# PRISM-BENCH: DIAGNOSING GROUNDED AND FAITHFUL MULTIMODAL REASONING BEYOND FINAL-ANSWER ACCURACY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Final-answer accuracy reports whether a vision-language model was right, but not whether its answer was grounded in the image, whether its stated reasoning reflects that evidence, or what the answer cost to produce. We present PRISM-BENCH, a benchmark of 1,000 expert-curated high quality questions spanning five under-represented reasoning categories, in which every item carries an idealized reasoning trace. Because the traces supply a per-item ground truth for the reasoning *process*, they enable measurements that an accuracy-only benchmark cannot support. Our central contribution is CORE, a cost-aware metric that charges each response against the length a human actually needed for *that* item rather than against a global token budget. CORE reorders 24.2% of all model pairs relative to accuracy while remaining uncorrelated with response length ($\rho = -0.05$), so it measures a construct neither accuracy nor mean length recovers. The traces further support two diagnostics: a perception probe that attributes roughly 85% of frontier-model failures to misperception, placing the bottleneck at grounding rather than inference; and trace alignment, which predicts answer correctness for all 25 systems tested (macro AUROC 0.66) and survives length control, so it captures reasoning content rather than verbosity. Across 25 open and closed systems including GPT-5 and Claude Opus 4.6, the best accuracy is 66.0% against an 83.9% human baseline, recent frontier gains fall on recognition rather than spatial grounding, and thinking variants generate 2.2–2.5$\times$ as many tokens as their instruct counterparts for lower accuracy. The optional CORE-PRM variant is critic-agnostic by design, and we validate it with two answer-blind frontier judges. The benchmark, traces, tags, and evaluation code will be released upon acceptance.

## 1 INTRODUCTION

Vision-language models are moving from describing a scene to acting within it, in robotics, driver assistance, and clinical decision support. That shift changes what an evaluation has to report. Correctness remains necessary, but a deployed system must also be **grounded**, so that its answer follows from what the image contains and not from a linguistic prior; **faithful**, so that its stated reasoning reflects the evidence it actually used; and **affordable**, now that reasoning-heavy inference dominates serving cost.

Standard multimodal benchmarks report none of these properties, because they score only the final answer. A model that arrives at the right option through a hallucinated causal story is indistinguishable from one that reasons soundly, and a model that spends ten thousand tokens looks identical to one that spends a hundred. In deployment these are different systems: the first fails silently once the input distribution shifts, and the second fails the latency budget.

**Why a new resource is needed.** These properties cannot be recovered by re-scoring an existing benchmark, because the necessary supervision does not exist in one. Measuring whether reasoning is *faithful* requires a per-item reference for the reasoning process, not just the answer. Measuring whether reasoning is *affordable* requires knowing how much reasoning the item actually demands, which requires the same reference. And measuring *grounding* requires items whose answers cannot be reached from a caption or a linguistic prior, which requires questions written for that purpose.



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.

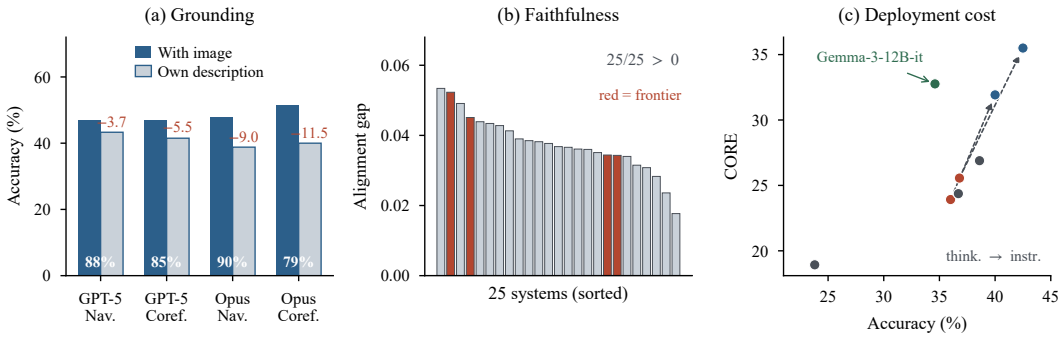


**Figure 1:** Three deployment-relevant measurements that final-answer accuracy cannot provide. **(a)** Accuracy *falls* when a model must answer from its own image description instead of the image (drop annotated in red); the white figure inside each bar is the share of that model's baseline failures a separate judge attributes to perception rather than inference. **(b)** For all 25 evaluated systems, correct responses align more closely with the human reasoning trace than incorrect ones (macro AUROC 0.66, and 0.665 after regressing out response length). **(c)** CORE reorders models relative to accuracy; dashed arrows connect matched thinking/instruct pairs, which spend more tokens for less accuracy.

PRISM-BENCH is built to supply all three: 1,000 items, each with a human-authored idealized reasoning trace, multi-label reasoning tags, and a question designed so that surface cues do not determine the answer.

Our contributions are:

- **A cost-aware metric, CORE**, that normalizes each response against the human reference length for *that item* rather than a global budget. It reorders 24.2% of model pairs relative to accuracy while remaining uncorrelated with raw length, and we show it is not recoverable from accuracy and mean length together.
- **The benchmark.** 1,000 questions across five categories, each with a ground-truth answer, a human-authored reasoning trace, and multi-label reasoning tags, built by a four-stage pipeline in which no annotator validates an item they authored.
- **A grounding diagnostic.** A perception probe attributes roughly 85% of frontier failures to misperception; accuracy falls by up to 11.5 points when a model must work from its own description of the image.
- **A faithfulness diagnostic.** Trace alignment separates correct from incorrect responses for all 25 systems tested, at macro AUROC 0.66 and 0.665 after length control, giving a critic-free signal that uses the traces for content rather than length.
- **A broad evaluation.** 25 open and closed systems including GPT-5 and Claude Opus 4.6, with an expanded human baseline, showing that the frontier deficit is concentrated in spatial grounding.

## 2 Related Work

**Multimodal reasoning benchmarks.** A rich ecosystem probes visual mathematics (Lu et al., 2024; Zhang et al., 2024a; Qiao et al., 2024), spatial and relational understanding (Jia et al., 2025; Daxberger et al., 2025), abstraction (Cai et al., 2025), and domain-specific settings (Luo et al., 2025; Bi et al., 2025; Hao et al., 2025). Broader suites widen task and modality coverage and reduce shortcut exploitation, including MMMU-Pro (Yue et al., 2024), MegaBench (Chen et al., 2025), and the tri-modal OmniBench (Li et al., 2024). Individual capabilities in our categories are also covered by dedicated benchmarks: occlusion perception (Liu et al., 2025), reliability of medical multimodal models (Sepehri et al., 2025), visual text localization and reasoning (Fu et al., 2025), embodied navigation (Qiao et al., 2025), and creative constraint reasoning (Seely et al., 2025). We do not claim any single category is unexplored, and several of these suites probe their target capability in more depth than we do. Step-level references exist too, notably VisualProcessBench in the VisualPRM suite (Wang et al., 2025b); ours differ in being *idealized minimal* human solutions

rather than correctness labels on model-generated steps, which is what lets them serve as a per-item efficiency floor.

**Cost-aware evaluation.** Recent work has recognized that accuracy alone misprices reasoning models. OckBench (Du et al., 2025) scores accuracy jointly with token usage, and OptimalThinking-Bench (Aggarwal et al., 2026) evaluates over- and under-thinking by computing accuracy under a range of thinking-token budgets. Both are text-only, and both compare a response against a *global* budget: a fixed token constant in the first case, a swept threshold in the second. CORE differs in the reference it charges against. Each response is normalized by the length a human annotator actually needed on *that item*, so verbosity is penalized in proportion to how much reasoning the item genuinely demands (Section 4). This per-item normalization is only possible because the benchmark ships human traces, and Section 5.4 shows it is what makes CORE irreducible to accuracy and mean length. To our knowledge CORE is the first cost-aware reasoning metric to normalize against a per-item human reference.

**Process supervision and automatic critics.** Chain-of-thought prompting (Zhang et al., 2024c) and RL-style reasoning training (DeepSeek-AI et al., 2025; Muennighoff et al., 2025) have been extended to multimodal models via GRPO-style and rule-based optimization (Tan et al., 2025; Wan et al., 2025; Meng et al., 2025), self-reflection (Wang et al., 2025a), and structured captioning (Xia et al., 2025). Process reward models score intermediate steps, including VisualPRM (Wang et al., 2025b), DreamPRM (Cao & Xie, 2025), and the vision-language PRM suite ViLBench (Tu et al., 2025), though current PRMs label unstably and generalize poorly out of distribution (Sun et al., 2025). Our trace-alignment measure (Section 5.3) is deliberately *critic-free*: it compares a response against a stored human reference rather than against a learned scorer, which sidesteps the reliability question entirely. Test-time compute scaling (Snell et al., 2024) makes cost-aware evaluation increasingly necessary, and verbose reasoning has been linked to hallucination (Zhang et al., 2024b).

## 3 PRISM-BENCH

PRISM-BENCH contains 1,000 manually curated questions, 200 in each of five categories (Figure 2), chosen to probe what we call the *Perception–Cognition Disconnect*: cases where a model perceives visual elements yet fails to build a coherent world model. We deliberately exclude saturated categories such as chart and math QA, where memorized templates are exploitable.

1. **Visual Coreference / Occlusion** (200): resolving references to partially occluded targets among distractors, requiring object identity and spatial logic under incomplete evidence.
2. **Handwriting / OCR Reasoning** (200): recognizing handwritten text and reasoning over it, where one misread character cascades downstream.
3. **Navigation / Map-based Reasoning** (200): interpreting schematic maps and composing multi-step spatial instructions.
4. **Multi-hop Reasoning** (200): chaining visual and textual clues across image regions.
5. **Scientific / Medical Reasoning** (200): interpreting medical visual cues with domain causal knowledge, where models often perceive correctly but hallucinate causal explanations.

Each item carries optional answer choices, a ground-truth answer, a human-authored idealized reasoning trace, and multi-label reasoning tags (visual, spatial, textual, commonsense). Categories are balanced at 200 items each by construction; two items additionally carry a second category label, since the tagging scheme permits an item to exercise more than one reasoning type. Table 1 summarizes the resulting resource.

