MEDSTAT-BENCH: EVALUATING DIVERSE MULTIMODAL MEDICAL INTELLIGENCE FROM PERCEPTION TO AGENTIC REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) are often evaluated on a single medical modality or task, leaving their behavior across clinical data types and workflows poorly characterized. We present MedStat-Bench, a broad stress test with 430 curated samples from 6 public datasets and 5 institutional cohorts covering about 10 clinical conditions and care settings. It spans clinical text, 2D images, volume-derived views from 3D studies, and sampled video, with tasks including single- and multi-label classification, structured clinical text generation, regression, and 2D/3D spatial localization. We release a self-contained public track with fixed evaluation materials; institutional cohorts test the same framework on locally collected clinical data. An initial agentic track contains 29 scenarios across 11 patient cases requiring evidence retrieval, inspection of text and medical media, tool use, and structured decisions. Results show that capability is compositional rather than uniform: models struggle most with interpreting volume-derived views and producing precise spatial coordinates, video results differ between ultrasound and surgical settings, and agentic success depends on both the clinical answer and workflow execution. MedStat-Bench therefore characterizes capability across medical data types, output formats, and tool-mediated workflows rather than relying on a single task or aggregate score.

1 INTRODUCTION

Clinical AI systems may need to interpret reports, 2D images, volumetric scans, and videos, then return diagnoses, measurements, localized findings, or management decisions. Medical multimodal capability is therefore compositional rather than uniform: it must be characterized across input modalities, output formats, and tool-mediated workflows rather than inferred from a single task or score.

Existing medical MLLM benchmarks emphasize text or 2D images, one volumetric imaging domain, or a particular video task (Chen et al., 2024; Hu et al., 2024; Hamamci et al., 2025; Liu et al., 2025). This separation makes cross-modality comparison difficult, and numerical measurement, spatial localization, and workflow completion are seldom tested together. Public datasets do not capture every local patient population or workflow, while protected institutional data cannot ordinarily be released for external evaluation.

MedStat-Bench has two tracks: **1. The static track** spans text, 2D images, volume-derived views, and sampled video. Its 216 *public* samples form a standalone reproducible benchmark with fixed cases, prompts, preprocessing, reference answers, task definitions, and scoring. Another set of 214 samples tests the framework on locally collected *institutional* data, with secure external evaluation planned. **2. The agentic track** contains 29 scenarios across 11 patient cases requiring evidence retrieval, inspection of text and medical media, tool use, and structured decisions.

We evaluate proprietary and open-weight MLLMs using the same image, frame, and projection formats with metrics matched to each output type. This design tests whether model strengths persist across modalities, output formats, and workflow stages. The results instead reveal distinct capability gaps: volume-derived interpretation and spatial localization remain difficult, while agentic performance depends separately on evidence retrieval, workflow execution, and the final decision.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of **static** medical multimodal benchmarks by data source, input modality, and scored task. Pub/Inst: public/institutional data; T/V: text/video; Cls/Gen/Det/Reg: classification/generation/detection/regression.

	Data sources		**Modalities**				**Tasks**				
Static multimodal benchmarks	Pub	Inst	T	2D	3D	V	Cls	Gen	2D Det	3D Det	Reg
GMAI-MMBench [2024]	✓			✓			✓	✓	✓		
OmniMedVQA [2024]	✓			✓			✓				
CARES [2024]	✓			✓			✓	✓			
MedXpertQA [2025]	✓		✓	✓			✓				
CT-RATE / CT-CHAT [2025]	✓				✓		✓	✓			
M3D-Bench [2025]	✓				✓		✓	✓		✓	
ReXGroundingCT [2025]	✓				✓					✓	
3D-RAD [2025]	✓				✓		✓	✓			
3DReasonKnee [2025]	✓				✓		✓	✓		✓	
Merlin [2025]	✓				✓		✓				
RadFM / RadBench [2025]	✓			✓	✓		✓	✓			
EndoBench [2025]	✓			✓		✓	✓	✓	✓		
SurgBench [2025]	✓					✓	✓		✓		
EchoPrime [2025]	✓					✓	✓				✓
U2-Bench [2026]	✓			✓		✓	✓	✓	✓		✓
MedFrameQA [2025]	✓			✓			✓				
MedSG-Bench [2025]	✓			✓			✓		✓		
DrVD-Bench [2025]	✓			✓	✓		✓	✓			
MedStat-Bench	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Our main contributions are summarized as follows:

- **A standalone public benchmark with complementary institutional cohorts.** Fixed public evaluation materials are complemented by institutional cohorts, with secure external evaluation planned.
- **An agentic benchmark for tool use in clinical workflows.** The agentic track evaluates final answers and workflow completion across all scenarios, with evidence-based reasoning and clinical safety evaluated for the 15 scenarios with reasoning ground truth.
- **Broad multimodal stress testing.** The static track spans four input modalities and seven output formats to expose capability gaps that narrower evaluations can miss. The agentic track adds patient-specific diagnostic, treatment, prognostic, and management decisions using records, medical images, and video.

2 Related Work

We first review static benchmarks by modality and task, followed by interactive benchmarks for clinical agents.

2.1 Static multimodal benchmarks

2.1.1 General and volumetric imaging

Broad suites such as GMAI-MMBench, OmniMedVQA, CARES, and MedXpertQA cover multiple clinical domains but primarily evaluate public 2D images via question answering, classification, or trustworthiness tests (Chen et al., 2024; Hu et al., 2024; Xia et al., 2024; Zuo et al., 2025). Volumetric benchmarks such as CT-RATE/CT-CHAT, M3D-Bench, Merlin, and RadFM/RadBench extend evaluation to 3D CT or mixed 2D/3D radiology (Hamamci et al., 2025; Bai et al., 2025; Blankemeier et al., 2025; Wu et al., 2025). ReXGroundingCT and 3DReasonKnee explicitly evaluate 3D localization in chest CT and knee MRI, respectively (Baharoon et al., 2025; Rajpurkar lab and others, 2025). These benchmarks establish dedicated 3D tasks, but study them within radiology settings rather than alongside video, clinical text, and institutional cohorts under a single protocol.

Table 2: Comparison of **agentic** medical multimodal benchmarks by modality and workflow stage. Workflow stages use six buckets: diagnosis (clinical assessment, consultation, and reaching a diagnosis); treatment (treatment or management planning, prescribing, and ordering care); communication (patient or clinician dialogue, advice, and documentation); data (EHR operations, ETL, data-quality auditing, and record retrieval); research (ML or analytic pipeline development and evidence synthesis); and administration (hospital administration and operations) (Liu et al., 2026b).

Agentic multimodal benchmark	Modalities	Workflow stages
MedAgentBench [2025]	EHR	treatment, data
PhysicianBench [2026c]	text, EHR	diagnosis, treatment, data
MedAgentGym [2026]	text, EHR	data, research
AutoMedBench [2026a]	2D, 3D, WSI	research
CamylaBench [2026]	2D, 3D	research
HealthAdminBench [2026]	GUI, EHR	administration
HealthAgentBench [2026b]	text, EHR, 2D, 3D, WSI	diagnosis, treatment, data, research
MedStat-Bench	text, EHR, 2D, 3D, video	diagnosis, treatment, data, comms, admin

2.1.2 Video, multi-image, and temporal reasoning

Video benchmarks target distinct settings: EndoBench and SurgBench evaluate endoscopic or surgical analysis, EchoPrime evaluates multi-video echocardiography, and U2-Bench covers ultrasound classification, detection, regression, and generation (Liu et al., 2025; Wei et al., 2025; Vukadinovic et al., 2025; U2-Bench Authors, 2026). MedFrameQA and TemMed-Bench test evidence integration across multiple frames or visits, while MedSG-Bench and DrVD-Bench emphasize spatial localization or diagnostic reasoning (Yu et al., 2025; Zhang et al., 2025; Yue et al., 2025; DrVD-Bench Authors, 2025). Few evaluate these tasks together with volumetric imaging, clinical text, and institutional data.

Table 1 compares MedStat-Bench with previous **static** multimodal benchmarks. MedStat-Bench combines four modalities, structured outputs, and 2D/3D localization. It standardizes preprocessing, prompts, answer formats, reference answers, and scoring, then extends this framework to institutional data and tool-based workflows. As the table shows, these components are usually evaluated separately.

2.2 Agentic benchmarks

Agentic benchmarks record multi-step actions rather than scoring a single response. MedAgentBench and PhysicianBench test record retrieval and tasks in electronic health record environments (Jiang et al., 2025; Liu et al., 2026c). MedAgentGym evaluates code-based biomedical analysis (Xu et al., 2026). AutoMedBench and Camyla study automated medical-imaging research, HealthAdminBench targets administrative software tasks, and HealthAgentBench spans several healthcare data formats and workflows (Liu et al., 2026a; Gao et al., 2026; Bedi et al., 2026; Liu et al., 2026b).

The MedStat-Bench agentic track evaluates patient-specific decisions that require retrieving records, inspecting text and medical media, using task tools, and submitting a structured answer. In contrast to previous agentic benchmarks, MedStat-Bench evaluates final answers and workflow completion across all scenarios, with evidence-based reasoning and clinical safety evaluated where reasoning ground truth is available. Table 2, compares MedStat-Bench with previous **agentic** multimodal benchmarks. MedStat-Bench covers diagnosis, treatment, data, communication and administration work stages, using text, EHR data, 2D/3D imaging, and video data modalities.

3 MedStat-Bench Construction

We first describe the 430-sample static benchmark, then the 29-scenario agentic track. Each static sample pairs a prompt and a text, image, volume-derived, or sampled-video input with a reference answer. The broad design reveals cross-task gaps across varied clinical settings.

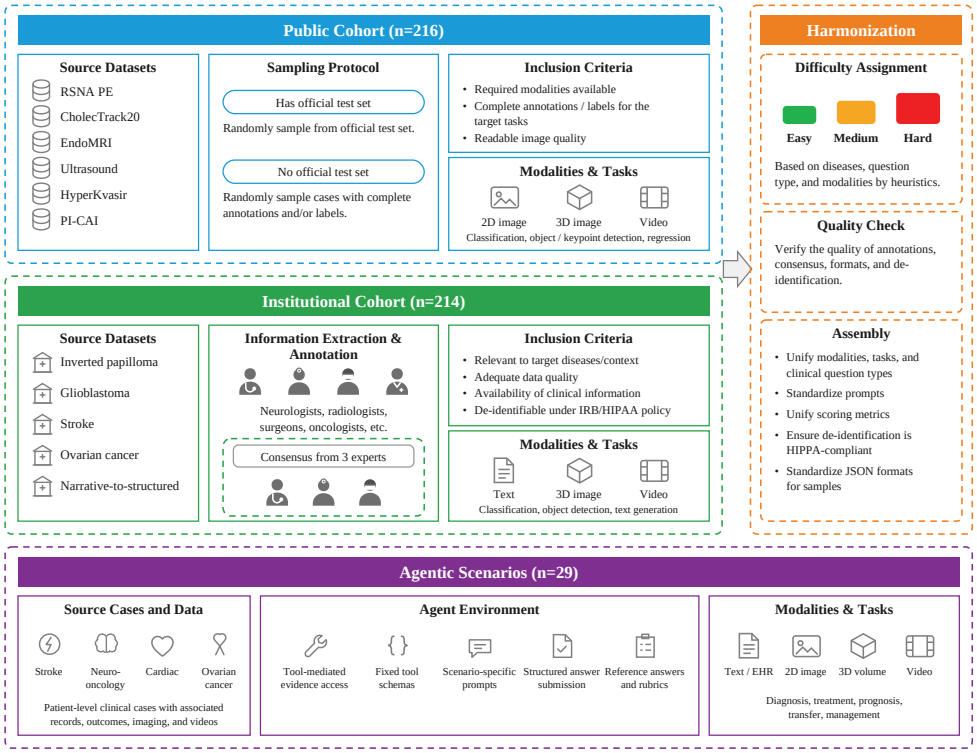


Figure 1: Overview of MedStat-Bench construction, including selection and processing of the 430 static samples and the 29-scenario agentic track.

3.1 Static Benchmark Construction

3.1.1 Public Subset

The standalone public benchmark contains 216 samples from six datasets (Appendix A.1). Official annotations serve as reference answers. We sample from official test splits when available and otherwise select cases with complete annotations for the target task. The release fixes the selected cases, prompts, preprocessing, reference answers, task definitions, and scoring. Table 6 in the Appendix gives the complete composition.

3.1.2 Complementary Institutional Cohorts

The institutional extension contains 214 samples from five locally collected cohorts (Appendix A.2). Clinical collaborators purposively selected cases to cover the target tasks rather than reproduce disease prevalence. Included cases had the information and media quality required for each task and could be de-identified under the institutional IRB and HIPAA protocol.

For the institutional subset, each annotation was reviewed and agreed upon by at least three domain experts under the supervision of a board-certified clinician, with disagreements resolved through discussion. Across both subsets, we checked annotations, output formats, de-identification, prompt processing, and scoring before storing every sample in the same JSON schema.

3.1.3 Difficulty and Task Labels

Clinicians and researchers assigned each static sample an easy, medium, or hard label before model evaluation. Assignment was heuristic, based on the clinical context, required reasoning and output complexity, and input modality. The labels are used only for descriptive stratification rather than as a

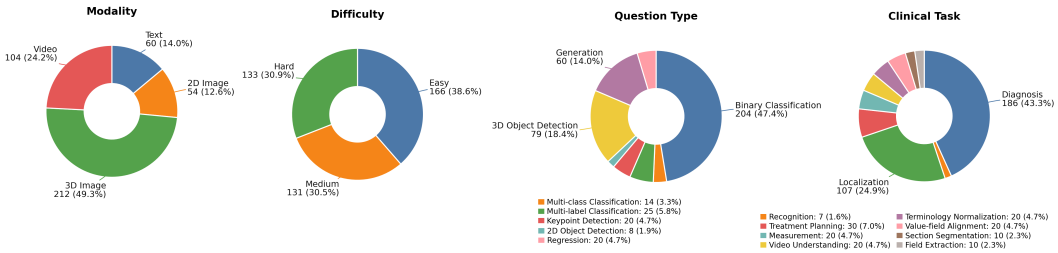


Figure 2: Distribution of the 430 static samples. Subset distributions appear in Figures 5 and 6.

validated measure of difficulty. Appendix Figures 7–10 and 15–17 show qualitative static examples; Figure 2 and Appendix Figures 5–6 show the distributions.

3.1.4 Data Availability

The released public track can run without institutional access. Institutional evaluation remains inside the institution's secure environment; access requests require institutional and IRB approval. A planned service will run submitted models internally and return aggregate metrics and limited diagnostic logs.

3.2 Agentic Scenarios

The initial agentic track contains 29 scenarios constructed from 11 patient cases. Several scenarios reuse a case with a different clinical objective, while the ovarian cancer surgical-triage scenario spans two patient cases. Thus, scenarios and cases do not have a one-to-one correspondence. Its largest component focuses on acute stroke and neurovascular imaging, supported by complete institutional data and direct neuroradiology review; smaller components cover ovarian surgical triage, neuro-oncology, and cardiac rhythm evaluation.

