EVA-REASON: LEARNING ADAPTIVE EVIDENCE ACQUISITION FOR NEUROIMAGING REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Strong multimodal models often struggle to answer difficult neuroimaging questions directly from MRI. We show that these failures are not simply explained by a lack of predictive signal. On the same controlled task, an end-to-end CNN reaches 0.962 balanced accuracy and a Qwen2.5-VL post-projector linear probe reaches 0.897, while direct generation is at chance (0.333). Language-side fine-tuning with frozen vision encoder/projector raises direct generation accuracy to 0.743, narrowing but not closing the representation-to-generation gap. We examine whether explicit quantitative evidence can improve this reasoning bottleneck. We evaluate nine frontier and open-weight multimodal models on the same questions with and without access to quantitative measurement tools. On the hardest questions (L3, the most challenging reasoning level), Tool-OFF accuracy ranges from 36.4% to 56.4%. Four frontier models improve significantly with tools, reaching 75.6% to 80.0% L3 accuracy. Other models improve little or degrade, showing that tool access alone is insufficient. Motivated by these findings, we introduce **EVA-Reason**, a framework for learning adaptive evidence acquisition in neuroimaging. We construct multi-step evidence-seeking trajectories using Claude Sonnet 4.6 and Gemini 2.5 Pro, and use them to fine-tune a 3B VLM; frontier comparators use the same interface zero-shot. We evaluate EVA-Reason on NeuroQA questions organized into three reasoning levels. On the held-out test set, under normalized exact-match scoring, EVA-Reason achieves 96.4% overall accuracy and 89.3% L3 accuracy. Matched Qwen-3B controls reach 64.5/82.2% overall/L3 without quantitative evidence and 94.6/88.9% when a prespecified relevant-evidence set is supplied upfront; EVA-Reason selects evidence autonomously at inference, without a prespecified per-item evidence set. Under zero-shot exact match, EVA-Reason reaches 89.3% L3 versus 80.0/79.6/79.1% for Sonnet 5/Gemini 2.5 Pro/GPT-5.5; several gaps narrow to non-significance under few-shot prompting or option-prefix scoring. EVA-Reason requires no frontier-model calls at inference and runs locally on a single GPU. These results show that a small open-weight model can learn to actively acquire and use quantitative neuroimaging evidence across multiple decision steps.

1 INTRODUCTION

Recent vision-language models (VLMs) have substantially expanded the capabilities of multimodal AI, enabling models to jointly process images and language for tasks such as visual question answering, report generation, and multimodal reasoning. This progress has increasingly extended to medicine through generalist biomedical and domain-adapted models, including Med-PaLM M, LLaVA-Med, Med-Flamingo, BiomedCLIP, RadFM, M3D-LaMed, and MedGemma (Tu et al., 2024; Li et al., 2023; Moor et al., 2023; Zhang et al., 2025; Wu et al., 2025; Bai et al., 2024; Sellergren et al., 2025). Medical visual question answering benchmarks such as VQA-RAD, SLAKE, and PMC-VQA have further enabled systematic evaluation of models that connect visual observations with medical questions (Lau et al., 2018; Liu et al., 2021; Zhang et al., 2024). However, most existing work focuses on 2D images, report-level generation, or questions that can be answered through relatively direct visual inspection, while complex reasoning over three-dimensional medical imaging remains less explored (Bai et al., 2024; Abbasi et al., 2026).

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.

Brain MRI presents a demanding setting because many meaningful questions depend on quantitative relationships rather than a single conspicuous visual feature. Answering them may require comparing the size of a structure with a normative range, reasoning across multiple anatomical regions, distinguishing age-related variation from abnormality, or combining several quantitative findings. Automated segmentation methods, including neuroimaging-specific pipelines such as FreeSurfer and SynthSeg and general biomedical segmentation frameworks such as nnU-Net, provide a practical basis for deriving quantitative anatomical measurements from medical images (Fischl et al., 2002; Billot et al., 2023; Isensee et al., 2021). NeuroQA recently exposed the difficulty of this setting at scale by evaluating multimodal models on 3D brain MRI questions spanning anatomical, quantitative, and clinical reasoning (Abbasi et al., 2026). Yet an important question remains: when VLMs fail on such questions, is the necessary information missing from their visual representations, or is the model failing to use information that is already present?

We investigate this question through linear probing, which provides a simple test for whether task-relevant information remains accessible from intermediate representations without modifying the underlying model (Alain & Bengio, 2017). On a controlled neuroimaging task involving the same 279 subjects, labels, folds, and out-of-fold evaluation protocol, an end-to-end CNN achieves **0.962 balanced accuracy**. For Qwen2.5-VL (Bai et al., 2025), a linear probe on the projected visual tokens reaches **0.897**, while the corresponding probe for MedGemma-4B-IT (Sellergren et al., 2025) reaches **0.965**. Yet direct generation from either model on the same task remains at chance (**0.333**). These results reveal a *representation-to-generation gap*: substantial predictive information remains linearly accessible in the representation passed to the language model without being reliably reflected in its generated answer (see Appendix A for full probing details).

On the same controlled task, providing five raw volumetric measurements upfront substantially improves zero-shot balanced accuracy across three frontier models (Appendix A). These findings motivate a complementary question: can a model learn to request the quantitative evidence it needs during reasoning? Neuroimaging pipelines can provide structured measurements such as regional volumes, normalized quantities, and anatomical statistics that a model can inspect directly. More broadly, prior work on tool-augmented language models has shown that reasoning can be interleaved with external actions such as information retrieval and API calls (Yao et al., 2023; Schick et al., 2023; Patil et al., 2024). However, access to a tool is not sufficient by itself. A model must determine whether additional evidence is needed, which evidence to request, whether to continue acquiring evidence or answer, and how to integrate the returned information.

To study this problem directly, we evaluate nine multimodal models spanning frontier and open-weight systems under two conditions: answering without quantitative tools and answering with access to quantitative measurement tools. On the hardest L3 questions, Tool-OFF accuracy ranges from **36.4% to 56.4%**. For models that reliably operate the evidence-acquisition loop, access to quantitative measurements increases accuracy to **75.6%–80.0%**, with gains of up to **36.4 percentage points**. Tool access, however, does not consistently improve performance. MedGemma-4B-IT, for example, achieves **40.9% L3 accuracy** without tools but falls to **31.6%** with tools; its answer rate also decreases from 99.9% to 88.7%, and it never reaches the predefined evidence-sufficiency condition on scorable L3 episodes. Notably, MedGemma's projected visual representation remains highly predictive in the probing analysis, showing that strong linear decodability does not by itself imply effective tool-mediated reasoning. Together, these findings distinguish three capabilities that are often conflated: **retaining linearly decodable predictive visual information, translating that information into an answer, and actively acquiring and integrating additional evidence**.

Motivated by this gap, we introduce **EVA-Reason**, a framework for learning adaptive evidence acquisition in neuroimaging. Rather than requiring a model to answer immediately, EVA-Reason learns a sequential process in which it can request quantitative measurements, inspect returned evidence, and decide whether to acquire additional evidence or provide a final answer. We construct multi-step evidence-seeking trajectories using Claude Sonnet 4.6 as the trajectory generator and Gemini 2.5 Pro as an LLM judge for candidate trajectory selection. These trajectories are used to train a 3B vision-language model to request and integrate quantitative evidence across multiple reasoning steps. At inference time, EVA-Reason operates independently and requires no frontier-model calls.

We evaluate EVA-Reason on questions drawn from the public NeuroQA benchmark (Abbasi et al., 2026), organized into three levels of increasing reasoning complexity. L1 contains the most direct

questions, L2 requires intermediate reasoning, and L3 contains the most challenging questions, emphasizing multi-step quantitative reasoning. On the frozen held-out test set, EVA-Reason achieves **96.4% overall accuracy and 89.3% L3 accuracy**. On the same frozen test items and under the same task and tool-interface protocol, Claude Sonnet 5 achieves 93.6% overall and 80.0% L3, Gemini 2.5 Pro 94.2% and 79.6%, and GPT-5.5 94.1% and 79.1%. EVA-Reason reaches 89.3% L3 versus 80.0/79.6/79.1% zero-shot; few-shot prompting narrows several frontier gaps to non-significance (Appendix D.13).

Our contributions are threefold:

1. We identify a **representation-to-generation gap** in neuroimaging VLMs: predictive information remains strongly linearly accessible, and supervised language-side adaptation substantially improves generation but does not close the gap to the post-projector probe.
2. We systematically study **quantitative evidence access** in neuroimaging reasoning, showing that measurement tools can substantially improve difficult reasoning for models that use them effectively, while tool availability alone is insufficient.
3. We introduce **EVA-Reason**, a framework that trains a compact 3B VLM to adaptively select, request, and integrate quantitative neuroimaging evidence across multiple reasoning steps, without an external mechanism that specifies the relevant evidence set for each test item.

2 Related Work

Medical VQA and 3D neuroimaging models. Medical visual question answering has traditionally been formulated as single-shot prediction from an image and question, from radiology benchmarks such as VQA-RAD (Lau et al., 2018) and SLAKE (Liu et al., 2021) to instruction-tuned biomedical assistants such as LLaVA-Med (Li et al., 2023). Recent work extends multimodal modeling to volumetric medical images. M3D-LaMed introduces image–text and instruction data for 3D medical tasks (Bai et al., 2024), while RadFM targets generalist reasoning across 2D and 3D radiological data (Wu et al., 2025). Decipher-MR develops a vision–language foundation model for 3D MRI (Yang et al., 2026), and NeuroQA evaluates multimodal models on large-scale 3D brain MRI questions spanning anatomical, quantitative, and clinical reasoning (Abbasi et al., 2026). EVA-Reason addresses a complementary problem: whether a model can identify missing quantitative evidence, acquire it over multiple steps, and use it to answer difficult neuroimaging questions.

From visual representations to generated answers. Task-relevant information can remain accessible in internal representations without being reflected in a model's output. Linear probing provides a way to test such accessibility (Alain & Bengio, 2017). Recent multimodal analyses have examined how visual information evolves across the vision encoder, projector, and language model (Verma et al., 2024), and shown that information absent from generated responses may remain recoverable internally (Zhao et al., 2024; Fu et al., 2025). VLMs can also attend to correct visual evidence while answering incorrectly (Liu et al., 2026), while projection layers may alter representation geometry and reduce recoverability of fine-grained visual information (Li et al., 2025). Related architectural work emphasizes the importance of the visual language interface (McKinzie et al., 2024). Our neuroimaging analysis compares an end-to-end image predictor, linear probes of VLM representations, and direct generation using identical subjects, targets, and evaluation folds. Rather than treating probing as an interpretability objective, we use it to motivate evidence acquisition when predictive information remains accessible but is not reliably translated into an answer.

Tool-augmented and programmatic visual reasoning. Language models can extend their capabilities by interleaving generation with external actions. ReAct combines reasoning and actions (Yao et al., 2023), Toolformer learns to invoke external APIs (Schick et al., 2023), and Gorilla studies interaction with large API collections (Patil et al., 2024). In vision, VisProg generates modular programs for compositional reasoning (Gupta & Kembhavi, 2023), while ViperGPT composes vision and language modules through code generation (Surís et al., 2023). V* uses LLM-guided visual search to retrieve relevant image regions (Wu & Xie, 2024), and Visual Program Distillation uses LLM-generated tool programs to train VLMs for complex visual reasoning (Hu et al., 2024). EVA-

Reason similarly uses intermediate actions, but acquires named quantitative neuroimaging measurements and learns to integrate them sequentially rather than relying on a single-shot visual response.

Adaptive evidence and feature acquisition. Active feature acquisition (AFA) studies instance-specific, sequential selection of which unobserved feature to acquire next from a partially observed input. Prior work learns acquisition policies using cost-sensitive sequential prediction, reinforcement learning, and non-greedy acquisition strategies (Contardo et al., 2016; Valancius et al., 2024; Norcliffe et al., 2025). EVA-Reason instantiates active feature acquisition in neuroimaging reasoning: the acquirable features are interpretable quantitative measurements exposed through tools, acquisition depends on the question and previously observed evidence, and the model chooses between requesting another measurement and producing a final answer.

Learning from model-generated reasoning trajectories. Large models are increasingly used to supervise smaller models. Self-Instruct uses model-generated instructions and responses for instruction tuning (Wang et al., 2023), while Distilling Step-by-Step uses LLM-generated rationales to train smaller models (Hsieh et al., 2023). In multimodal reasoning, Visual Program Distillation converts LLM-generated programs into VLM supervision (Hu et al., 2024), and strong language models can serve as judges for selecting model outputs (Zheng et al., 2023). EVA-Reason follows this teacher-to-student paradigm to learn an interactive evidence-acquisition policy: a frontier model generates evidence-seeking trajectories, an LLM judge selects candidates, and additional supervision targets next-action decisions at student-visited states. Additional repair trajectories address difficult cases involving valid None answer options. The combined supervision fine-tunes a compact 3B VLM to make tool-call and final-answer decisions without the teacher or judge at inference time.

3 METHOD

3.1 EVA-REASON OVERVIEW

EVA-Reason treats neuroimaging question answering as a sequential evidence-acquisition problem. Given an MRI input I, question q, and evidence history H_t, the model chooses between

$$a_t \in \{\texttt{TOOL_CALL}(m_t),\ \texttt{FINAL}(y)\},$$

where m_t denotes a requested quantitative measurement. A successful tool call returns evidence that is appended to H_t before the next decision. Thus, the model can acquire multiple measurements before producing a final answer. During training, a trajectory-generating orchestrator (LLM 1) proposes candidate trajectories, a judge (LLM 2) selects candidates, and Qwen2.5-VL-3B-Instruct (Bai et al., 2025) serves as the student. The orchestrator and judge are used only to construct supervision (see Section 3.3). At inference time, the trained student interacts directly with the quantitative evidence interface (Section 3.2) and requires no frontier-model calls.

3.2 VISUAL INPUT AND QUANTITATIVE EVIDENCE

All evaluated models receive the same rendered T1-weighted MRI inputs. Each image is preprocessed and registered to MNI space using a structural MRI processing pipeline (Abbasi & Adeli, 2025), and rendered as one or two fixed 12-slice contact sheets according to the anatomical concepts in the question. Full preprocessing and rendering details are provided in Appendix B.1. The evidence interface exposes nine tools covering 83 quantitative neuroimaging measurements, primarily FreeSurfer-derived regional volumes and cortical thickness (Fischl et al., 2002). It additionally exposes normalized measurements and cohort-specific normative quantities, including PCNtoolkit brain-chart references where applicable (Rutherford et al., 2022a;b). A rule-based semantic router maps model requests to canonical measurements; the router itself uses no LLM or learned retrieval model. The complete tool schema and routing rules are provided in Appendix B.2.

