772 on Qwen3-VL-30B decomposed VQA steps ($n = 30$). Reviewer 2 vs. judge is 0.628 and 0.664 respectively, while inter-human agreement is 0.855 and 0.782. Cross-model agreement is similar in rank and magnitude, indicating the judge comparison is not driven by a single model family.

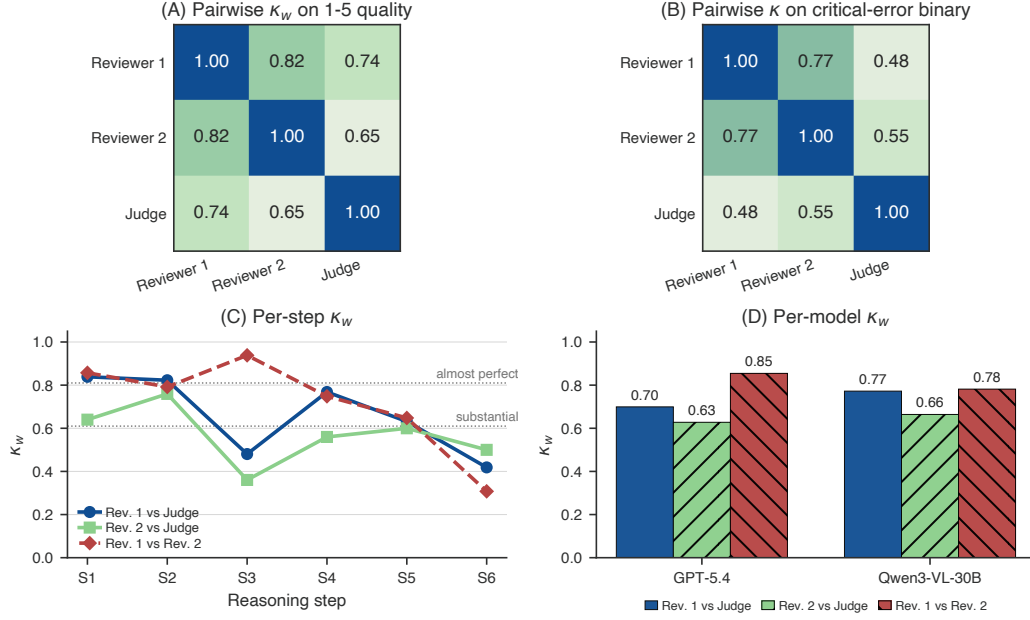


Figure 9: Three-way inter-annotator agreement analysis evaluates the factual-audit judge as a scoring instrument rather than treating it as an unchecked automatic metric. The audit covers the case-reasoning track ($n = 60$ decomposed VQA steps). (A,B) Pairwise agreement on the 1–5 quality scale and the critical-error binary. (C) Per-step κ_w . (D) Per-model κ_w .

C.5.5 Scope of the validation

The validation is scoped to two reviewers and 60 decomposed VQA steps from the case-reasoning track. We position the result as an agreement audit of the scoring instrument, not as an independent de novo radiologist adjudication of the full benchmark. The full per-step expert grades and judge raw fields are released so that external auditors may extend the validation to additional reviewers, items, or scoring specification revisions.

C.6 Cross-family judge sensitivity audit

The primary open-ended evaluator in the main experiments is gpt-4o. To examine whether this choice affects reported open-ended pass rates, we ran a small cross-family sensitivity analysis on stratified broad open-ended items. The audit uses 52 items per cell, sampled evenly across L1.6, L2.6, L3.6, and L4.4 (13 items per family), and compares release-scored judgments from the GPT judge with Gemini 2.5 Pro and Claude Sonnet 4.6 on outputs from GPT-5.4, Claude Opus 4.6, and Gemini 2.5 Pro. All pass rates and quality scores in this subsection are recomputed with the released scoring implementation rather than ad hoc parsing of judge records.

The audit is intentionally narrow: it checks whether the released rubric behaves differently when the judge family changes on a stratified subset. The main sensitivity is localized to GPT-5.4 outputs. The release judge passes 32.7% of the sampled GPT-5.4 items, whereas both cross-family judges pass 15.4%; the discordance is mostly release-only passes. For Claude Opus 4.6 and Gemini 2.5 Pro outputs, pass rates are closer across judges, and Gemini-output agreement is high ($\kappa = 0.898$ for pass decisions). We therefore treat exact open-ended Pass % values as rubric-and-judge measurements, while the central empirical conclusion remains stable in this audit: clinically acceptable open-ended answers are rare across judge families.