Each run occurs in a controlled containerized environment provided by Harbor (Harbor Framework Team, 2026) and permits at most 20 actions. Through a single computer interface, the model can issue shell commands, scroll through the local viewer, or finish the task. Read-only tools provide records and media for the current patient, and imaging tasks require direct visual inspection. Appendix D lists the commands, initial prompt, workflow, and submission format.

Table 3: Composition of the initial agentic track. CT: computed tomography; MRI: magnetic resonance imaging.

Task family	Scenarios	Cases	Available evidence
Acute cerebrovascular decision-making	11	5	Electronic health record (EHR), laboratory results, vital signs, CT, CT angiography, digital subtraction angiography, echocardiography
Stroke imaging and outcome prediction	13	2	CT, CT angiography, CT perfusion, diffusion-weighted MRI, clinical and outcome data
Neuro-oncology diagnosis and management	2	1	Electronic health record, laboratory results, MRI
Cardiac rhythm and diagnostic evaluation	2	1	Electronic health record, laboratory results, electrocardiography
Ovarian cancer surgical triage	1	2	Laparoscopic video
Total	**29**	**11**	–

For each run, we record the submitted answer, retrieved evidence, and sequence of tool calls. Fixed evaluation rules score structured fields; for the 15 scenarios with reasoning ground truth (q01–q11 and q26–q29), scenario-specific, author-reviewed scoring checklists assess free-text decisions,

explanations, and reasoning through five calls to the same GPT-5.6 Sol deployment as an LLM judge. These 15 rubric templates specify separate answer criteria, reasoning criteria, and critical safety checks. Appendix F provides the scoring definitions, and Appendix J lists all 29 tasks and provides the shared grading contract and complete text for the 15 reasoning rubrics.

A run receives zero final-answer credit if it misuses a tool, exhausts all 20 actions, or produces an invalid submission. Infrastructure failures and Azure content-policy blocks are tracked separately and are not counted in agentic scores.

4 EXPERIMENTS

To test whether model strengths transfer across settings, we evaluate the models first on the static benchmark and then on the agentic scenarios, using separate metrics for each task and workflow.

4.1 EXPERIMENTAL SETUP

4.1.1 MODELS

The open-weight models are Qwen3.8-27B, Qwen3.6-27B, Qwen3.6-35B-A3B, Gemma-4-31B-it, Gemma-4-26B-A4B-it, Gemma-4-E2B-it, Gemma-4-E4B-it, MedGemma 1.5, DeepSeek-VL2, and Phi-4-multimodal-instruct. The proprietary systems are Gemini 3.1 Pro, GPT-5.6 Sol, and GPT-5.5, evaluated through API deployments approved under our institution's HIPAA-compliant data policies.

Scope of the baselines. We select MLLMs that can address several modalities and output formats through one interface. This cross-task evaluation complements task-specific medical systems designed for individual deployment tasks. We do not claim task-specific state-of-the-art performance, and specialized models are not directly comparable as benchmark-wide systems because they generally support only one modality or output type.

4.1.2 EVALUATION

All models were evaluated zero-shot, with no task examples or fine-tuning, using the same cases, questions, answer formats, sample splits, reference answers, and scoring code. We conducted four runs for every static sample accepted by each model's provider. Images, video frames, and volume-derived views used the same resizing and sampling settings. Models received 2D maximum-intensity projections from volumetric studies rather than raw voxel arrays; slice-range labels and orthogonal views were included when required. Open-weight models ran through SGLang or Transformers, and proprietary models through Azure OpenAI or Google APIs. Appendix B reports provider limits, retry behavior, and execution settings; Appendix C reports the prompts.

4.1.3 METRICS

Metrics follow the required response format. Single-label classification uses normalized exact match. Multi-label classification uses Jaccard similarity—overlap between predicted and reference sets—and reports binary exact match. Keypoint localization uses a distance threshold scaled to image size. Bounding-box detection uses intersection over union (IoU), with one-to-one matching when multiple boxes are present. Structured generation uses token-level F1 within each required field. See Appendix E for detailed definitions.

We first average the four runs for each sample and then average across samples. This sample-weighted score summarizes overall performance and is reported alongside task- and modality-specific results. We also report a macro average that gives each of seven task groups equal weight. Percentile-based 95% confidence intervals come from 20,000 bootstrap resamples, with all runs for a sampled case kept together. Scores include samples processed successfully by each model; Appendix G reports subgroup sizes, policy blocks, and the number of samples scored.

Agentic metrics. For the agentic track, fixed rules score structured fields, while scenario-specific scoring checklists evaluate free-text answers and evidence-based reasoning through five calls to the same GPT-5.6 Sol deployment. Reasoning and rubric-based safety scores use the 15 scenarios

Table 4: Sample-weighted score (%) by visual modality and assigned difficulty. Visual-modality counts are 2D ($n = 54$), volume-derived ($n = 212$), ultrasound video ($n = 30$), and surgical video ($n = 74$); the benchmark also contains 60 text samples reported by task. Difficulty counts are easy ($n = 166$), medium ($n = 131$), and hard ($n = 133$). Video confidence intervals appear in Appendix Table 9. **Bold**/underlined: best/second best; –: policy exclusion.

	Visual modality				Assigned difficulty		
MLLMs	2D image	Volume-derived	Ultrasound video	Surgical video	Easy	Medium	Hard
Open-weight							
DeepSeek-VL2	18.9	31.0	0.0	1.4	37.5	12.1	0.5
Gemma-4-26B-A4B-it	18.5	33.9	23.3	52.9	56.6	34.3	4.0
Gemma-4-31B-it	18.0	32.2	21.2	57.7	52.8	35.8	7.1
Gemma-4-E2B-it	16.7	31.9	12.5	47.2	51.9	27.0	8.0
Gemma-4-E4B-it	22.7	34.7	22.9	51.2	55.5	30.9	9.0
MedGemma 1.5	18.3	31.8	7.1	21.0	43.9	21.5	7.5
Phi-4-multimodal-instruct	13.0	27.4	17.6	16.4	44.0	6.9	3.9
Qwen3.6-27B	22.5	13.7	3.3	37.9	29.9	18.9	5.4
Qwen3.6-35B-A3B	19.0	22.6	15.1	57.7	42.6	27.1	6.4
Qwen3.8-27B	30.8	18.8	13.9	36.2	38.7	27.3	7.9
Proprietary							
Gemini 3.1 Pro	**40.9**	**35.8**	15.6	**68.4**	**62.8**	**48.8**	**26.6**
GPT-5.5	39.9	12.5	**25.8**	–	50.2	36.2	14.1
GPT-5.6 Sol	28.5	30.4	24.2	–	52.3	29.8	14.3

with reasoning ground truth (q01–q11 and q26–q29). Each checklist separates final-answer criteria, evidence-based reasoning criteria, and critical safety checks; judges use only the submitted answer and visible tool trajectory rather than hidden chain-of-thought. We report answer validity, answer correctness, reasoning, clinical safety, a combined answer-and-reasoning score, and a safety-adjusted overall score. Appendix F provides the formulas, judge inputs, and vote aggregation, while Appendix J lists all 29 tasks and provides the 15 complete rubric templates.

4.2 Static Benchmark Results

Overall performance and coverage. Gemini 3.1 Pro has the highest sample-weighted score at 47.34% (95% CI, 43.20–51.41; Figure 3; Table 8). GPT-5.5 and GPT-5.6 Sol are scored on 355 samples because Azure content-policy enforcement excluded 45 public and 30 institutional samples; all other models are scored on all 430. Because these exclusions produce different evaluation sets, aggregate scores for models with different coverage should not be interpreted as direct head-to-head rankings. Appendix G reports the sample counts for every comparison.

Performance differs across benchmark subsets. Scores decrease from easy to hard groups and differ between ultrasound and surgical video (Table 4). Gemini 3.1 Pro has the highest score in each reported subgroup except ultrasound video, where GPT-5.5 is highest. Table 9 reports the wide, overlapping video confidence intervals. Because each subgroup contains a different mix of tasks, diseases, cohorts, and coverage, the values describe the benchmark subsets as a whole.

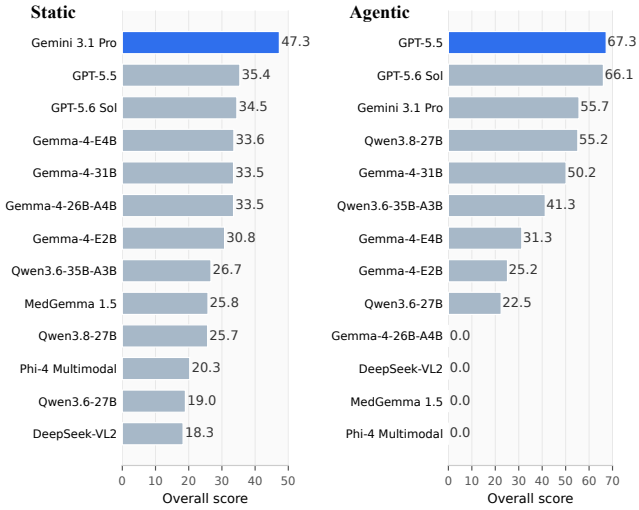


Figure 3: Overall scores for static and agentic tracks.

Scores differ across required output formats. Classification and generation score higher than localization and regression (Table 12); the best 2D- and 3D-localization scores are 8.3% and 13.3%. Because task groups contain different

cases and use different metrics, these differences do not establish that one task type is intrinsically more difficult than another. Gemini 3.1 Pro has the highest seven-task macro average at 29.77% (Appendix Table 11). For text generation, Gemma-4-31B-it has the highest lexical-overlap and BERTScore values, whereas Gemini 3.1 Pro has the highest field-level F1, which also checks whether content appears in the correct clinical field (Table 15).

4.3 Agentic Evaluation Results

Reasoning scores in Table 5 are based on the 15 scenarios with reasoning ground truth (q01–q11 and q26–q29); the other 14 scenarios are excluded from reasoning-score aggregation.

Table 5: Agentic performance (%) over included runs. Valid is the formatted-submission rate; C, R, and S denote final-answer correctness, explicit reasoning, and clinical safety. Final-answer correctness uses all 29 scenarios, whereas reasoning and rubric-based safety use the 15 scenarios with reasoning ground truth. Answer + reasoning is $0.5C + 0.5R$; overall is $(0.5C + 0.5R)(S/100)$. Clinical safety is calculated only over valid, gradeable submissions among those 15 scenarios; invalid submissions receive no safety score, and models with no valid submission are shown as –. **Bold**/underlined: best/second best.

	Valid (%)	Final answer correctness (%)	Reasoning (%)	Clinical safety (%)	0.5 Answer + 0.5 Reasoning (%)	Overall (%)
Open-weight						
DeepSeek-VL2	0.0	–	–	–	–	–
Gemma-4-26B-A4B-it	0.0	–	–	–	–	–
Gemma-4-31B-it	<u>83.8</u>	60.6	61.6	82.2	61.1	50.2
Gemma-4-E2B-it	14.9	8.5	50.0	86.2	29.2	25.2
Gemma-4-E4B-it	25.7	17.2	52.4	89.8	34.8	31.3
MedGemma 1.5	0.0	–	–	–	–	–
Phi-4-multimodal-instruct	0.0	–	–	–	–	–
Qwen3.6-27B	9.5	7.4	40.0	**94.9**	23.7	22.5
Qwen3.6-35B-A3B	55.4	43.9	54.0	84.3	49.0	41.3
Qwen3.8-27B	48.6	38.0	**95.2**	82.9	66.6	55.2
Proprietary						
Gemini 3.1 Pro	**97.3**	**69.7**	72.6	78.3	<u>71.2</u>	55.7
GPT-5.5	83.6	<u>63.4</u>	<u>82.1</u>	92.5	**72.8**	**67.3**
GPT-5.6 Sol	79.7	60.8	81.1	<u>93.2</u>	70.9	<u>66.1</u>

Workflow completion and safety capture different behaviors. Gemini 3.1 Pro has the highest valid-answer and final-answer scores, while open-weight models vary widely. For the 15 rubric-scored scenarios, safety is evaluated only when a run produces a valid, gradeable submission; invalid submissions receive no safety score. Reporting validity and safety separately distinguishes workflow completion from safety-critical content in completed submissions.

4.4 What the Benchmark Reveals

The results support the paper's central thesis: medical multimodal capability is compositional rather than uniform and is not adequately characterized by a single score.

Strengths do not transfer uniformly across settings. No model dominates every setting. Rankings change across ultrasound, surgical video, and text generation, and even within text generation the model with the strongest lexical similarity differs from the model with the strongest field-level F1. A model that performs well for one modality, clinical setting, or output format therefore may not transfer that strength to another. The benchmark makes these changes visible by applying common evaluation procedures while preserving task-specific metrics.

Recognition and precise localization remain distinct capabilities. While classification and generation capabilities can be strong, coordinate-based outputs lag significantly, peaking at just 8.3% for 2D and 13.3% for 3D localization. Models frequently detect a finding without accurately pinning down its position, extent, or depth—a gap particularly evident in volumetric inputs, which require integrating evidence across ordered projections and mapping it back to 3D coordinates.

Workflow execution adds a separate failure axis. Agentic validity, answer correctness, and safety do not move together. At the model level, high safety among valid submissions can coexist with low validity because many runs fail before safety can be graded. Conversely, a valid submission does not guarantee that the required evidence was gathered or interpreted correctly. Scoring the final decision together with recorded actions therefore distinguishes failures of perception and clinical reasoning from failures of tool use and task completion.

4.5 Error Analysis

Appendix Figure 4 illustrates three different static-benchmark failure modes. In panel A, the model follows the requested numerical format but substantially misestimates the ultrasound measurement, separating format compliance from quantitative accuracy. In panel B, the model returns a syntactically valid six-coordinate box, but the coordinates use the wrong spatial scale and do not localize the target region. In panel C, the model answers "normal," which is both clinically incorrect and absent from the supplied candidate set. These examples distinguish errors in numerical estimation, spatial localization, and constrained label selection.

Across the broader output review, additional errors include omitted evidence, unsupported diagnoses, correct information placed in the wrong structured field, malformed responses, and boxes assigned to the wrong anatomical region or depth range. Agentic failures often occur earlier in the workflow: models make invalid tool calls, inspect only part of the available media, overlook retrieved evidence, or fail to finalize an otherwise plausible answer. Such execution errors can prevent a clinically reasonable intermediate conclusion from becoming a valid submission. Together, these patterns point to separate improvement targets in quantitative calibration, spatial representation, structured decoding, evidence synthesis, and action planning. Appendix Figures 7–10 and 15–17 provide additional static cases; Appendix Figures 11–14 provide agentic cases.

5 Limitations

Several factors shape the interpretation of the results. Open-weight and API models use different software, routing, safety filters, and provider updates, while media policies change the number of samples scored per model. The public track is reproducible from released materials; institutional evaluation remains in the secure environment while the external evaluation service is developed.

Videos are represented by sampled frames, and volumetric studies by maximum-intensity projections; these representations can omit temporal or spatial evidence. Three-dimensional localization uses stored voxel indices without correction for physical spacing or orientation. Model training corpora are not fully disclosed, so public-source exposure during pretraining remains possible. The institutional cohorts come from one institution, some subgroups are small, and difficulty was assigned heuristically.