**Construction.** Images come from RefCOCO (Kazemzadeh et al., 2014) (coreference, multi-hop), schematic layouts derived from OpenStreetMap (navigation), handwriting scans and IAM (OCR), and MedXpertQA (Zuo et al., 2025) (medical). A four-stage pipeline enforces separation of duties so that no annotator both authors and validates the same item (Figure 5): (1) experts design questions that decouple surface cues from the reasoning target; (2) annotators author idealized minimal traces, mandatorily audited and edited by experts, so that no trace enters the benchmark as raw model output; (3) annotators assign multi-label tags; (4) a *separate* annotator validates answer correctness, trace

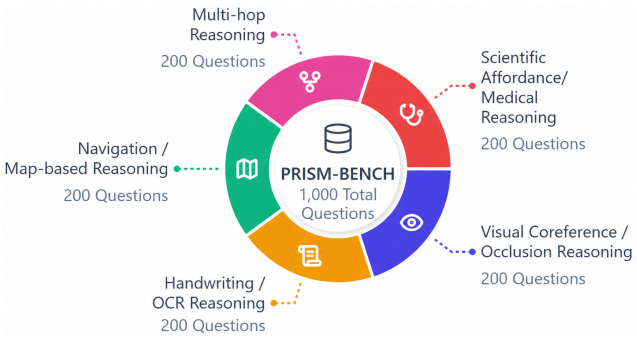


**Figure 2:** PRISM-BENCH distribution across the five complementary reasoning categories, balanced at 200 items each. The categories mix functional reasoning patterns, such as multi-hop composition, with domain-specific grounding requirements, such as medical interpretation, because deployment demands both. The equal weighting is a design decision, so our aggregate accuracy is an unweighted mean over a portfolio we selected.

**Table 1:** PRISM-BENCH dataset statistics. The three answer-choice rows partition all 1,000 items; most are 4-way or wider multiple choice, and 11 items are open-ended or binary.

| Statistic | Value |
|---|---|
| Total questions | 1,000 |
| Categories | 5 (200 each) |
| Average question length | 24.7 words |
| Average reasoning trace length | 76 words |
| Questions with 0–2 answer choices | 11 |
| Questions with 3–4 answer choices | 245 |
| Questions with 4+ answer choices | 744 |
| Annotation time per question | 20–30 min |
| Image sources | RefCOCO, OpenStreetMap, MedXpertQA, handwriting volunteers, IAM |

validity, and clarity, with failing items returned to their author for revision and re-validation, or removed. What is released is therefore the fixed point of the loop, not its first pass.

**Answer reliability.** Two measurements support the labels. First, five independent annotators answered 50 items per category and converged on the gold answer at 91.2% (OCR), 86.0% (navigation), 84.0% (coreference) and 74.4% (multi-hop), averaging 83.9%; high convergence among independent annotators is direct evidence that the answers are well defined, and the residual is genuine difficulty rather than label noise, since models score far lower on the same items (Section D). Second, we applied a strict single-defensible-answer criterion to the 401 items in the two categories where referent and orientation ambiguity is structurally most likely, navigation and coreference/occlusion: 393 (98%) passed unchanged, and the 8 exceptions are listed with identifiers and shipped with per-item audit labels (Section G).

**Contamination.** Question, option, answer, and trace text is newly authored for 807 of 1,000 items (80.7%); only source images are reused, and internal 5-gram Jaccard overlap across questions is low (mean and median maximum pairwise similarity 0.21 and 0.11), so items are not near-duplicates. Because the source images are public, we also tested exposure behaviorally rather than relying on provenance alone: if image familiarity leaked answers, models would perform best on the most reused sources. The opposite holds. GPT-5 and Claude Opus 4.6 score 60.6 and 61.3 on COCO/RefCOCO images and 46.8 and 47.8 on OpenStreetMap schematics, making the two most public sources the two hardest, while handwriting scans, least likely to have been memorized, are easiest at 82.1 and 83.6 (Section F). Difficulty tracks the reasoning demanded, not the familiarity of the pixels, which is what one expects when questions ask about occluded content and multi-step spatial composition that no public caption contains.

# 4 CORE: CHARGING REASONING AGAINST A HUMAN REFERENCE

Two models with equal accuracy can differ by an order of magnitude in tokens, and on a served system that difference is the whole cost story. The difficulty in scoring it is choosing what to compare a response against. A global token budget treats every item as equally deserving of reasoning, which it plainly is not: a map question a person answers at a glance and a medical differential requiring four inferential steps should not carry the same allowance. PRISM-BENCH can do better because it ships a human reference for every item, so we charge each response against the length a careful person actually needed on *that* item:

$$CORE(a, \mathbf{l}) = \frac{a}{\frac{1}{N}\sum_{i=1}^{N} \frac{\log l_i}{\log L_i^{\mathrm{gt}}}}, \tag{1}$$

where $a$ is accuracy, $l_i$ the generated length, and $L_i^{\mathrm{gt}}$ the human trace length for item $i$. The denominator is the mean log-length ratio, so it is a unitless inflation factor: 1 when responses match their references, above 1 when they are longer, below 1 when they are shorter. When $l_i = L_i^{\mathrm{gt}}$ for all $i$, CORE reduces exactly to accuracy. We normalize in log space because length distributions are heavy-tailed, so a single runaway response cannot dominate the average, and because the log makes the penalty proportional rather than absolute. Both lengths are floored at two units so the logs stay well defined for degenerate outputs such as a bare option letter. Note also that $l_i$ counts model tokens while $L_i^{\mathrm{gt}}$ counts words of the human trace, so the ratio reflects tokenizer granularity as well as verbosity.

**A worked example.** Consider two items: a navigation question whose human reference is 20 words, and a medical question whose reference is 80 words. Model A spends 210 words on each; Model B spends 20 on the navigation item and 400 on the medical one. Both average 210 words, so any statistic built on mean length treats them identically. CORE does not. Model A's ratios are $\log 210/\log 20 = 1.78$ and $\log 210/\log 80 = 1.22$, averaging $1.50$. Model B's are $\log 20/\log 20 = 1.00$ and $\log 400/\log 80 = 1.37$, averaging $1.18$. At equal accuracy, Model B scores about 27% higher, because it spent its tokens where the item warranted them and was terse where it did not. The distinction is invisible to mean length, and it is exactly the behaviour a cost-aware metric should reward.

**The reference is a floor, not a unique path.** $L_i^{\mathrm{gt}}$ is the length of an *idealized minimal* trace: the shortest logically complete chain to the answer. It is not a claim that the human route is the only valid one, and CORE does not compare routes. The log ratio is deliberately tolerant, so an alternative correct derivation of moderately different length is barely penalized while genuine redundancy is; a response twice the reference length on a 20-word item incurs a ratio of $1.23$, not $2.0$. Section 5.3 provides the corresponding empirical check that the references are representative rather than idiosyncratic.

**A critic-augmented variant.** For completeness we also define a variant that folds in step-level process reward scores $v_i \in [0, 1]$:

$$CORE - PRM(\mathbf{a}, \mathbf{v}, \mathbf{l}) = \frac{1}{2N}\sum_{i=1}^{N} \frac{(a_i + v_i)\log L_i^{\mathrm{gt}}}{\log l_i}. \tag{2}$$

CORE is critic-free and is our primary metric: every result we report rests on accuracy and CORE. CORE-PRM is a critic-agnostic *template*, since its validity is inherited from whichever scorer is supplied, and we verify that this framing is substantive rather than a hedge. On 200 traces balanced across categories and outcomes, we compare three scorers on the weakest question a critic should pass: does it rate valid reasoning above invalid reasoning, *without* being shown the answer? An off-the-shelf 8B process reward model separates the two by only $+0.012$ (AUROC $0.537$) and is inverted in two of five categories, whereas answer-blind frontier judges reach $+0.145$ (Gemini 2.5 Pro, AUROC $0.643$) and $+0.181$ (GPT-5, AUROC $0.663$) with no inverted category (Section E). Critic choice, not the metric's form, is what determines whether CORE-PRM is informative.

**Table 2:** Accuracy (%) of all 25 evaluated systems across the five PRISM-BENCH categories, against the human baseline (5 annotators × 50 items per category). No system exceeds 66% overall, or 84% in any single category. The newest frontier systems lead on OCR yet remain near 47% on navigation, where humans reach 86%: recent gains concentrate on recognition rather than spatial grounding.