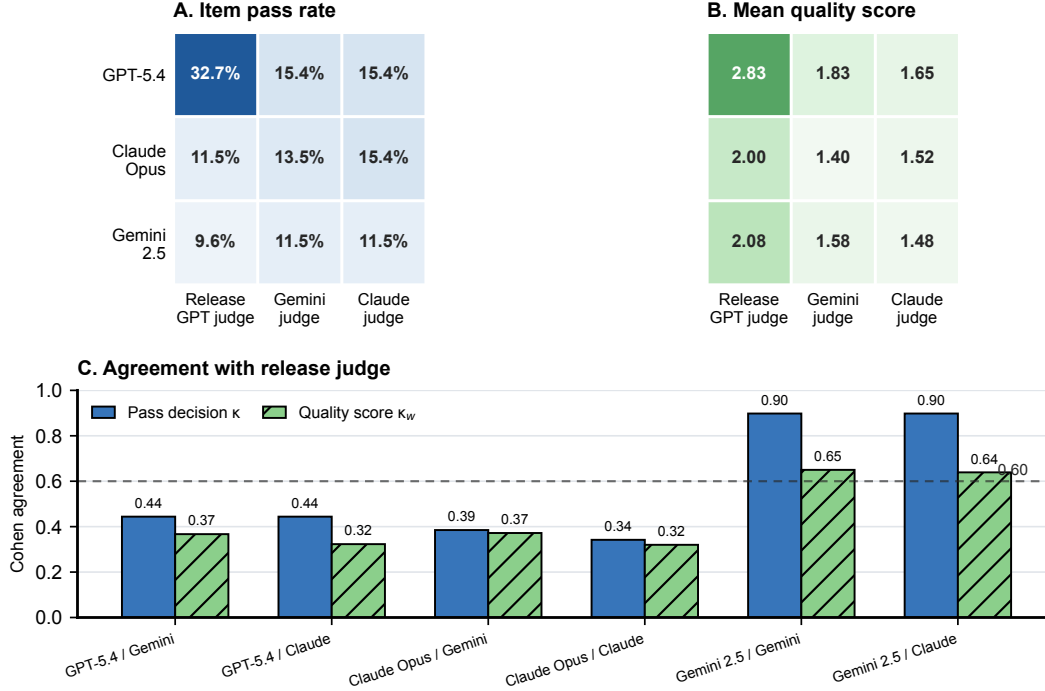


Figure 10: Cross-family judge audit measures sensitivity to the judge family under the released factual scoring rule. Each cell contains the same stratified sample size ($n = 52$). Alternative Gemini and Claude judges preserve low absolute pass rates, while the agreement panel shows where the release judge aligns most strongly with cross-family alternatives.

C.7 Evaluation records and reproducibility

The public scoring specification is released as docs/SCORING.md in the companion repository: <https://github.com/lanqz7766/BrainTRACE/blob/main/docs/SCORING.md>. The repository contains the rendering specifications, evaluation interfaces, scoring implementation, canonical item lists, exclusion registries, and audit records needed to recompute the reported scores. Closed-form scoring uses exact matching after answer normalization and removal of visible `<think>...</think>` text. Open-ended and case-level answers use the factual-audit rule defined in Appendix C.1.

For each track, the release records the evaluated item set, the model answer, and the corresponding scoring record. Closed-form records contain the canonical item identifier, the normalized answer, and the exact-match decision. Open-ended records contain the rubric slots, required-fact coverage, critical-error flag, unsupported-claim flag, quality score, and judge identity. Case-reasoning records contain the six ordered step decisions for each longitudinal case, which determine Step Pass and Case Success under the formulas above. The release also includes the exclusion registries used after rendering and reference checks. These records define the denominators in the reported tables; this paragraph documents reproducibility rather than adding another scale statistic.

Figure 11 consolidates uncertainty estimates across headline metrics. It is a visual uncertainty summary and replaces separate appendix-only CI tables for the case-reasoning and evidence-channel ablation analyses.

D Additional empirical diagnostics

This appendix chapter collects diagnostics that should not be interpreted as a second leaderboard. Each analysis isolates a different reading of the same benchmark: performance under available 3D