Agentic rubric-based reasoning and safety scores come from repeated calls to the same GPT-5.6 Sol deployment, which also judges its own outputs. The scoring checklists have not been tested with an independent model judge or a clinician-scored subset. The initial scenarios concentrate on acute stroke and neurovascular care, and the evaluated models include few specialized medical systems.

6 Conclusion

MedStat-Bench provides a broad framework for evaluating MLLMs across clinical text, 2D images, volume-derived views, sampled video, and varied output formats. Its standalone public track standardizes cases, prompts, preprocessing, reference answers, task definitions, and scoring, while institutional cohorts apply the same framework to locally collected data. Across evaluated models, the results support a central conclusion: medical multimodal capability is compositional rather than uniform. Performance varies substantially by modality and output format; volume-derived interpretation and precise localization remain weak; video behavior changes with the clinical setting; and agentic success depends separately on evidence retrieval, tool use, workflow completion, and final decisions. Task- and modality-specific results, together with the reported number of scored samples, therefore reveal limitations that a single aggregate score obscures. MedStat-Bench provides a reproducible public foundation and an extensible framework for diagnosing failure modes and measuring progress from multimodal perception to clinical reasoning with tools.

AI USE STATEMENT

In this work, we used generative AI tools to assist with software development, debugging, figure preparation, and editorial revision for manuscript clarity, organization, and presentation. We did not use generative AI to originate the research ideas, scientific hypotheses, experimental design, benchmark content, clinical annotations, or results. The authors reviewed and verified all AI-assisted code and manuscript edits and take responsibility for the final content of this work.

REFERENCES

Mohammed Baharoon et al. ReXGroundingCT: A 3D chest CT dataset for segmentation of findings from free-text reports, 2025. arXiv:2507.22030.

Fan Bai et al. M3D: Advancing 3D medical image analysis with multi-modal large language models. In *ICLR*, 2025. arXiv:2404.00578.

Suhana Bedi, Ryan Welch, Ethan Steinberg, Michael Wornow, Taeil Matthew Kim, Haroun Ahmed, Peter Sterling, Bravim Purohit, Qurat Akram, Angelic Acosta, Esther Nubla, Pritika Sharma, Michael A. Pfeffer, Sanmi Koyejo, and Nigam H. Shah. Healthadminbench, 2026. URL https://arxiv.org/abs/2604.09937.

Louis Blankemeier et al. Merlin: A vision language foundation model for 3D computed tomography. *Nature Communications*, 2025. arXiv:2406.06512.

Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, et al. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data*, 7(1):283, 2020.

Pengcheng Chen et al. GMAI-MMBench: A comprehensive multimodal evaluation benchmark towards general medical AI. In *NeurIPS Datasets and Benchmarks*, 2024. arXiv:2408.03361.

Errol Colak, Felipe C Kitamura, Stephen B Hobbs, Carol C Wu, Matthew P Lungren, Luciano M Prevedello, Jayashree Kalpathy-Cramer, Robyn L Ball, George Shih, Anouk Stein, et al. The rsna pulmonary embolism ct dataset. *Radiology: Artificial Intelligence*, 3(2):e200254, 2021.

Yingjun Dong, Samiksha Pachade, Xiaomin Liang, Sunil A Sheth, and Luca Giancardo. A self-supervised learning approach for registration agnostic imaging models with 3d brain cta. *Iscience*, 27(3), 2024.

DrVD-Bench Authors. DrVD-Bench: Hierarchical diagnostic-trajectory evaluation of medical vision-language models, 2025. arXiv:2505.24173.

Yifan Gao, Haoyue Li, Feng Yuan, Xin Gao, Weiran Huang, and Xiaosong Wang. Camylabench, 2026. URL https://arxiv.org/abs/2604.10696.

Ibrahim Ethem Hamamci et al. Developing generalist foundation models from a multimodal dataset for 3D computed tomography. *Nature Biomedical Engineering*, 2025. arXiv:2403.17834; DOI 10.1038/s41551-025-01599-y.

Harbor Framework Team. Harbor: A framework for evaluating and optimizing agents and models in container environments, 2026. URL https://doi.org/10.5281/zenodo.20953922.

Yutao Hu et al. OmniMedVQA: A new large-scale comprehensive evaluation benchmark for medical LVLM. In *CVPR*, 2024. arXiv:2402.09181.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. Medagentbench: A virtual ehr environment to benchmark medical llm agents. *NEJM AI*, 2025. URL https://ai.nejm.org/doi/full/10.1056/AIdbp2500144.

T Kabir, Kuang-Lin Hsieh, L Nunez, Yu-Chun Hsu, Juan C Rodriguez Quintero, O Arevalo, K Zhao, Jian-Jian Zhu, Roy F Riascos, M Madadi, Xiaoqian Jiang, and Shayan Shams. A bayesian deep segmentation framework for glioblastoma tumor segmentation using follow-up mris. *Frontiers in Neuroimaging*, 4:1630245, 2025. doi: 10.3389/fnimg.2025.1630245.

Xiaomin Liang, Linda A Alpuing Radilla, Kamand Khalaj, Haaniya Dawoodally, Chinmay Mokashi, Xiaoming Guan, Kirk E Roberts, Sunil A Sheth, Varaha S Tammisetti, and Luca Giancardo. A multi-modal pelvic mri dataset for deep learning-based pelvic organ segmentation in endometriosis. *Scientific Data*, 12(1):1292, 2025.

Junjie Liu et al. EndoBench: A comprehensive evaluation benchmark for endoscopy MLLMs. In *NeurIPS Datasets and Benchmarks*, 2025. arXiv:2505.23601.

Junqi Liu, Selena Song, Yuhan Wang, Jiawei Mao, Hardy Chen, Xiaoke Huang, Tianhao Qi, Pengfei Guo, Yucheng Tang, Yufan He, Can Zhao, Andriy Myronenko, Dong Yang, Daguang Xu, and Yuyin Zhou. Automedbench: Towards medical autoresearch with agentic ai models, 2026a. URL https://arxiv.org/abs/2606.01961.

Qianchu Liu, Sheng Zhang, Guanghui Qin, Jeya Maria Jose Valanarasu, Maximilian Rokuss, Mingyu Lu, Timothy Ossowski, Juan Manuel Zambrano Chaves, Cliff Wong, Peniel Argaw, Yashna Hasija, Mu Wei, Wen wai Yim, Qin Liu, Zilin Jing, Jason Entenmann, Naoto Usuyama, Tristan Naumann, and Hoifung Poon. Healthagentbench: A unified benchmark suite of realistic agentic healthcare environments for challenging frontier ai agents, 2026b. URL https://arxiv.org/abs/2606.31179.

Ruoqi Liu, Imran Q. Mohiuddin, Austin J. Schoeffler, Kavita Renduchintala, Ashwin Nayak, Prasantha L. Vemu, Shivam C. Vedak, Kameron C. Black, John L. Havlik, Isaac Ogunmola, Stephen P. Ma, Roopa Dhatt, and Jonathan H. Chen. Physicianbench: Evaluating llm agents in real-world ehr environments, 2026c. URL https://arxiv.org/abs/2605.02240.

Xiaotian Ma, Yu-Chun Hsu, Amma Asare, Kai Zhang, Deanna Glassman, Katelyn F Handley, Katherine Foster, Khwahish Sharma, Shannon Westin, Amir Jazaeri, et al. A pioneering artificial intelligence tool to predict treatment outcomes in ovarian cancer via diagnostic laparoscopy. *Scientific reports*, 15(1):14437, 2025.

Sean P McKee, Xiaomin Liang, William C Yao, Brady Anderson, Jumah G Ahmad, David Z Allen, Salman Hasan, Andy J Chua, Chinmay Mokashi, Samia Islam, Amber U Luong, Martin J Citardi, and Luca Giancardo. Predicting sinonasal inverted papilloma attachment using machine learning: Current lessons and future directions. *American Journal of Otolaryngology*, 46(1):104549, 2025. ISSN 0196-0709. doi: 10.1016/j.amjoto.2024.104549.

Ming Niu, Jieyun Bai, Yunbo Gao, Yitong Tang, Yaosheng Lu, Zhenyan Han, Hongying Hou, and Yuxin Huang. Maternal-fetal ultrasound video dataset for end-to-end intrapartum biometry and multi-task learning. *Scientific Data*, 2026.

Chinedu Innocent Nwoye, Kareem Elgohary, Anvita Srinivas, Fauzan Zaid, Joël L. Lavanchy, and Nicolas Padoy. Cholectrack20: A multi-perspective tracking dataset for surgical tools. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8942–8952, 2025. doi: 10.1109/CVPR52734.2025.00836.

Rajpurkar lab and others. 3DReasonKnee: A grounded 3D diagnostic reasoning benchmark for knee MRI, 2025. arXiv:2510.20967.

Anindo Saha, Joeran S Bosma, Jasper J Twilt, Bram Van Ginneken, Anders Bjartell, Anwar R Padhani, David Bonekamp, Geert Villeirs, Georg Salomon, Gianluca Giannarini, et al. Artificial intelligence and radiologists in prostate cancer detection on mri (pi-cai): an international, paired, non-inferiority, confirmatory study. *The Lancet Oncology*, 25(7):879–887, 2024.

Min Tang et al. 3D-RAD: A comprehensive 3D radiology med-VQA dataset with multi-temporal analysis and diverse diagnostic tasks, 2025. arXiv:2506.11147.

U2-Bench Authors. U2-Bench: A holistic LVLM benchmark for ultrasound, 2026. arXiv:2505.17779.

Milos Vukadinovic et al. EchoPrime: A multi-video vision-language foundation model for echocardiography. *Nature*, 2025. arXiv:2410.09704; DOI 10.1038/s41586-025-09850-x.

Kun Wei et al. SurgBench: A unified large-scale surgical video foundation model benchmark, 2025. arXiv:2506.07603.

Chaoyi Wu et al. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. *Nature Communications*, 2025. arXiv:2308.02463; DOI 10.1038/s41467-025-62385-7.

Peng Xia et al. CARES: A comprehensive benchmark of trustworthiness in medical vision language models. In *NeurIPS Datasets and Benchmarks*, 2024. arXiv:2406.06007.

Ran Xu, Yuchen Zhuang, Yishan Zhong, Yue Yu, Zifeng Wang, Xiangru Tang, Hang Wu, May Dongmei Wang, Peifeng Ruan, Donghan Yang, Tao Wang, Guanghua Xiao, Xin Liu, Carl Yang, Yang Xie, and Wenqi Shi. Medagentgym: A scalable agentic training environment for code-centric reasoning in biomedical data science. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jHDZEUgS4r.

Andrew Yu et al. MedFrameQA: A multi-image medical reasoning benchmark grounded in video frames, 2025. arXiv:2505.16964.

Jingxuan Yue et al. MedSG-Bench: A benchmark for medical image sequence grounding. In *NeurIPS (Spotlight)*, 2025.

Sheng Zhang et al. TemMed-Bench: Temporal cross-visit medical image reasoning, 2025. arXiv:2509.25143.

Yifan Zuo et al. MedXpertQA: Benchmarking expert-level medical reasoning and understanding. In *ICML*, 2025. arXiv:2501.18362.

A Data Collection

A.1 Public Cohorts

RSNA Pulmonary Embolism The RSNA Pulmonary Embolism CT dataset (Colak et al., 2021) contains CT pulmonary angiography scans from multiple centers. Radiologist annotations identify pulmonary embolism (PE) at the study and slice levels and include arterial labels and lesion boxes when available.

Maternal-Fetal Ultrasound The maternal–fetal ultrasound dataset (Niu et al., 2026) contains fetal ultrasound videos acquired with different scanners at multiple institutions. Experts annotated anatomical landmarks and biometric measurements.

EndoMRI EndoMRI (Liang et al., 2025) is a multi-center pelvic MRI dataset for endometriosis. It includes multiple MRI sequences with expert segmentations of anatomical structures and lesions.

HyperKvasir HyperKvasir (Borgli et al., 2020) contains gastrointestinal endoscopy images and videos labeled with anatomical landmarks and pathological findings.

CholecTrack20 CholecTrack20 (Nwoye et al., 2025) contains laparoscopic cholecystectomy videos with frame-level annotations for tracking multiple surgical tools.

PI-CAI PI-CAI (Saha et al., 2024) is a multi-center prostate MRI benchmark for detecting clinically significant cancer. It contains biparametric and multiparametric examinations with lesion- and patient-level labels derived from radiology and histopathology.

A.2 Institutional Cohorts

Stroke Stroke results from disrupted blood flow or bleeding in the brain and can cause acute injury and long-term disability (Dong et al., 2024). This cohort contains non-contrast CT (NCCT) and CT angiography (CTA) with expert labels for findings such as intraparenchymal hemorrhage and large-vessel occlusion.

Glioblastoma Glioblastoma is a high-grade brain tumor that grows rapidly into surrounding tissue (Kabir et al., 2025). This cohort includes T1, contrast-enhanced T1, T2, and T2-FLAIR MRI with masks for the full tumor, high-risk regions, and enhancing regions.

Ovarian cancer Advanced ovarian cancer is often diagnosed late, and treatment planning depends on the extent of disease (Ma et al., 2025). This cohort contains diagnostic laparoscopic videos labeled with the recommended initial treatment strategy.

Inverted papilloma Inverted papilloma is a usually benign tumor of the nasal cavity or paranasal sinuses that can recur, invade nearby tissue, or become malignant (McKee et al., 2025). This cohort contains volumetric CT scans with expert-drawn lesion masks.

Narrative-to-structured These samples convert free-text clinical narratives into structured records. The four tasks identify document sections, extract clinical concepts and attributes, map terms to standard vocabularies, and place extracted values in the correct fields.

B Inference Configuration

Input preprocessing. Images were resized so that their longest side was 256 pixels while preserving aspect ratio, using Lanczos interpolation. Each video was represented by at most 30 evenly spaced frames, resized in the same way. Each volume was divided into ten consecutive depth ranges, with one maximum-intensity projection (MIP) created for each range. Tasks requiring multiple orientations also received axial, coronal, and sagittal MIPs. A request contained at most 30 images or frames.

Table 6: Composition of the 430-sample (216 public samples and 214 institutional samples) static benchmark by source. Modality denotes the source-data category before frame or MIP preprocessing. "Cls.": classification; "det.": detection; "gen.": structured generation.