For supervision and analysis, we use a frozen evidence audit that derives required measurable concepts from the question text and answer options only, without consulting the gold answer or hidden benchmark rule. Let R denote the required concepts and $c_r(s_t) \in \{0, 0.5, 1\}$ the evidence credit for concept r. We define

$$C(s_t) = \frac{1}{|R|} \sum_{r \in R} c_r(s_t),$$

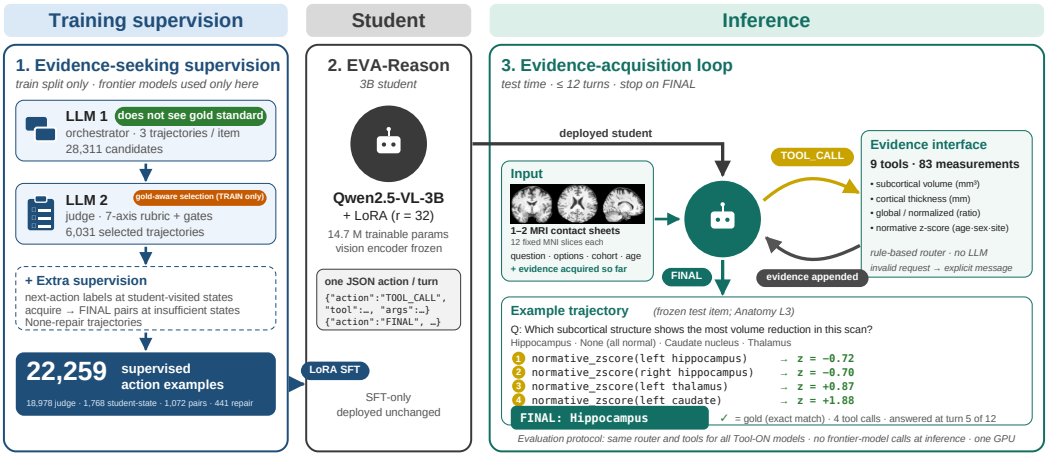


Figure 1: **Overview of EVA-Reason.** *Left:* LLM 1 generates candidate trajectories without gold answers, and LLM 2 selects them using a fixed training-only gold-aware rubric; additional supervision yields 22,259 action examples. *Center:* Qwen2.5-VL-3B-Instruct is LoRA-tuned with a frozen vision encoder to emit one `TOOL_CALL` or `FINAL` per turn. *Right:* At inference, the student interacts with the shared nine-tool, 83-measurement interface for up to 12 turns, with returned evidence appended after each call. EVA-Reason requires no frontier-model calls at inference.

and call a state sufficient when $C(s_t) = 1$. A tool call is evidence-improving when it strictly increases C. Detailed laterality and normative scoring rules are provided in Appendix B.3.

3.3 Evidence-Seeking Supervision

For each of the 9,437 training items, LLM 1 generates three candidate multi-step trajectories, yielding 28,311 candidates. LLM 1 sees only the student-visible state, including the MRI, question, answer options when present, cohort, disclosed age, available tools, and evidence history, and does not receive the gold answer or hidden benchmark rule. LLM 2 then selects training trajectories using the gold answer and a fixed quality rubric. Eligible trajectories must produce a valid final answer, satisfy fixed rubric thresholds, and pass the training-eligibility criteria. After balancing high-frequency question strata, 6,031 selected trajectories form the primary supervision source. LLM 2 is used only for training-data construction and never for test scoring. We augment these trajectories with three additional sources of supervision. First, following the motivation of dataset aggregation (Ross et al., 2011), LLM 1 provides next-action labels at states encountered during earlier student rollouts. Second, we construct paired acquisition-to-answer examples at difficult L3 states where additional evidence is needed before a final answer. Third, we add targeted repair trajectories for failures involving valid *None* answers. Privileged teacher-side information used to construct these labels is programmatically verified to be absent from every student-visible prompt. After terminal-action weighting and prompt-level conflict resolution, the final supervised corpus contains 22,259 examples: 18,978 from judge-selected trajectories, 1,768 from student-visited states, 1,072 from paired reach-sufficiency supervision, and 441 from None-repair trajectories. Full construction counts, filtering rules, and corpus arithmetic are provided in Appendix B.9.

3.4 Student Training and Inference

We fine-tune Qwen2.5-VL-3B-Instruct using LoRA with rank 32, $\alpha = 64$, and dropout 0.05 on the language-model attention projections. The vision encoder, multimodal projector, and all other non-LoRA parameters remain frozen. Training uses assistant-only supervised loss with a 4,096-token context window. The final model is trained for two epochs with AdamW and contains 14.7M trainable parameters. Complete optimization details are reported in Appendix B.10. At inference time, decoding is greedy. The model alternates between `TOOL_CALL` and `FINAL` for at most 12 turns, with early termination after four consecutive unproductive turns. Episodes ending without a valid `FINAL` are scored as incorrect. The same tool interface, router, and execution loop are used for

all Tool-ON comparison models (see Section 4.2). Representative L3 questions and EVA-Reason tool-call trajectories are shown in Appendix B.12. All supervision uses the frozen training split, with zero subject overlap across train, validation, and `test_public`. Within each SFT run, checkpoints are selected using an item-disjoint TRAIN monitor; the raw-corpus SFT variant is selected on validation L3 accuracy, then frozen before a single `test_public` evaluation. Full details are provided in Appendix B.13.

4 Experimental Setup

4.1 Benchmark and Evaluation Split

We evaluate on the frozen NeuroQA `test_public` split, containing 1,283 questions from 590 subjects across five cohorts. Primary accuracy analyses use the 939 closed-form questions, comprising 517 yes/no and 422 multiple-choice items; the remaining 344 open-ended questions are run but excluded from normalized exact-match accuracy. Among the closed-form questions, 373 are L1, 341 are L2, and 225 are L3. The L3 set contains 57 Anatomy, 93 Reasoning, and 75 Presence questions. The 225 L3 items instantiate twelve recurring question forms, listed in Appendix C.3 (Table 11). All model comparisons use this frozen evaluation set. Training and model-selection details are provided in Section 3.4.

4.2 Models and Evaluation Conditions

For supervision construction, LLM 1 is Claude Sonnet 4.6 and LLM 2 is Gemini 2.5 Pro. We compare EVA-Reason with Claude Sonnet 4.6, Claude Sonnet 5, Claude Opus 5, Gemini 2.5 Pro, Gemini 2.5 Flash, GPT-4o, GPT-5.5, Llama-4 Maverick, and MedGemma-4B-IT (Sellergren et al., 2025). EVA-Reason uses the frozen Qwen2.5-VL-3B-Instruct student described in Section 3. Exact model identifiers and completed evaluation conditions are reported in Appendix C.1. All nine comparison models are evaluated on the frozen `test_public` set under matched Tool-ON and Tool-OFF conditions. In Tool-OFF, the model receives the frozen MRI panel or panels, question, answer options when applicable, cohort, and disclosed age, and must answer directly. In Tool-ON, the model additionally receives the common evidence interface described in Section 3.2. All Tool-ON models use the same tool schemas, semantic router, backend, availability rules, 12-turn limit, and termination protocol. Sex and site are not included directly in the evaluation prompt; when available, they can only be requested through the `subject_context` tool. We additionally evaluate matched Qwen-3B controls with no quantitative evidence and with a prespecified relevant-evidence set supplied upfront. For the latter, a frozen deterministic mapping identifies relevant measurements from the question and answer options only, without access to the gold answer or hidden benchmark rule. The model then produces a single final answer without tool calling or multi-turn interaction. An all-evidence static condition is reported in Appendix D.5. No evaluation-time model receives the gold answer, hidden benchmark rule, question category, evidence audit, or sufficiency label. EVA-Reason makes no frontier-model calls at inference time.

4.3 Scoring and Behavioral Metrics

Our primary outcome is normalized exact-match accuracy. Predictions and answer options are lower-cased, stripped of leading option-letter prefixes, and normalized for punctuation and whitespace; a prediction is correct only when the resulting text exactly matches the gold option. Overall accuracy is computed over the 939 closed-form items and L3 accuracy over the 225 closed-form L3 items. As a formatting sensitivity analysis, we also report option-prefix accuracy, which additionally accepts a response that begins with the complete normalized text of the unique gold option, allowing trailing explanatory text. The same rule is applied to every model. Full scoring rules are given in Appendix C.4. We additionally characterize evidence-seeking behavior using answer rate, reach-sufficiency, tool calls, redundant calls, and execution errors. Reach-sufficiency is evaluated over the 200 sufficiency-scorable L3 items and records whether the model reaches $C(s) = 1$ before termination. It is a behavioral diagnostic rather than an accuracy criterion: a model may answer correctly without satisfying the audit, or acquire sufficient evidence and still answer incorrectly. Because the audit measures complete coverage of predefined measurable concepts, $C(s) = 1$ need not coincide

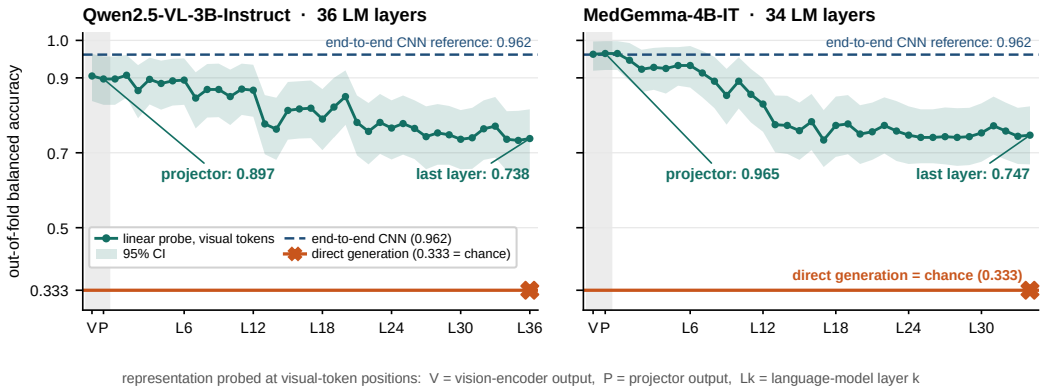


Figure 2: Layer-wise linear probes on the controlled 279-subject task (out-of-fold balanced accuracy, 95% bootstrap CI; chance = 0.333). Projected visual tokens reach 0.897 (Qwen2.5-VL-3B) and 0.965 (MedGemma-4B-IT), declining to 0.738 and 0.747 at the final layer while remaining well above chance. The end-to-end CNN reaches 0.962; direct generation from both VLMs is at chance.

with the minimal evidence sufficient to distinguish among answer options; a model may therefore answer correctly before reaching audit sufficiency.

4.4 Statistical Analysis

All accuracy comparisons are paired. To account for multiple questions per subject, primary inference uses a subject-cluster bootstrap with 4,000 resamples (seed 0) and a two-sided paired sign-flip test over subject-level accuracy differences with 20,000 Monte Carlo permutations. We additionally report exact two-sided McNemar tests and paired item-level bootstrap intervals. Subject-level inference remains primary. For the nine EVA-Reason versus Tool-ON comparisons, secondary item-level p-values are Holm-corrected; all remain significant. Tool-ON versus Tool-OFF comparisons use the same paired analyses. Full definitions are given in Appendix C.5. All model development precedes test evaluation, and the selected checkpoint is frozen before its single `test_public` evaluation.

5 Results

5.1 Predictive Information Is Linearly Decodable but Poorly Read Out

On the controlled 279-subject task, an end-to-end CNN reaches 0.962 balanced accuracy (95% CI [0.927, 0.990]), while direct generation from both VLMs remains at chance (0.333). We therefore probe their visual-token representations to determine where task-relevant information remains linearly accessible.

Figure 2 shows that linear decodability declines across later language-model layers. Language-side LoRA fine-tuning on the same folds, with the vision encoder and projector frozen, raises Qwen2.5-VL generation from 0.333 to 0.743, below the 0.897 post-projector probe. Separately, five raw measurements raise zero-shot balanced accuracy from 0.623/0.496/0.342 to 0.894/0.776/0.686 for GPT-5.5/Gemini 2.5 Pro/ Claude Sonnet 5, respectively (Appendix A).

5.2 Quantitative Evidence Helps Only When Models Can Use It

Figure 3A compares Tool-OFF and Tool-ON performance across nine models. Tool-OFF L3 accuracy ranges from 36.4% (GPT-4o) to 56.4% (Gemini 2.5 Flash), with strongly model-dependent effects of quantitative evidence.

Four models improve significantly on L3: Claude Sonnet 5 from 43.6% to 80.0% ($+36.4$, 95% CI $[+29.3, +43.6]$, $p < 0.0001$), GPT-5.5 from 53.3% to 79.1% ($+25.8$, $[+16.9, +34.2]$, $p < 0.0001$), Gemini 2.5 Pro from 56.0% to 79.6% ($+23.6$, $[+16.4, +31.1]$, $p < 0.0001$), and Claude Sonnet 4.6 from 55.1% to 75.6% ($+20.4$, $[+13.3, +28.0]$, $p < 0.0001$).

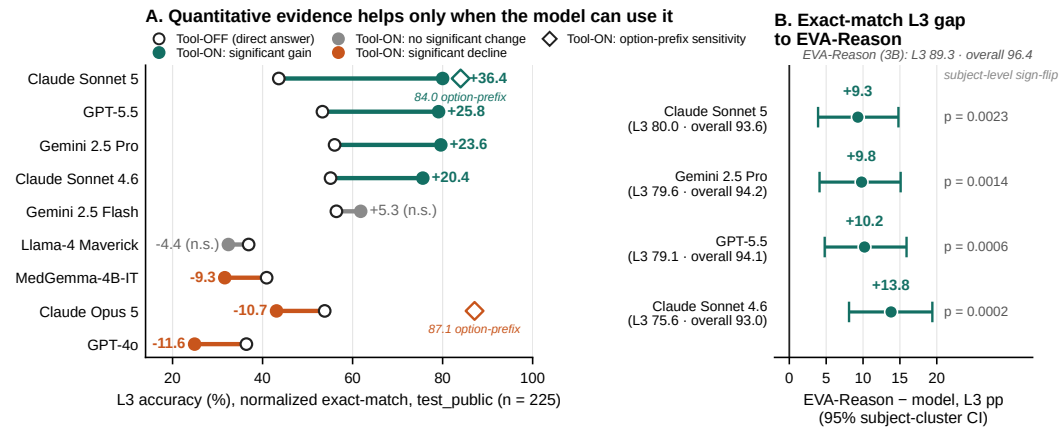


Figure 3: **Quantitative evidence access and EVA-Reason on frozen `test_public`.** *(A)* L3 accuracy for nine models without tools (hollow) and with the common evidence interface (filled); colors denote significant gains, no change, or declines, and diamonds show option-prefix sensitivity. *(B)* EVA-Reason minus the four strongest Tool-ON models on L3, with 95% subject-cluster CIs and sign-flip p-values; labels report L3/overall accuracy.