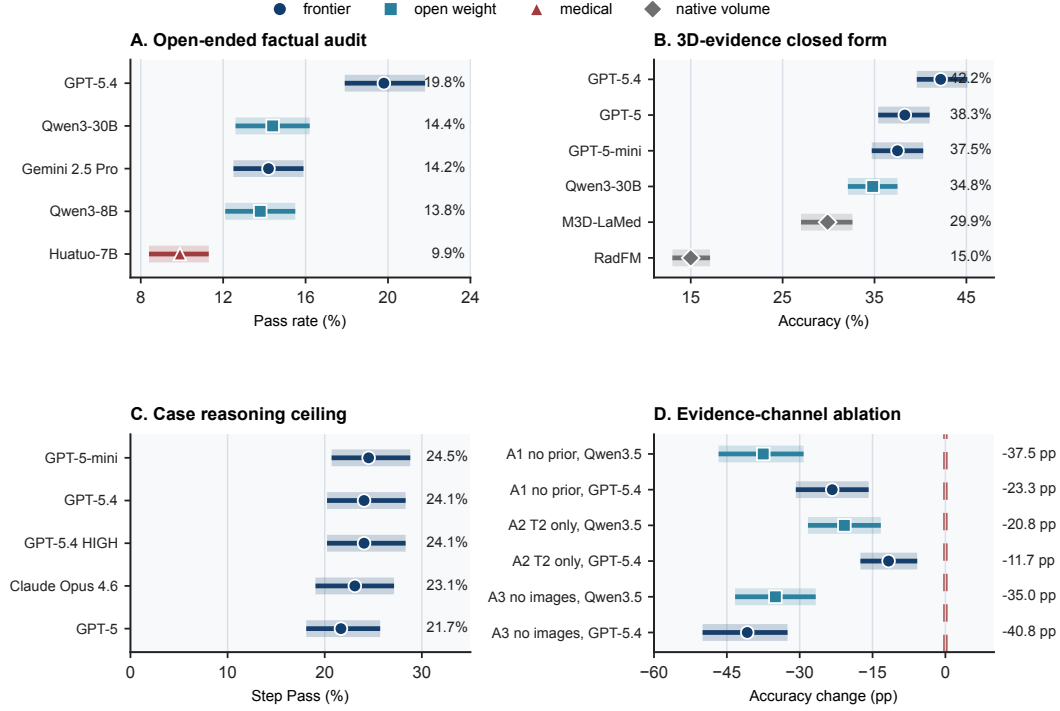


Figure 11: Bootstrap confidence intervals summarize uncertainty in the headline empirical pattern without adding a second leaderboard narrative. Thick bands show 95% bootstrap confidence intervals and points show estimates. Panel D uses paired bootstrap intervals on matched ablation samples; all intervals remain below zero. Intervals use nonparametric percentile bootstrap over raw scored records or judge-step records, matching the protocol summarized in Appendix C.7.

input interfaces, rank instability across evaluation axes, and difficulty after correcting for level-specific chance floors.

D.1 Paper-reported 3D-evidence results

Table 8 is the numeric companion to the main-text 3D column and to the representative 3D interval panel in Figure 11. It is retained for confidence interval provenance and interface transparency rather than as a separate leaderboard. The clinical targets are shared, but the input pathways are not controlled equivalents: non-native-volume VLMs receive the 3D-evidence questions through 16-slice axial renderings for each study-sequence, whereas M3D-LaMed and RadFM use their supported native-volume pathways. The table therefore compares available model interfaces on shared clinical targets, not the isolated value of one 3D representation. Open-ended Pass is shown only for configurations with released factual-audit records; unavailable cells are marked with dashes rather than inferred from closed-form scores.

D.2 Ranking sensitivity across evaluation axes

Figure 12 summarizes how model order changes across evaluation axes. Ranks are computed column-wise over eligible main-paper rows, with missing entries excluded. The diagnostic point is that BrainTRACE should not be reduced to a single aggregate score: closed-form recognition, open-ended clinical acceptability, 3D-evidence performance, Step Pass, and Case Success reward different aspects of evidence use.

Table 8: Paper-reported 3D-evidence results provide compact numeric provenance for the main 3D columns and the representative CI visualization. Closed-form entries give exact accuracy (%) with bootstrap confidence intervals over the 1,123 closed-form items. Open-ended entries report Pass %; dashes indicate that released factual-audit records are unavailable for that configuration. Non-native-volume VLMs use 16-slice axial renderings; M3D-LaMed and RadFM use their native-volume interfaces, so this table is an interface comparison on shared clinical targets.

Ranks 1–10				Ranks 11–20			
Rank	Model	Closed [95% CI]	Pass	Rank	Model	Closed [95% CI]	Pass
1	GPT-5.4	42.2 [39.6, 45.1]	15.6	11	Qwen2.5-VL-7B	33.2 [30.5, 36.0]	3.9
2	GPT-5	38.3 [35.4, 41.0]	11.7	12	Gemini 2.5 Pro	32.0 [29.3, 34.7]	13.0
3	GPT-5-mini	37.5 [34.7, 40.3]	10.4	13	M3D-LaMed	29.9 [27.0, 32.6]	0.0
4	Qwen3-VL-4B	35.4 [32.6, 37.9]	5.2	14	HuatuogPT-Vision-7B	29.1 [26.4, 31.7]	0.0
5	Lingshu-7B	35.4 [32.6, 38.1]	3.9	15	MedGemma-4B-it	18.9 [16.7, 21.4]	3.9
6	InternVL3-8B	35.3 [32.3, 37.9]	–	16	Janus-Pro-7B	17.6 [15.5, 19.9]	0.0
7	Qwen3-VL-30B	34.8 [32.1, 37.5]	11.7	17	RadFM	15.0 [13.0, 17.1]	2.6
8	InternVL3-14B	34.0 [31.4, 36.8]	–	18	Qwen3.5-27B	12.6 [10.5, 14.6]	3.9
9	Qwen3-VL-8B	33.4 [30.6, 36.1]	9.1	19	Claude Opus 4.6	1.2 [0.6, 1.9]	6.5
10	InternVL3-78B	33.3 [30.6, 36.3]	–	20	LLaVA-Med-v1.5	0.0 [0.0, 0.0]	1.3

Evaluation choice changes the apparent model ranking

Rows are grouped by model family; gold boxes mark axis winners; side bars show max-min rank.