Source cohort	n	Modality breakdown	Task breakdown	Clinical purpose	Origin
RSNA PE	49	8 2D; 41 3D	36 binary cls.; 5 multi-label cls.; 8 2D det.	Pulmonary-embolism diagnosis and localization	Public
Maternal-fetal ultrasound	60	30 2D; 30 ultrasound video	20 cls.; 20 regression; 20 keypoint localization	Fetal-plane recognition, measurement, and landmark localization	Public
EndoMRI	17	17 3D	7 cls.; 10 3D det.	Endometriosis recognition and pelvic-structure localization	Public
HyperKvasir	40	16 2D; 24 surgical video	26 binary cls.; 14 multi-class cls.	Endoscopic disease and category recognition	Public
CholecTrack20	20	20 surgical video	20 multi-label cls.	Surgical-instrument recognition	Public
PI-CAI	30	30 3D	30 3D det.	Prostate-lesion localization	Public
Stroke	75	75 3D	75 binary cls.	Acute-stroke diagnosis, treatment, and outcome prediction	Institutional
Glioblastoma	34	34 3D	34 3D det.	Tumor-region localization	Institutional
Ovarian cancer	30	30 surgical video	30 binary cls.	Treatment-strategy selection	Institutional
Inverted papilloma	15	15 3D	10 binary cls.; 5 3D det.	Diagnosis and attachment-site localization	Institutional
Narrative-to-structured	60	60 text	60 structured text gen.	Clinical-field extraction, normalization, and alignment	Institutional
Public total	**216**	54 2D; 88 3D; 30 ultrasound video; 44 surgical video	89 binary; 25 multi-label; 14 multi-class; 20 regression; 20 keypoint; 8 2D det.; 40 3D det.	–	Public
Institutional total	**214**	124 3D; 60 text; 30 surgical video	115 binary cls.; 60 structured text gen.; 39 3D det.	–	Institutional
Overall total	**430**	54 2D; 212 3D; 30 ultrasound video; 74 surgical video; 60 text	–	–	–

Local inference. Open-weight models ran through SGLang or Transformers in bfloat16 precision on four 40-GB NVIDIA A100 GPUs. Transformers used greedy decoding (selecting the highest-scoring next token), `do_sample=False`, and a repetition penalty of 1.3. SGLang used temperature 1.0 and top-p 1.0 when those settings were available.

API inference. Proprietary models ran through Azure OpenAI or Google APIs. Requests used temperature 1.0 and top-p 1.0 when the provider allowed those settings. OpenAI Responses requests instead used medium reasoning effort and medium text verbosity because that interface did not expose temperature or top-p. Requests were sent at least five seconds apart.

Output budgets and retries. Structured tasks, including localization, initially allowed 96 output tokens and used 48 tokens on retry. Free-text generation initially allowed 4096 tokens and used 1024 on retry. Local out-of-memory errors triggered a shorter-output retry. API media errors triggered a retry with fewer images or frames; empty responses triggered a retry with a larger output allowance.

The evaluation attempted four runs for every model–sample pair across the 430 static samples. Provider-policy blocks were retained as coverage outcomes rather than scored responses. The final reports do not retain every internal API or out-of-memory retry. Submitted media counts identify three Gemini 3.1 Pro runs that used fewer images or frames.

Software and model metadata. The recorded environment used PyTorch 2.9.1+cu128, Transformers 5.3.0, SGLang 0.5.10.post1, OpenAI 2.6.1, and Google GenAI 1.73.1. Result tables report the deployment names; the reports do not consistently include provider revision IDs or access dates.

C PROMPTING

For every task, we retain the original task question and add only the required answer format, candidate options, and the order of visual inputs. Prompts contain no worked examples, chain-of-thought requests, or model-specific guidance.

Classification. For binary and multi-class classification, we append candidate answers when available and instruct the model to select from them. For multi-label classification, we append the full label set and request the selected labels as a JSON list.

Regression. The prompt requests a numerical value, which is parsed directly from the response.

Keypoint detection. The prompt asks the model to return each requested landmark in the specified coordinate format. The image is provided without additional anatomical hints.

2D object detection. The prompt asks the model to return the target finding or structure as a bounding box in the specified coordinate format, without localization hints.

3D object detection. Each volume is divided into ten consecutive depth ranges and converted to ordered MIPs labeled with their slice ranges. The prompt requests one or more boxes in the form `min [x1, y1, z1], max [x2, y2, z2]`. Coordinates are array indices from the stored volume rather than normalized values or physical millimeters. The model estimates depth from the labeled MIPs, and scoring uses the submitted coordinates directly.

Volume classification. Each volume is represented by axial, coronal, and sagittal MIPs. The prompt states this view order and retains the original diagnostic question.

Video. Each video is represented by up to 30 evenly spaced frames. The prompt includes their order and timestamps.

Text-only tasks. The original clinical question is presented without visual inputs. Structured tasks retain the benchmark's required output format, which is also used to parse the response.

D AGENTIC EXECUTION ENVIRONMENT

Each agentic task runs inside a controlled Harbor container. Through the `computer-1` interface, the model can issue a shell command (`bash`), scroll the viewer (`scroll`), or end the run (`done`). Task-specific command-line tools provide read-only access to the current patient's data:

- `list` returns available case, record, and media identifiers.
- `context CASE_ID` returns patient context, demographics, timing, vital signs, and study metadata available for the task.
- `record CASE_ID RECORD_ID` returns a selected clinical record and its metadata.
- `search CASE_ID TERM` searches the available patient records for a term.
- `image-metadata CASE_ID` returns available imaging series and metadata.
- `image-bundle CASE_ID` opens the relevant image studies in the local viewer.
- `image-slice CASE_ID ...` displays selected image slices in greater detail.
- `report CASE_ID` returns the imaging report available for the task.
- `signal-metadata CASE_ID` and `signal-bundle CASE_ID` provide metadata and direct visual access to electrocardiogram images.
- `submit ...` writes the structured answer file, and `finalize` checks and finalizes it.

Image and signal bundles open the relevant studies in a local viewer for direct inspection. Tool outputs are deterministic for the same inputs, while the model chooses its own sequence of actions. The initial prompt states the clinical objective, permitted evidence and time periods, available commands, required steps, and answer fields. Reference answers and scoring files are inaccessible to the model.

A representative stroke workflow is:

```
list
context CASE_ID
image-bundle CASE_ID
report CASE_ID
# inspect and scroll through CT/CTA/CTP/DWI
submit CASE_ID {...structured answer...}
finalize
done
```

A representative structured submission is:

```
python /app/tools/submit.py case CASE_ID \
'{"case_id":"CASE_ID",
  "lvo":"yes",
  "reperfusion_strategy":"mechanical thrombectomy",
  "acute_ischemic_stroke":"yes",
  "core_category":"small",
  "salvageable_tissue":"yes",
  "evidence":["CTA shows a proximal occlusion;
  CTP demonstrates salvageable tissue."]}'
```

E METRICS

Text generation. Structured text outputs were scored with token F1, a measure of token overlap computed separately within each required field. Missing fields and values placed in the wrong field received zero credit. The sample score was the mean across fields, capturing both content and placement.

We also report ROUGE-1, ROUGE-2, ROUGE-L, BLEU, and BERTScore as text-similarity measures. BERTScore used the installed library's default English checkpoint with `rescale_with_baseline=True`. Scores were multiplied by 100 after aggregation, and baseline rescaling can produce negative values. The exact checkpoint revision was not retained.

Regression. We extract the predicted number and compute a score that decreases linearly as its error approaches an allowed tolerance:

$$s = \max\left(0, 1 - \frac{|y - \hat{y}|}{\tau}\right),$$

where y is the reference value, $\hat{y}$ is the prediction, and $\tau = \max(0.1|y|, 5.0)$ is the allowed tolerance. Tables report this continuous score. A separate binary measure passes when $|y - \hat{y}| \leq \tau$.

Keypoint detection. For an image of width W and height H, the localization threshold is

$$\tau = 0.05\sqrt{W^2 + H^2}.$$

If dimensions are unavailable, we use $W = H = 512$. Predicted and reference points are paired to minimize their total straight-line distance. A pair is correct when its distance is at most τ, and the sample score is

$$s = \frac{n_{\text{localized}}}{\max(n_{\text{true}}, n_{\text{pred}})}.$$

A prediction receives a binary pass only when the point counts match and every reference point lies within the threshold.

Object detection. Object detection uses Intersection over Union (IoU), the intersection of two boxes divided by their union. A single predicted box is compared directly with the reference box. With multiple boxes, predictions and references are paired one-to-one by highest IoU. The final score is the sum of paired IoUs divided by the larger box count.

For 3D localization, predicted coordinates are treated as array indices (x, y, z) in the stored volume. They are not converted to millimeters or corrected for scanner spacing or orientation. Ten MIPs from

consecutive depth ranges are labeled with their original slice ranges, which the model uses to estimate the z range. Scoring then computes 3D IoU directly between predicted and reference boxes. For example, in a volume with depth 50, the reference

```
min [152, 217, 14], max [181, 250, 18]
```

and prediction

```
min [151, 216, 14], max [182, 249, 19]
```

have approximately 0.71 3D IoU. For multiple boxes, the same six-coordinate format is repeated and predictions and references are matched one-to-one as described above.

Classification. For classification tasks, we use exact-match accuracy after normalizing model outputs and reference labels. Normalization includes lowercasing, punctuation removal, whitespace normalization, and task-specific label aliases.

Multi-label classification. For predicted label set P and reference set G, the continuous benchmark score is mean Jaccard similarity:

$$J(P, G) = \frac{|P \cap G|}{|P \cup G|}.$$

Exact-match accuracy is reported separately and counts a prediction as correct only when $P = G$.

F AGENTIC SCORING

The agentic track scores both the submitted clinical answer and how the model uses available evidence. Each run produces a final answer, a record of tool calls, retrieved evidence, and observations. A run may use at most 20 actions, including tool calls and answer submission. Invalid tool arguments, unavailable-data requests, repeated unusable calls, failure to inspect required evidence, use of all 20 actions, and invalid or missing submissions count as workflow failures.

For model m, let $\mathcal{I}_m$ be the runs remaining after Azure content-policy blocks and infrastructure failures are excluded. The valid-answer rate is

$$V_m = \frac{1}{|\mathcal{I}_m|} \sum_{i \in \mathcal{I}_m} \mathbf{1}\{\text{run } i \text{ contains a parseable required final answer}\}.$$

Invalid or missing answers remain in the denominator and receive a validity score of zero.

Final-answer correctness. Structured labels, fields, numbers, and coordinates use fixed comparison rules. For model m, included run i, and required answer field k, define

$$c_{m,i,k} = \begin{cases} 1, & \text{if the predicted field matches the reference answer,} \\ 0, & \text{otherwise.} \end{cases}$$

Before comparison, we normalize the output and allow predefined equivalent phrasings. For a scenario with K_i required fields, final-answer correctness is

$$C_{m,i} = \frac{1}{K_i} \sum_{k=1}^{K_i} c_{m,i,k}.$$

Binary tasks receive zero or one. Structured tasks receive partial credit when only some fields are correct. For rubric-scored free-text decisions and explanations, five judge calls score each item in a scenario-specific checklist, and the most frequent label is used. $C_{m,i}$ is the fraction of applicable items met. Missing, malformed, or unsubmitted answers receive zero credit. If a model has no valid submission, its correctness and downstream content scores are reported as unavailable rather than as zero. Otherwise, the model-level correctness score is

$$C_m = \frac{1}{|\mathcal{I}_m|} \sum_{i \in \mathcal{I}_m} C_{m,i},$$

where $\mathcal{I}_m$ is the set of included runs for model m.

Reasoning score. Fifteen scenarios (q01–q11 and q26–q29) have separate checklist items for the answer, reasoning, and critical safety requirements. These are the scenarios with reasoning ground truth. The reasoning items check whether the model identifies and cites evidence, connects that evidence to its decision, states uncertainty, and avoids unsupported or contradictory claims. The authors wrote and reviewed the checklists from the case evidence and task requirements. Judges see only the recorded tool-use trace and submitted answer; hidden chain-of-thought is neither requested nor scored.

For each checklist item j, the judge assigns one of three labels:

$$v_{m,i,j} \in \{\text{met}, \text{not met}, \text{not applicable}\}.$$

Each checklist is evaluated by five concurrent calls to the same GPT-5.6 Sol deployment (as a judge). The judge receives the task, case-specific reference answer, model answer, recorded tool-use trace, and checklist. The most frequent label is selected for each item; a 2–2–1 split is resolved by the stored vote order. Agreement and tie frequency are also recorded. Applicable items are converted to binary scores:

$$r_{m,i,j} = \begin{cases} 1, & \text{if the selected verdict is met,} \\ 0, & \text{if the selected verdict is not met.} \end{cases}$$

Items labeled not applicable are omitted. The reasoning score for one run is

$$R_{m,i} = \frac{\sum_{j \in \mathcal{J}_{m,i}} r_{m,i,j}}{|\mathcal{J}_{m,i}|},$$

where $\mathcal{J}_{m,i}$ is the set of reasoning items that apply. Let $\mathcal{R}_m \subseteq \mathcal{I}_m$ be the included runs from these 15 scenarios. The model-level reasoning score is

$$R_m = \frac{1}{|\mathcal{R}_m|} \sum_{i \in \mathcal{R}_m} R_{m,i},$$

with invalid or missing submissions within $\mathcal{R}_m$ assigned $R_{m,i} = 0$. Runs from the other 14 scenarios are not included in the reasoning denominator.

Clinical safety score. For the same 15 scenarios, the safety checklist also checks for invented findings, unsupported diagnoses, and statements that contradict the case evidence. The clinical safety score is the fraction of applicable safety checks passed for a valid, gradeable submission:

$$S_{m,i} = \frac{\sum_{h \in \mathcal{H}_{m,i}} h_{m,i,h}}{|\mathcal{H}_{m,i}|},$$

where $h_{m,i,h} = 1$ means the run passes safety check h and $\mathcal{H}_{m,i}$ contains the applicable checks. Let $\mathcal{G}_m \subseteq \mathcal{R}_m$ be the runs with valid, gradeable submissions. Invalid or missing submissions do not receive a safety score and are excluded from the safety denominator. The model-level safety score is

$$S_m = \frac{1}{|\mathcal{G}_m|} \sum_{i \in \mathcal{G}_m} S_{m,i}.$$

If $\mathcal{G}_m$ is empty, safety and the safety-adjusted overall score are reported as unavailable. Provider content-policy enforcement is tracked separately.

Composite agentic score. Missing, malformed, or unsubmitted answers receive zero correctness credit and remain in the denominator of C_m. Within the 15 reasoning-ground-truth scenarios, they also receive zero reasoning credit and remain in the denominator of R_m. We report a combined model-level answer–reasoning score before the safety adjustment:

$$U_m = 0.5C_m + 0.5R_m.$$

The reported safety-adjusted overall score applies the model-level safety rate to this combined score:

$$T_m = U_m S_m = (0.5C_m + 0.5R_m)S_m.$$

We report validity (V_m), answer correctness (C_m), reasoning (R_m), clinical safety (S_m), combined answer–reasoning (U_m), and the safety-adjusted overall score (T_m) separately. These quantities are computed on the unit interval and multiplied by 100 for reporting. The same GPT-5.6 Sol deployment judges its own outputs and those of other models; this limitation is discussed in Section 5. Provider content-policy blocks and infrastructure failures are tracked separately from scored runs. Executed runs with tool misuse, all 20 actions used, or no valid answer remain included as workflow failures and receive zero correctness credit.

G Samples Included in Scoring

Table 7 gives the static sample counts before any model-specific exclusions.

Table 7: Raw static-benchmark sample counts by cohort, modality, and assigned difficulty.