5.3 EVA-Reason vs. Matched Controls and Frontier Models

As summarized in Figure 3B, EVA-Reason reaches 96.4% accuracy over all 939 closed-form test items and 89.3% on the 225 L3 questions, while answering every item. On L3, it exceeds Claude Sonnet 5 by 9.3 percentage points (95% subject-cluster CI $[+3.9, +14.8]$, subject-level sign-flip $p = 0.0023$), Gemini 2.5 Pro by 9.8 points ($[+4.1, +15.1]$, $p = 0.0014$), GPT-5.5 by 10.2 points ($[+4.8, +15.9]$, $p = 0.0006$), and Claude Sonnet 4.6 by 13.8 points ($[+8.1, +19.4]$, $p = 0.0002$). In the secondary item-level analysis, all four headline comparisons remain significant after Holm step-down correction across all nine EVA-Reason versus Tool-ON L3 comparisons (exact two-sided McNemar tests; maximum Holm-adjusted $p = 0.0052$). Under option-prefix sensitivity, EVA-Reason remains at 89.3% L3, while Sonnet 5 and Opus 5 reach 84.0% and 87.1%; the corresponding gaps are not significant ($p = 0.088$ and 0.52; Appendix D.3). With three tool-use demonstrations, the gaps also become nonsignificant for Sonnet 5 and GPT-5.5 (Appendix D.13). Matched Qwen-3B controls further separate supervised adaptation and evidence selection. Static Relevant-Evidence provides a prespecified relevance reference: it reaches 88.9% L3 accuracy versus 89.3% for EVA-Reason (p=1.00), but requires a fixed deterministic question-to-measurement mapping. EVA-Reason instead selects evidence autonomously and reaches 96.4% overall accuracy versus 94.6% for the static control (+1.8 points, subject-level sign-flip p=0.036). The static mapping yields no measurement for 75/939 items. On L3, the models disagree on 33 items: EVA-Reason alone is correct on 17 and Static Relevant-Evidence alone on 16; 16 of the 17 EVA-only successes include evidence outside the prespecified static plan, and nine add a required concept under the frozen audit (Appendix D.5). EVA-Reason's L3 performance is not explained solely by template-associated answer priors. On the 57 L3 items whose gold answers oppose the TRAIN-template majority, EVA-Reason reaches 68.4% accuracy versus 38.6% for Direct-SFT (subject-level p=0.0058; Appendix D.5). EVA-Reason also runs locally on one L40S with lower observed latency and fewer output tokens than frontier comparators (Appendix D.9).

5.4 Evidence Acquisition and Failure Modes

EVA-Reason returns a valid `FINAL` on every L3 episode; the sufficiency audit is reached in 76/200 (38.0%) scorable episodes, with 94.7% accuracy (72/76) after sufficiency versus 86.3% (107/124) otherwise. Consistent with its role as a behavioral coverage diagnostic rather than an accuracy criterion, reach-sufficiency varies substantially across Tool-ON models with similar L3 accuracy (Appendix Fig. 5). Failure analysis separates acquisition/stopping from integration errors (Appendix D.11). Evidence-feedback interventions probe evidence use and sequential adaptation. Re-

Table 1: Normalized exact-match accuracy (%) on frozen `test_public`. Overall: 939 closed-form items; L3: 225.

Model	Overall	**L3**	Anat.	Reason.	Pres.
EVA-Reason (3B)	**96.4**	**89.3**	**100.0**	81.7	90.7
Qwen-3B Direct-SFT	64.5	82.2	86.0	73.1	90.7
Qwen-3B Static Relevant-Evidence	94.6	88.9	98.2	75.3	**98.7**
Claude Sonnet 5	93.6	80.0	68.4	86.0	81.3
Gemini 2.5 Pro	94.2	79.6	49.1	**91.4**	88.0
GPT-5.5	94.1	79.1	70.2	88.2	74.7
Claude Sonnet 4.6	93.0	75.6	49.1	83.9	85.3

placing returned values with same-cohort donor values while fixing the tool-call sequence reduces L3 accuracy from 89.3% to 69.8%. Withholding values during acquisition and revealing them only afterward reduces L3 to 76.9% (-12.4 points, 95% subject-cluster CI $[-16.9, -7.6]$, $p = 5 \times 10^{-5}$), despite exposing 4.80 versus 3.5 scalar values per item. Acquisition sets have mean Jaccard similarity 0.709, consistent with intermediate evidence influencing subsequent selection and stopping (Appendix D.6). Removing the rendered MRI panels while retaining subject-specific MRI-derived quantitative evidence produces little L3 change for EVA-Reason and the strongest Tool-ON models, whereas several Tool-OFF models remain sensitive to image removal (Appendix D.7). Thus, under Tool-ON conditions, performance is driven primarily by MRI-derived quantitative evidence rather than direct dependence on the rendered panels, isolating evidence acquisition and integration as the primary capability evaluated here; the setting is nevertheless not imaging-free because the retained evidence is itself MRI-derived. Excluding ABCD from training yields 97.8% overall and 90.3% L3 on 271 held-out items (62 L3), versus 98.5% and 93.5% for standard EVA-Reason (Appendix D.10).

6 Discussion and Limitations

Our results separate three capabilities that are often conflated in multimodal systems: retaining linearly decodable visual information, translating that information into a generated answer, and acquiring additional quantitative evidence. The representation probe shows that predictive information can remain linearly accessible even when direct generation is at chance. In contrast to natural-image evaluations (Fu et al., 2025), linear decodability degrades across language-model layers rather than remaining constant (Figure 2). EVA-Reason demonstrates that a compact 3B model can learn to acquire and use such evidence, reaching 89.3% L3 accuracy without frontier-model calls at inference.

Counter-majority analysis further shows that EVA-Reason's L3 performance is not explained solely by template-associated answer priors (Appendix D.5). Performance is largely preserved when ABCD is excluded from training (Appendix D.10). The evidence interface uses FreeSurfer-derived measurements and cohort-specific normative resources, and the sufficiency audit applies to 200 of 225 L3 items and is designed to assess measurement coverage rather than enumerate all valid reasoning paths. Notably, the PCNtoolkit normative z-scores exposed through the interface are not used to construct NeuroQA labels, and benchmark decision thresholds are not exposed to evaluated models (Appendix C.3). The headline L3 comparisons also remain significant after excluding all 41 "NOT present" Presence items (Appendix D.2). Three-shot tool-use demonstrations raise Sonnet 5/GPT-5.5/Gemini 2.5 Pro to 86.7/86.2/82.2% L3, showing that few demonstrations can substantially improve evidence acquisition (Appendix D.13). The present evaluation focuses on structural T1 MRI and a quantitative neuroimaging evidence interface; extension to other imaging modalities and measurement pipelines remains future work.

Finally, Direct-SFT, Static Relevant-Evidence, and EVA-Reason reach 64.5/82.2%, 94.6/88.9%, and 96.4/89.3% overall/L3, respectively. EVA-Reason matches prespecified relevant evidence on L3 while learning to select evidence autonomously and adapt acquisition to intermediate evidence. Full-Evidence reaches 94.7% L3 using 138.6 values/item versus 3.5 for EVA. The goal of selective acquisition is therefore not to outperform exhaustive evidence exposure, but to identify a compact, interpretable evidence subset while avoiding large amounts of irrelevant context. Withholding intermediate feedback reduces L3 to 76.9%, further supporting selective, sequential evidence use (Appendix D.5; Appendix D.6).

AI USE STATEMENT

Generative AI tools were used for language and grammar editing of the manuscript and for assistance with code debugging. All scientific content, research questions, methodological decisions, experimental design, analyses, interpretation of results, and conclusions were developed and verified by the authors. The authors reviewed all AI-assisted output and take full responsibility for the content of this paper.

ETHICS STATEMENT

This work uses existing de-identified neuroimaging data and collects no new data from human participants. The datasets underlying the NeuroQA benchmark are used under their applicable access conditions, licenses, and data-use agreements. The study does not attempt to re-identify participants, and no personally identifying information is provided to the evaluated models.

EVA-Reason is developed and evaluated as a research system for studying multimodal neuroimaging reasoning and adaptive evidence acquisition. It is not intended for clinical diagnosis, treatment selection, or other patient-level medical decision making. The quantitative evidence interface relies primarily on FreeSurfer-derived anatomical measurements, normalized quantities, and cohort-specific normative references. These resources may reflect biases associated with source cohorts, demographic composition, imaging protocols, preprocessing methods, and normative reference populations. Performance reported in this work should therefore not be interpreted as evidence of clinical validity or guaranteed generalization to populations, institutions, diseases, or imaging protocols not represented in the evaluation.

Frontier models are used during supervision construction, as described in Section 3. Claude Sonnet 4.6 generates candidate evidence-seeking trajectories and Gemini 2.5 Pro selects training trajectories. These models are not used by EVA-Reason at inference time.

REPRODUCIBILITY STATEMENT

Section 3 describes the EVA-Reason formulation, quantitative evidence interface, supervision construction, student training, inference procedure, and split-integrity protocol. Section 4 specifies the frozen evaluation set, comparison models, Tool-ON and Tool-OFF conditions, scoring rules, behavioral metrics, and statistical analysis.

The appendix provides additional implementation details. MRI preprocessing and rendering are described in Appendix B.1; the quantitative evidence interface and semantic routing procedure in Appendix B.2; and the evidence-sufficiency audit in Appendix B.3. Supervision construction, filtering, and corpus accounting are reported in Appendix B.9. Optimization settings are provided in Appendix B.10. Checkpoint selection, variant selection, and split integrity are documented in Appendix B.13. Exact evaluated model versions and inference configurations are given in Appendix C.1, while scoring procedures and statistical tests are specified in Appendix C.4 and Appendix C.5. The controlled representation-probing experiment is reported in Appendix A. All frozen `test_public` trajectories of EVA-Reason and of every Tool-ON comparison model, including tool calls, returned observations, and final answers, are released as supplementary material.

REFERENCES

Mohammad H. Abbasi and Ehsan Adeli. sMRI Processing Pipeline: A Lightweight, End-to-End Workflow for Structural Brain MRI Preprocessing and Quality Control. Zenodo, v1.0.0. doi: 10.5281/zenodo.17503175, 2025.

Mohammad H. Abbasi, Favour Nerrise, Shaurnav Ghosh, Ridvan Yesiloglu, Yuncong Mao, Bailey Trang, Mohammad Asadi, Merryn Daniel, Gustavo Chau Loo Kung, Ken Chang, Pavan Pinkesh Shah, Adam Turnbull, Kyan Younes, Seena Dehkharghani, and Ehsan Adeli. NeuroQA: A large-scale image-grounded benchmark for 3D brain MRI understanding. *arXiv preprint arXiv:2605.20525*, 2026.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (Workshop Track)*, 2017. arXiv:1610.01644.

Fan Bai, Yuxin Du, Tiejun Huang, Max Q.-H. Meng, and Bo Zhao. M3D: Advancing 3D medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Benjamin Billot, Douglas N. Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V. Dalca, and Juan Eugenio Iglesias. SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Medical Image Analysis*, 86:102789, 2023. doi: 10.1016/j.media.2023.102789.

Gabriella Contardo, Ludovic Denoyer, and Thierry Artières. Sequential cost-sensitive feature acquisition. In *Advances in Intelligent Data Analysis XV*, volume 9897 of *Lecture Notes in Computer Science*, pp. 284–294. Springer, 2016. doi: 10.1007/978-3-319-46349-0_25.

Bruce Fischl, David H. Salat, Evelina Busa, Marilyn Albert, Megan Dieterich, Christian Haselgrove, Andre van der Kouwe, Ron Killiany, David Kennedy, Shuna Klaveness, Albert Montillo, Nikos Makris, Bruce Rosen, and Anders M. Dale. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355, 2002. doi: 10.1016/S0896-6273(02)00569-X.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: VLMs overlook their visual representations. In *Proceedings of the Second Conference on Language Modeling (COLM)*, 2025.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14953–14962, 2023.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.507.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9590–9601, 2024.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. doi: 10.1038/s41592-020-01008-z.

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5:180251, 2018. doi: 10.1038/sdata.2018.251.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, pp. 28541–28564, 2023. doi: 10.52202/075280-1240.

Wenyan Li, Raphael Tang, Chengzu Li, Caiqi Zhang, Ivan Vulić, and Anders Søgaard. Lost in embeddings: Information loss in vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 22676–22693. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1235.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1650–1654, 2021. doi: 10.1109/ISBI48211.2021.9434010.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Jingying Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Hanqing Lu, Benoit Dumoulin, and Hanghang Tong. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in VLMs. In *International Conference on Learning Representations*, 2026.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Samuel Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. MM1: Methods, analysis & insights from multimodal LLM pre-training. In *Computer Vision – ECCV 2024*, volume 15087 of *Lecture Notes in Computer Science*, pp. 304–323. Springer, 2024. doi: 10.1007/978-3-031-73397-0_18.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-Flamingo: A multimodal medical few-shot learner. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pp. 353–367. PMLR, 2023.

Alexander Luke Ian Norcliffe, Changhee Lee, Fergus Imrie, Mihaela van der Schaar, and Pietro Lio. Stochastic encodings for active feature acquisition. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 46784–46814. PMLR, 2025.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive APIs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 126544–126565, 2024. doi: 10.52202/079017-4020.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635. PMLR, 2011.