| Model | Size | Overall | Coref. | OCR | Nav. | Multi-hop | Sci./Med. |
|---|---|---|---|---|---|---|---|
| *Closed-source* | | | | | | | |
| Gemini 2.5 Pro | – | **66.0** | **57.5** | 77.0 | **58.2** | 69.2 | **68.5** |
| GPT-5 | – | 63.6 | 47.0 | 82.0 | 47.0 | **73.6** | 68.0 |
| Claude Opus 4.6 | – | 62.4 | 51.5 | **83.5** | 47.8 | 69.7 | 59.3 |
| OpenAI o3 | – | 57.2 | 39.5 | 75.5 | 46.8 | 65.2 | 59.5 |
| Claude Sonnet 4 | – | 39.6 | 39.0 | 46.5 | 39.8 | 38.3 | 34.5 |
| *Open-weight (instruct)* | | | | | | | |
| Qwen3-VL-30B-A3B-Inst. | 30B | 42.5 | 41.0 | 52.5 | 35.3 | 48.3 | 35.5 |
| Qwen3-VL-8B-Inst. | 8B | 40.0 | 39.5 | 50.5 | 42.3 | 40.3 | 27.5 |
| Gemma-3-12B-it | 12B | 34.6 | 39.0 | 40.5 | 27.4 | 38.3 | 28.0 |
| InternVL3.5-8B | 8B | 30.2 | 29.5 | 36.5 | 27.4 | 30.9 | 27.0 |
| Gemma-3-4B-it | 4B | 29.0 | 36.0 | 30.5 | 23.9 | 32.8 | 22.0 |
| Qwen2.5-VL-7B-Inst. | 7B | 27.4 | 30.0 | 31.5 | 32.8 | 22.9 | 20.0 |
| InternVL3-8B-Inst. | 8B | 26.7 | 30.5 | 25.0 | 26.9 | 30.9 | 20.5 |
| Qwen2.5-VL-3B-Inst. | 3B | 22.1 | 24.5 | 19.0 | 28.9 | 19.4 | 18.5 |
| InternVL3-2B-Inst. | 2B | 20.7 | 31.0 | 16.0 | 22.4 | 23.9 | 10.0 |
| Molmo-7B-D-0924 | 7B | 20.1 | 25.5 | 21.0 | 20.9 | 19.4 | 14.0 |
| *Open-weight (thinking / RL-trained)* | | | | | | | |
| Skywork-R1V3-38B | 38B | 38.6 | 42.5 | 42.5 | 32.3 | 44.3 | 31.5 |
| Qwen3-VL-30B-A3B-Think. | 30B | 36.8 | 34.0 | 57.5 | 36.8 | 38.8 | 17.0 |
| Skywork-R1V2-38B | 38B | 36.7 | 40.0 | 32.5 | 30.9 | 48.8 | 31.5 |
| Ovis2.5-9B | 9B | 36.1 | 36.5 | 40.5 | 37.8 | 42.8 | 22.5 |
| Qwen3-VL-8B-Think. | 8B | 36.0 | 32.0 | 54.0 | 33.3 | 33.3 | 27.5 |
| VL-Rethinker | 7B | 29.1 | 32.0 | 35.0 | 31.3 | 30.4 | 16.5 |
| Kimi-VL-A3B-Think. | 16B | 29.0 | 34.5 | 33.5 | 32.8 | 26.9 | 17.5 |
| Ovis2.5-2B | 2B | 28.4 | 32.5 | 30.5 | 32.3 | 29.4 | 17.0 |
| Visionary-R1 | 7B | 26.2 | 30.5 | 23.0 | 31.8 | 24.4 | 21.5 |
| GLM-4.1V-9B-Think. | 9B | 23.8 | 29.0 | 18.5 | 24.4 | 27.4 | 19.5 |
| *Human baseline* | – | *83.9* | *84.0* | *91.2* | *86.0* | *74.4* | n/a |

## 5 EXPERIMENTS

Models generate an answer plus a step-by-step rationale, via chain-of-thought prompting or native thinking mode. Open-weight models use greedy decoding, giving zero run-to-run variance, and all models share a fixed prompt template and answer-extraction cascade (Section A).

### 5.1 FRONTIER MODELS REMAIN FAR BELOW HUMANS

Table 2 reports all 25 systems across the five categories. No model exceeds 66% overall, against a human average of 83.9%. The two aggregates cover different category sets, since the human study excludes Scientific/Medical, so the like-for-like comparison is over the four categories both measure: there the best system reaches 65.5% against 83.9%, a gap of 18.4 points. The *shape* of that gap matters more than its size. The newest frontier systems post the strongest OCR scores, above Gemini's, yet their navigation scores sit near 47%, below Gemini's 58.2% and far below the 86% humans reach. Recent progress has concentrated in recognition, while spatial grounding, the capability most relevant to embodied and autonomous settings, has not moved.

Figure 3 makes this explicit. None of the evaluated systems approaches the human baseline, and the remaining deficit is almost entirely spatial. On multi-hop, where reasoning is largely textual once the scene has been read, the best system comes within 0.8 points of humans; on coreference and navigation it trails by 26.5 and 27.8 points. A system can look competitive on an aggregate multimodal leaderboard and still be unusable for the embodied tasks that motivate deployment. Two further patterns hold across the table. Scale is not decisive, with Ovis2.5-9B beating several 30B+ systems on navigation; and instruct variants match or exceed their thinking counterparts at matched

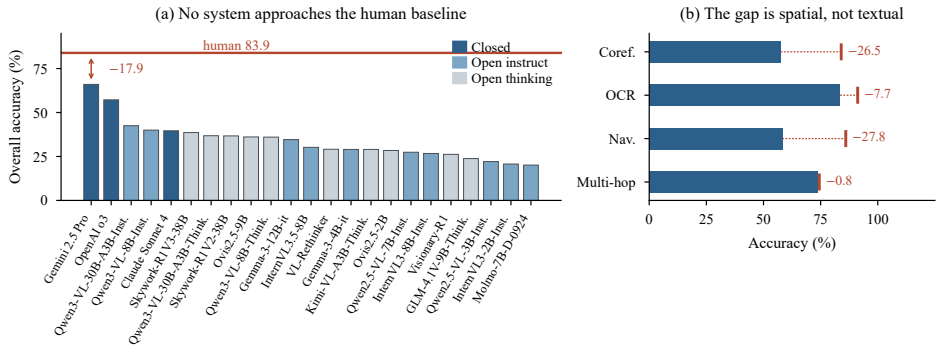


**Figure 3:** **(a)** Overall accuracy for all evaluated systems, shaded by family, against the 83.9% human baseline (red). Restricted to the four categories where both are measured, the best system trails humans by 18.4 points, and the open/closed gap is larger than any difference within the open families. **(b)** Best system of any kind versus the human baseline (red tick) per category. The deficit is concentrated in the two spatially grounded categories and nearly vanishes on multi-hop.

**Table 3:** Perception-vs-reasoning probe on navigation and coreference. Accuracy *falls* when models answer from their own descriptions, and the large majority of failures are attributed to perception.

| Model | Category | Baseline acc. | From own description | % failures perceptual |
|---|---|---|---|---|
| GPT-5 | Navigation | 47.0 | 43.3 | **88** |
| GPT-5 | Coreference | 47.0 | 41.5 | 85 |
| Claude Opus 4.6 | Navigation | 47.8 | 38.8 | **90** |
| Claude Opus 4.6 | Coreference | 51.5 | 40.0 | 79 |
| GPT-5 | *both categories* | *47.0* | *42.4* | *86* |
| Claude Opus 4.6 | *both categories* | *49.6* | *39.4* | *85* |

scale (40.0% against 36.0% for Qwen3-VL-8B, 42.5% against 36.8% at 30B), the accuracy-side shadow of the cost result in Section 5.4.

### 5.2 Grounding: Most Failures Are Perceptual

To locate where models break, we run a two-stage probe on the two categories that depend most on grounding. The model first describes the image, without seeing the question. It then answers from that description alone, with the visual input removed. Separately, a judge labels each baseline failure as perceptual or inferential (Section B).

Two independent signals converge (Table 3, plotted in Figure 1a). The judge attributes about 85% of failures to perception. Independently of any judge, accuracy *drops* when a model must work from its own description, by 3.7 points for GPT-5 on navigation and 11.5 for Opus on coreference, which means the descriptions omit information the model needed. Were the failures inferential, re-reading a faithful self-description would cost nothing. We report approximately 85% rather than one exact figure because the two measurements have different weaknesses: the attribution step inherits its judge's bias, while the accuracy comparison depends on no judge at all.

The practical consequence is that adding reasoning tokens to a model that has already mis-seen the scene cannot recover the answer. Effort is better spent on visual grounding than on longer chains, which is consistent with thinking variants underperforming instruct variants at matched scale.

### 5.3 Faithfulness: Trace Alignment Predicts Correctness

Does a model's stated reasoning track the evidence, or is it a narrative attached to a guess? We segment each gold trace and model response into steps, embed them, and for each gold step take the maximum cosine similarity to any model step, yielding a semantic overlap score. This uses the traces

**Table 4:** Trace alignment predicts answer correctness. AUROC is computed per system over the items with a parseable answer; the length-controlled column regresses out response length. Representative systems shown; all 25 exceed chance in both columns (Section C).

| Model | AUROC (align → correct) | AUROC (length-controlled) |
|---|---|---|
| OpenAI o3 | 0.71 | 0.71 |
| GPT-5 | 0.69 | 0.69 |
| Claude Opus 4.6 | 0.68 | 0.67 |
| Qwen3-VL-8B-Instruct | 0.67 | 0.68 |
| Qwen3-VL-8B-Thinking | 0.67 | 0.68 |
| Skywork-R1V3-38B | 0.63 | 0.64 |
| **Macro-average** | **0.66** | **0.665** |

**Table 5:** Accuracy, normalized reasoning length, and CORE for representative open-weight models. Thinking variants spend markedly more tokens for no accuracy return; concise models rise sharply under a cost-aware view.

| Model | Acc. (%) | Norm. length | CORE | CORE-PRM |
|---|---|---|---|---|
| Qwen3-VL-30B-A3B-Instruct | 42.5 | 1.20 | **35.50** | 46.95 |
| Qwen3-VL-30B-A3B-Thinking | 36.8 | 1.44 | 25.56 | 38.52 |
| Qwen3-VL-8B-Instruct | 40.0 | 1.25 | 31.92 | 46.26 |
| Qwen3-VL-8B-Thinking | 36.0 | 1.51 | 23.92 | 36.40 |
| Gemma-3-12B-it | 34.6 | 1.06 | 32.76 | **49.58** |
| Skywork-R1V3-38B | 38.6 | 1.44 | 26.89 | 38.58 |
| Skywork-R1V2-38B | 36.7 | 1.51 | 24.37 | 35.84 |
| GLM-4.1V-9B-Thinking | 23.8 | 1.26 | 18.93 | 20.05 |

for *content*, not length: it asks whether the response covered the reference reasoning, not whether it matched its wording or its length (Section C).

Correct responses align more closely with the human trace than incorrect ones for every one of the 25 evaluated systems (Table 4; Figure 1b shows the full spread). The macro AUROC of 0.66 is essentially unchanged at 0.665 once response length is regressed out, so the effect belongs to reasoning content and not to verbosity.

This result does double duty. It gives a critic-free faithfulness signal requiring no process reward model, and it validates the reference traces themselves: a single annotator's idiosyncratic path could not predict correctness across 25 heterogeneous systems, so the traces encode reasoning structure that is shared rather than personal. The boundary is worth stating: an AUROC of 0.66 is a real but moderate effect, adequate as evidence that the construct exists and insufficient to use alignment alone as a per-item verdict.