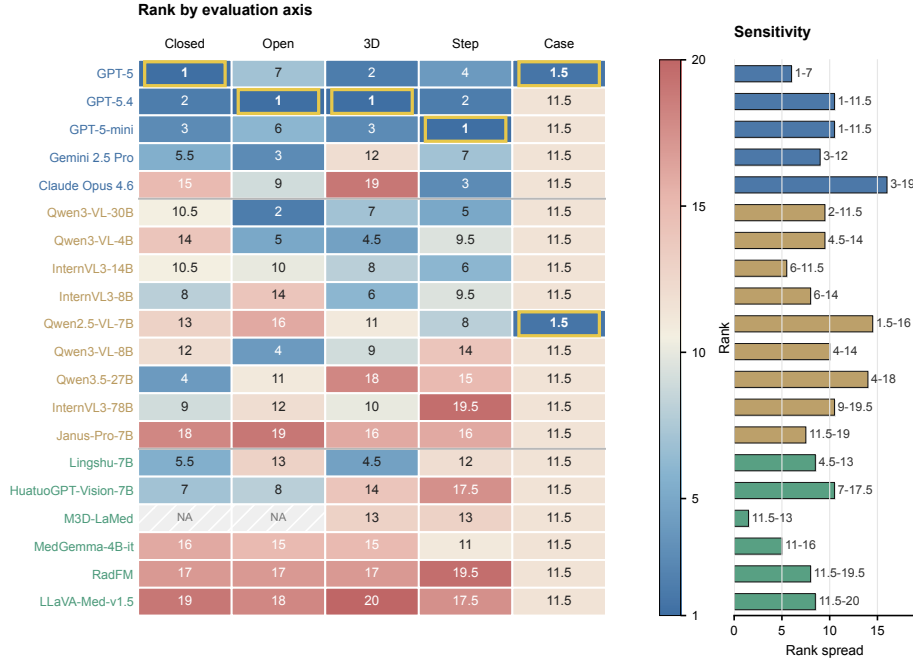


Figure 12: Evaluation choice changes the apparent model ranking. Each cell reports a model’s rank on one evaluation axis, with lower ranks better; the side bars summarize each model’s rank spread across axes. The figure is a sensitivity diagnostic, not a replacement for the per-axis results.

D.3 Calibrated level difficulty

Figure 13 reports a calibrated difficulty score (CDS) for closed-form L1–L4 items. Raw accuracy is not directly comparable across levels because the multiple-choice chance floor differs by template mix. For model m and level ℓ , we define

$$\text{CDS}_{m,\ell} = 100 \times \frac{\text{Acc}_{m,\ell} - \text{Chance}_{\ell}}{100 - \text{Chance}_{\ell}}, \quad \text{Chance}_{\ell} = \frac{1}{|\mathcal{Q}_{\ell}|} \sum_{q \in \mathcal{Q}_{\ell}} \frac{100}{|\mathcal{A}_q|},$$

where Q_ℓ is the set of closed-form questions at level ℓ and $|\mathcal{A}_q|$ is the number of answer options for question q . Thus 0 is the level-specific random baseline, 100 is perfect accuracy, and negative values indicate below-chance behavior.