Saige Rutherford, Charlotte Fraza, Richard Dinga, Seyed Mostafa Kia, Thomas Wolfers, Mariam Zabihi, Pierre Berthet, Amanda Worker, Serena Verdi, Derek Andrews, Laura K. M. Han, Johanna M. M. Bayer, Paola Dazzan, Phillip McGuire, Roel T. Mocking, Aart Schene, Chandra Sripada, Ivy F. Tso, Elizabeth R. Duval, Soo-Eun Chang, Brenda W. J. H. Penninx, Mary M. Heitzeg, S. Alexandra Burt, Luke W. Hyde, David Amaral, Christine Wu Nordahl, Ole A. Andreassen, Lars T. Westlye, Roland Zahn, Henricus G. Ruhe, Christian F. Beckmann, and Andre F. Marquand. Charting brain growth and aging at high spatial precision. *eLife*, 11:e72904, 2022a. doi: 10.7554/eLife.72904.

Saige Rutherford, Seyed Mostafa Kia, Thomas Wolfers, Charlotte Fraza, Mariam Zabihi, Richard Dinga, Pierre Berthet, Amanda Worker, Serena Verdi, Henricus G. Ruhe, Christian F. Beckmann, and Andre F. Marquand. The normative modeling framework for computational psychiatry. *Nature Protocols*, 17:1711–1734, 2022b. doi: 10.1038/s41596-022-00696-5.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68539–68551, 2023. doi: 10.52202/075280-2997.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via Python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11888–11898, 2023. doi: 10.1109/ICCV51070.2023.01092.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical AI. *NEJM AI*, 1(3):AIoa2300138, 2024. doi: 10.1056/AIoa2300138.

Michael Valancius, Maxwell Lennon, and Junier Oliva. Acquisition conditioned oracle for non-greedy active feature acquisition. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48957–48975. PMLR, 2024.

Gaurav Verma, Minje Choi, Kartik Sharma, Jamelle Watson-Daniels, Sejoon Oh, and Srijan Kumar. Cross-modal projection in multimodal LLMs doesn't really project visual attributes to textual space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 657–664. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-short.60.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-Instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.754.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. *Nature Communications*, 16(1):7866, 2025. doi: 10.1038/s41467-025-62385-7.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024. doi: 10.1109/CVPR52733.2024.01243.

Zhijian Yang, Noel DSouza, Istvan Megyeri, Xiaojian Xu, Amin Honarmandi Shandiz, Farzin Haddadpour, Krisztian Koos, Laszlo Rusko, Emanuele Valeriano, Bharadwaj Swaminathan, Lei Wu, Parminder Bhatia, Taha Kass-Hout, and Erhan Bas. Decipher-MR: A vision-language foundation model for 3D MRI representations. *npj Digital Medicine*, 9:419, 2026. doi: 10.1038/s41746-026-02596-4.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1):AIoa2400640, 2025. doi: 10.1056/AIoa2400640.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*, 4:277, 2024. doi: 10.1038/s43856-024-00709-2.

Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. The first to know: How token distributions reveal hidden knowledge in large vision-language models? In *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science, pp. 127–142. Springer, 2024. doi: 10.1007/978-3-031-73195-2_8.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020.

Table 2: Balanced accuracy of the controlled CNN reference, frozen representation probes, fine-tuned generation, and direct generation on the same 279-subject task.

Representation / readout	Qwen2.5-VL-3B	MedGemma-4B-IT
Reference and CNN-study controls		
End-to-end CNN reference	0.962 [0.927, 0.990]	
ICV-only control	0.459	
Whole-brain-R0 control	0.411	
Frozen VLM representation probes		
Vision encoder	0.905 [0.838, 0.964]	0.964 [0.919, 0.997]
Post-projector / LLM input	0.897 [0.828, 0.957]	0.965 [0.921, 0.998]
Middle LLM layer (18 / 17)	0.790 [0.707, 0.870]	0.735 [0.662, 0.814]
Final LLM layer (36 / 34)	0.738 [0.653, 0.816]	0.747 [0.669, 0.824]
Fine-tuned generation (Qwen, LM-side LoRA)	0.743 [0.668, 0.822]	–
Direct generation	0.333	0.333

A Visual Representation Probe

We test whether predictive neuroimaging information remains linearly decodable from a VLM's visual representations even when the model does not express that information in its generated answer.

The controlled task uses 279 HCP-YA subjects (one session each), labelled with a three-class bilateral-concordant lateral-ventricle normative band (LOW 22, NORMAL 217, HIGH 40). The 12 fixed-geometry R1-V ventricular tiles of each subject are the image input. An end-to-end CNN trained on these 12 tiles as input channels serves as the reference and is evaluated by out-of-fold balanced accuracy over five subject-level folds (57/57/55/55/55). Every probe below uses the same subjects, labels, folds, and metric; chance balanced accuracy is 0.333. No metadata variables (age, sex, or intracranial volume) are supplied to any representation probe.

The primary probe presents the 12 R1-V tiles as 12 separate 256×256 RGB images, producing 972 visual tokens for Qwen2.5-VL-3B and 3,072 for MedGemma-4B-IT. We use Qwen2.5-VL-3B-Instruct at frozen revision `66285546` and MedGemma-4B-IT at revision `290cda5e`.

Frozen features are mean-pooled over the visual-token positions and extracted from:

1. the vision-encoder output before the projector;
2. the projector output, i.e. visual tokens entering the language model;
3. a middle language-model layer (layer 18 of 36 for Qwen2.5-VL-3B; layer 17 of 34 for MedGemma-4B-IT); and
4. the final language-model layer (36 / 34, after final normalization).

For the layer-wise analysis in Figure 2, the same procedure is applied to every language-model layer of the 12-tile input. Layer 0 denotes the embedding output, which is identical to the projector output, and layer k denotes the output of decoder block k.

Each probe uses a `StandardScaler` followed by multinomial logistic regression with L2 regularization. The regularization parameter is selected from

$$C \in \{0.001, 0.01, 0.1, 1\}$$

by inner stratified five-fold cross-validation using balanced accuracy (shuffle seed 1234), performed only within the training fold of each outer fold. Class weights are balanced, the solver is L-BFGS, and no PCA or other dimensionality reduction is used. All 95% confidence intervals below, including the CNN reference, are percentile intervals from 2,000 subject-level bootstrap resamples of the out-of-fold predictions (seed 0).

For direct zero-shot generation, each model receives the same 12 images with a system prompt identifying them as MRI slices through the lateral ventricles of an HCP Young Adult subject and asking the model to judge ventricular volume relative to age- and sex-matched norms. The user

Table 3: Zero-shot balanced accuracy on the controlled task with MRI alone or with five raw volumetric measurements provided upfront.

Model	MRI-only	+ Raw measurements	Δ
GPT-5.5	0.623	0.894	+0.271
Gemini 2.5 Pro	0.496	0.776	+0.280
Claude Sonnet 5	0.342	0.686	+0.344

prompt provides cohort, age in whole years, and sex and asks whether bilateral lateral-ventricular volume is *LOW*, *NORMAL*, or *HIGH*, with an instruction to answer with exactly one of these labels. Decoding is greedy with at most 16 new tokens, and every response parses to one of the three labels. Qwen2.5-VL predicts *HIGH* for all 279 subjects, while MedGemma predicts *NORMAL* for 278 of 279, yielding balanced accuracy 0.333 for both. This protocol is separate from the benchmark evaluation.

Fine-tuned generation. To test whether supervised adaptation improves generated readout, we fine-tune Qwen2.5-VL with language-side LoRA using the same five subject-level folds, while keeping the vision encoder and multimodal projector frozen. Each fold uses a fresh language-side LoRA ($r = 32$, $\alpha = 64$, dropout $= 0.05$; $\mathtt{q/k/v/o}$ projections), three epochs (96 optimizer steps), micro-batch 1 with gradient accumulation 16, AdamW at 10^{-4}, class-balanced oversampling, seed 20260914, and the fixed end-of-training checkpoint; decoding is greedy with 16 new tokens. Pooled out-of-fold balanced accuracy reaches 0.743 (95% CI [0.668, 0.822]), compared with 0.333 for frozen direct generation and 0.897 for the frozen post-projector linear probe. The difference from the frozen final-layer probe is small and not significant ($\Delta = +0.005$, 95% paired-bootstrap CI $[-0.101, +0.108]$, sign-flip $p = 0.127$). In contrast, fine-tuned generation remains significantly below the post-projector probe ($\Delta = -0.154$, $[-0.254, -0.050]$, $p = 0.0114$). All 279 responses are parseable; per-class recall is 0.409 for LOW, 0.945 for NORMAL, and 0.875 for HIGH, with residual error concentrated in the minority LOW class ($n = 22$). We run this adaptation control only for Qwen2.5-VL, the EVA-Reason student backbone; MedGemma is retained as a frozen-representation comparison.

Explicit quantitative evidence. We additionally evaluate GPT-5.5, Gemini 2.5 Pro, and Claude Sonnet 5 on the same 279-subject task with five raw volumetric measurements provided upfront: left and right lateral-ventricle volume, total ventricular volume, intracranial volume, and whole-brain volume. No normative z-score, target band, or gold label is provided, and no training, tool calling, or sequential interaction is used.

Raw measurements improve balanced accuracy for all three frontier models. This controlled analysis tests whether explicit quantitative information can improve readout; it does not test adaptive versus static evidence acquisition.

We additionally repeat the probe using a single 768×256 contact-sheet input. The qualitative pattern remains: post-projector balanced accuracy is 0.860 for Qwen and 0.868 for MedGemma; middle/final language-model probes are 0.692/0.670 and 0.764/0.707, respectively. Direct generation remains at 0.333 for Qwen and 0.330 for MedGemma.

These findings show that predictive information remains linearly decodable after visual projection, is not reflected in zero-shot generated answers, and is partially recovered by supervised language-side adaptation. Two caveats apply. First, the linear probe and fine-tuned generator are different supervised readouts, so the numerical proximity between fine-tuned generation and the frozen final-layer probe should be interpreted descriptively rather than as evidence of equivalence. Second, the decrease in later-layer probe performance indicates reduced linear accessibility inside the frozen language model, not that the information is destroyed.

Table 4: Quantitative evidence tools available to EVA-Reason and Tool-ON comparison models.

Tool	Evidence provided	Units
`subcortical_volume`	Bilateral subcortical structure volume	mm^3
`cortical_thickness`	Regional cortical thickness	mm
`cortical_volume`	Regional cortical gray-matter volume	mm^3
`global_measure`	Global brain-volume measurements	mm^3
`laterality`	Bilateral values and asymmetry	mm^3/mm, index
`normalized_volume`	Volume normalized by a global measure	ratio / %
`normative_zscore`	Normative deviation for a measurement	z-score
`subject_context`	Age, sex, or site	–
`mri_panel`	Requested anatomical MRI panel	–

B Additional Method and Reproducibility Details

This appendix provides implementation details omitted from the main paper for space. Unless otherwise stated, all values below are taken from the frozen data-construction, training, and evaluation artifacts used for the final EVA-Reason model.

B.1 MRI Rendering and Visual Inputs

All models are evaluated using the same rendered T1-weighted MRI inputs. Source volumes are registered to MNI nonlinear space, z-score normalized, and spatially cropped before rendering.

The frozen renderer extracts 12 slices at predefined world-coordinate locations. Slices are intensity-windowed using the 1st–99th percentile and combined into a 6×2 contact sheet with a final resolution of 768×256 pixels.

Two predefined panel configurations, R1HB and R1VH, provide complementary anatomical views. Depending on the anatomical concepts referenced by a question, each benchmark item receives one or two contact sheets. For a given item, all evaluated models receive exactly the same rendered PNG input.

B.2 Quantitative Evidence Interface

The evidence interface contains nine canonical tools spanning 83 FreeSurfer-derived quantitative measurement keys (Fischl et al., 2002). These comprise 59 volumetric measurements in mm^3 and 24 cortical-thickness measurements in mm, with 41 core measurements available across all FreeSurfer cohorts.

Normative z-scores are obtained from frozen PCNtoolkit lifespan models conditioned on age, sex, and site (Rutherford et al., 2022a;b). Only cohort–measurement pairs passing the frozen calibration criteria are available through the normative tool.

Semantic router. Tool requests are resolved by a deterministic rule-based router using fixed concept mappings and cohort-specific tool availability. No LLM, embedding retrieval, or learned routing model is used. Invalid or unavailable tool requests are returned as explicit error observations. The same evidence interface and router are used for EVA-Reason and all Tool-ON comparison models.

B.3 Evidence Sufficiency

The evidence audit determines whether the observations collected at a state cover the measurable requirements expressed by the question. Requirements are derived deterministically from the frozen question text and answer options, without access to the gold answer, benchmark rule, or hidden decision threshold.

For item i, the requirement set R_i is constructed from measurable concepts identified by a fixed concept dictionary and restricted to measurements available through the item's cohort-specific tools.

Table 5: Orchestrator generation configuration.

Model	`claude-sonnet-4-6`
Candidates per item	3
Temperature	1.0
Thinking budget	2,048 tokens
Maximum output tokens	4,096
Maximum interaction turns	12
Early stop	4 consecutive unproductive turns

Morphological concepts outside the quantitative evidence interface are excluded. Items without a non-empty explicit requirement set are not scored for sufficiency.

Each required concept maps to canonical measurement keys and a laterality mode: *bilateral*, *one-side*, *aggregate*, or *aggregate-or-bilateral*. Let l_r, r_r, and a_r indicate whether left, right, and aggregate evidence has been observed for requirement r. Concept credit is

$$c_r(s) = \begin{cases} (l_r + r_r)/2, & \text{bilateral}, \\ \max(l_r, r_r), & \text{one-side}, \\ a_r, & \text{aggregate}, \\ \max(a_r, (l_r + r_r)/2), & \text{aggregate-or-bilateral}. \end{cases}$$

Thus $c_r(s) \in \{0, 0.5, 1\}$. For normative requirements, `normative_zscore` provides credit when a calibrated normative reference is available; otherwise, the corresponding descriptive measurement can provide credit.

Overall evidence credit is

$$C(s) = \frac{1}{|R_i|} \sum_{r \in R_i} c_r(s),$$

and a state is considered sufficient iff $C(s) = 1$. The sufficiency turn is the first successful tool-call index after which this condition holds. A call is *credit-raising* if it strictly increases $C(s)$. Only evidence returned by successful tool calls is counted.

The same frozen evidence-audit implementation is used for paired supervision, reach-sufficiency evaluation, and all reported sufficiency diagnostics. The audit itself is never exposed to the model at inference time.