### 5.4 Cost: Longer Reasoning Does Not Pay

Table 5 shows the deployment picture accuracy hides. At matched scale, Qwen3-VL-8B-Thinking generates $2.5\times$ as many tokens as its instruct counterpart (1,812 versus 737) for *lower* accuracy, and CORE separates them by 8 points; the same holds at 30B. Gemma-3-12B-it, ranked 8th of 20 by accuracy, ranks 2nd by CORE on the strength of disciplined reasoning (full ranks in Table 12). Section J shows the mechanism from a second direction: reflective phrases track verbosity rather than quality, so the tokens CORE charges for are largely unproductive oscillation rather than useful deliberation.

**CORE is not a repackaging of accuracy and length.** The natural objection to any composite metric is that it restates its inputs, and specifically that accuracy plus mean response length would carry the same information. It does not. Across the 20 open-weight models with exact token counts, the Spearman correlation between CORE and normalized length is $\rho = -0.050$ ($p = 0.834$), so CORE is not an inverse-length score; against accuracy it is $\rho = +0.674$ ($p = 0.001$), high enough that correctness dominates, as any sound metric requires, but well short of the 1.0 redundancy would

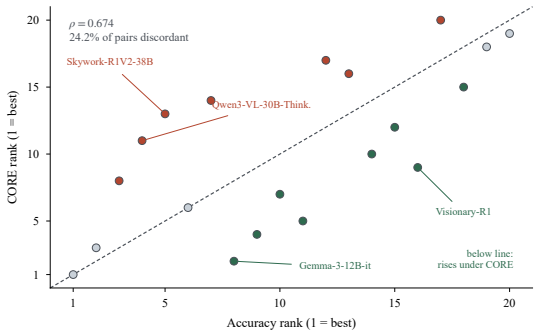


**Figure 4:** Accuracy rank versus CORE rank for all 20 open-weight models. Points on the dashed line are ordered identically by the two metrics; points below rise under CORE and points above fall. If CORE were a repackaging of accuracy the points would lie on the line; if it were an inverse-length score they would be uncorrelated with it. Neither holds.

imply. Concretely, **24.2% of model pairs (46 of 190) are ranked discordantly by accuracy and CORE** (Figure 4). The mechanism is the per-item normalization worked through in Section 4: a marginal mean cannot reconstruct it, both because $\mathrm{mean}(\log l_i) \neq \log(\mathrm{mean}\, l_i)$ and because any aggregate discards which items the tokens were spent on. This is also what distinguishes CORE from cost-aware metrics built on global budgets (Du et al., 2025; Aggarwal et al., 2026), which are computable from marginal statistics by construction.

## 6 Limitations

**The critic-augmented variant needs a strong critic, and the strong critics are closed.** Our critic comparison (Section E) resolves the question of *whether* CORE-PRM can be instantiated reliably, but it leaves two practical constraints. First, the two scorers that work are proprietary frontier systems, so reproducing CORE-PRM at their quality currently costs API access, whereas the open 8B process reward model we tested is close to non-discriminative on this data. Second, even the frontier judges reach only $\sim$0.65 AUROC against correctness, which supports aggregate comparison but not per-item verdicts on reasoning validity. We therefore keep CORE, which is critic-free and fully reproducible, as the primary metric, and no conclusion in this paper depends on $v_i$. Closing this gap needs open multimodal critics competitive with frontier judges, for example scorers trained for general vision-language rather than mathematical reasoning (Tu et al., 2025).

**English-only.** All questions, options, answers, and traces are in English. Whether the category difficulty ordering or the perception-dominated failure profile transfers to other languages and scripts, particularly for the handwriting category, is untested.

**Future training on the benchmark.** PRISM-BENCH is released openly, which makes eventual contamination inevitable: once items enter training corpora, scores will rise for reasons unrelated to reasoning ability. Our behavioral source test (Section 3) establishes that public image reuse confers no advantage *today*, but it cannot protect against future ingestion of the questions and traces themselves. We recommend treating reported numbers as anchored to the evaluation date, and we intend to maintain a held-out extension set so that the gap between public and held-out performance remains measurable.

## 7 Conclusion

PRISM-BENCH evaluates multimodal reasoning along the axes that determine whether a system can be deployed: grounding, faithfulness, and cost. Pairing every question with a human-authored reasoning trace is what makes those axes measurable. It yields CORE, a cost-aware metric that charges each response against the reasoning the item actually demands and reorders a quarter of all model pairs relative to accuracy while remaining uncorrelated with length; a perception probe that

attributes most frontier failures to grounding rather than inference; and a length-controlled alignment measure that predicts correctness for all 25 systems. Across 25 systems the results show that frontier models remain far below human performance, that recent gains have concentrated in recognition while spatial grounding has stalled, and that additional reasoning tokens do not repair a scene the model has already mis-seen.

### Reproducibility Statement

The benchmark, human-authored traces, reasoning tags, per-item audit labels, and full evaluation code will be released upon acceptance. Section A specifies the shared prompt template, the per-model decoding configuration and thinking modes, and the answer-extraction cascade. Sections B to E and G give the protocol for each experiment, including sampling designs and run counts. Open-weight models use greedy decoding, so their results are deterministic and exactly reproducible, and Section H reports per-model accuracy under four prompt and decoding variants. The metric definitions are given in closed form in Equations (1) and (2).

### AI Use Statement

In preparing this manuscript, large language models were used to aid and polish the writing. The research questions, experimental design, analyses, and conclusions are the authors' own, and the authors verified every reported number against the underlying results files. Separately, and as described in Section 3, a language model was used during dataset construction to suggest initial drafts of some reasoning traces; every such draft was mandatorily audited and edited by an expert annotator before acceptance, and no trace enters the benchmark as raw model output.

## References

Pranjal Aggarwal, Seungone Kim, Jack Lanchantin, Sean Welleck, Jason Weston, Ilia Kulikov, and Swarnadeep Saha. OptimalThinkingBench: Evaluating over and underthinking in LLMs. In *International Conference on Learning Representations (ICLR)*, 2026.

Jing Bi, Junjia Guo, Susan Liang, Guangyu Sun, Luchuan Song, Yunlong Tang, Jinxi He, Jiarui Wu, Ali Vosoughi, Chen Chen, and Chenliang Xu. VERIFY: A benchmark of visual explanation and reasoning for investigating multimodal reasoning fidelity, 2025.

Huanqia Cai, Yijun Yang, and Winston Hu. MM-IQ: Benchmarking human-like abstraction and reasoning in multimodal models. *arXiv preprint arXiv:2502.00698*, 2025.

Qi Cao and Pengtao Xie. DreamPRM-1.5: Unlocking the potential of each instance for multimodal process reward model training, 2025. URL https://arxiv.org/abs/2509.05542.

Jiacheng Chen, Tianhao Liang, Sherman Siu, Zhengqing Wang, Kai Wang, Yubo Wang, Yuansheng Ni, Wang Zhu, Ziyan Jiang, Bohan Lyu, Dongfu Jiang, Xuan He, Yuan Liu, Hexiang Hu, Xiang Yue, and Wenhu Chen. MEGA-Bench: Scaling multimodal evaluation to over 500 real-world tasks. In *International Conference on Learning Representations (ICLR)*, 2025.

Erik Daxberger, Nina Wenzel*, David Griffiths*, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, and Peter Grasch. MM-Spatial: Exploring 3D spatial understanding in multimodal LLMs. In *ICCV*, 2025. URL https://arxiv.org/abs/2503.13111.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang,

Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Zheng Du, Hao Kang, Song Han, Tushar Krishna, and Ligeng Zhu. OckBench: Measuring the efficiency of LLM reasoning. *arXiv preprint arXiv:2511.05722*, 2025.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. OCRBench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can MLLMs reason in multimodality? EMMA: An enhanced MultiModal ReAsoning benchmark. *arXiv preprint arXiv:2501.05444*, 2025.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. OmniSpatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*, 2025.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 787–798, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1086. URL https://aclanthology.org/D14-1086.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Xiangzhou Wang, Zhaoxiang Zhang, Zachary Liu, Emmanouil Benetos, Wenhao Huang, and Chenghua Lin. OmniBench: Towards the future of universal omni-language models, 2024. URL https://arxiv.org/abs/2409.15272.

Zhaochen Liu, Kaiwen Gao, Shuyi Liang, Bin Xiao, Limeng Qiao, Lin Ma, and Tingting Jiang. Beyond the visible: Benchmarking occlusion perception in multimodal large language models. *arXiv preprint arXiv:2508.04059*, 2025.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

Junyu Luo, Zhizhuo Kou, Liming Yang, Xiao Luo, Jinsheng Huang, Zhiping Xiao, Jingshu Peng, Chengzhong Liu, Jiaming Ji, Xuanzhe Liu, et al. FinMME: Benchmark dataset for financial multi-modal reasoning evaluation. *arXiv preprint arXiv:2505.24714*, 2025.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. MM-Eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling, 2025. URL https://arxiv.org/abs/2501.19393.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-Math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.

Yanyuan Qiao, Haodong Hong, Wenqi Lyu, Dong An, Siqi Zhang, Yutong Xie, Xinyu Wang, and Qi Wu. NavBench: Probing multimodal large language models for embodied navigation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Jeffrey Seely, Yuki Imajuku, Tianyu Zhao, Edoardo Cetin, and Llion Jones. Sudoku-Bench: Evaluating creative reasoning with sudoku variants. *arXiv preprint arXiv:2505.16135*, 2025.

Mohammad Shahab Sepehri, Zalan Fabian, Maryam Soltanolkotabi, and Mahdi Soltanolkotabi. MediConfusion: Can you trust your AI radiologist? probing the reliability of multimodal medical foundation models. In *International Conference on Learning Representations (ICLR)*, 2025.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Linzhuang Sun, Hao Liang, Jingxuan Wei, Bihui Yu, Tianpeng Li, Fan Yang, Zenan Zhou, and Wentao Zhang. MM-Verify: Enhancing multimodal reasoning with chain-of-thought verification, 2025. URL https://arxiv.org/abs/2502.13383.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-RFT: Reinforcement fine-tuning for visual reasoning of vision language models, 2025. URL https://arxiv.org/abs/2503.20752.