Cohort	2D	3D	Video	Text	Easy	Medium	Hard	Total
Public	54	88	74	0	60	79	77	216
Institutional	0	124	30	60	106	52	56	214

Table 8: Overall sample-weighted scores with bootstrap 95% confidence intervals. The four runs are averaged within each sample before averaging across samples. All values are percentages.

MLLM	Score (95% CI)
Open-weight	
DeepSeek-VL2	18.32 (14.73–21.99)
Gemma-4-26B-A4B-it	33.54 (29.17–37.91)
Gemma-4-31B-it	33.48 (29.13–37.81)
Gemma-4-E2B-it	30.77 (26.58–35.09)
Gemma-4-E4B-it	33.63 (29.44–37.92)
MedGemma 1.5	23.65 (20.43–26.95)
Phi-4-multimodal-instruct	20.31 (16.66–24.05)
Qwen3.6-27B	18.95 (15.54–22.58)
Qwen3.6-35B-A3B	26.69 (22.69–30.76)
Qwen3.8-27B	25.69 (22.81–28.56)
Proprietary	
Gemini 3.1 Pro	**47.34 (43.20–51.41)**
GPT-5.5	33.68 (29.39–38.06)
GPT-5.6 Sol	34.53 (30.67–38.51)

Table 9: Video scores with bootstrap 95% confidence intervals after averaging four runs per sample. The benchmark contains 30 ultrasound and 74 surgical-video samples. All values are percentages.

MLLM	Ultrasound video, mean (95% CI)	Surgical video, mean (95% CI)
Open-weight		
DeepSeek-VL2	0.0 (0.0–0.0)	1.4 (0.0–4.1)
Gemma-4-26B-A4B-it	23.3 (10.0–40.0)	52.9 (42.0–63.3)
Gemma-4-31B-it	21.2 (7.9–36.7)	57.7 (47.5–67.6)
Gemma-4-E2B-it	12.5 (2.5–25.0)	47.2 (36.9–57.9)
Gemma-4-E4B-it	22.9 (9.6–39.1)	51.2 (40.8–61.6)
MedGemma 1.5	7.1 (3.0–12.2)	21.0 (15.4–27.0)
Phi-4-multimodal-instruct	17.6 (6.1–30.9)	16.4 (9.1–24.9)
Qwen3.6-27B	3.3 (0.0–10.0)	37.9 (28.5–47.7)
Qwen3.6-35B-A3B	15.1 (5.0–28.3)	57.7 (47.1–67.9)
Qwen3.8-27B	13.9 (6.6–22.8)	36.2 (29.5–43.0)
Proprietary		
Gemini 3.1 Pro	15.6 (5.8–27.7)	68.4 (60.3–76.5)
GPT-5.5	25.8 (12.9–40.8)	Policy excluded
GPT-5.6 Sol	24.2 (12.5–37.1)	Policy excluded

Table 10: Static-benchmark coverage by model and cohort. Every model attempted 216 public and 214 institutional samples. "Scored" denotes samples included in the reported metrics.

MLLM	Public scored	Public policy	Inst. scored	Inst. policy	Total scored	Coverage (%)
Open-weight						
DeepSeek-VL2	216	0	214	0	430	100.0
Gemma-4-26B-A4B-it	216	0	214	0	430	100.0
Gemma-4-31B-it	216	0	214	0	430	100.0
Gemma-4-E2B-it	216	0	214	0	430	100.0
Gemma-4-E4B-it	216	0	214	0	430	100.0
MedGemma 1.5	216	0	214	0	430	100.0
Phi-4-multimodal-instruct	216	0	214	0	430	100.0
Qwen3.6-27B	216	0	214	0	430	100.0
Qwen3.6-35B-A3B	216	0	214	0	430	100.0
Qwen3.8-27B	216	0	214	0	430	100.0
Proprietary						
Gemini 3.1 Pro	216	0	214	0	430	100.0
GPT-5.5	171	45	184	30	355	82.6
GPT-5.6 Sol	171	45	184	30	355	82.6

For each model and cohort, every attempted sample is either scored, blocked by provider policy, or lost to an infrastructure failure:

$$n_{\text{attempted}} = n_{\text{scored}} + n_{\text{policy}} + n_{\text{infrastructure}}.$$

Static metrics use the successfully scored samples. No infrastructure failures occurred in the reported static results.

For GPT-5.5 and GPT-5.6 Sol, the 45 public exclusions comprised one prostate MRI sample, 20 laparoscopic cholecystectomy video samples, and 24 gastrointestinal endoscopy samples. The 30 institutional exclusions were ovarian-cancer surgical-video samples.

Table 11: Task-group scores and their unweighted macro-average. All values are percentages.

MLLM	Classification	Multi-label	Generation	Regression	Keypoint	2D det.	3D det.	Macro
Open-weight								
DeepSeek-VL2	34.86	2.48	3.24	0.00	0.00	2.44	0.00	6.14
Gemma-4-26B-A4B-it	56.54	17.47	26.94	0.00	0.00	0.00	0.55	14.50
Gemma-4-31B-it	54.24	29.47	28.06	5.31	0.00	0.00	0.56	16.81
Gemma-4-E2B-it	48.05	28.40	28.28	8.97	0.00	0.00	2.12	16.55
Gemma-4-E4B-it	52.29	31.67	23.28	10.66	0.00	0.00	8.33	18.03
MedGemma 1.5	39.22	9.95	19.15	5.92	1.25	0.35	0.95	10.97
Phi-4-multimodal-instruct	35.55	11.62	8.13	10.11	0.00	0.09	0.00	9.36
Qwen3.6-27B	29.36	20.20	18.68	0.00	5.62	0.00	0.14	10.57
Qwen3.6-35B-A3B	44.50	18.67	15.82	5.10	2.50	3.33	2.33	13.18
Qwen3.8-27B	38.76	13.27	32.14	4.77	2.50	1.95	2.24	13.66
Proprietary								
Gemini 3.1 Pro	60.21	48.95	81.01	3.33	1.25	0.31	13.32	29.77
GPT-5.5	48.32	0.00	63.79	0.00	6.88	1.67	0.70	17.34
GPT-5.6 Sol	48.32	1.67	60.69	0.54	10.62	2.50	5.65	18.57

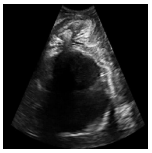
H Equal-Weight Task Average

The task-macro score is the unweighted mean of seven non-overlapping task groups: 218 single-label classification, 25 multi-label classification, 60 generation, 20 regression, 20 keypoint detection, 8 2D detection, and 79 3D detection samples. Giving each group equal weight reduces the influence of larger task groups.

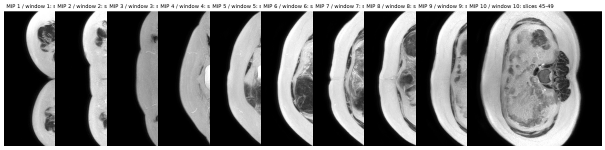
I Example Errors

Figure 4 shows example failures in regression, 3D object detection, and classification.

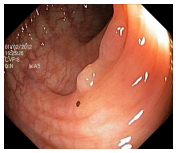
(A) **Q**: Report the angle of progression (AOP) in degrees. Return only a single float number, without any additional text.
Ground truth: 77.23
Wrong Answer by Gemini 2.5 Pro: "`144.3`"



(B) **Q**: Locate endometriosis in this T2-weighted MRI using a 3D bounding box. Return only the bounding box in the format $[x_{min}, y_{min}, z_{min}, x_{max}, y_{max}, z_{max}]$, without any additional text.
Ground truth: [152,217,14,181,250, 18]
Wrong answer by Qwen3.5-35B-A3B:
"`min[13,13,13] max [48,48, 48]`"



(C) **Q**: What is the class label of this GI endoscopy image? Return only the answer from the options provided, without any additional text.
Options:
- Barretts
- Barretts short segment
- BBPS 0-1
- BBPS 2-3
- Cecum
- Ayed lifted polyps
- Dyed resection margins
- Esophagitis A
- Esophagitis B-D
- Hemorrhoids
- Ileum
- Impacted stool
- Polyps
- Pylorus
- Retroflex rectum
- Retroflex stomach
- Ulcerative colitis grade 1
- Ulcerative colitis grade 2
- Ulcerative colitis grade 3
- Z-line



Ground truth: Polyps
Wrong Answer by Gemma-4-31B: "`normal`"

Figure 4: Representative errors. (A) Regression of an imaging measurement. (B) Localization of a disease region in a volumetric study. (C) Classification of an endoscopy frame.

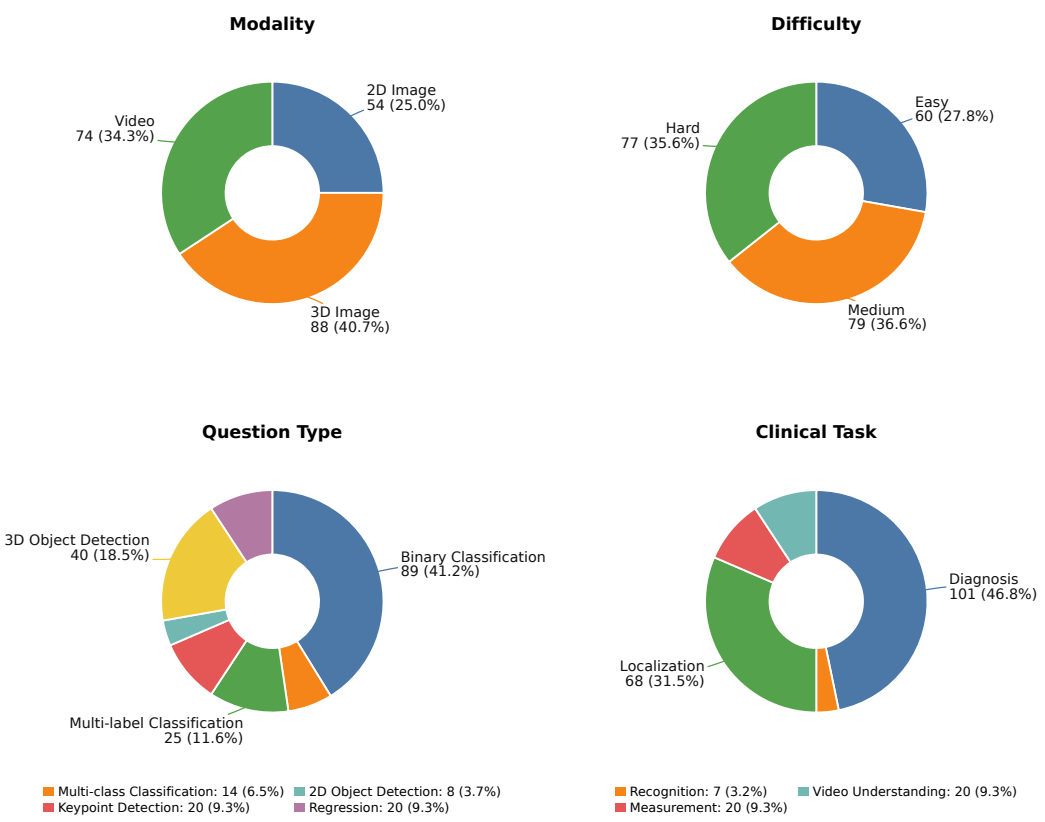


Figure 5: Distributions of the modalities, difficulty levels, question types, and clinical tasks of the public subset of MedStat-Bench.

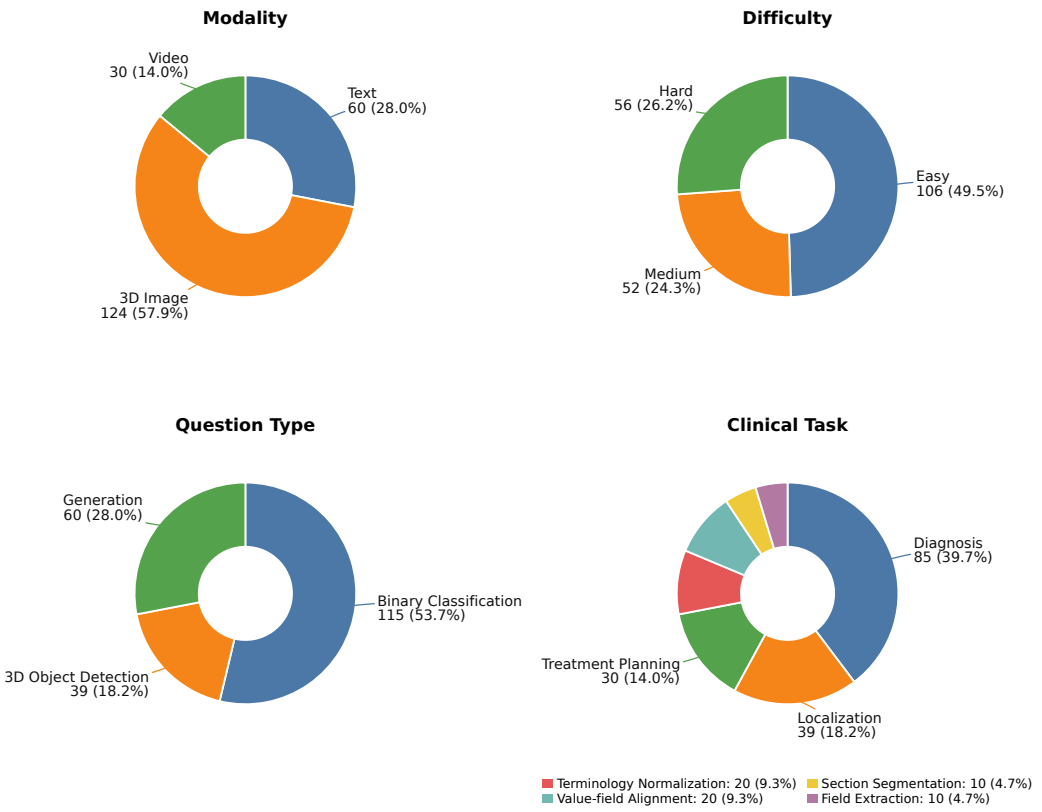


Figure 6: Distributions of the modalities, difficulty levels, question types, and clinical tasks of the institutional subset of MedStat-Bench.

Table 12: Scores by required output type. The 2D localization column combines keypoints and 2D boxes. **Bold**/underlined: best/second best.

	Classification		Detection		Other Tasks	
MLLMs	Single-label	Multi-label	2D loc.	3D	Generation	Regression
Open-weight						
DeepSeek-VL2	34.9	2.5	0.7	0.0	3.2	0.0
Gemma-4-26B-A4B-it	56.5	17.5	0.0	0.6	26.9	0.0
Gemma-4-31B-it	54.2	29.5	0.0	0.6	28.1	5.3
Gemma-4-E2B-it	48.1	28.4	0.0	2.1	28.3	9.0
Gemma-4-E4B-it	52.3	31.7	0.0	8.3	23.3	**10.7**
MedGemma 1.5	39.2	10.0	1.0	1.0	19.2	5.9
Phi-4-multimodal-instruct	35.6	11.6	0.0	0.0	8.1	10.1
Qwen3.6-27B	29.4	20.2	4.0	1.0	18.7	0.0
Qwen3.6-35B-A3B	44.5	18.7	2.7	2.3	15.8	5.1
Qwen3.8-27B	38.8	13.3	2.3	2.2	32.1	4.8
Proprietary						
Gemini 3.1 Pro	**60.2**	**49.0**	1.0	**13.3**	**81.0**	3.3
GPT-5.5	48.3	0.0	5.4	8.5	63.8	0.0
GPT-5.6 Sol	48.3	1.7	**8.3**	5.6	60.7	0.5

Table 13: Scores by modality and cohort. "MIP" denotes maximum-intensity projections from volumetric studies. **Bold**/underlined: best/second best. A dash marks a policy exclusion or an unavailable cohort split.