B.4 Trajectory-Generating Orchestrator

Claude Sonnet 4.6 is used as the trajectory-generating orchestrator. For each of the 9,437 TRAIN items, three independent candidate trajectories are generated:

$$9{,}437 \times 3 = 28{,}311.$$

The orchestrator sees only student-visible information: the question, answer options when present, cohort, disclosed age, available tool schemas, MRI panels, evidence accumulated so far, and the turn counter. It does not receive the gold answer, benchmark category, rule family, hidden rule, or hidden threshold.

The exact frozen system instruction is:

> You are an expert neuroimaging analyst answering one question about one subject's brain MRI.
>
> You work step by step. On each turn you either call ONE tool to acquire evidence, or give your final answer.
>
> Respond with a single valid JSON object and nothing else.

Table 6: Frozen Gemini judge rubric.

Axis	Weight	Definition
`relevance`	0.15	Selected tools are those an expert would consider informative.
`evidence_coverage`	0.20	Evidence required by the question was acquired.
`sufficiency`	0.25	Acquired evidence supports the conclusion drawn.
`efficiency`	0.05	Evidence was acquired directly without unnecessary wandering.
`answer_consistency`	0.20	Reasoning entails the final answer and the answer is correct.
`premature_stopping`	0.10	The trajectory continued acquiring evidence until it could answer.
`tool_economy`	0.05	No redundant, repeated, or uninformative tool calls.

To acquire evidence:

```
{"thought": "<why this evidence, in scientific terms>",
"action": "TOOL_CALL", "tool": "<tool name>",
"args": {<arguments>}}
```

To answer:

```
{"thought": "<how the evidence you gathered supports this
conclusion>",
"action": "FINAL", "answer": "<your answer>"}
```

Guidance:

- Acquire the evidence you actually need to answer THIS question, and stop when you have it.
- A tool returns a measured quantity, never a judgement. Interpretation is yours.
- If a tool reports that something is unavailable, adapt rather than repeat the same call.
- Do not write any text outside the JSON object.

The user turn contains the question; answer options, or an explicit open-ended marker; subject context including cohort and disclosed age; available tool schemas; evidence acquired so far; the current turn count out of 12; and an instruction to make one tool call or return the final answer.

B.5 Gemini Judge and Candidate Selection

Candidate trajectories are evaluated by `gemini-2.5-pro` with temperature 0. Unlike the orchestrator, the judge receives the gold answer and benchmark category because its role is training-data construction rather than deployed inference.

Candidates must satisfy the following hard requirements:

- the episode produced a final answer;
- the judge output is parseable;
- no backend failure occurred; and
- closed-form items are answered correctly.

Eligible candidates are evaluated on seven axes, each scored from 1 to 5, where 5 is best.

The judge additionally produces `answer_correct` (true/false, or null for open-ended items) and an `sft_eligible` flag. The normalized weighted score is

$$J(\tau) = \frac{\sum_i w_i s_i - 1}{4},$$

mapping the weighted 1–5 rubric scale to $[0, 1]$.

A candidate is eligible for supervised training only if

$$\texttt{sft_eligible} = \texttt{true}, \qquad J(\tau) \geq 0.70, \qquad s_i > 2 \; \forall i.$$

Overall, 9,086 of the 28,311 candidate trajectories are selected (32.1%). To reduce over-representation of high-frequency strata, L1-Measurement is capped from 2,500 to 700 trajectories and L2-Location from 2,055 to 800, yielding 6,031 trajectories for the main teacher component.

B.6 Student-Visited-State Supervision

To expose the student to states induced by its own behavior, we collect teacher next-action labels at states visited by earlier student policies, following the motivation of dataset aggregation (Ross et al., 2011).

Two collection rounds are used:

- **Iteration 1:** greedy rollouts of a partially supervised student on 800 difficult TRAIN items produce 1,002 states and 992 teacher labels.
- **Iteration 2:** rollouts of the updated student on 1,000 TRAIN items under fixed sampling quotas produce 1,248 additional candidate states.

Across both rounds, Claude produces 2,240 next-action labels, including 1,020 *KEEP* labels for states where the teacher agrees with the student's proposed action. The resulting supervision therefore contains both corrections and successful student actions.

For labeling, the teacher receives the same student-visible context together with a supervisor-only gold-information block. Programmatic checks verify that privileged information is absent from the student-visible prompt.

After corpus-level prompt-conflict deduplication, student-visited-state supervision contributes 1,768 training examples.

B.7 Paired Reach-Sufficiency Supervision

Paired acquisition-to-answer supervision is constructed from earlier student rollouts at evidence-insufficient L3 states. Source states include premature final answers, unavailable requests, unproductive stalls, and error-recovery states.

The source contains 485 ACQUIRE→FINAL pairs comprising 1,317 supervised states. For labeling, Claude receives the student-visible context together with a supervisor-only block containing the gold answer, cohort-specific tool availability, and the current evidence audit. Programmatic checks verify that this privileged information is absent from the student-visible prompt.

Targets are objectively gated. A `TOOL_CALL` is accepted only if it is executable and strictly increases evidence credit. A `FINAL` is accepted only when the state is sufficient and the answer matches the gold answer; otherwise, the target is replaced according to the frozen pair-construction rule.

After prompt-conflict deduplication, this source contributes 1,072 training examples across 400 TRAIN items.

B.8 None-Repair Supervision

A targeted repair set addresses a recurring failure mode involving valid *None* answer options. We identify 128 TRAIN items displaying this failure pattern and generate 624 candidate repair trajectories using a generic repair instruction that does not expose the gold answer or hidden threshold.

A candidate is accepted only when:

1. its final answer is correct;
2. all required candidates have been measured;
3. no redundant request occurs;
4. no unavailable request occurs;
5. the stopping criterion is satisfied; and
6. at most one repair trajectory is retained per candidate item.

Table 7: Construction ledger for the final raw SFT corpus.

Stage	Teacher	Student-state	Paired	None-repair	Total
Initial rows	19,252	2,240	1,317	380	23,189
FINAL duplication	+1,935	0	0	+61	+1,996
Invalid-action exclusion	-1	0	0	0	-1
Pre-deduplication	21,186	2,240	1,317	441	25,184
Prompt-conflict dedupe	$-2,208$	-472	-245	0	$-2,925$
Final `raw.jsonl`	18,978	1,768	1,072	441	22,259

This process yields 61 accepted trajectories. Before terminal weighting they contain 380 supervised turns. Duplicating each of the 61 final-answer turns produces 441 rows for corpus assembly.

B.9 Full Corpus Construction Ledger

Table 7 summarizes the construction of the final supervised fine-tuning corpus.

Prompt-conflict deduplication retains a single training target for prompts associated with conflicting actions while preserving intentional terminal-action weighting.

The final corpus contains 22,259 examples: 18,978 from selected teacher trajectories, 1,768 from student-visited states, 1,072 from paired acquisition-to-answer supervision, and 441 targeted *None*-repair examples.

Examples are partitioned by deterministic item hash into 21,249 fitting rows and 1,010 monitor rows. Monitor evaluation collapses identical prompts, yielding 929 unique monitor prompts; this deduplication does not alter the fitting partition.

B.10 SFT Configuration

Table 8 summarizes the optimization, adaptation, and hardware configuration of the final EVA-Reason SFT run.

Checkpoint selection. The 929-prompt TRAIN monitor is evaluated at step 0, every 200 optimizer steps, and at the final step. Monitor score is

$$M = A_{\text{action}} + A_{\text{toolarg}} + A_{\text{final}}.$$

A checkpoint becomes the new best checkpoint only when M strictly exceeds the previous best; ties retain the earlier checkpoint. The final step (2,656) achieves the maximum score, $M = 2.243$, and is therefore selected for the raw-SFT run.

B.11 Inference Protocol

At inference time, decoding is greedy with at most 512 newly generated tokens per turn. Episodes are limited to 12 turns and terminate early after four consecutive unproductive turns. Unproductive events include parse errors, duplicate requests, and tool errors.

Malformed calls, unknown measurements, unavailable requests, and repeated requests are handled by the frozen execution layer. An episode ending without a valid `FINAL` answer is scored as incorrect.

The same router, evidence interface, turn limit, and error-handling rules are used for EVA-Reason and all Tool-ON comparison models.

B.12 Representative L3 Questions and Tool-Call Trajectories

Figure 4 shows three representative L3 questions from the frozen `test_public` split together with EVA-Reason's actual tool-call trajectories. Panels A and B show correctly answered Anatomy

Table 8: Complete SFT configuration for the final EVA-Reason run.

Backbone	Qwen2.5-VL-3B-Instruct
Frozen revision	`66285546`
Adaptation	LoRA
LoRA rank	32
LoRA α	64
LoRA dropout	0.05
LoRA targets	q/k/v/o projections
Trainable parameters	14.7 M
Vision encoder	Frozen
Multimodal projector	Frozen
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	0.01
Scheduler	Constant
Warmup	None
Micro-batch size	1
Gradient accumulation	16
Effective accumulation batch	16
Epochs	2
Optimizer steps	2,656
Maximum sequence length	4,096
Truncation	Right
Loss masking	Assistant targets only
Precision	bfloat16
Gradient clipping	1.0
Seed	20260914
Hardware	1$\times$ NVIDIA L40S
Wall-clock time	12.9 h

and Reasoning examples, respectively. Panel C shows a representative Presence failure in which the model acquires several quantitative measurements but produces an incorrect final answer. All questions, answer options, tool calls, returned values, and final answers are reproduced from the frozen benchmark manifest and frozen evaluation trajectories. Free-text thought fields are omitted from the figure for brevity.

B.13 Model Selection and Split Integrity

All trajectory generation, judge selection, student-state collection, teacher labeling, paired-supervision construction, and *None*-repair generation use TRAIN items only. No validation or `test_public` item, subject, answer, or privileged supervision block enters model training. Subject overlap between frozen splits is zero.

Within each SFT run, checkpoint selection uses the TRAIN-only monitor described in Appendix B.10. Validation data are used only for final model-variant selection.

The raw-corpus SFT model is selected based on validation performance. After model selection, the checkpoint hash is frozen before any `test_public` evaluation. The selected model is then evaluated once on the frozen `test_public` split.

The final reported EVA-Reason model is trained with supervised fine-tuning; no reinforcement-learning stage is used in the reported model.

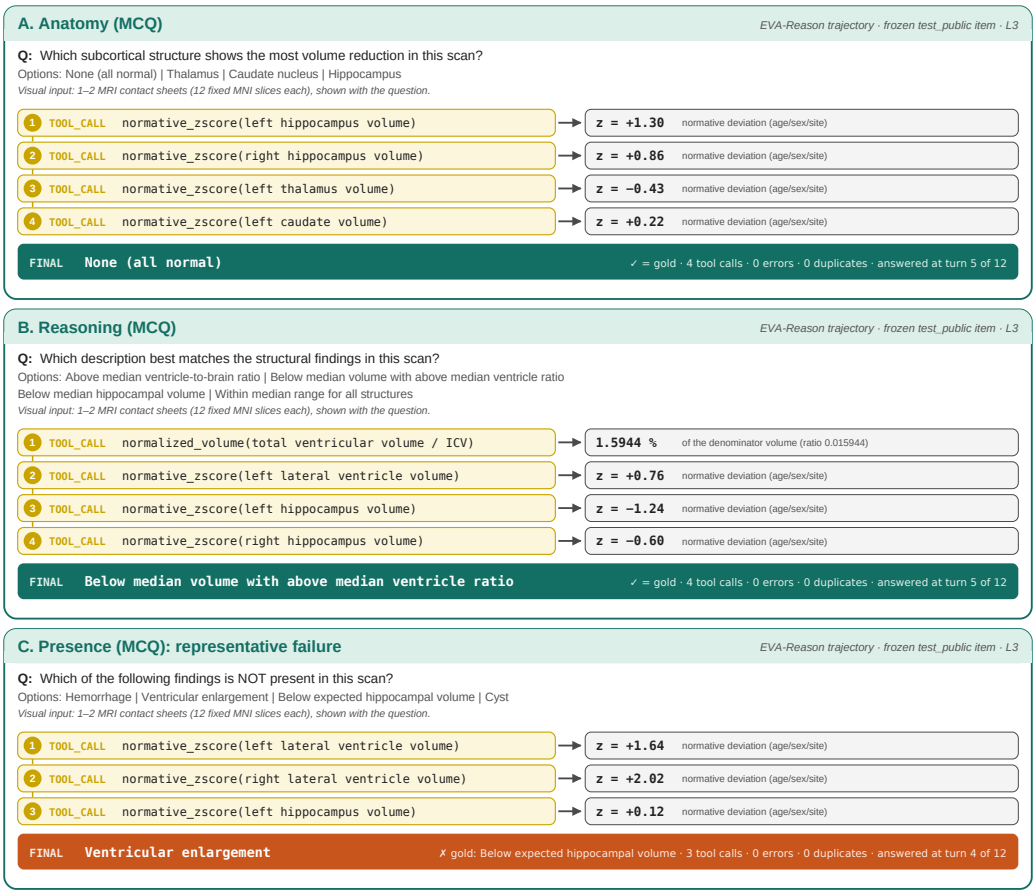


Figure 4: **Representative L3 questions and EVA-Reason tool-call trajectories on frozen `test_public`.** Each panel shows the question and answer options, the exact sequence of TOOL_CALL requests and returned evidence, and the final answer. Free-text thought fields are omitted for brevity. Panels A and B show correctly answered Anatomy and Reasoning trajectories, respectively. Panel C shows a representative Presence failure in which the model acquires several quantitative measurements but produces an incorrect final answer; this example is not interpreted as having sufficient measurement coverage of all answer options.

C Additional Experimental Details

C.1 Evaluated Models and Frozen Versions

Table 9 lists the frozen model identifiers and completed `test_public` evaluation conditions. The same prompt is used across models within each evaluation condition; no model-specific prompt variants are used. API-model decoding parameters follow the frozen model registry used for the evaluation runs.

All models share the same tool interface and 12-turn limit, but inference compute is not matched across providers. Claude Sonnet 4.6 uses extended thinking with a 2,048-token budget and provider-required $T = 1$; Gemini 2.5 Pro/Flash use `thinkingBudget`=2048; Sonnet 5 and Opus 5 gateway runs do not request extended thinking; and GPT-5.5 omits temperature, leaving the provider default. EVA-Reason and MedGemma use greedy decoding with at most 512 new tokens per turn, whereas API models allow up to 4,096.

Table 9: Frozen models and completed `test_public` evaluation conditions.