Haoqin Tu, Weitao Feng, Hardy Chen, Hui Liu, Xianfeng Tang, and Cihang Xie. ViLBench: A suite for vision-language process reward modeling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6764–6779, 2025.

Zhongwei Wan, Zhihao Dou, Che Liu, Yu Zhang, Dongfei Cui, Qinjian Zhao, Hui Shen, Jing Xiong, Yi Xin, Yifan Jiang, Chaofan Tao, Yangfan He, Mi Zhang, and Shen Yan. SRPO: Enhancing multimodal LLM reasoning via reflection-aware reinforcement learning, 2025. URL https://arxiv.org/abs/2506.01713.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. VL-Rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning, 2025a. URL https://arxiv.org/abs/2504.08837.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, Lewei Lu, Haodong Duan, Yu Qiao, Jifeng Dai, and Wenhai Wang. VisualPRM: An effective process reward model for multimodal reasoning, 2025b. URL https://arxiv.org/abs/2503.10291.

Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. Visionary-R1: Mitigating shortcuts in visual reasoning with reinforcement learning, 2025. URL https://arxiv.org/abs/2505.14677.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024a.

Yusen Zhang, Sarkar Snigdha Sarathi Das, and Rui Zhang. Verbosity ≠ veracity: Demystify verbosity compensation behavior of large language models, 2024b. URL https://arxiv.org/abs/2411.07858.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models, 2024c. URL https://arxiv.org/abs/2302.00923.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. MedXpertQA: Benchmarking expert-level medical reasoning and understanding, 2025. URL https://arxiv.org/abs/2501.18362.

## A Reproducibility: Prompts, Decoding, and Extraction

**Prompt.** All models receive a fixed instruction prefix asking the model to reason using the image and text provided, and to place the final answer in a `\boxed{\text{...}}` field. Multiple-choice items append the option list and an instruction to explain with reasoning; open-ended items append a request to answer based on the image.

**Decoding.** Open-weight models use deterministic greedy decoding (`do_sample=False`), giving zero run-to-run variance. Maximum generation length ranges from 512 to 64,000 tokens by model, set to each model's practical limit so that no response is truncated mid-reasoning. Native thinking modes are enabled where supported. OpenAI o3 uses `reasoning_effort=low` with `max_completion_tokens=16000`.

**Answer extraction.** A case-insensitive cascade is applied in order: `\boxed{\text{...}}`, then `\boxed{...}`, then "Final answer:", then "the answer is option ...", followed by normalization. The same cascade is applied to every model, and items where extraction fails are recorded rather than silently dropped.

**Verbatim templates.** Because generation and judging are sensitive to exact phrasing, we give the templates in full. The evaluation prompt is the system line, the question, and then a format instruction that differs for multiple-choice and open-ended items:

```
You are a helpful assistant designed to answer questions and
reason, think using the image and text provided. {question}
(multiple choice) Options: {options} Explain with reasoning.
Put your answer like this in the end (latex format):
\(\boxed{\text{answer}}\), where answer is the final answer
from the options. Do not put anything else in latex in the
answer.
(open ended) Please answer the question based on the image
provided. Put your answer like this in the end (latex
format): \(\boxed{\text{answer}}\), where answer is the
final answer.
```

The perception-probe judge (Section B) receives:

```
You analyze why a vision-language model answered a
multimodal question incorrectly. Inputs: the QUESTION, the
CORRECT answer, the GOLD human reasoning, and the MODEL'S
OWN DESCRIPTION of the image (its perception). Decide
the PRIMARY cause of the model's failure: 'perception':
the model's description OMITTED or got WRONG a key visual
fact needed to answer (it failed to see/read the image
correctly). 'reasoning': the description CONTAINED the
needed visual facts, but the reasoning/answer was still
wrong. Reply with exactly one word inside a box.
```

The critic judges (Section E) receive the image, question, options and candidate reasoning, and are instructed to grade each step on two criteria only, consistency with the image and logical entailment from earlier steps, with an explicit instruction not to guess the answer or reward a step for matching what they believe the answer to be, returning one validity flag per step.

## B Perception–Reasoning Probe Protocol

The probe runs in four steps on the two most grounding-dependent categories.

1. **Baseline.** The model answers with the image present, under the standard prompt and extraction cascade.
2. **Describe.** The model is shown the same image and asked for a thorough description, *with no access to the question*. Withholding the question is what makes the stage informative: a description written with the question in hand would selectively include exactly the details needed and omit the rest.
3. **Answer from description.** A fresh context receives the question and the model's own description, with the image removed, and must answer from text alone.
4. **Failure attribution.** For each baseline failure, a separate judge (GPT-5) sees the question, the gold answer, the gold reasoning trace, and the model's own description, and labels the primary cause of failure *perceptual* (the description omitted or misstated a key visual fact needed to answer) or *inferential* (the description contained the needed facts but the model reasoned from them incorrectly).

The two signals are deliberately redundant because each has a weakness the other lacks. The attribution step depends on a judge and inherits its bias; the accuracy comparison between steps 1 and 3 depends on no judge at all, since a self-description that preserved the needed information could not reduce accuracy. Both point the same way, which is why we report $\sim 85\%$ rather than one exact figure. One inferential limit should be stated: because the description in step 2 is written without sight of the question, it may omit fine-grained details that a question-guided pass would have attended to. The accuracy drop therefore establishes that a task-agnostic description is a lossy channel for the information the item needs; it does not by itself prove the model failed to register those details at baseline. The judge-based attribution, which inspects the baseline response directly, is what supplies that second piece of evidence. The probe covers two categories and two frontier systems, chosen because they are the settings where grounding is most clearly the binding constraint; we do not claim the figure transfers to OCR or medical items, where the failure structure differs.

## C Trace Alignment Protocol

Each gold trace and each model response is segmented into steps at sentence and enumeration boundaries. Steps are embedded with `BAAI/bge-small-en-v1.5` using normalized embeddings, and for every gold step we take the maximum cosine similarity to any step of the model response; the item score is the mean of those per-step maxima. Because the score anchors on the *gold* steps, it measures coverage (recall) of the reference reasoning rather than precision: a response that covers every reference step but also adds unrelated or hallucinated material is not penalized for the additions. We therefore read alignment as evidence that the reference reasoning was reproduced, not as a guarantee that nothing else was. The measure therefore asks whether the response *covered* the reference reasoning, not whether it reproduced its wording or its length, so a response reaching the same conclusion by a different valid route scores well provided it touches the same intermediate facts.

Per model we compute AUROC of this score against binary correctness over the $\sim 1{,}000$ items. The length-controlled variant regresses response length out of the alignment score and recomputes AUROC on the residual, which is what distinguishes a content effect from a verbosity artifact. Table 6 reports means and gaps for all 25 systems; the correct-minus-incorrect gap is positive in 25 of 25 cases, ranging from $+0.0177$ to $+0.0534$. Accuracy in that table is implicit in the $n$ correct/incorrect columns and is computed over items with a parseable answer, so it differs marginally from Table 2, which uses all 1,000 items as the denominator. Absolute similarities are uniformly high because any on-topic response about the same image shares vocabulary with the reference, so the gap and not

**Table 6:** Trace alignment for all 25 evaluated systems, sorted by gap. "Sim." is the mean over gold trace steps of the maximum cosine similarity to any step of the model response. The gap is positive for every variant: responses that reach the correct answer track the human reasoning trace more closely than responses that do not.

| Model | $n$ correct | $n$ incorrect | Sim. (correct) | Sim. (incorrect) | Gap |
|---|---|---|---|---|---|
| Qwen2.5-VL-7B-Instruct | 274 | 719 | 0.8159 | 0.7625 | +0.0534 |
| OpenAI o3 | 570 | 408 | 0.8038 | 0.7515 | +0.0523 |
| VL-Rethinker | 291 | 701 | 0.8148 | 0.7657 | +0.0491 |
| GPT-5 | 633 | 361 | 0.8095 | 0.7644 | +0.0451 |
| Kimi-VL | 290 | 538 | 0.8368 | 0.7929 | +0.0439 |
| InternVL3.5-8B | 302 | 692 | 0.8138 | 0.7704 | +0.0434 |
| Visionary-R1 | 261 | 680 | 0.8023 | 0.7595 | +0.0428 |
| Claude Sonnet 4 | 396 | 541 | 0.8200 | 0.7787 | +0.0413 |
| Qwen2.5-VL-3B-Instruct | 213 | 695 | 0.7959 | 0.7569 | +0.0390 |
| Qwen3-VL-8B-Instruct | 400 | 555 | 0.8225 | 0.7840 | +0.0385 |
| Molmo-7B | 201 | 622 | 0.7936 | 0.7554 | +0.0382 |
| InternVL3-8B-Instruct | 267 | 694 | 0.8109 | 0.7732 | +0.0377 |
| Qwen3-VL-30B-A3B-Thinking | 367 | 506 | 0.8386 | 0.8018 | +0.0368 |
| InternVL3-2B-Instruct | 206 | 637 | 0.8028 | 0.7662 | +0.0366 |
| Qwen3-VL-30B-A3B-Instruct | 424 | 544 | 0.8220 | 0.7859 | +0.0361 |
| Gemma-3-12B-it | 346 | 650 | 0.8107 | 0.7747 | +0.0360 |
| Qwen3-VL-8B-Thinking | 360 | 584 | 0.8421 | 0.8070 | +0.0351 |
| Gemini 2.5 Pro | 658 | 336 | 0.8200 | 0.7856 | +0.0344 |
| Claude Opus 4.6 | 622 | 375 | 0.8279 | 0.7936 | +0.0343 |
| Gemma-3-4B-it | 290 | 706 | 0.8020 | 0.7680 | +0.0340 |
| Ovis2.5-9B | 361 | 625 | 0.8201 | 0.7886 | +0.0315 |
| Ovis2.5-2B | 284 | 706 | 0.8046 | 0.7738 | +0.0308 |
| Skywork-R1V2-38B | 366 | 583 | 0.8317 | 0.8034 | +0.0283 |
| Skywork-R1V3-38B | 385 | 603 | 0.8321 | 0.8085 | +0.0236 |
| GLM-4.1V-9B-Thinking | 237 | 734 | 0.7678 | 0.7501 | +0.0177 |

the level carries the signal. Per category, alignment is strongest for OCR and weakest for multi-hop, matching the intuition that compositional reasoning admits more divergent valid routes.