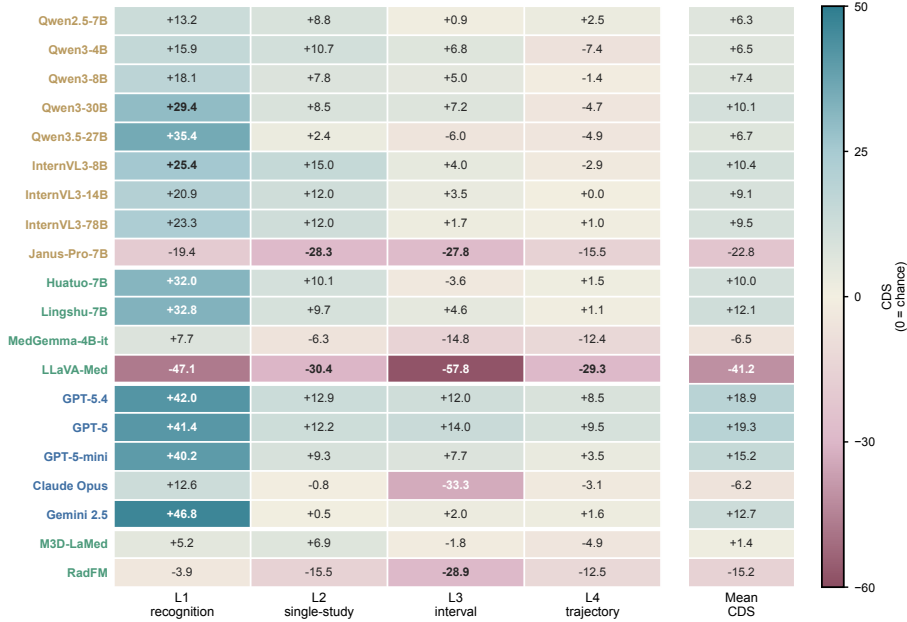


Figure 13: Chance-calibrated difficulty shows the drop from recognition to longitudinal reasoning after correcting for level-specific answer floors. CDS sets each level’s random baseline to 0 and perfect accuracy to 100; plum cells are below chance, neutral cells are near chance, and teal cells are above chance. Model names use the same family colors as Figure 12.

The calibration prevents levels with higher answer-option floors from appearing easier than they are. For example, GPT-5 has higher raw L3 accuracy than raw L2 (45.6% vs. 32.9%), but the corresponding CDS values are close (+14.0 vs. +12.2) because L3 has a higher chance floor. Conversely, Qwen3.5-27B has a seemingly moderate raw L3 accuracy of 33.0%, but its L3 CDS is -6.0 because this falls below the L3 random baseline; the same model remains clearly above chance on L1 (+35.4 CDS). The calibrated view therefore supports the main empirical interpretation: recognition-level strength does not imply reliable longitudinal reasoning.

E Evidence-channel ablation details

This appendix clarifies the matched evidence-channel comparisons used in Section 4.3. These comparisons are diagnostic: they test whether accuracy on selected closed-form items depends on longitudinal context, multi-sequence information, and visual evidence when the question text, answer options, permitted non-image fields, and scoring rule remain fixed. They are not intended as population-wide performance estimates over all BrainTRACE items.

E.1 Evidence retained in each comparison

The latest-timepoint condition retains the current study but removes prior longitudinal studies. Its paired accuracy drop measures dependence on serial context, including interval change and trajectory evidence. The T2-only condition retains the longitudinal T2 series but removes non-T2 sequences. Its drop measures dependence on multi-sequence contrast and sequence-specific findings, including enhancement or susceptibility cues when those are needed by the reference. The no-image condition withholds visual inputs while keeping the same question and answer space. Its drop measures residual dependence on question text, answer priors, or permitted non-image cues when rendered MRI evidence is absent.

E.2 Finding and interpretation

Figure 11 gives the paired bootstrap intervals for these evidence-channel drops. The intervals are computed within each matched comparison, so the appendix uses them to support the direction and stability of the drop rather than to introduce another leaderboard. Across the retained comparisons, the paired drops remain negative for both audited models. This supports the paper’s main interpretation that the selected BrainTRACE items require the retained evidence channels, not only answer priors or generic question text.

The matched item sets were constructed to share a common full-evidence reference accuracy of 50.0%, making the paired drops directly comparable within this diagnostic design. Because the ablation set is calibrated to this 50.0% baseline, the results should be interpreted as paired diagnostic drops rather than population-level estimates. Abstentions in the no-image condition are counted as incorrect under the same closed-form exact-match rule used throughout the benchmark. The companion release provides the evaluated items for each comparison, ensuring that the denominators underlying the paired intervals remain auditable.

F Qualitative failure cases

This appendix chapter is a compact case atlas for the factual-audit protocol. It shows why fluent answers can fail when they reverse laterality, miss or mislocalize a lesion, reverse interval direction, or introduce an unsupported diagnostic category. The examples complement the aggregate results by making the scoring rule inspectable.

The section presents paired GPT-5.4 success and failure examples across the five clinical levels. Each pair uses the same task template and illustrates a clinically material failure mode. Annotation legend: blue = matched required facts between reference and response; green = correct clinical fact in reference; red = erroneous claim in model response. Images are reproduced from the MR-RATE corpus under its CC BY-NC-SA 4.0 license.

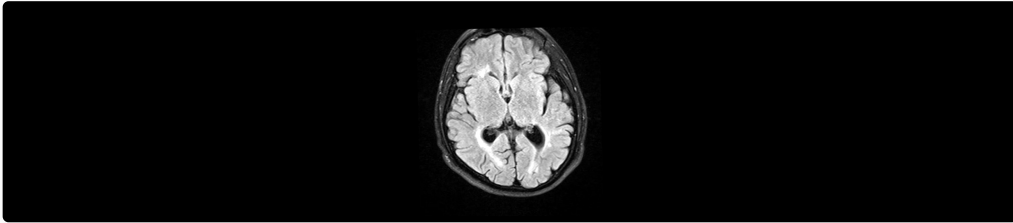
Table 9 summarizes the failure mode illustrated at each clinical level.