Model	Frozen identifier	Provider	Decoding	Tool-ON	Tool-OFF
EVA-Reason	Qwen2.5-VL-3B-Instruct@`66285546` + LoRA (`0e29de5d`)	Local	Greedy, 512 tok/turn	✓	–
Claude Sonnet 4.6	`claude-sonnet-4-6`	Vertex AI, `us-east5`	T=1, thinking 2048, 4096 tok	✓	✓
Claude Sonnet 5	`claude-sonnet-5`	institutional API gateway	T=0, 4096 tok	✓	✓
Claude Opus 5	`claude-opus-5`	institutional API gateway	T=0, 4096 tok	✓	✓
Gemini 2.5 Pro	`gemini-2.5-pro`	Vertex AI, `us-central1`	T=0, thinking 2048, 4096 tok	✓	✓
Gemini 2.5 Flash	`gemini-2.5-flash`	Vertex AI, `us-central1`	T=0, thinking 2048, 4096 tok	✓	✓
GPT-4o	`gpt-4o`	institutional API gateway	T=0, 4096 tok	✓	✓
GPT-5.5	`gpt-5.5`	institutional API gateway	Provider-default T, 4096 tok	✓	✓
Llama-4 Maverick	`meta/llama-4-maverick-17b-128e-instruct-maas`	Vertex AI, us-east5	T=0, 4096 tok	✓	✓
MedGemma-4B-IT	`google/medgemma-4b-it` @`290cda5e`	Local	Greedy, 512 tok/turn	✓	✓

C.2 Evaluation Inputs and Protocol

For every benchmark item, all models receive the same frozen MRI PNG input or inputs, question text, answer options when applicable, cohort, and disclosed age. The age representation follows the benchmark's frozen age policy, including the rounded-age caveat when applicable.

Sex and site are not directly included in the benchmark evaluation prompt. In the Tool-ON condition, they may be obtained only by calling `subject_context`. The separate visual-representation probe in Appendix A uses a different direct-generation protocol and should not be conflated with the benchmark evaluation.

In Tool-ON, the model additionally receives the canonical tool schemas, evidence accumulated so far, and the current turn counter. The common system protocol is the one reproduced in Appendix B.4. All Tool-ON models share the same:

- semantic router and quantitative backend;
- cohort-specific availability mask;
- maximum of 12 interaction turns;
- termination after four consecutive unproductive turns;
- handling of malformed, duplicate, unknown, and unavailable calls; and
- rule that an episode ending without a valid `FINAL` answer is incorrect.

Tool-OFF uses a direct-answer protocol requiring a single `FINAL` JSON response.

Model-visible evaluation state is constructed from an explicit allow-list. The gold answer, benchmark category, hidden rule, evidence audit, and sufficiency labels are maintained separately and are never rendered into an evaluation-time model prompt.

C.3 Frozen Evaluation Set

Open-ended items are executed but are not included in the reported closed-form accuracy. Overall normalized exact-match accuracy therefore uses 939 items, and L3 accuracy uses 225 items. Family-level L3 accuracies use their corresponding denominators in Table 10.

Table 10: Composition of the frozen `test_public` split.

Items	1,283
Unique subjects	590
Yes–No	517
Multiple-choice	422
Open-ended	344
Closed-form scored items	939
L1	373
L2	341
L3	225
L3 Anatomy	57
L3 Reasoning	93
L3 Presence	75
L3 with *None* gold option	49
L3 multiple-choice / Yes–No	199 / 26
L3 sufficiency-scorable	200

Table 11: L3 question forms in frozen `test_public` (225 closed-form items). Numeric slots (subject age) are shown as n. Option sets are fixed per form; the *None* / "no significant findings" option is the gold answer on 49 items.

Category	Format	Question form	n
Anatomy	MCQ	Which subcortical structure in this scan shows the most volume deviation from expected range?	47
Anatomy	MCQ	Which subcortical structure shows the most volume reduction in this scan?	9
Anatomy	Y/N	Are the ventricles normal in size relative to brain size in this scan?	1
Presence	MCQ	Which of the following findings is NOT present in this scan?	41
Presence	MCQ	Which of the following findings is present in this scan?	21
Presence	MCQ	Which best describes the structural findings in this scan?	2
Presence	Y/N	Does this subject's brain show any structural deviations from age-related changes (subject age: n)?	11
Reasoning	MCQ	Which description best matches the structural findings in this scan?	70
Reasoning	MCQ	Which structural pattern best describes the findings in this scan?	9
Reasoning	Y/N	Are there any structural findings that deviate from what is expected for a subject of this age (subject age: n)?	8
Reasoning	Y/N	Are there any structural patterns consistent with findings that may suggest mild cognitive impairment (MCI) in this subject?	4
Reasoning	Y/N	Are there any structural abnormalities that suggest atypical development for a n-year-old subject?	2

Reach-sufficiency uses only the 200 L3 items whose question form admits a finite measurable evidence requirement. The remaining 25 L3 items are unscorable for this metric and are excluded from its denominator.

Benchmark-label provenance. The L3 labels are inherited unchanged from the frozen NeuroQA benchmark. Quantitative question forms use cohort-level rules applied to FreeSurfer-derived quantities; PCNtoolkit normative z-scores are not used to generate labels, and hidden benchmark thresholds and rules are not exposed to evaluated models. For the 41 L3 "NOT present" items, gold answers are generated only from measurable rule-derived findings. Non-quantitative options such as hemorrhage, cyst, tumor mass, and cortical malformation may appear in the option set but are never assigned as gold by benchmark construction.

Table 11 lists the twelve L3 question forms and their frequencies; multiple-choice forms draw from fixed option sets.

C.4 Scoring Rules

Normalized exact-match scoring. Predictions and answer options are normalized by lowercasing, removing a leading option-letter prefix such as "B)", and normalizing punctuation and whitespace. A prediction is correct only when its normalized form equals the normalized gold option.

A bare *None* response is mapped to an option beginning with *None* when exactly one such option exists; this mapping is independent of which option is gold. Yes–No and multiple-choice questions otherwise use the same exact normalized matching rule. Additional explanatory text, paraphrases, and bare option letters are incorrect. Missing or malformed `FINAL` outputs are also incorrect.

No predictions are manually corrected. Applying the normalization operations above leaves all reported frozen `test_public` accuracies unchanged.

Option-prefix sensitivity scoring. As a secondary sensitivity analysis, a response is also counted correct when its normalized text begins with the complete normalized text of exactly one answer option and that option is the gold answer. Extra trailing text is therefore tolerated, whereas a bare option letter or paraphrase remains incorrect. The rule is applied identically to every model.

This scoring change particularly affects Claude Opus 5, whose L3 accuracy changes from 43.1% under normalized exact-match scoring to 87.1% under option-prefix scoring, with overall accuracy changing from 75.4% to 95.8%. Claude Sonnet 5 increases by 4.0 percentage points on L3. The corresponding change is at most 0.4 percentage points for Claude Sonnet 4.6, GPT-4o, and MedGemma-4B-IT, while the remaining evaluated models, including EVA-Reason, are unchanged.

C.5 Statistical Tests

All accuracy comparisons are paired because models are evaluated on the same frozen items. At the item level, let b denote the number of items answered correctly only by the first model and c the number answered correctly only by the second. We use the exact two-sided McNemar test,

$$p = \min\{1,\, 2P\,[X \le \min(b, c)]\}\,, \qquad X \sim \mathrm{Binomial}(b + c, \tfrac{1}{2}).$$

Items for which the two models have the same correctness outcome do not contribute to the discordant-pair statistic.

Because multiple questions can originate from the same subject, we also perform cluster-aware inference. For confidence intervals, subjects are sampled with replacement for 4,000 bootstrap resamples using fixed seed 0, with all scored questions belonging to each sampled subject retained together. The 2.5th and 97.5th percentiles of the paired accuracy-difference distribution form the 95% subject-cluster confidence interval. The L3 evaluation contains 225 questions from 162 unique subjects, with a mean of 1.39 and a maximum of 5 L3 questions per subject.

For cluster-aware significance testing, we use a two-sided paired sign-flip test over subject-level summed accuracy differences with 20,000 Monte Carlo sign flips. Subject-cluster and item-level analyses yield the same significance conclusions for the headline comparisons.

In the secondary item-level analysis, family-wise error across the nine EVA-Reason versus Tool-ON comparisons is controlled using the Holm procedure; all nine remain significant after adjustment.

For completeness, paired item-level bootstrap intervals are also computed using 4,000 resamples with fixed seed 0. Confidence intervals are reported for paired accuracy differences rather than individual absolute accuracies.

C.6 Behavioral Metrics

Answer rate. The fraction of evaluated episodes that terminate with a valid `FINAL` response.

Reach-sufficiency. The fraction of the 200 sufficiency-scorable L3 episodes in which the collected evidence reaches $C(s) = 1$ under the frozen sufficiency audit (Appendix B.3). Only successfully executed, non-errored evidence contributes. Because emitting `FINAL` terminates an episode, sufficiency must be reached before termination.

Table 12: Frozen Tool-ON `test_public` results under normalized exact-match scoring. ΔL3 denotes EVA-Reason minus the comparison model on the 225 L3 items. p-values shown in this table are from exact two-sided item-level McNemar tests; subject-level cluster-aware inference is reported in the main text and Appendix C.5.

Model	Overall	L3	ΔL3 vs. EVA	p
EVA-Reason (3B, local)	96.4	89.3	–	–
Claude Sonnet 5	93.6	80.0	+9.3	0.0031
Gemini 2.5 Pro	94.2	79.6	+9.8	0.0026
GPT-5.5	94.1	79.1	+10.2	0.0014
Claude Sonnet 4.6	93.0	75.6	+13.8	< 0.0001
Gemini 2.5 Flash	88.1	61.8	+27.6	< 0.0001
Claude Opus 5	75.4	43.1	+46.2	< 0.0001
Llama-4 Maverick	73.7	32.4	+56.9	< 0.0001
MedGemma-4B-IT	42.7	31.6	+57.8	< 0.0001
GPT-4o	72.9	24.9	+64.4	< 0.0001

Tool calls. The number of `TOOL_CALL` actions emitted by the model, including calls that subsequently return an execution error or duplicate a previous request.

Redundant calls. The number of `TOOL_CALL` actions minus the number of distinct calls in the episode, i.e. exact repeats of an earlier request.

Tool and backend errors. Tool errors are calls whose execution returns an error through the evidence interface. Backend errors refer to transport- or provider-level execution failures tracked separately by the evaluation system.

Unproductive stops. Episodes terminated by the frozen four-consecutive-unproductive-turn rule rather than by a valid `FINAL` response.

Unless otherwise stated, behavioral summaries in the main results are computed over the 225 closed-form L3 episodes; reach-sufficiency uses the 200-item scorable subset defined above.

C.7 Test-Set Integrity

The final EVA-Reason checkpoint is selected before `test_public` evaluation and its checkpoint hash is frozen. The selected final model is evaluated once on `test_public`; no test result is used for checkpoint or model-variant selection.

All comparison models are evaluated on the same frozen test items and associated MRI images. Incomplete frontier runs are excluded rather than combined with completed test results. Model development, supervision construction, and validation-based variant selection are described in Appendix B.13.

D Additional Results

D.1 Full Tool-ON Comparison

Table 12 reports the complete Tool-ON comparison on the frozen `test_public` split. All models are evaluated on the same 939 closed-form items, including 225 L3 questions. Normalized exact-match scoring is used throughout this table. Paired differences are EVA-Reason minus the comparison model on the 225 L3 items. The confidence intervals reported below are paired item-level bootstrap intervals (4,000 resamples); the corresponding p-values are from exact two-sided item-level McNemar tests.

For the four comparison models with the highest normalized exact-match L3 accuracy, EVA-Reason exceeds Claude Sonnet 5 by 9.3 L3 percentage points (95% CI $[+3.6, +15.1]$, $p = 0.0031$), Gemini 2.5 Pro by 9.8 points ($[+3.6, +15.6]$, $p = 0.0026$), GPT-5.5 by 10.2 points ($[+4.4, +16.0]$, $p = 0.0014$), and Claude Sonnet 4.6 by 13.8 points ($[+7.6, +20.0]$, $p < 0.0001$).

Table 13: Normalized exact-match L3 accuracy (%) by question family.

Model	Anatomy	Reasoning	Presence
EVA-Reason	100.0	81.7	90.7
Claude Sonnet 5	68.4	86.0	81.3
Gemini 2.5 Pro	49.1	91.4	88.0
GPT-5.5	70.2	88.2	74.7
Claude Sonnet 4.6	49.1	83.9	85.3
Gemini 2.5 Flash	24.6	75.3	73.3
Llama-4 Maverick	7.0	51.6	28.0
MedGemma-4B-IT	8.8	29.0	52.0
GPT-4o	10.5	44.1	12.0

Option-prefix scoring sensitivity, including the large formatting effect for Claude Opus 5, is reported in Appendix D.3.

D.2 L3 Performance by Question Family

Table 13 decomposes normalized exact-match L3 accuracy by question family. The 225 L3 items contain 57 Anatomy, 93 Reasoning, and 75 Presence questions.

EVA-Reason's largest remaining family-level error concentration is on Reasoning questions, where Gemini 2.5 Pro (91.4%), GPT-5.5 (88.2%), and Claude Sonnet 5 (86.0%) exceed EVA-Reason (81.7%).

Claude Opus 5 is omitted from Table 13 because its normalized exact-match scores are strongly affected by the output-format behavior described in Appendix D.3; its aggregate strict and uniform-lenient results are reported there.

Sensitivity to the "NOT present" benchmark convention. To test whether the L3 comparisons are driven by the 41 Presence items whose gold answers follow the benchmark's "NOT present" construction, we repeat the analysis after excluding this entire question form. On the remaining 184 L3 items, EVA-Reason achieves 90.2% accuracy, compared with 82.6% for Claude Sonnet 5, 77.7% for Gemini 2.5 Pro, 78.3% for GPT-5.5, and 76.1% for Claude Sonnet 4.6. The corresponding EVA-Reason gaps remain significant under the primary subject-level analysis: +7.6 points (95% CI $[+1.7, +13.7]$, $p = 0.0191$), +12.5 ($[+6.4, +18.6]$, $p = 0.00025$), +12.0 ($[+6.1, +18.4]$, $p = 0.00020$), and +14.1 ($[+7.9, +20.5]$, $p = 0.00010$), respectively. On the excluded form itself, EVA-Reason achieves 85.4% accuracy, compared with 87.8% for Gemini 2.5 Pro and 82.9% for GPT-5.5. Together with the exclusion analysis, this shows that EVA-Reason's overall L3 advantage is not driven by this benchmark convention.