## D Human Study Protocol

Five independent annotators each answered 50 items per category, giving 250 responses per category. Convergence on the gold answer is 91.2% (handwriting/OCR), 86.0% (navigation), 84.0% (coreference/occlusion) and 74.4% (multi-hop), averaging 83.9%. Multi-hop has both the lowest human convergence and the smallest human–model gap, which is consistent: it is the category hardest for everyone, not the category where labels are worst. Scientific/medical is excluded because reliable annotation requires licensed clinicians; its gold answers remain anchored to expert-validated external curation, which is a different and weaker guarantee than five-annotator convergence.

## E Critic Validation

We sampled 200 model reasoning traces balanced across 5 categories, 8 open-weight models, and both outcomes (100 correct, 100 incorrect), and scored each with three candidate critics. The test asks the weakest question a critic should pass: does it rate reasoning that reached the right answer above reasoning that did not?

**The critics must be answer-blind.** A process reward model scores a chain *without* access to the answer key, so any fair comparison must withhold it. For VisualPRM-8B (Wang et al., 2025b) this is automatic: it scores each reasoning step and we take the mean per-step score as the response-level $v_i \in [0, 1]$, exactly as CORE-PRM consumes it. For the frontier judges we enforce it explicitly. Each judge receives the image, the question, and the options together with the candidate reasoning, and is instructed to label every reasoning step valid or invalid on two criteria only, consistency with what the image shows and logical entailment from earlier steps; $v_i$ is the fraction of steps labelled valid. The judge never sees the gold answer or the gold trace, and is told not to reward a step for agreeing with what it believes the answer to be. Showing the judge the answer key would make discrimination trivial and the comparison circular.

**Table 7:** Critic discrimination between correct and incorrect reasoning on the same 200 balanced traces. A usable critic shows a positive gap in every category. The open 8B process reward model is close to non-discriminative and inverts where grounding matters most; both answer-blind frontier judges separate the two classes an order of magnitude more strongly and invert nowhere.

| Critic | mean $v_i$ (corr.) | mean $v_i$ (incorr.) | gap | AUROC | inverted cat. |
|---|---|---|---|---|---|
| VisualPRM-8B | 0.608 | 0.596 | $+0.012$ | 0.537 | 2 / 5 |
| Gemini 2.5 Pro (judge) | 0.552 | 0.407 | $+0.145$ | 0.643 | **0 / 5** |
| GPT-5 (judge) | 0.662 | 0.481 | $\mathbf{+0.181}$ | **0.663** | **0 / 5** |

**Table 8:** Per-category discrimination gap (mean $v_i$ on correct minus incorrect) and AUROC against correctness, for each critic. VisualPRM's two inversions fall in exactly the categories that demand the most compositional and spatial grounding; both frontier judges correct them.

| | VisualPRM-8B | | Gemini 2.5 Pro | | GPT-5 | |
|---|---|---|---|---|---|---|
| **Category** | gap | AUROC | gap | AUROC | gap | AUROC |
| Scientific / Medical | $+0.061$ | 0.667 | $+0.073$ | 0.615 | $+0.156$ | 0.661 |
| Handwriting / OCR | $+0.053$ | 0.552 | $+0.177$ | 0.680 | $+0.307$ | 0.745 |
| Navigation / Map | $+0.038$ | 0.579 | $+0.124$ | 0.647 | $+0.228$ | 0.746 |
| Multi-hop | $-0.037$ | 0.490 | $+0.173$ | 0.671 | $+0.033$ | 0.515 |
| Coreference / Occlusion | $-0.056$ | 0.453 | $+0.178$ | 0.651 | $+0.183$ | 0.654 |
| **Overall** | $+0.012$ | 0.537 | $+0.145$ | 0.643 | $+0.181$ | 0.663 |

**Findings.** VisualPRM-8B separates the two classes by $+0.012$ overall and is *inverted* in multi-hop and coreference/occlusion, scoring invalid reasoning higher than valid reasoning in exactly the categories that demand the most compositional and spatial grounding; the pattern suggests it rewards locally fluent chains rather than valid ones. This is not an artifact of one checkpoint: we selected VisualPRM over DreamPRM (Cao & Xie, 2025) during protocol design because DreamPRM showed substantially higher variance across categories, so the more stable of two open candidates still failed this test. Both frontier judges, by contrast, produce a positive gap in all five categories and roughly an order of magnitude more separation overall. They also agree with each other ($\rho = +0.453$) more than either agrees with VisualPRM ($+0.381$ and $+0.400$), which is what one expects if the frontier judges track reasoning validity and the small PRM is the outlier.

**A structural note on Equation (2).** Because the length term multiplies $(a_i + v_i)$, an item on which a model is both incorrect and judged invalid contributes zero however many tokens it consumed, so CORE-PRM does not charge for verbose failures. CORE does, since its denominator aggregates lengths over all items irrespective of outcome. Anyone adopting CORE-PRM should be aware that the cost signal it carries is weaker than CORE's on exactly the items where waste is most expensive.

**What this establishes, and what it does not.** It establishes that CORE-PRM is critic-agnostic in a substantive sense: the metric's form is sound and its usefulness is bounded by the scorer plugged into it, so a weak result with one critic is not evidence against the design. It does not establish that any available critic is reliable enough for per-item judgements. An AUROC near $0.65$ supports aggregate comparison between systems, not a verdict on an individual trace, and the two critics that achieve it are proprietary. CORE therefore remains our primary metric, critic-free and exactly reproducible, and no conclusion in this paper depends on $v_i$.

## F Contamination Evidence

**Limits of the inference.** This test is indirect. It cannot exclude leakage; it shows that public reuse confers no measurable advantage here, which is a weaker but checkable claim. Difficulty and familiarity are also not fully separable: a benchmark whose public-source items happened to be easy would produce the opposite pattern without any leakage. We therefore report the behavioral test as a

**Table 9:** Accuracy (%) by image source for two frontier models. Counts are per *image source* and sum to the full 1,000 items; they are not in one-to-one correspondence with the 200-item reasoning categories, since a single source can supply more than one category (COCO/RefCOCO images serve both coreference and multi-hop) and a category can draw on more than one source. The most public, most-reused sources are the *hardest*, the opposite of the contamination prediction.

| Source | Image status | $n$ | GPT-5 | Claude Opus 4.6 |
|---|---|---|---|---|
| COCO / RefCOCO | public, reused | 406 | 60.6 | 61.3 |
| OpenStreetMap schematic | public, reused | 201 | 46.8 | 47.8 |
| Handwriting (scans, IAM) | image only | 195 | 82.1 | 83.6 |
| MedXpertQA | expert source | 193 | 67.9 | 57.5 |

companion to provenance, not a replacement, and recommend it as a default because it is nearly free whenever items carry source labels.

## G Label Validity Criterion

An item passes only if a competent person with full context could not defensibly select a different option. We applied this to the 401 items in navigation (201) and coreference/occlusion (200), the two categories where referent and orientation ambiguity is structurally most likely. 393 items (98%) passed unchanged. The 8 failures fall into three patterns, and we give identifiers so the claim is checkable against the release. Four are odd-one-out items whose discriminating criterion was never stated, so more than one option satisfies some reading (UIDs 586, 593, 616, 1061); two are none-of-the-above items where the closest correct option was not listed (605, 635); two use a vague superlative with no stated comparison basis (558, 626). All 8 are flagged for correction, and the release ships a per-item audit label (`pass`, `reworded`, `options-fixed`, `removed`) with the rubric.

The distinction doing the work here is between *ill-posedness*, which we fix, and *difficulty*, which is single-answer and retained by design. Fine-grained counting under occlusion and subtle medical differentials are hard and models do poorly on them, but they admit one defensible answer; relabelling hard-but-well-posed items as ambiguous would delete the signal the benchmark exists to measure. Those 8 items had already passed Stage 4, so this re-audit measures what an independent validator misses under a stricter criterion, not what the pipeline never checked. The audit covers 401 of 1,000 items, chosen as the worst case rather than sampled at random; handwriting answers are objectively transcribed text and medical answers are anchored to external expert curation, both cleaner by construction, but neither has been audited under this criterion.

## H Protocol Stability

To confirm that reported accuracies are properties of the models rather than of our prompt, we evaluated three frontier systems under four defensible prompt and decoding variants on a 150-item stratified subset (30 per category), giving 1,799 valid runs with 0 API or parsing errors. The conditions are **P0**, the default prompt used for every result in the paper; **P1**, a terser instruction that removes the explicit output-formatting scaffold; **P2**, an instruction explicitly requesting step-by-step reasoning; and **P0+**, the default prompt at raised temperature.

**Table 10:** Accuracy (%) under four benign protocol variations; "own range" is a model's max−min across conditions. Each model's score is stable to within 4.8 points of its own mean.

| Model | P0 default | P1 terse | P2 CoT | P0+ temp | Own range |
|---|---|---|---|---|---|
| GPT-5 | 62.0 | **71.3** | 64.7 | 68.0 | 9.3 |
| Gemini 2.5 Pro | **68.7** | 65.7 | **68.3** | **70.5** | 4.8 |
| Claude Opus 4.6 | 58.7 | 60.4 | 63.1 | 59.1 | 4.4 |

Measurement noise is low: no model deviates from its own mean by more than 4.8 points, so the harness is reproducible and each per-model score is meaningful in isolation. Open-weight models are

excluded from this study because they run under greedy decoding with zero run-to-run variance, so the temperature condition is undefined for them. One caveat is worth recording for readers comparing closely matched systems: GPT-5 and Gemini 2.5 Pro sit within roughly two points of each other on this subset and exchange the top position depending on the prompt, while Claude Opus 4.6 is last under every condition. Adjacent positions among systems separated by less than the protocol-induced spread should therefore be read as unresolved, which is one reason we report grounding, faithfulness, and cost alongside accuracy rather than relying on a single scalar.