Table 9: The case atlas covers distinct clinically material failure modes across the five task levels.

Level	Task template	Failure mode	Error type	Images
L1	one_sentence_image_description	Reversed laterality (left → right)	wrong_laterality	1
L2	morphology_and_boundary	Missed lesion / false negative	missing_finding	1
L3	open_interval_change_at_site	Decrease reported as increase	wrong_direction	6
L4	open_trajectory_summary	Mislocated + wrong trajectory	wrong_anatomy, wrong_direction	3
L5	full_impression_3to5_sentences	Wrong diagnostic category	wrong_diagnosis	6

L1 Image and acquisition recognition — L1.6 one_sentence_image_description

✓ SUCCESS CASE



QUESTION

In one concise sentence, state what this image shows (include modality, plane, and main observation).

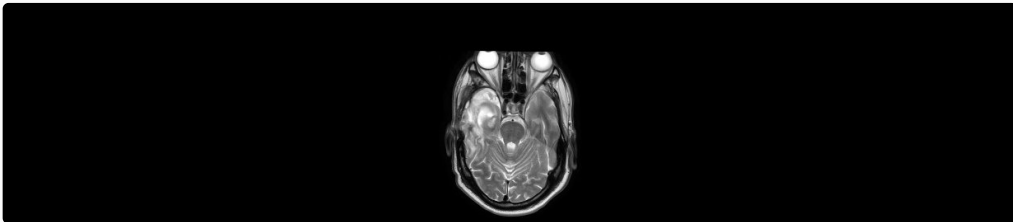
REFERENCE

Axial FLAIR MRI demonstrating a deep white matter focal abnormality.

GPT-5.4 RESPONSE

Axial FLAIR brain MRI shows periventricular white matter hyperintense lesions, including a left frontal pericallosal focus, consistent with demyelinating plaques.

✗ FAILED CASE



QUESTION

Describe this image in one sentence, covering modality, plane, and any visible abnormality.

REFERENCE

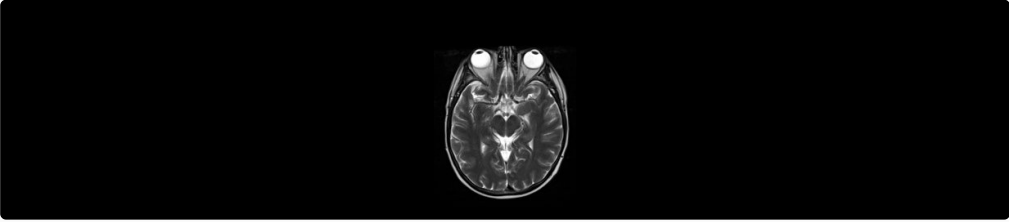
Axial T2-weighted MRI demonstrating a left temporal focal abnormality.

GPT-5.4 RESPONSE

Axial T2-weighted brain MRI shows a large heterogeneous T2-hyperintense mass centered in the right temporal/middle cranial fossa with associated local mass effect.

Figure 14: L1: Image and acquisition recognition (L1.6 one_sentence_image_description). The model matches the required facts in the success case but reverses left/right laterality in the failure case. This is a clinically material contradiction under the factual-audit rule even when the rest of the description is fluent.

✓ SUCCESS CASE



QUESTION

In one or two sentences, describe the dominant lesion's shape, border definition, and internal texture. Respond "no lesion present" if there is no lesion.

REFERENCE

no lesion present

GPT-5.4 RESPONSE

no lesion present

✗ FAILED CASE



QUESTION

Please characterize the dominant lesion in 1–2 sentences (shape / boundary / texture). If no lesion is visible, answer "no lesion present".

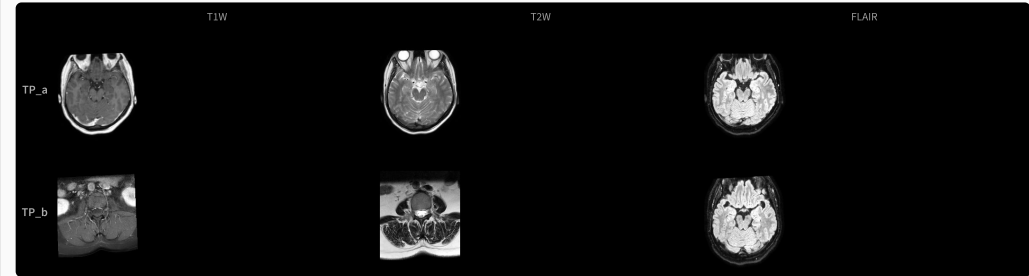
REFERENCE

In the **left peritrigonal area**, a significant increase in the size of the lesion, which had **ill-defined borders** and T2-weighted hyperintensity, was observed; the lesion's long axis was measured at **52mm**, and an increase in mass effect was also observed.