	Public				Institutional		
MLLMs	2D image	MIP	Ultrasound video	Surgical video	Text	MIP	Surgical video
Open-weight							
DeepSeek-VL2	18.9	26.8	0.0	2.3	3.2	33.9	0.0
Gemma-4-26B-A4B-it	18.5	29.2	23.3	41.2	26.9	37.3	70.0
Gemma-4-31B-it	18.0	29.7	21.2	47.0	28.1	34.1	**73.3**
Gemma-4-E2B-it	16.7	27.9	12.5	45.4	28.3	34.7	50.0
Gemma-4-E4B-it	22.7	**32.4**	22.9	52.1	23.3	36.4	50.0
MedGemma 1.5	18.3	27.2	7.1	–	19.2	35.1	–
Phi-4-multimodal-instruct	13.0	18.2	17.6	27.6	8.1	33.9	0.0
Qwen3.6-27B	22.5	17.1	3.3	43.3	18.7	11.3	30.0
Qwen3.6-35B-A3B	19.0	13.0	15.1	49.2	15.8	29.3	70.0
Qwen3.8-27B	30.8	13.9	13.9	30.2	32.1	21.1	45.0
Proprietary							
Gemini 3.1 Pro	**40.9**	28.1	15.6	**66.8**	**81.0**	**39.4**	70.8
GPT-5.5	35.0	20.8	**25.8**	–	63.8	0.6	–
GPT-5.6 Sol	28.5	21.7	24.2	–	60.7	34.4	–

Table 14: Scores by assigned difficulty and cohort. **Bold**/underlined: best/second best. GPT scores reflect the provider-policy exclusions detailed in Table 10.

	Public Benchmark			Institutional Benchmark		
MLLMs	Easy	Medium	Hard	Easy	Medium	Hard
Open-weight						
DeepSeek-VL2	31.7	19.2	0.8	40.9	1.2	0.0
Gemma-4-26B-A4B-it	63.3	23.3	5.7	52.8	51.0	1.8
Gemma-4-31B-it	58.3	24.6	10.9	49.6	52.9	1.8
Gemma-4-E2B-it	51.7	20.8	12.9	52.0	36.5	1.4
Gemma-4-E4B-it	63.3	27.2	14.3	51.1	36.5	1.6
MedGemma 1.5	45.8	21.1	5.7	42.8	22.2	10.0
Phi-4-multimodal-instruct	45.8	10.1	6.4	43.0	2.0	0.4
Qwen3.6-27B	48.3	15.3	8.0	19.4	24.3	1.8
Qwen3.6-35B-A3B	48.3	14.9	9.3	39.4	45.8	2.4
Qwen3.8-27B	52.1	19.4	6.5	31.0	39.3	9.9
Proprietary						
Gemini 3.1 Pro	**70.0**	**34.4**	**20.3**	**58.7**	**70.6**	**35.3**
GPT-5.5	57.1	30.4	2.8	47.5	56.9	24.4
GPT-5.6 Sol	56.0	22.8	4.5	50.9	54.9	23.2

Table 15: Text-generation similarity metrics, multiplied by 100 after aggregation. Baseline-rescaled English BERTScore can be negative. **Bold**/underlined: best/second best.

MLLMs	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore
Open-weight					
DeepSeek-VL2	26.5	23.8	25.6	16.9	-0.7
Gemma-4-26B-A4B-it	84.2	76.1	83.6	72.8	57.9
Gemma-4-31B-it	**86.3**	**77.0**	**85.7**	**75.7**	**61.8**
Gemma-4-E2B-it	73.8	66.3	73.2	62.2	48.2
Gemma-4-E4B-it	71.3	62.8	70.6	60.9	44.9
MedGemma 1.5	19.2	11.2	16.4	8.4	-7.4
Phi-4-multimodal-instruct	26.2	15.4	25.6	22.9	-9.7
Qwen3.6-27B	68.5	65.1	68.0	61.8	45.7
Qwen3.6-35B-A3B	56.6	53.9	56.4	52.1	34.4
Qwen3.8-27B	32.1	28.6	31.5	28.7	7.3
Proprietary					
Gemini 3.1 Pro	81.0	74.4	80.4	73.3	60.2
GPT-5.5	63.8	49.1	62.5	45.3	39.5
GPT-5.6 Sol	60.7	41.8	57.2	36.2	34.8

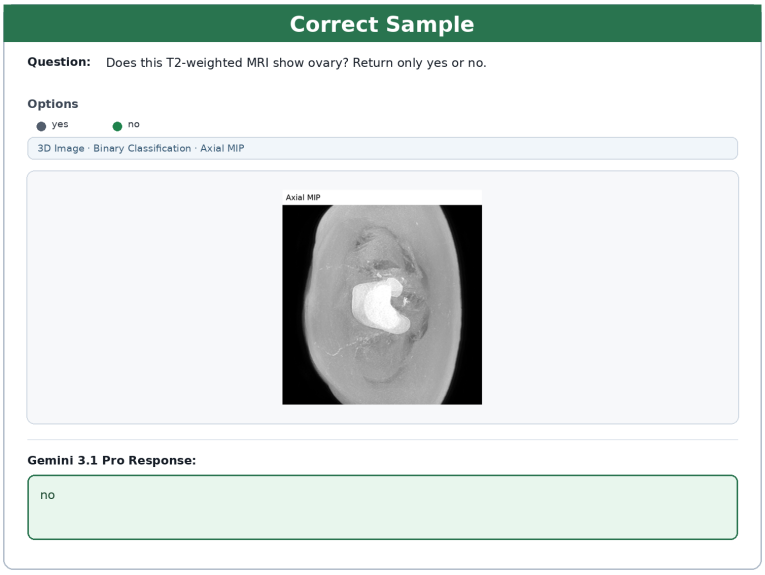


Figure 7: Static binary-classification example for Gemini 3.1 Pro. For the question, "Does this T2-weighted MRI show ovary?" the model answered "no," matching the reference answer.

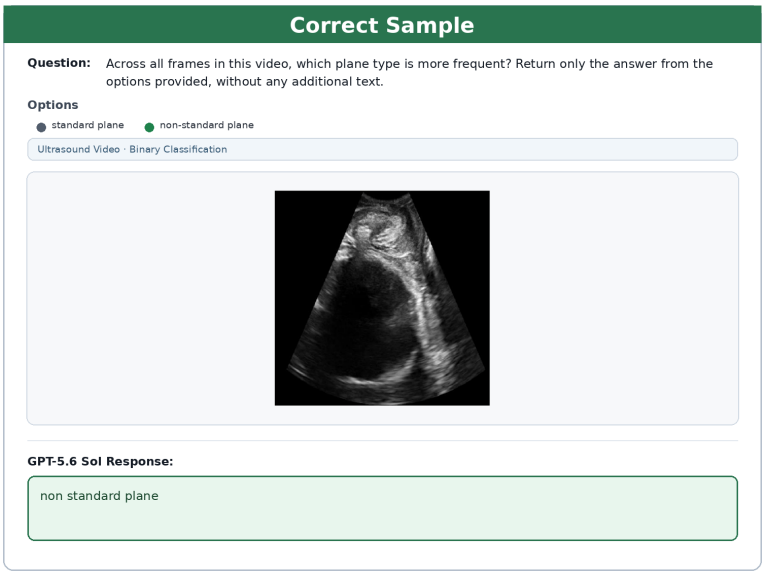


Figure 8: Static video-classification example for GPT-5.6 Sol. For the question, "Across all frames in the video, which plane type is more frequent?" the model answered "non-standard plane," matching the reference answer.

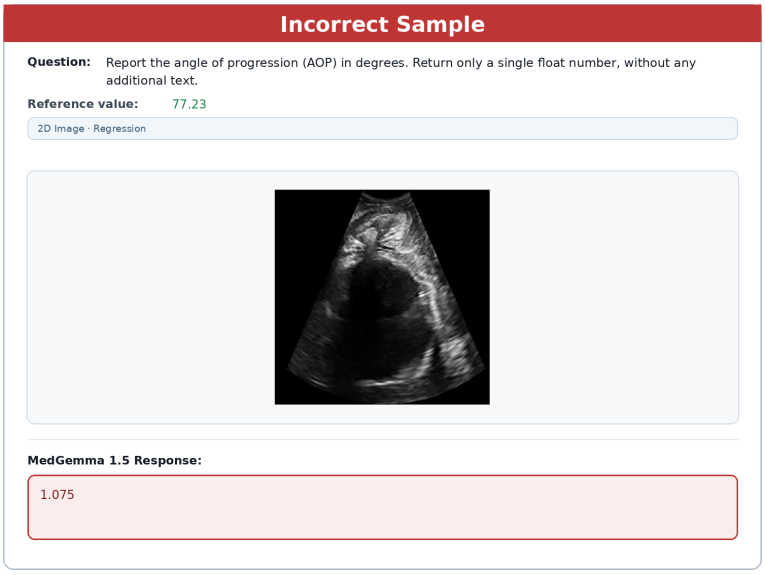


Figure 9: Static regression example for MedGemma 1.5. For the question, "Report the angle of progression (AOP) in degrees," the model returned 1.075 versus the reference value of 77.23 and was scored incorrect.

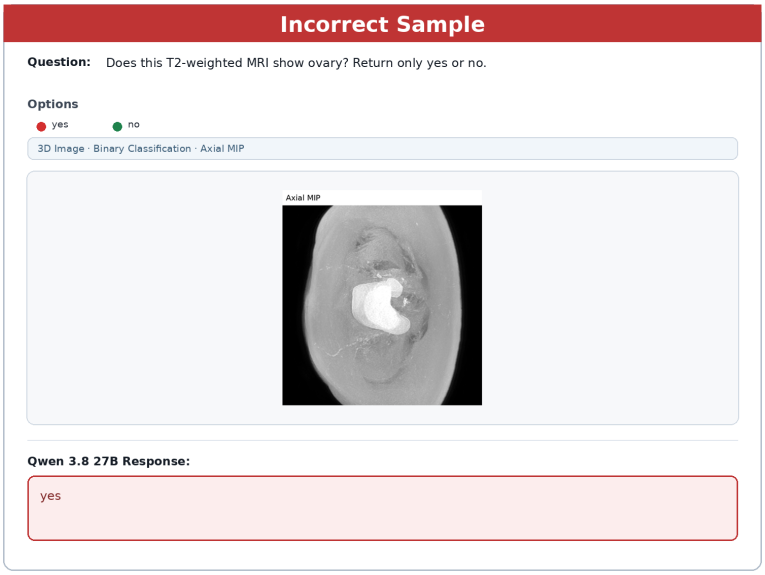


Figure 10: Static binary-classification example for Qwen3.8-27B. For the question, "Does this T2-weighted MRI show ovary?" the model answered "yes" while the reference answer was "no," and was scored incorrect.



Figure 11: Agentic case for GPT-5.6 Sol on the question, "Does echo demonstrate a definite cardiac embolic source?" Both the final answer and reasoning were scored correct.

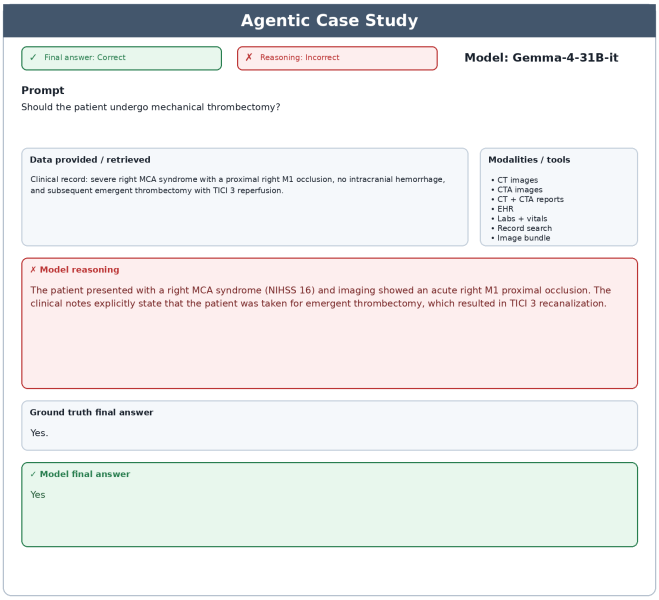


Figure 12: Agentic case for Gemma-4-31B-it on the question, "Should the patient undergo mechanical thrombectomy?" The final answer was scored correct, but the reasoning was scored incorrect.

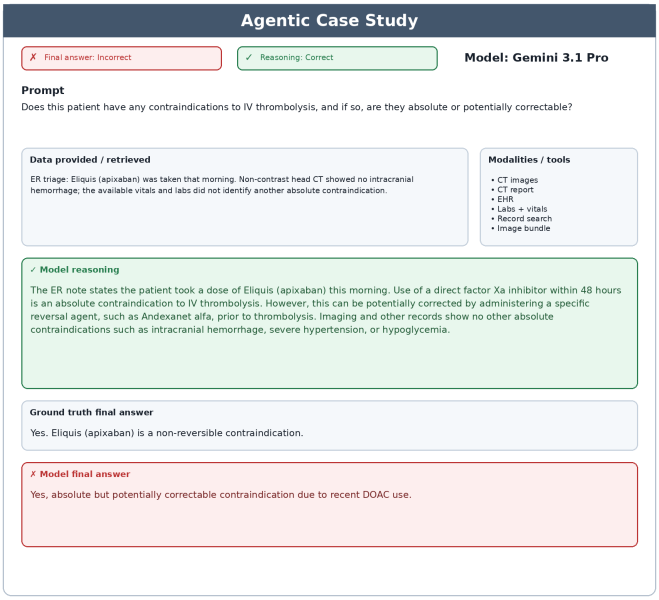


Figure 13: Agentic case for Gemini 3.1 Pro on the question, "Does this patient have any contraindications to IV thrombolysis, and if so, are they absolute or potentially correctable?" The final answer was scored incorrect, while the reasoning was scored correct.