## I THE ANNOTATION PIPELINE

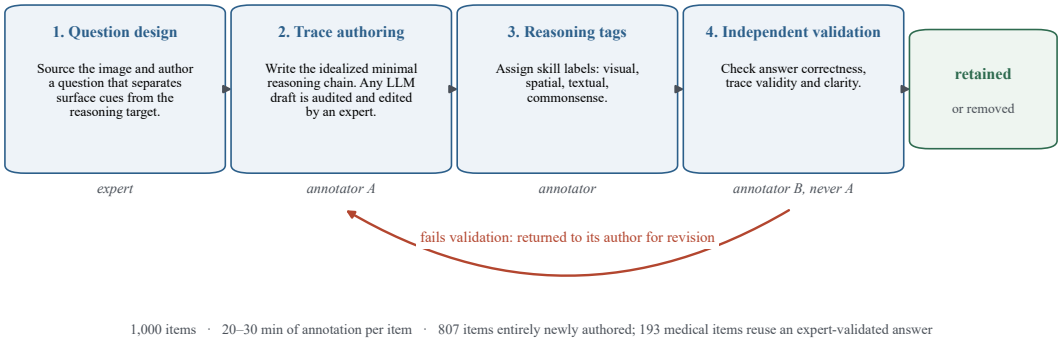


**Figure 5:** The four-stage annotation loop. Duties are separated so that no annotator validates an item whose reasoning trace they wrote, and an item failing Stage 4 returns to its author for revision instead of being silently accepted or discarded. Stage 2 produces the human reference length $L_i^{\text{gt}}$ that CORE uses and the reference content that trace alignment uses.

Each item requires 20–30 minutes of total annotation effort across the four stages. Two points about Stage 2 affect how the reference traces should be read. First, where a language model suggested an initial draft, that draft was mandatorily audited and edited by an expert annotator before acceptance; no trace enters the benchmark as raw model output. Second, the trace is an *idealized minimal* chain, the shortest logically complete path to the answer, which is what makes its length usable as an efficiency floor in Equation (1).

## J REASONING LENGTH AND REFLECTIVE TOKENS

Table 11 counts hesitation and backtracking phrases per response for 12 representative models, matched against a fixed phrase list released with the evaluation code ("Wait", "Actually" and "Let me reconsider" are examples from it rather than the whole list), supporting Section 5.4 from a second direction: reflection tracks verbosity rather than quality. Qwen3-VL-8B-Thinking averages 17.0 reflective phrases and 1,812 tokens to reach 36.0%, while its instruct counterpart averages 2.6 phrases and 737 tokens to reach 40.0%. Gemma-3 models reach 29.0–34.6% on roughly 200 tokens with essentially no reflective phrases. This identifies what CORE is rewarding: if extended reflection were productive deliberation, penalizing it would be a defect in the metric, but the evidence is that it is largely unproductive oscillation.

## K COST-AWARE RESULTS FOR ALL OPEN-WEIGHT MODELS

Table 12 gives the cost-aware view for the 20 open-weight models where generated tokens can be counted exactly, including both ranks. Closed systems are excluded because their thinking tokens are not exposed, so their reasoning cost cannot be measured; this is a limitation of provider APIs, not of the metric.

**Table 11:** Reflective-token analysis for 12 representative models. "Reflective tokens" counts hesitation and backtracking phrases ("Wait", "Actually", "Let me reconsider") per response. More self-correction does not yield higher accuracy: Qwen3-VL-8B-Thinking averages 17.0 reflective phrases and 1,812 tokens to reach 36.0%, while its instruct counterpart averages 2.6 phrases and 737 tokens to reach 40.0%.

| Model | Acc. (%) | Avg. tokens | Reflective tokens |
|---|---|---|---|
| Gemma-3-4B-it | 29.0 | 195.5 | 0.1 |
| Gemma-3-12B-it | 34.6 | 198.4 | 0.1 |
| Qwen3-VL-8B-Inst. | 40.0 | 736.7 | 2.6 |
| Qwen3-VL-8B-Think. | 36.0 | 1811.7 | 17.0 |
| Qwen3-VL-30B-A3B-Inst. | 42.5 | 581.5 | 1.6 |
| Qwen3-VL-30B-A3B-Think. | 36.8 | 1253.2 | 12.0 |
| Ovis2.5-2B | 28.4 | 557.0 | 9.2 |
| Ovis2.5-9B | 36.1 | 658.6 | 11.9 |
| Skywork-R1V2-38B | 36.7 | 1380.9 | 9.8 |
| Skywork-R1V3-38B | 38.6 | 928.7 | 7.9 |
| Kimi-VL-A3B-Think. | 29.0 | 715.0 | 8.2 |
| GLM-4.1V-9B-Think. | 23.8 | 470.7 | 12.1 |

**Table 12:** Cost-aware metrics for the 20 open-weight models with exact token counts. "Norm. len." is mean response length divided by the human reference length. Ranks are computed within this table; 46 of the 190 model pairs (24.2%) are ordered differently by accuracy and CORE. CORE is unavailable for closed systems that hide thinking tokens.

| Model | Size | Acc. (%) | Norm. len. | CORE | VisualPRM $v$ | CORE-PRM | Acc. rank | CORE rank |
|---|---|---|---|---|---|---|---|---|
| Gemma-3-4B-it | 4B | 29.0 | 1.04 | 28.01 | 0.64 | 45.48 | 11 | 5 |
| Gemma-3-12B-it | 12B | 34.6 | 1.06 | 32.76 | 0.69 | 49.58 | 8 | 2 |
| GLM-4.1V-9B-Think. | 9B | 23.8 | 1.26 | 18.93 | 0.26 | 20.05 | 17 | 20 |
| InternVL3-2B-Inst. | 2B | 20.7 | 1.04 | 19.83 | 0.57 | 37.60 | 19 | 18 |
| InternVL3-8B-Inst. | 8B | 26.7 | 1.06 | 25.18 | 0.65 | 43.80 | 15 | 12 |
| InternVL3.5-8B | 8B | 30.2 | 1.05 | 28.76 | 0.66 | 45.90 | 9 | 4 |
| Kimi-VL-A3B-Think. | 16B | 29.0 | 1.34 | 21.61 | 0.70 | 37.76 | 12 | 17 |
| Molmo-7B-D-0924 | 7B | 20.1 | 1.04 | 19.38 | 0.56 | 36.91 | 20 | 19 |
| Ovis2.5-2B | 2B | 28.4 | 1.26 | 22.62 | 0.55 | 34.33 | 13 | 16 |
| Ovis2.5-9B | 9B | 36.1 | 1.31 | 27.61 | 0.59 | 37.27 | 6 | 6 |
| Qwen2.5-VL-3B-Inst. | 3B | 22.1 | 0.97 | 22.85 | 0.55 | 40.68 | 18 | 15 |
| Qwen2.5-VL-7B-Inst. | 7B | 27.4 | 1.04 | 26.32 | 0.62 | 43.11 | 14 | 10 |
| Qwen3-VL-8B-Think. | 8B | 36.0 | 1.51 | 23.92 | 0.70 | 36.40 | 7 | 14 |
| Qwen3-VL-8B-Inst. | 8B | 40.0 | 1.25 | 31.92 | 0.72 | 46.26 | 2 | 3 |
| Qwen3-VL-30B-A3B-Think. | 30B | 36.8 | 1.44 | 25.56 | 0.72 | 38.52 | 4 | 11 |
| Qwen3-VL-30B-A3B-Inst. | 30B | 42.5 | 1.20 | 35.50 | 0.67 | 46.95 | 1 | 1 |
| Skywork-R1V2-38B | 38B | 36.7 | 1.51 | 24.37 | 0.69 | 35.84 | 5 | 13 |
| Skywork-R1V3-38B | 38B | 38.6 | 1.44 | 26.89 | 0.71 | 38.58 | 3 | 8 |
| Visionary-R1 | 7B | 26.2 | 0.99 | 26.52 | 0.56 | 42.15 | 16 | 9 |
| VL-Rethinker | 7B | 29.1 | 1.06 | 27.54 | 0.63 | 44.00 | 10 | 7 |

**Rank correlations.** Over these 20 models: CORE against accuracy, $\rho = +0.674$ ($p = 0.001$); CORE against normalized length, $\rho = -0.050$ ($p = 0.834$); accuracy against normalized length, $\rho = +0.657$ ($p = 0.002$); CORE-PRM against accuracy, $\rho = +0.313$ ($p = 0.179$). The accuracy–length correlation is a confound rather than a mechanism: more capable families happen to emit longer outputs, but within a matched-scale family the thinking variant generates 2.2–2.5× as many tokens for lower accuracy (2.16× at 30B, 2.46× at 8B; Table 11).

## L Category Failure Modes

**Visual coreference / occlusion.** Strong systems enumerate candidate entities, apply the spatial constraint, and eliminate. Weaker systems show a saliency bias, latching onto the most visually prominent entity regardless of the referring expression, and a spatial phrase inversion in which relations such as "the man behind the woman in red" have figure and ground swapped.

**Handwriting / OCR.** A single misrecognized character propagates through otherwise valid downstream arithmetic, producing a confidently wrong answer with an internally consistent chain. This is the clearest case where a correct-looking trace is actively misleading.

**Navigation / map reasoning.** Landmark identification is generally reliable; path composition is not. Failures concentrate in off-by-one directional steps and left/right confusion at intersections. A second pattern is linguistic shortcutting, inferring cardinal orientation from street names rather than from the layout.

**Multi-hop.** The characteristic pattern is first-hop success with second-hop fabrication. Longer chains make this worse, not better, consistent with Section J.

**Scientific / medical.** Models frequently perceive the visual feature correctly and then hallucinate a plausible but incorrect causal attribution. This category has the highest rate of confidently wrong explanations and is the clearest safety concern for high-stakes deployment, since the stated reasoning reads as authoritative.

## M QUALITATIVE EXAMPLES

The figures below show one item per category with the image, the options, the ground-truth rationale, and the verbatim reasoning of four systems with their generated token counts.