GPT-5.4 RESPONSE

no lesion present

Figure 15: L2: Anatomical localization and single-study finding recognition (L2.6 morphology_and_boundary). Both cases use a single axial T2W input and receive the same model response. The success case is compatible with the reference, whereas the failure case omits a 52 mm left peritrigonal lesion with ill-defined borders; the omission is a false negative rather than a wording difference.

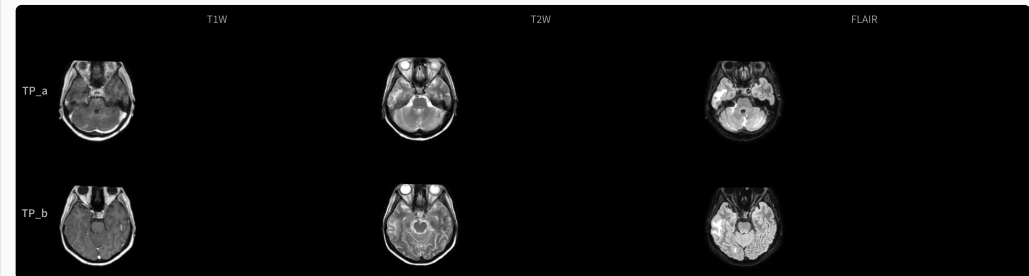


Summarize, in 1-2 sentences, the most important change at the supratentorial white matter from TP_a to TP_b.

Stable inactive MS plaques in the supratentorial white matter, subcortical, periventricular, brainstem, and corpus callosum.

No significant interval change is evident in the supratentorial white matter. Bilateral symmetric FLAIR hyperintense white matter abnormality appears grossly **stable** from TP_a to TP_b.

X



Please describe in 1-2 sentences the dominant interval change observed at the right cerebellar between the two timepoints.

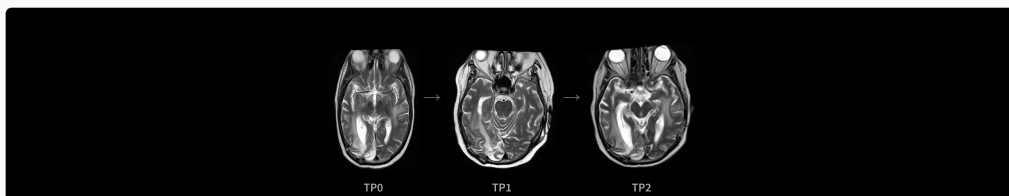
The size of the 13 mm diameter right inferior cerebellar metastasis has **decreased to 8 mm**, and edema has clearly **decreased**. No newly added cerebellar metastases.

GPT-5.4 RESPONSE

There is interval **increase in conspicuity/size** of the right cerebellar lesion, with more evident nodular enhancement on the later study. Mild associated T2/FLAIR hyperintensity is also **more apparent**.

Figure 16: L3: Disease-specific interpretation and interval change (L3.6 `open_interval_change_at_site`). With longitudinal multi-sequence inputs (2 timepoints $\times$ 3 sequences), the model identifies stable lesions in the success case but reverses the interval direction in the failure case, reporting increased size when the metastasis decreased from 13 mm to 8 mm.

✓ SUCCESS CASE



QUESTION

Describe the multi-timepoint trajectory of the most prominent abnormal signal area in 2–3 sentences. Cover trajectory direction, any turning point, and the current state.

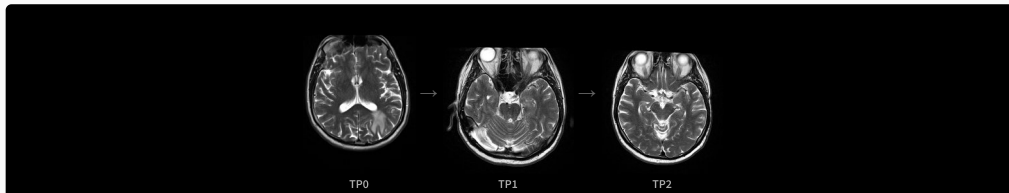
REFERENCE

The dominant abnormality in the parietal is **decreasing** across the shown timepoints. The key interval is between **TP0 and TP1**. At the most recent timepoint, the finding is currently **absent or not measurably visible**.

GPT-5.4 RESPONSE

The dominant abnormality in the right posterior temporal/occipital region **decreases from TP0 to TP1**, with less conspicuous surrounding T2 hyperintensity. It then turns slightly upward by TP2, with re-expansion of the CSF-like cavity. Currently, a sizable chronic-appearing encephalomalacic/cystic abnormality **persists**.