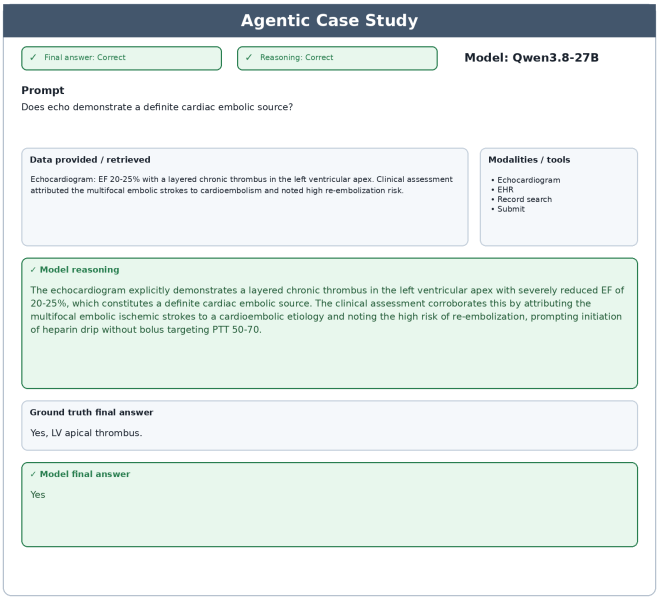


Figure 14: Agentic case for Qwen3.8-27B on the question, "Does echo demonstrate a definite cardiac embolic source?" Both the final answer and reasoning were scored correct.

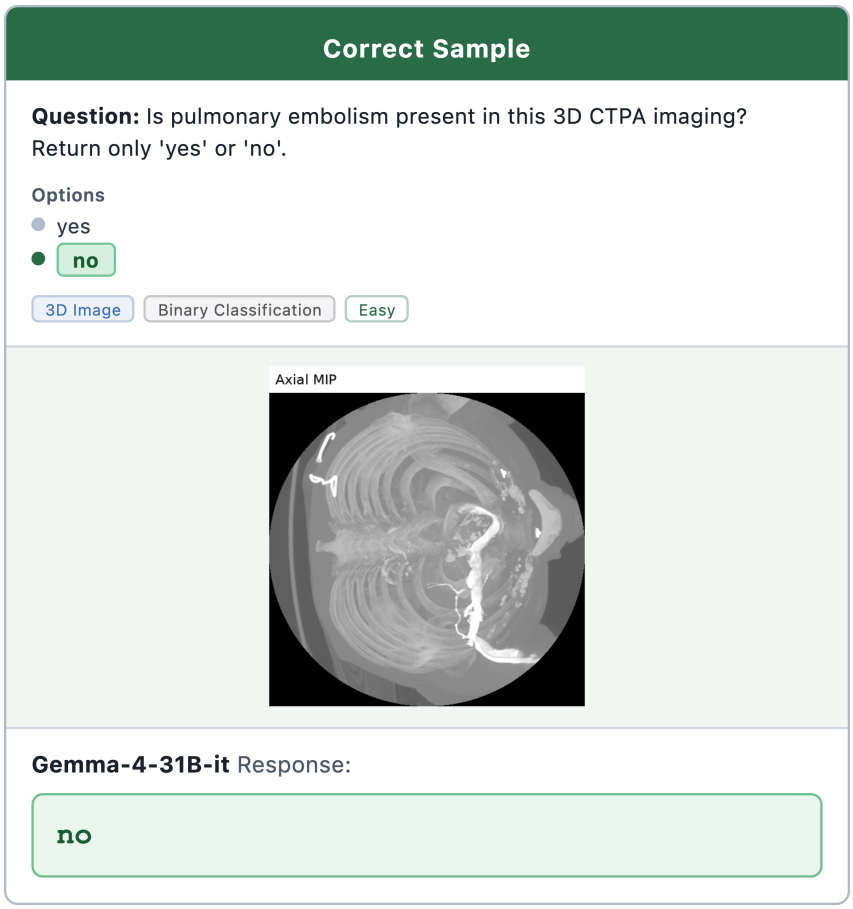


Figure 15: Example of an easy binary classification sample answered correctly by Gemma-4-31B-it.

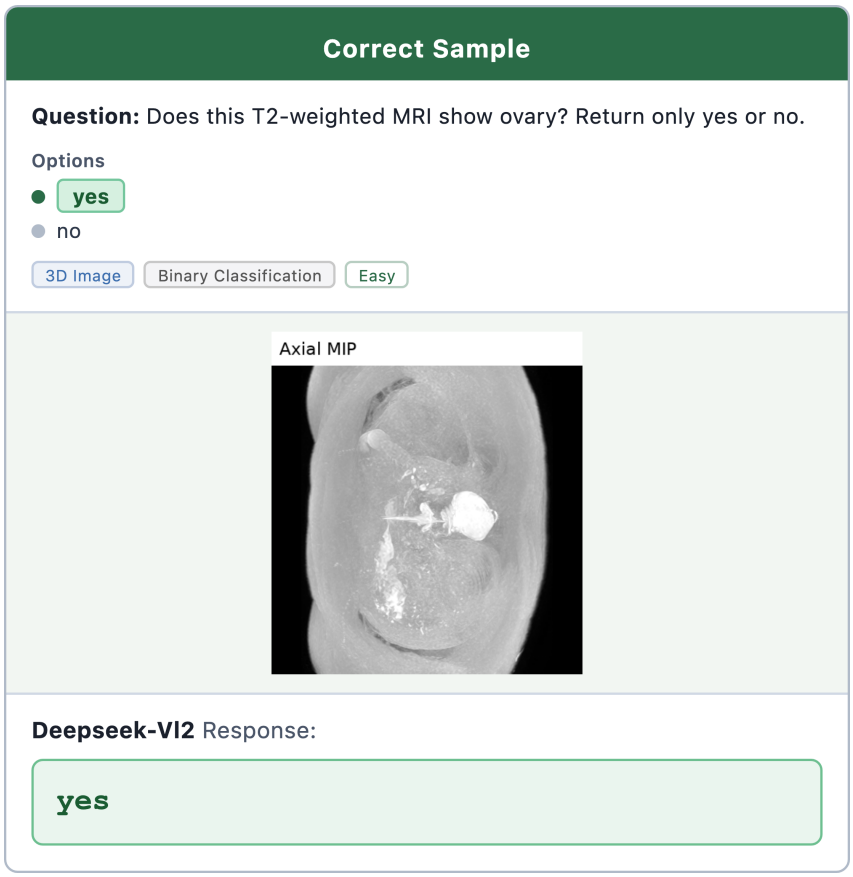


Figure 16: Example of an easy binary classification sample answered correctly by DeepSeek-VL2.

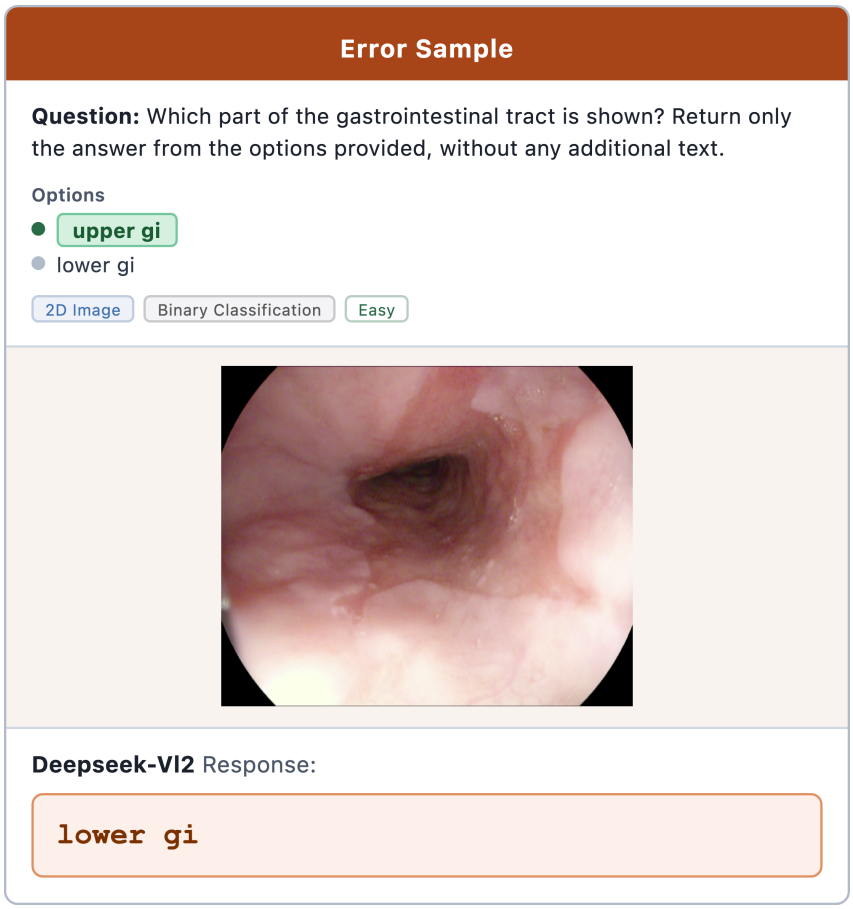


Figure 17: Example of an easy binary classification sample answered incorrectly by DeepSeek-VL2.

J AGENTIC REASONING EVALUATION RUBRICS

The agentic track contains 29 tasks, listed in Table 16. Complete reasoning rubrics are available for 15 scenarios: q01–q11 and q26–q29, which have reasoning ground truth.

J.1 SHARED GRADING CONTRACT

Judges evaluate only the model's submitted answer and recorded tool trajectory. The rubrics do not request or attempt to reconstruct hidden chain-of-thought.

- `met` means the proposition is explicitly supported by the submitted answer or visible trajectory.
- `not_met` means the proposition is contradicted, omitted despite the needed case evidence being available, or asserted without the required support.
- `not_applicable` means the proposition does not apply to the patient or the requested source is unavailable. It does not excuse omission of an available finding.
- Answer criteria are compared with the case-specific reference answer and its accepted equivalents. Clinically equivalent wording should receive credit.
- A reasoning criterion is met only when the model states the relevant case fact and uses it in the direction required by the conclusion. Generic medical knowledge, a bare conclusion, or a tool call without interpretation is insufficient.
- A source criterion is met only when the model identifies the source at the level requested by the item, such as the noncontrast CT image, CT report, CTA, echocardiogram, medication list, or laboratory record. Merely saying "the data" is insufficient.
- For a criterion conditional on treatment or a finding, use `not_applicable` only when the condition is absent in the case. If the condition is present and the model omitted the criterion, use `not_met`.
- Critical checks beginning with "Does not" are positive safety checks. They are `met` only when the prohibited error is absent from the submitted answer and visible trajectory.
- Do not award reasoning credit merely because the final answer is correct. A correct answer with unsupported, contradictory, or hallucinated reasoning loses the affected reasoning and critical-check items.

Use only the case records, images, task instructions, and visible trajectory supplied to the judge. Do not credit facts imported from another patient, later clinical events, or external sources that are not cited or visible in the run. When the case reference permits more than one answer, use the task's accepted-equivalent rule rather than exact string matching.

The rubrics define item-level judgments. Score aggregation, validity, vote aggregation, and safety adjustment follow Appendix F.

J.2 RUBRIC COVERAGE

Table 16 lists all 29 agentic tasks. Complete item-level rubric templates are provided below only for the 15 scenarios with reasoning ground truth.

Table 16: All 29 agentic tasks. Complete reasoning rubric templates are provided for the 15 scenarios with reasoning ground truth.

Rubric	Task
q01	IV thrombolysis and dose
q02	IV thrombolysis with CT report
q03	thrombolysis contraindications
q04	mechanical thrombectomy decision
q05	mechanical thrombectomy with reports
q06	antiplatelet treatment
q07	stroke etiology
q08	definite cardiac embolic source
q09	discharge antithrombotic strategy
q10	peri-procedural blood-pressure management

Continued on next page

Table 16 – continued

Rubric	Task
q11	pre-procedure issue requiring action
q12	primary surgery versus neoadjuvant chemotherapy
q13	integrated acute-stroke treatment decision
q14	thrombectomy candidacy
q15	transfer to a thrombectomy-capable center
q16	treatment strategy and reperfusion prediction
q17	hemorrhagic-transformation prediction
q18	discharge modified Rankin Scale prediction
q19	90-day functional independence
q20	hospital length-of-stay prediction
q21	reliability of perfusion processing
q22	thrombectomy target and final reperfusion grade
q23	transfer decision
q24	discharge functional independence
q25	intracranial-hemorrhage prediction
q26	next definitive management
q27	integrated MRI diagnosis
q28	ECG rhythm
q29	next diagnostic step for recurrent palpitations
Total	**29 agentic tasks**

J.3 Complete Rubric Templates

q01: IV thrombolysis and dose

Task family: Acute cerebrovascular decision-making.

Scenario-specific grading guidance. Apply the shared grading contract. Compare the patient-specific decision and dose with the q01 reference answer. For a treatment-positive case, the answer must identify tenecteplase/TNK and the reference dose; for a treatment-negative case, do not require a dose. The reasoning must use the patient's noncontrast CT, symptom timing, blood pressure, laboratory results, and documented contraindication status as applicable. Do not treat a generic eligibility checklist as evidence that the patient met it.

Answer criteria.

1. States whether IV thrombolysis should be given.
2. Names the thrombolytic agent when treatment is recommended.
3. States the dose when treatment is recommended.
4. Matches the case-specific reference decision.

Reasoning criteria.

5. Uses the noncontrast CT finding.
6. Uses the symptom-timing information.
7. Uses the blood-pressure information.
8. Uses the laboratory information.
9. Identifies a documented contraindication when present.
10. Links eligibility evidence to the treatment decision.
11. Links patient weight or other dosing input to the dose when available.
12. Identifies the source of each key finding.

Critical safety checks.

13. Does not invent a hemorrhage.
14. Does not invent a contraindication.
15. Does not give a dose for an agent it did not name.
16. Does not contradict the case reference.

q02: IV THROMBOLYSIS WITH CT REPORT

Task family: Acute cerebrovascular decision-making.

Scenario-specific grading guidance. Apply the shared grading contract. Compare the patient-specific treatment and dose decision with the q02 reference answer. This task intentionally provides both the CT image and its report. Award the corresponding criteria only when the model interprets each source without substituting one for the other. The decision must be tied to the patient's symptom timing, blood pressure, laboratory results, and contraindication evidence; do not credit unsupported guideline recitation.

Answer criteria.

1. States whether IV thrombolysis should be given.
2. Names the thrombolytic agent when treatment is recommended.
3. States the dose when treatment is recommended.
4. Matches the case-specific reference decision.

Reasoning criteria.

5. Uses the direct CT-image finding.
6. Uses the CT-report finding.
7. Uses the symptom-timing information.
8. Uses the blood-pressure information.
9. Uses the laboratory information.
10. Identifies a documented contraindication when present.
11. Links the evidence to the treatment decision.
12. Identifies the source of each key finding.

Critical safety checks.

13. Does not treat the CT report as evidence of a finding absent from the image or record.
14. Does not invent a contraindication.
15. Does not give an unsupported dose.
16. Does not contradict the case reference.

q03: THROMBOLYSIS CONTRAINDICATIONS

Task family: Acute cerebrovascular decision-making.

Scenario-specific grading guidance. Apply the shared grading contract. Compare the contraindication classification with the patient-specific q03 reference answer. A positive answer must name the documented contraindication and explain its clinical classification; a negative answer must not manufacture one from a non-specific risk factor. Distinguish a documented intracranial hemorrhage or nonreversible anticoagulant exposure from a potentially correctable blood-pressure or laboratory issue.