D.3 Normalized Exact-Match versus Option-Prefix Scoring

Normalized exact-match scoring requires the normalized prediction to exactly equal the normalized gold option. Some models, most notably Claude Opus 5, append measured values to the selected option text. We therefore perform a secondary uniform sensitivity analysis in which a prediction is counted correct when it begins with the complete normalized text of exactly one option and that option is the gold answer. The same rule is applied to every model.

Claude Opus 5 changes most strongly under the uniform rule, from 43.1% to 87.1% on L3 and from 75.4% to 95.8% overall. Under this sensitivity analysis, EVA-Reason exceeds Opus 5 by only 2.2 L3 percentage points (95% CI $[-3.1, +7.6]$, $p = 0.52$), and the difference is not statistically significant.

Claude Sonnet 5 rises from 80.0% to 84.0% on L3. At that point its gap to EVA-Reason is 5.3 points (95% CI $[-0.0, +10.7]$, $p = 0.088$) and is also not statistically significant. EVA-Reason, Gemini 2.5 Pro, GPT-5.5, Gemini 2.5 Flash, and Llama-4 Maverick are unchanged, while the remaining models change by at most 0.4 L3 percentage points.

Table 14: Normalized exact-match versus option-prefix accuracy (%). The lenient rule is a secondary sensitivity analysis; normalized exact-match scoring remains the primary evaluation.

Model	Overall exact / prefix	L3 exact / prefix
EVA-Reason	96.4 / 96.4	89.3 / 89.3
Claude Sonnet 5	93.6 / 94.8	80.0 / 84.0
Gemini 2.5 Pro	94.2 / 94.2	79.6 / 79.6
GPT-5.5	94.1 / 94.1	79.1 / 79.1
Claude Sonnet 4.6	93.0 / 93.1	75.6 / 76.0
Gemini 2.5 Flash	88.1 / 88.1	61.8 / 61.8
Claude Opus 5	75.4 / 95.8	43.1 / 87.1
Llama-4 Maverick	73.7 / 73.7	32.4 / 32.4
MedGemma-4B-IT	42.7 / 42.8	31.6 / 32.0
GPT-4o	72.9 / 73.1	24.9 / 25.3

Table 15: Matched Tool-OFF and Tool-ON `test_public` evaluations (939 closed items / 225 L3), under normalized exact-match scoring. Δ denotes Tool-ON minus Tool-OFF. Confidence intervals are paired item-level bootstrap intervals and p-values are from exact two-sided McNemar tests. Answer rates are over the 939 closed items; reach-sufficiency and calls are Tool-ON L3 quantities, with reach evaluated over the 200 scorable L3 items.

Model	OFF ov.	ON ov.	Δ ov.	OFF L3	ON L3	ΔL3 [95% CI]	p	Overall AR ON/OFF	Reach ON	Calls ON
Claude Sonnet 5	54.1	93.6	+39.5	43.6	80.0	+36.4 [+29.3, +43.6]	< 0.0001	100/100	2.0	2.78
GPT-5.5	57.5	94.1	+36.6	53.3	79.1	+25.8 [+16.9, +34.2]	< 0.0001	100/100	58.0	3.80
Gemini 2.5 Pro	56.1	94.2	+38.1	56.0	79.6	+23.6 [+16.4, +31.1]	< 0.0001	99.7/99.7	81.5	5.04
Claude Sonnet 4.6	51.8	93.0	+41.2	55.1	75.6	+20.4 [+13.3, +28.0]	< 0.0001	100/100	31.0	3.61
Gemini 2.5 Flash	57.2	88.1	+30.9	56.4	61.8	+5.3 [−4.4, +15.6]	0.33	98.3/97.8	73.5	4.91
Llama-4 Maverick	46.8	73.7	+26.9	36.9	32.4	−4.4 [−11.1, +1.8]	0.24	96.9/100	27.5	7.25
MedGemma-4B-IT	51.9	42.7	−9.2	40.9	31.6	−9.3 [−17.3, −1.3]	0.031	88.7/99.9	0.0	1.57
Claude Opus 5	51.9	75.4	+23.5	53.8	43.1	−10.7 [−19.1, −1.8]	0.021	100/86.4	35.0	3.66
GPT-4o	49.6	72.9	+23.3	36.4	24.9	−11.6 [−19.1, −4.0]	0.0043	82.6/99.8	31.0	4.56

D.4 Tool-ON versus Tool-OFF

All nine comparison models have complete matched Tool-OFF and Tool-ON evaluations on the same frozen `test_public` items. Table 15 reports normalized exact-match accuracy and paired Tool-ON minus Tool-OFF differences.

Tool-OFF L3 accuracy ranges from 36.4% for GPT-4o to 56.4% for Gemini 2.5 Flash. Four models show statistically significant positive L3 effects from quantitative evidence access: Claude Sonnet 5 ($+36.4$ points), GPT-5.5 ($+25.8$), Gemini 2.5 Pro ($+23.6$), and Claude Sonnet 4.6 ($+20.4$). Their Tool-ON L3 accuracies range from 75.6% to 80.0%.

Gemini 2.5 Flash ($+5.3$, $p = 0.33$) and Llama-4 Maverick (-4.4, $p = 0.24$) show no statistically significant L3 change. MedGemma-4B-IT declines by 9.3 points (95% CI $[-17.3, -1.3]$, $p = 0.031$), and GPT-4o declines by 11.6 points ($[-19.1, -4.0]$, $p = 0.0043$).

Claude Opus 5 requires separate interpretation because of output formatting. Its Tool-OFF answer rate is 86.4%: 128 of 939 direct responses are not parseable `FINAL` objects. Under normalized exact-match scoring its L3 result changes from 53.8% Tool-OFF to 43.1% Tool-ON. Under the option-prefix sensitivity rule, however, its Tool-ON L3 accuracy is 87.1% versus 53.8% Tool-OFF, a gain of 33.3 points ($p < 0.0001$). Its apparent decline under exact-match scoring is therefore attributable to output-format sensitivity; it disappears under option-prefix sensitivity scoring.

D.5 Matched Supervised and Static-Evidence Controls

To distinguish gains from evidence-acquisition training from those obtainable through supervised adaptation alone, we train a matched Direct-SFT control using Qwen2.5-VL-3B-Instruct. The control uses the same frozen backbone revision (`66285546`), LoRA configuration ($r = 32$, $\alpha = 64$, dropout 0.05 on q/k/v/o projections), optimizer and learning rate, two-epoch schedule, 4,096-token

Table 16: Matched supervised and static-evidence controls on frozen `test_public`.

Model	Overall	L1	L2	L3	Anatomy	Reasoning	Presence
Direct-SFT	64.5	43.2	76.2	82.2	86.0	73.1	90.7
Static Relevant-Evidence	94.6	97.6	95.0	88.9	98.2	75.3	98.7
EVA-Reason	96.4	98.7	98.5	89.3	100.0	81.7	90.7
Full-Evidence	97.7	97.3	100.0	94.7	98.2	89.2	98.7

context length, assistant-only loss, random seed, and frozen vision encoder and multimodal projector as EVA-Reason (Appendix B.10).

Training uses the same 9,437 TRAIN items, with one example per item. Each input contains the same MRI contact sheet(s), question, answer options when present, cohort, and disclosed age as the Tool-OFF condition. The target is only the gold final answer, represented as `"action":"FINAL","answer":<gold>`. The control receives no tool schema, quantitative evidence, `TOOL_CALL` supervision, teacher trajectory, or evidence-acquisition supervision.

Checkpoint selection uses an item-disjoint 5% TRAIN monitor and never uses `test_public`. The selected checkpoint is step 1080 of 1,122 (monitor accuracy 0.705). It is then evaluated once on the same frozen `test_public` split under normalized exact-match scoring.

We additionally train a matched static relevant-evidence control using the same Qwen2.5-VL-3B-Instruct backbone, TRAIN split, LoRA configuration, optimization, and training schedule. A frozen deterministic mapping derives relevant measurable concepts from the question and answer options only, without access to the gold answer, hidden benchmark rule, or test labels. Available measurements for those concepts are supplied upfront in a single prompt. The model receives no tool-calling or multi-turn supervision and produces only a final answer. The selected checkpoint is step 880 (monitor accuracy 0.955).

The static relevant-evidence control reaches 94.6% overall and 88.9% L3 accuracy while receiving 1.9 scalar evidence values per item on average (3.4 on L3); for 75/939 items the mapping yields no measurement and the control answers from the image and question alone. EVA-Reason reaches 96.4% overall and 89.3% L3 while selecting its evidence autonomously. The paired L3 difference (EVA-Reason minus the static relevant-evidence control) is +0.4 points (95% subject-cluster CI $[-4.8, +5.7]$, subject-level sign-flip $p = 1.00$). The corresponding overall difference (EVA-Reason minus control) is +1.8 points (95% subject-cluster CI $[+0.2, +3.5]$, $p = 0.036$). On L3, the two models disagree on 33 items: 17 are correct only for EVA-Reason and 16 only for Static Relevant-Evidence. Among the 17 EVA-only successes, 16 include calls outside the prespecified static plan, and in nine cases those calls add a required concept under the frozen audit. Conversely, among the 16 Static-only successes, EVA-Reason fails to reach audit sufficiency in 12.

The Full-Evidence condition supplies all successfully available evidence upfront and reaches 97.7% overall and 94.7% L3 accuracy. It exposes 138.2 scalar values per closed-form item on average and 138.6 per L3 item, compared with 2.5 and 3.5, respectively, for EVA-Reason. Here a scalar value is one numeric or context field printed in the evidence block, rather than one canonical measurement key or tool call. Derived tools can therefore contribute multiple values: for example, a successful laterality request prints left, right, smaller-side, and asymmetry-index values, while normative and normalized variants can re-express measurements already available in raw form. Subject-context fields are also counted. This comparison highlights the compact evidence subset selected by EVA-Reason relative to exhaustive static exposure. The low Direct-SFT L1 accuracy is concentrated in quantitative lookup questions. All 373 L1 items are marked by the frozen benchmark metadata as requiring a FreeSurfer-derived measurement, and 336/373 are Measurement questions. Direct-SFT answers only 129/336 (38.4%) of these correctly, whereas EVA-Reason answers 333/336 (99.1%). In contrast, Direct-SFT reaches 32/37 (86.5%) on the remaining L1 items. This pattern indicates that direct supervision can exploit answer and template regularities but does not recover the quantitative values required by the benchmark's measurement questions.

Replacing the MRI input with a constant blank image produces little change for Direct-SFT: overall accuracy changes from 64.5% to 63.2%, L1 remains 43.2%, L2 changes from 76.2% to 73.6%, and L3 from 82.2% to 80.4%. None of the matched changes is statistically significant. This indicates that

most Direct-SFT performance is sustained by question, option, and template regularities rather than informative image content. This blank-image control is distinct from the image-omission protocol in Appendix D.7, which removes the image entirely rather than replacing it with a blank input. Option-prefix scoring does not change the Direct-SFT L3 result (185/225).

A TRAIN-only per-template majority baseline reaches 74.7% on L3, confirming strong answer-structure priors in this subset. Direct-SFT reaches 82.2%, but 163 of its 185 correct L3 predictions are majority-label answers. On the 57 L3 items whose gold answer differs from the TRAIN-derived template majority, Direct-SFT achieves 38.6% (22/57), whereas EVA-Reason reaches 68.4% (39/57). Thus, EVA-Reason's advantage is concentrated on cases where template-level answer priors are insufficient.

D.6 Evidence-Feedback Interventions

We test whether EVA-Reason's predictions depend on the quantitative values returned by the evidence interface rather than only on its learned acquisition pattern. For each L3 episode, returned values are replaced with values from another subject in the same cohort while preserving the requested measurement identity.

In the frozen-replay condition, the original tool-call sequence is held fixed and only the returned values are replaced. L3 accuracy decreases from 89.3% to 69.8%; 64/225 (28.4%) answers change, including 51/201 (25.4%) of originally correct answers. Overall accuracy decreases from 96.4% to 61.0%.

In the live condition, the model may adapt subsequent tool choices after observing the perturbed evidence. L3 accuracy is 70.2%, overall accuracy is 61.2%, and 62/225 (27.6%) L3 answers change; 51/201 (25.4%) originally correct answers become incorrect. The close agreement between the frozen-replay and live interventions shows that EVA-Reason's predictions depend strongly on the returned quantitative values rather than only on the acquisition sequence.

Open-loop evidence acquisition. We next isolate the role of intermediate evidence feedback using the same frozen EVA-Reason checkpoint. During acquisition, requested measurements are executed normally but their values are withheld; the model receives only an acknowledgment of each request. After acquisition terminates, all successfully acquired values are revealed together and the model produces a final answer with further tool calls disabled.

Withholding intermediate feedback reduces L3 accuracy from 89.3% (201/225) to 76.9% (173/225), a difference of -12.4 points (95% subject-cluster CI $[-16.9, -7.6]$, subject-level sign-flip $p = 5 \times 10^{-5}$; exact McNemar $p = 2.5 \times 10^{-5}$). Final answers change on 49/225 (21.8%) items. Open-loop acquisition exposes 4.80 scalar values per item, compared with 3.5 for standard EVA-Reason, yet the requested measurement sets have mean Jaccard similarity 0.709 and are identical on only 29.3% of items.

Withholding feedback also changes acquisition behavior: 35.1% of episodes terminate through the unproductive-turn rule and 30.3% of attempted calls repeat an already requested measurement identity. Episodes that stop early with a FINAL action reach 72.2% L3 accuracy (104/144), whereas those that continue to the unproductive-turn limit reach 86.1% (68/79), suggesting that without feedback the policy tends to stop prematurely. We therefore treat this as an intervention on sequential feedback rather than a matched one-shot baseline. Among the 68 items with the same successfully acquired measurement set as standard EVA-Reason, the accuracy gap is smaller (92.6% vs. 88.2%; $+4.4$ points, 95% subject-cluster CI $[-1.5, +11.6]$, McNemar $p = 0.375$). Together, these results indicate that intermediate evidence feedback primarily affects subsequent evidence selection and stopping rather than only final integration given a fixed evidence set.