✗ FAILED CASE



QUESTION

In 2–3 sentences, summarise the imaging course of the most prominent abnormal signal area across these 3 sequential studies, including direction, any pivotal change, and current state.

REFERENCE

The dominant abnormality in the **right occipital** is **increasing** across the shown timepoints. The key interval is between TP0 and TP1. At the most recent timepoint, the finding is currently absent or not measurably visible.

GPT-5.4 RESPONSE

The most conspicuous abnormality is a chronic-appearing T2 hyperintense encephalomalacic/gliotic area in the **right cerebellar hemisphere**. It becomes clearly evident by TP1 and then shows **no meaningful further interval change** on TP2, remaining a **stable** chronic posterior fossa lesion.

Figure 17: L4: Longitudinal trajectory or quantitative change assessment (L4.4 open_trajectory_summary). Across three serial T2W studies, the model traces a decreasing trajectory in the success case. In the failure case, it mislocalizes the finding to the cerebellar hemisphere rather than the right occipital region and misreads an increasing trajectory as stable.

✓ **SUCCESS CASE**

T2W

SWI

QUESTION

Write the impression section of a radiology report for the most recent study, using the prior studies as comparison. 3–5 sentences.

REFERENCE

No significant pathology was detected in the supra-infratentorial neural parenchyma. On longitudinal review across the 3 sequential studies, the dominant finding has been **stable**. **No new interval finding** is identified. Overall, the intracranial status is **stable**.

GPT-5.4 RESPONSE

No significant interval change compared with the prior examinations. **Stable** small susceptibility focus in the pons/brainstem without appreciable associated mass effect or surrounding T2 signal abnormality. **No new** hemorrhage or other acute intracranial abnormality is evident on these images.

✗ **FAILED CASE**

T2W

SWI

QUESTION

Provide a 3–5 sentence diagnostic impression covering the findings, their longitudinal behavior, and the current overall status for this case.

REFERENCE

Cerebral atrophy; no parenchymal or leptomeningeal metastasis was detected. Diffuse bone metastases in the skull base and calvarial bones. On longitudinal review, the dominant finding has been **progressive**. **No definite additional interval finding** is described. Overall, the intracranial status is **progressive**.

GPT-5.4 RESPONSE

There is a hemorrhagic posterior fossa/brainstem lesion with prominent susceptibility blooming, most compatible with a cavernous malformation with evolving blood products. It appears more extensive on the intermediate study than at baseline. On the latest study, the dominant hemorrhagic abnormality is less conspicuous with residual chronic susceptibility. Overall, the current appearance favors **evolving/resolving hemorrhage without clear ongoing progression**.

Figure 18: L5: Case-level and integrated clinical judgment (L5.1 full_impression_3to5_sentences). Both cases present 3 timepoints × 2 sequences (T2W + SWI). The model identifies stable status with no new findings in the success case, but assigns the wrong diagnostic category in the failure case by interpreting progressive bone metastases as resolving hemorrhage.

G Limitations and Responsible Use

BrainTRACE evaluates longitudinal reasoning over brain MRI follow-up studies in an MR-RATE-derived cohort. Its task families assume multi-sequence brain imaging, English language radiology reports, and cases with sufficient follow-up, report, and segmentation-derived evidence. It is not intended to represent other anatomies, other languages, one-off clinical referrals with sparse longitudinal context, or the full spectrum of neuroradiology indications. Its scores are designed for reproducible model comparison, not for estimating absolute clinical performance. Users should report per-level, per-attribute, and acquisition subgroup results rather than treating a single aggregate score as a clinical capability measure.

The benchmark is report-grounded rather than independently re-read from scratch. The references come from authentic MR-RATE clinical radiology reports and segmentation-derived measurements, and released questions, ground truth, and audit fields pass construction and consistency checks. The case reasoning track further receives item-level review by an ABR-certified radiation oncologist. Even with these controls, reports are clinical documents rather than pixel-level annotations. A report may compress timepoints, omit a visually apparent but clinically secondary finding, or express uncertainty, and segmentation-derived measurements can carry registration, acquisition, and algorithmic error. Scores should be read as agreement with the reviewed, report-grounded clinical references under the released benchmark specification, not as a new adjudicated imaging diagnosis or deployment certification. The primary open-ended scores use a GPT family factual audit judge, and a small cross-family audit in Appendix C.6 shows judge family sensitivity for GPT-5.4 outputs. The 3D evidence track tests a clinically important evidence condition, but its rendered 16-slice and native-volume pathways introduce an input interface confound. These results should be read as comparisons among available model interfaces on shared clinical targets, not as a controlled ablation of 3D representation alone. BrainTRACE releases evaluation records for audit, but it does not redistribute raw imaging. Users must obtain access to the upstream MR-RATE corpus [12] and follow the corresponding data-use terms.