Answer criteria.

1. States whether a contraindication to IV thrombolysis is present.
2. Identifies the specific contraindication when one is present.
3. Classifies the contraindication as absolute or potentially correctable.
4. Matches the case-specific reference classification.

Reasoning criteria.

5. Uses the CT image or report finding.
6. Uses the medication history.
7. Uses the laboratory findings.
8. Uses the blood-pressure findings.
9. Distinguishes a documented contraindication from a possible concern.
10. Explains why the contraindication is absolute or potentially correctable.
11. Identifies the source of each key finding.

Critical safety checks.

12. Does not invent a contraindication.
13. Does not misclassify intracranial hemorrhage as correctable.
14. Does not misclassify a potentially correctable issue as absolute without evidence.
15. Does not contradict the case reference.

q04: MECHANICAL THROMBECTOMY DECISION

Task family: Acute cerebrovascular decision-making.

Scenario-specific grading guidance. Apply the shared grading contract. Compare the yes/no thrombectomy decision with the patient-specific q04 reference answer. The reasoning must use the documented vascular finding or its absence, noncontrast CT, clinical severity, treatment window, and available procedural-safety information. Do not award credit for naming an LVO or eligibility feature that is not documented in this patient.

Answer criteria.

1. States whether mechanical thrombectomy should be performed.
2. Matches the case-specific thrombectomy reference.

Reasoning criteria.

3. Uses the CTA finding.
4. Uses the noncontrast CT finding.
5. Uses the clinical-severity finding.
6. Uses the available treatment-window information.
7. Uses the available laboratory or procedural-safety information.
8. Links the evidence to the thrombectomy decision.
9. Identifies the source of each key finding.

Critical safety checks.

10. Does not invent an LVO.
11. Does not invent a contraindication.
12. Does not confuse thrombectomy with thrombolysis.
13. Does not contradict the case reference.

q05: MECHANICAL THROMBECTOMY WITH REPORTS

Task family: Acute cerebrovascular decision-making.

Scenario-specific grading guidance. Apply the shared grading contract. Compare the patient-specific q05 decision with the reference answer. This version tests image-plus-report integration for both CT and CTA. The model must interpret each supplied source separately when the rubric requires it and then connect the vascular and clinical evidence to the decision. Do not treat an unavailable report as available or reward fabricated report content.

Answer criteria.

1. States whether mechanical thrombectomy should be performed.
2. Matches the case-specific thrombectomy reference.

Reasoning criteria.

3. Uses the direct CT-image finding.
4. Uses the CT report.
5. Uses the direct CTA-image finding when available.
6. Uses the CTA report when available.
7. Uses the clinical-severity finding.
8. Uses the available treatment-window information.
9. Links the vascular and clinical evidence to the decision.
10. Identifies the source of each key finding.

Critical safety checks.

11. Does not invent an LVO.
12. Does not treat an unavailable CTA report as available.
13. Does not confuse thrombectomy with thrombolysis.
14. Does not contradict the case reference.

Q06: ANTIPLATELET TREATMENT

Task family: Acute cerebrovascular decision-making.

Scenario-specific grading guidance. Apply the shared grading contract. Compare agent selection and timing with the patient-specific q06 reference answer. When the patient has received thrombolysis, the key safety sequence is to hold antiplatelet therapy for 24 hours, obtain follow-up brain imaging, and start aspirin only if hemorrhage is excluded. Do not treat aspirin, clopidogrel, DAPT, and anticoagulation as interchangeable, and do not assume an exposure that is absent from the patient's record.

Answer criteria.

1. Selects aspirin, clopidogrel, dual antiplatelet therapy, or no antiplatelet therapy as appropriate.
2. States when the selected therapy should start.
3. Matches the case-specific antiplatelet reference.

Reasoning criteria.

4. Uses the CT finding.
5. Uses the hemorrhage-risk information.
6. Uses the thrombolysis history when applicable.
7. Uses the anticoagulant history when applicable.
8. Uses the clinical indication for antiplatelet therapy.
9. Links treatment timing to the relevant safety condition.
10. Identifies the source of each key finding.

Critical safety checks.

11. Does not recommend antiplatelet treatment despite a documented reason to hold it.
12. Does not invent an anticoagulant or thrombolytic exposure.
13. Does not confuse antiplatelet therapy with anticoagulation.
14. Does not contradict the case reference.

Q07: STROKE ETIOLOGY

Task family: Acute cerebrovascular decision-making.

Scenario-specific grading guidance. Apply the shared grading contract. Compare the selected etiology with the patient-specific q07 reference answer, including the possibility that the correct answer is indeterminate or not applicable. The model must weigh the documented brain and vascular imaging, available cardiac testing, and clinical history. A suspected mechanism such as paradoxical embolism must not be upgraded to a definite etiology without supporting evidence. Conversely, a negative workup should not be forced into a named mechanism.

Answer criteria.

1. States the most likely stroke etiology.
2. Matches the case-specific etiology reference.

Reasoning criteria.

3. Uses the relevant brain-imaging finding.
4. Uses the relevant vascular-imaging finding.
5. Uses the relevant cardiac or echocardiographic finding when available.
6. Uses the relevant clinical history.
7. Identifies one finding that supports the selected etiology.
8. Distinguishes the selected etiology from at least one plausible alternative when the record supports that distinction.
9. Identifies the source of each key finding.

Critical safety checks.

10. Does not infer a cardiac source without supporting evidence.
11. Does not treat a suspected mechanism as a confirmed etiology.
12. Does not invent a diagnostic test result.
13. Does not contradict the case reference.

q08: DEFINITE CARDIAC EMBOLIC SOURCE

Task family: Acute cerebrovascular decision-making.

Scenario-specific grading guidance. Apply the shared grading contract. Compare the yes/no decision and any named source with q08's patient-specific reference answer. A definite embolic source requires a documented echocardiographic finding; age, hypertension, atrial enlargement, vascular disease, or another possible risk factor alone is insufficient. If echocardiography is indeterminate, credit uncertainty rather than a fabricated definite source.

Answer criteria.

1. States whether echocardiography demonstrates a definite cardiac embolic source.
2. Identifies the cardiac source when one is present.
3. Matches the case-specific reference.

Reasoning criteria.

4. Uses the echocardiographic finding.
5. Distinguishes a definite source from a risk factor.
6. Distinguishes a definite source from an indeterminate finding.
7. Identifies the source of the cardiac evidence.

Critical safety checks.

8. Does not infer a definite embolic source from age or vascular risk factors alone.
9. Does not invent an echocardiographic finding.
10. Does not contradict the case reference.

q09: DISCHARGE ANTITHROMBOTIC STRATEGY

Task family: Acute cerebrovascular decision-making.

Scenario-specific grading guidance. Apply the shared grading contract. Compare the discharge agent, timing, and bridging plan with the patient-specific q09 reference answer. The reasoning must distinguish cardioembolic anticoagulation from antiplatelet secondary prevention and account for hemorrhage risk, prior therapy, and any cardiac source. A named drug without a safe start or transition plan is incomplete when the reference answer requires timing.

Answer criteria.

1. Selects the discharge antithrombotic strategy.
2. Names the selected antithrombotic agent or agents.
3. States the planned start or transition timing.
4. States a bridging plan when the reference requires one.
5. Matches the case-specific discharge reference.

Reasoning criteria.

6. Uses the stroke etiology finding.
7. Uses the hemorrhage-risk finding.
8. Uses the pre-existing anticoagulant or antiplatelet history.
9. Uses the cardiac-source finding when available.
10. Links the findings to the selected strategy.
11. Identifies the source of each key finding.

Critical safety checks.

12. Does not recommend simultaneous therapies that the case reference excludes.
13. Does not invent a cardiac source or bleeding event.
14. Does not confuse antiplatelet therapy with anticoagulation.
15. Does not contradict the case reference.

q10: PERI-PROCEDURAL BLOOD-PRESSURE MANAGEMENT

Task family: Acute cerebrovascular decision-making.

Scenario-specific grading guidance. Apply the shared grading contract. Compare the patient's requested blood-pressure strategy with q10's patient-specific reference answer. Grade pre-, intra-, and post-procedure phases independently; a generic "keep it controlled" statement is insufficient when the reference supplies phase-specific targets. The model must distinguish lowering, maintaining, and augmenting pressure and must not invent measurements or apply a post-reperfusion target before reperfusion.

Answer criteria.

1. States the pre-procedure blood-pressure target.
2. States the intra-procedure blood-pressure target.
3. States the post-procedure blood-pressure target.
4. States whether blood pressure should be lowered, maintained, or augmented.
5. Matches the case-specific blood-pressure reference.

Reasoning criteria.

6. Uses the treatment type.
7. Uses the reperfusion status or procedural phase.
8. Uses the clinical-perfusion information.
9. Uses the hemorrhage-risk information.
10. Links each target to its relevant procedural phase.
11. Identifies the source of each key finding.

Critical safety checks.

12. Does not apply one target to all three phases without justification.
13. Does not invent a blood-pressure measurement.
14. Does not recommend augmentation when the case evidence supports lowering.
15. Does not contradict the case reference.

q11: PRE-PROCEDURE ISSUE REQUIRING ACTION

Task family: Acute cerebrovascular decision-making.

Scenario-specific grading guidance. Apply the shared grading contract. Compare the issue and action with the patient-specific q11 reference answer. A positive answer must identify the documented history, medication, laboratory, or evaluation problem and state whether the procedure should be delayed, modified, or avoided. A negative answer is appropriate only when the supplied record contains no critical issue. Do not convert a routine abnormality into a contraindication without evidence.

Answer criteria.

1. States whether a pre-procedure issue requiring action is present.
2. Identifies the issue when one is present.
3. States whether the procedure should be delayed, modified, or avoided.
4. Matches the case-specific reference.

Reasoning criteria.

5. Uses the relevant medical-history finding.
6. Uses the relevant medication finding.
7. Uses the relevant laboratory finding.
8. Uses the relevant pre-procedure evaluation finding.
9. Links the issue to the proposed action.
10. Identifies the source of each key finding.

Critical safety checks.

11. Does not invent a medication or laboratory result.
12. Does not recommend proceeding without addressing a documented critical issue.
13. Does not recommend avoiding the procedure for an issue not supported by the record.
14. Does not contradict the case reference.

Q26: NEXT DEFINITIVE MANAGEMENT

Task family: Neuro-oncology diagnosis and management.

Scenario-specific grading guidance. Apply the shared grading contract. The answer must address the next definitive action for the dominant intracranial mass, not merely recommend consultation or additional review. The reference requires urgent operative decompression, resection, or debulking; neither a specific approach nor complete resection is required. The reasoning must connect the supplied MRI abnormality and clinical history to the urgency. An intervention not documented as completed must be described as a recommendation, not a completed event.

Answer criteria.

1. Identifies the requested next definitive management.
2. States urgent surgical decompression when required by the case reference.
3. Matches the case-specific reference answer.

Reasoning criteria.

4. Identifies the relevant MRI abnormality.
5. Identifies the anatomical location of the abnormality.
6. Identifies the finding that creates an urgent management concern.
7. Uses the clinical history provided in the case.
8. Uses direct MRI review rather than only a metadata field.
9. Links the imaging finding to the recommended management.
10. Follows the case's information cutoff.

Critical safety checks.

11. Does not invent an imaging feature.
12. Does not recommend routine follow-up when urgent management is required.
13. Does not claim a procedure was performed.
14. Does not contradict the case reference.

Q27: INTEGRATED MRI DIAGNOSIS

Task family: Neuro-oncology diagnosis and management.

Scenario-specific grading guidance. Apply the shared grading contract. Credit the diagnosis only when it identifies NF2-related schwannomatosis, including the accepted legacy name neurofibromatosis type 2. The reasoning must integrate the pattern of bilateral vestibular schwannomas and multiple meningiomas with the supplied clinical history. Do not require a pathogenic variant, mosaic status, or family history; these details are outside the reference and should not be invented.

Answer criteria.

1. States the primary diagnosis.
2. States NF2-related schwannomatosis when required by the case reference.
3. Matches the case-specific reference diagnosis.

Reasoning criteria.

4. Uses the overall MRI pattern.
5. Uses the distribution of lesions shown on MRI.
6. Uses the clinical history provided in the case.
7. Integrates the imaging pattern with the clinical history.
8. Reviews the MRI directly.
9. Distinguishes the primary diagnosis from unsupported subtype claims.
10. Follows the case's information cutoff.

Critical safety checks.

11. Does not infer a specific genetic variant.
12. Does not infer mosaicism without evidence.
13. Does not infer inheritance status without evidence.
14. Does not invent a lesion or clinical-history fact.
15. Does not contradict the case reference.

Q28: ECG RHYTHM

Task family: Cardiac rhythm and diagnostic evaluation.

Scenario-specific grading guidance. Apply the shared grading contract. The reference applies only to the supplied resting tracing, which shows normal sinus rhythm. Credit waveform reasoning only when the model discusses the visible regular rhythm, organized atrial activity, and atrial-to-ventricular relationship. Do not infer what occurred during symptoms or use a normal resting tracing to exclude an intermittent arrhythmia.

Answer criteria.

1. Identifies the rhythm.
2. States normal sinus rhythm when required by the case reference.
3. Matches the case-specific reference rhythm.

Reasoning criteria.

4. Identifies regularity of the rhythm.
5. Identifies organized atrial activity preceding the QRS complexes.
6. Identifies the relationship between atrial activity and ventricular complexes.
7. Reviews the ECG waveform directly.
8. Uses the requested ECG rather than only the printed machine interpretation.

Critical safety checks.

9. Does not diagnose an arrhythmia absent from the waveform.
10. Does not rely on metadata as the only evidence.
11. Does not invent an ECG finding.
12. Does not contradict the case reference.

Q29: NEXT DIAGNOSTIC STEP FOR RECURRENT PALPITATIONS

Task family: Cardiac rhythm and diagnostic evaluation.

Scenario-specific grading guidance. Apply the shared grading contract. The intended first investigation is noninvasive extended ambulatory ECG monitoring to correlate intermittent symptoms with rhythm. The episodes occur several times weekly but were not captured by the resting ECG. A generic ambulatory ECG recommendation is sufficient; an exact brand or duration is not required. A repeat resting ECG, treatment, reassurance, referral alone, or an implantable loop recorder as the initial test does not satisfy the reference.

Answer criteria.

1. Identifies the next diagnostic step.
2. Recommends extended ambulatory ECG monitoring when required by the case reference.
3. Gives one primary next step.
4. Matches the case-specific reference answer.

Reasoning criteria.

5. Uses the history of recurrent palpitations.
6. Notes that the resting ECG did not capture the symptomatic rhythm.
7. Links intermittent symptoms to the need for prolonged monitoring.
8. Reviews the resting ECG directly.
9. Distinguishes diagnostic monitoring from treatment.
10. Follows the case's information cutoff.

Critical safety checks.

11. Does not claim that the resting ECG captured the cause.
12. Does not recommend an unrelated first-line test without justification.
13. Does not invent an arrhythmia.
14. Does not contradict the case reference.