D.7 Visual-Input Sensitivity

To isolate the contribution of the rendered MRI input, we repeat the frozen `test_public` evaluation with the image omitted entirely while keeping all remaining inputs and protocol components unchanged. The question, answer options, subject context, tool schemas, semantic router, quantitative backend, 12-turn interaction protocol, model-specific decoding configuration, and normalized

Table 17: Tool-ON visual-input sensitivity on frozen `test_public`. ΔL3 denotes normal-MRI minus no-image accuracy, so positive values indicate higher accuracy with the MRI present. Confidence intervals are subject-cluster bootstrap intervals; p-values are from subject-level paired sign-flip tests. EVA-Reason's previously completed no-image run is included for reference.

Model	Overall MRI / no-img	L3 MRI / no-img	ΔL3 [95% CI]	p
EVA-Reason	96.4 / 96.2	89.3 / 89.3	0.0	–
Claude Sonnet 5	93.6 / 94.8	80.0 / 84.4	-4.4 $[-9.5, +0.4]$	0.12
Gemini 2.5 Pro	94.2 / 94.8	79.6 / 82.2	-2.7 $[-7.2, +2.2]$	0.34
GPT-5.5	94.1 / 93.9	79.1 / 78.7	$+0.4$ $[-3.8, +4.6]$	1.00
Claude Sonnet 4.6	93.0 / 92.8	75.6 / 74.2	$+1.3$ $[-3.8, +6.3]$	0.74
Gemini 2.5 Flash	88.1 / 88.6	61.8 / 63.1	-1.3 $[-6.2, +3.3]$	0.71
Claude Opus 5	75.4 / 15.2	43.1 / 6.7	$+36.4$ $[+29.6, +43.1]$	< 0.0001
Llama-4 Maverick	73.7 / 72.7	32.4 / 24.9	$+7.6$ $[+0.9, +13.8]$	0.041
MedGemma-4B-IT	42.7 / 42.1	31.6 / 24.9	$+6.7$ $[+0.4, +13.4]$	0.056
GPT-4o	72.9 / 78.0	24.9 / 38.7	-13.8 $[-19.4, -8.3]$	< 0.0001

Table 18: Tool-OFF visual-input sensitivity on frozen `test_public`. ΔL3 denotes normal-MRI minus no-image accuracy. Confidence intervals are subject-cluster bootstrap intervals and p-values are from subject-level paired sign-flip tests.

Model	Overall MRI / no-img	L3 MRI / no-img	ΔL3 [95% CI]	p
Claude Sonnet 5	54.1 / 54.6	43.6 / 49.3	-5.8 $[-11.3, -0.5]$	0.058
Gemini 2.5 Pro	56.1 / 54.0	56.0 / 44.0	$+12.0$ $[+5.9, +18.0]$	0.0008
GPT-5.5	57.5 / 60.6	53.3 / 59.1	-5.8 $[-11.0, -0.9]$	0.040
Claude Sonnet 4.6	51.8 / 55.3	55.1 / 50.7	$+4.4$ $[+0.5, +8.4]$	0.042
Claude Opus 5	51.9 / 47.7	53.8 / 43.6	$+10.2$ $[+2.9, +17.4]$	0.006
Gemini 2.5 Flash	57.2 / 52.5	56.4 / 47.1	$+9.3$ $[+3.2, +15.6]$	0.005
GPT-4o	49.6 / 48.3	36.4 / 34.2	$+2.2$ $[-2.9, +7.0]$	0.51
Llama-4 Maverick	46.8 / 48.7	36.9 / 37.3	-0.4 $[-5.6, +4.7]$	1.00
MedGemma-4B-IT	51.9 / 47.5	40.9 / 28.0	$+12.9$ $[+7.2, +18.2]$	0.0001

exact-match scoring are identical to the corresponding MRI condition. No blank-image replacement or missing-image instruction is introduced.

Table 17 reports the primary Tool-ON sensitivity analysis. For EVA-Reason and the four strongest Tool-ON comparison models, removing the image produces no statistically detectable L3 decrease. EVA-Reason remains at 89.3% L3 accuracy; Claude Sonnet 5 and Gemini 2.5 Pro are numerically higher without the image, while GPT-5.5 and Claude Sonnet 4.6 change by less than two percentage points. Thus, the limited dependence on the rendered MRI is not specific to EVA-Reason: when explicit quantitative measurements can be requested, several models rely primarily on the evidence interface rather than on direct visual input.

Claude Opus 5 is a protocol-compliance exception in the no-image Tool-ON condition. Without image content, it produces a parseable final answer on 177/939 items (18.8%); 761 episodes terminate under the unproductive-turn rule after repeated parse failures, and one reaches the turn cap. Its large accuracy decrease is therefore interpreted as output-format collapse rather than evidence of visual dependence.

The secondary Tool-OFF analysis provides a useful contrast (Table 18). Without access to quantitative tools, removing the MRI substantially reduces L3 accuracy for several models, including Gemini 2.5 Pro, Gemini 2.5 Flash, Claude Opus 5, and MedGemma-4B-IT. Other models show smaller, null, or reversed effects. Notably, Claude Sonnet 5 and GPT-5.5 achieve higher Tool-OFF L3 accuracy without the image (49.3% vs. 43.6% and 59.1% vs. 53.3%, respectively), indicating substantial reliance on question and context priors in direct-answer evaluation rather than consistent visual read-out. Thus, the rendered MRI can contribute predictive signal in direct-answer evaluation, while its contribution is substantially diminished for the strongest models once explicit quantitative evidence is available.

Table 19: Selected Tool-ON behavioral metrics on L3. Answer rate is over the 225 L3 episodes; reach-sufficiency is over the 200 scorable L3 items.

Model	Answer rate (L3, $n = 225$)	Reach-suff.	Calls/ep.
EVA-Reason	100.0	76/200 (38.0)	3.30
Claude Sonnet 5	100.0	4/200 (2.0)	2.78
Gemini 2.5 Pro	98.7	163/200 (81.5)	5.04
GPT-5.5	100.0	116/200 (58.0)	3.80
Claude Sonnet 4.6	100.0	62/200 (31.0)	3.61
Gemini 2.5 Flash	94.7	147/200 (73.5)	4.91
Claude Opus 5	100.0	70/200 (35.0)	3.66
Llama-4 Maverick	88.0	55/200 (27.5)	7.25
MedGemma-4B-IT	96.0	0/200 (0.0)	1.57
GPT-4o	48.0	62/200 (31.0)	4.56

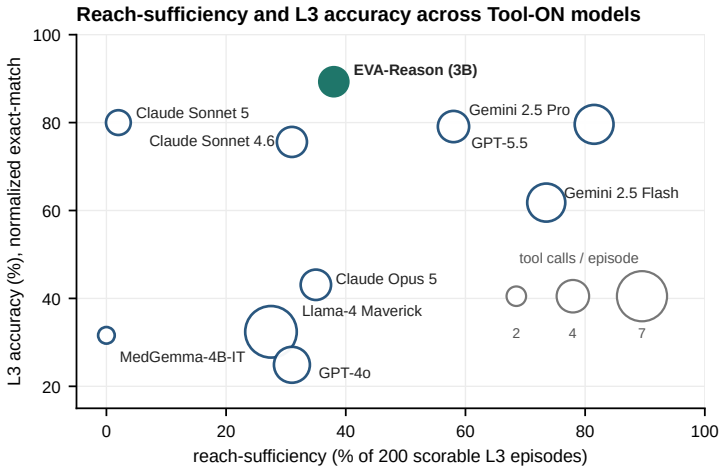


Figure 5: **Reach-sufficiency and L3 accuracy across Tool-ON models.** Each point represents one Tool-ON model on frozen `test_public`. The horizontal axis shows reach-sufficiency over the 200 scorable L3 episodes, the vertical axis shows normalized exact-match accuracy over all 225 L3 items, and marker area is proportional to mean tool calls per episode. Reach-sufficiency is interpreted as a diagnostic of explicit measurement coverage rather than a proxy for answer correctness.

D.8 Tool-Use Behavior

Table 19 summarizes acquisition behavior over the 225 closed-form L3 episodes. Reach-sufficiency is evaluated on the 200 L3 items for which the frozen sufficiency audit defines a finite evidence requirement.

EVA-Reason uses 3.30 tool calls per L3 episode and only 0.01 redundant calls per episode. Across the 225 L3 episodes it produces 10 tool errors, zero backend errors, and zero unproductive stops. GPT-5.5 uses 3.80 calls per episode with zero redundant calls, 8 tool errors, zero backend errors, and zero unproductive stops.

Figure 5 visualizes the relationship between reach-sufficiency and L3 accuracy across Tool-ON models.

D.9 Inference Efficiency

We additionally summarize inference efficiency on the 225 L3 episodes for EVA-Reason and the strongest frontier comparators. Model calls denote inference requests made during an episode, while tool calls denote quantitative-evidence requests. Observed model latency is the sum of per-call model latencies within an episode.

Table 20: Inference efficiency on the 225 L3 episodes. EVA-Reason runs locally on a single NVIDIA L40S with serialized greedy decoding. API output-token counts are provider-reported; EVA-Reason output-token counts are re-tokenized offline from the frozen generated outputs using the same processor and chat template. Observed model latencies reflect provider/server conditions at the time of evaluation and are not compute-matched across systems. Gemini output excludes an additional 4,818 reasoning tokens per item on average.

Model	Model calls/item	Tool calls/item	Output tok./item	Observed model latency (s/item)
EVA-Reason (3B, L40S)	4.3	3.3	338	12.1
Claude Sonnet 5	3.8	2.8	877	16.5
Gemini 2.5 Pro	6.4	5.0	1,181	66.8
GPT-5.5	4.8	3.8	1,198	25.3
Claude Opus 5	4.7	3.7	1,106	25.4

D.10 Held-Out Cohort Generalization

We evaluate transfer to a cohort that is completely excluded from training. All ABCD subjects are removed from the training supervision, and the resulting model is evaluated only on held-out ABCD validation and `test_public` items.

The protocol-selected checkpoint is step 1280. On the 271 closed-form held-out ABCD test items, it achieves 97.8% overall accuracy; on the 62 L3 items, it achieves 90.3%. The standard EVA-Reason model trained with ABCD included achieves 98.5% overall and 93.5% L3 accuracy on the same test subset. On the 264 held-out ABCD validation items, the cohort-holdout model achieves 97.7% overall and 90.7% L3 accuracy, compared with 98.5% and 96.3% for the standard model.

The cohort-holdout model produces a valid final answer for every held-out ABCD L3 item and averages 3.6 tool calls per L3 episode. These results show that excluding the entire ABCD cohort from training produces only a modest reduction in held-out test performance.

D.11 Failure Analysis of EVA-Reason on L3

EVA-Reason makes 24 errors among the 225 L3 items: 17 on Reasoning, 7 on Presence, and none on Anatomy. Twenty-one errors are multiple-choice and three are Yes–No. All 24 episodes terminate with a valid `FINAL` answer; there are no malformed or missing final answers and no unproductive-stop failures. None of the errors occurs on an item whose gold answer is a *None* option.

Under the frozen sufficiency audit, evidence at the final answer is classified as insufficient in 17 of the 24 errors, sufficient in 4, and unscorable in 3. Across all sufficiency-scorable L3 episodes, 72/76 answers produced after reaching audit sufficiency are correct (94.7%), compared with 107/124 answers produced while the audit still considers the evidence insufficient (86.3%). Thus, reaching explicit audit sufficiency is associated with higher accuracy, but most answers produced at insufficient states are nevertheless correct. A separate Full-Evidence diagnostic classifies 13 errors as acquisition/stopping failures, six as integration/final-answer failures, and five as failures even with Full-Evidence. These categories are distinct from audit sufficiency: the former asks whether omitted available evidence resolves the item, whereas the latter measures coverage of the frozen measurable-concept audit.

A post-hoc inspection of the trajectories shows a concentrated error pattern. Fourteen of the 17 Reasoning errors exchange the "within median range for all structures" option with an alternative containing a deviation in one or two structures. Six of the seven Presence errors exchange "below expected hippocampal volume" with "ventricular enlargement." Because no failure taxonomy was pre-registered, these patterns are descriptive rather than confirmatory.

D.12 Correct Answers Before Full Audit Sufficiency

Because the sufficiency audit measures complete coverage of predefined measurable concepts rather than the minimum evidence needed to distinguish among answer options, a correct answer can occur before $C(s) = 1$. The following test-set examples illustrate this distinction.

Evidence-form mismatch. In one EVA-Reason L3 Reasoning item, the model acquired bilateral hippocampal normative z-scores together with a ventricle-to-whole-brain ratio. The ventricular concept nevertheless received zero audit credit because the item's evidence role required normative z-scores. Thus, relevant ventricular evidence was available to the model in a form not credited by the audit.

Partial coverage, decisive evidence. For an HCP-Development L3 Reasoning item asking "Which description best matches the structural findings in this scan?", the answer options were: (i) below median volume with above median ventricle ratio, (ii) above median ventricle-to-brain ratio, (iii) below median hippocampal volume, and (iv) within median range for all structures. EVA-Reason sequentially acquired normative z-scores for total gray matter (-1.93), the left lateral ventricle ($+1.42$), and the left and right hippocampus (-1.10 and -1.02), then correctly selected option (i). The final audit score was $C(s) = 0.75$ because only one ventricular side had been acquired. The unmeasured right lateral ventricle was available and had $z = +1.15$, consistent with the evidence already obtained. The gold answer was also not the training-template majority.

These examples illustrate that failure to reach full audit sufficiency can reflect conservative coverage requirements rather than absence of usable evidence; they do not imply that prior-based answering cannot occur.

D.13 Three-Shot Tool-Use Sensitivity

We provide three complete TRAIN tool-use trajectories, one per L3 family, shared across models and frozen before evaluation, with zero subject overlap with `test_public`. L3 accuracy changes from 80.0 to 86.7% for Sonnet 5 ($+6.7$, $p = 0.013$), 79.1 to 86.2% for GPT-5.5 ($+7.1$, $p = 0.018$), and 79.6 to 82.2% for Gemini 2.5 Pro ($+2.7$, $p = 0.38$). Reach-sufficiency rises from 2.0 to 74.5% for Sonnet 5 and from 58.0 to 86.0% for GPT-5.5, while Gemini starts at 81.5%. Relative to EVA-Reason (89.3%), the remaining gaps for Sonnet 5 and GPT-5.5 are not significant ($p = 0.42, 0.30$); the Gemini gap remains significant ($p = 0.026$).