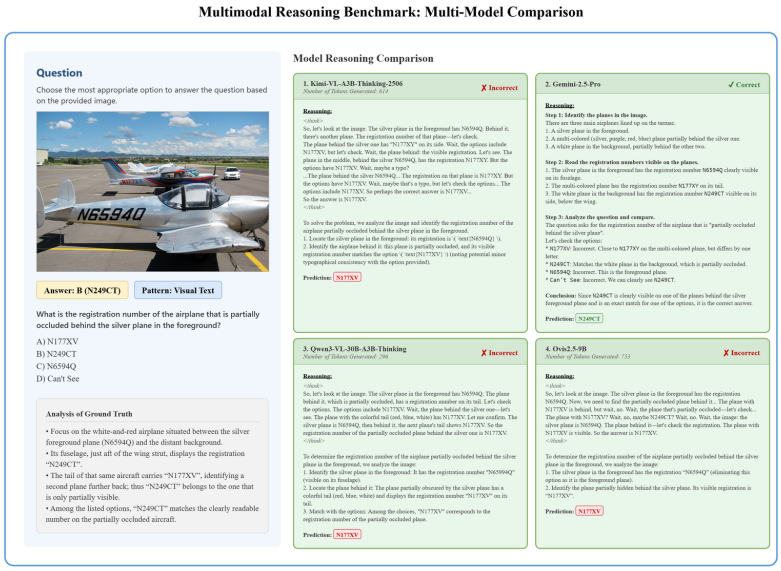


**Figure 6: Visual coreference under occlusion.** The question asks for the registration of the aircraft *partially occluded behind* the silver foreground plane. Gemini 2.5 Pro resolves it by enumerating the three visible aircraft and matching against the referring expression. Kimi-VL, Qwen3-VL-30B-Thinking and Ovis2.5-9B all answer with the *foreground* plane's registration. This is the saliency bias described in Section L: the referring expression is never applied. Token budgets of 614, 296 and 753 show this is not a matter of insufficient reasoning length. Note also that the wrong chains are step-numbered and cite specific visual evidence, which is why a step-scoring critic has little to work with here (Section E).

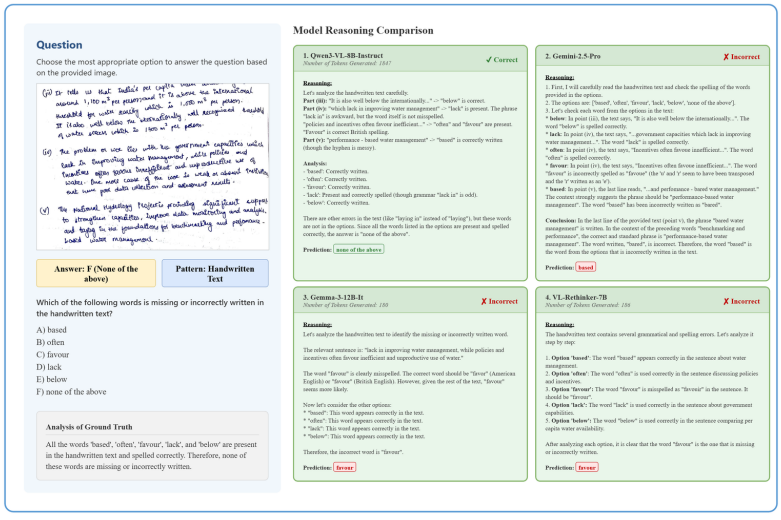


**Figure 7: Handwriting / OCR reasoning.** All five candidate words appear correctly in the passage, so the answer is "none of the above", which requires transcribing every option before concluding nothing is missing. Qwen3-VL-8B-Instruct does so. Gemini 2.5 Pro misreads one word and constructs an elaborate, internally consistent argument for "based". This illustrates both the cascade pattern and why handwriting items are cleaner by construction: the ground truth is decidable by inspection, so a reviewer can check the label without trusting our annotators. "Decidable" is not "easy": three of four systems fail.




**Figure 8: Navigation / map reading.** Gemma-3-12B-It answers correctly in **74 tokens**; systems generating 140, 172 and 316 tokens all answer incorrectly. This is the per-item mechanism behind Section 5.4: CORE normalizes each response against *this item's* human reference length, so verbosity on an item a person solves at a glance is charged more heavily than the same verbosity on a deep item. A marginal mean-length statistic cannot represent that distinction.

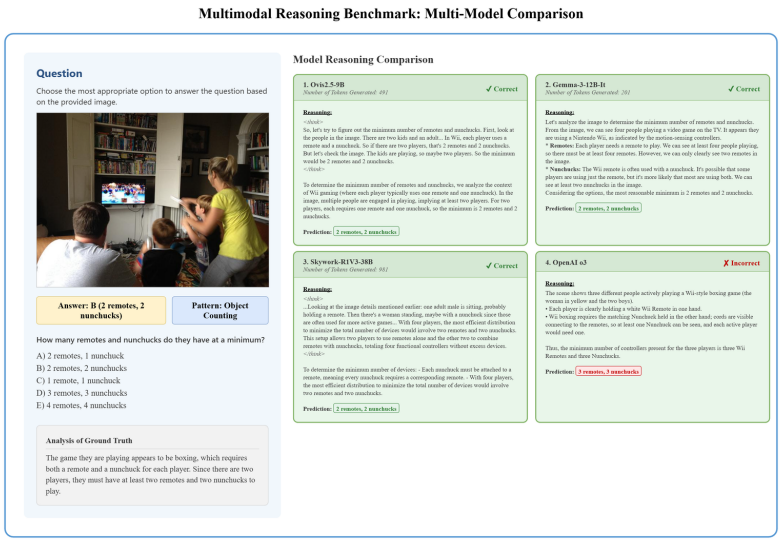


**Figure 9: Multi-hop reasoning.** Answering requires chaining what the scene shows to a fact about the game to a count. Ovis2.5-9B, Gemma-3-12B and Skywork-R1V3-38B chain correctly. OpenAI o3 miscounts players at the first hop and then reasons faithfully from that error, producing a well-structured chain wrong only at its root.

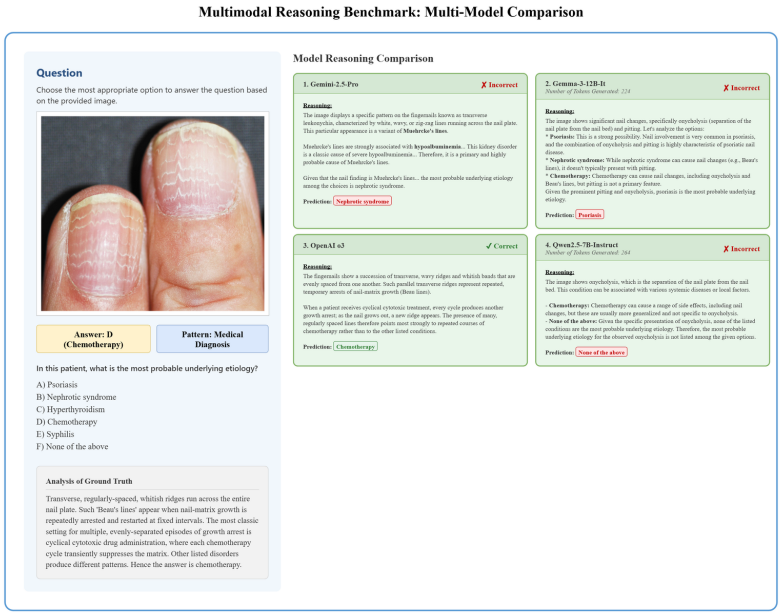


**Figure 10: Scientific / medical reasoning.** The finding is Beau's lines, implying cyclical chemotherapy. Only OpenAI o3 reaches this; Gemini reasons fluently to nephrotic syndrome, Gemma-3-12B to psoriasis. Each wrong answer arrives with a confident, clinically phrased causal chain. This is the category we exclude from the human study, and it shows why that exclusion is a real gap rather than a formality: a non-expert annotator could not adjudicate between these four answers.

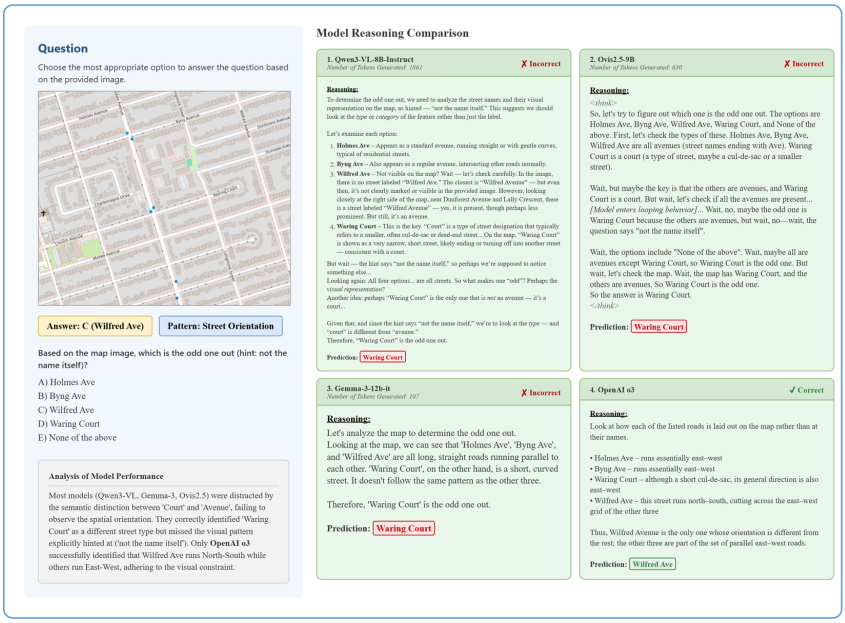


**Figure 11: Navigation / spatial orientation.** The question asks which street is the odd one out and states explicitly that the criterion is not the name. The answer is Wilfred Ave, the only street running north–south while the others run east–west. Only OpenAI o3 checks orientation. Qwen3-VL-8B-Instruct (1,861 tokens), Ovis2.5-9B and Gemma-3-12B all reason from the *linguistic* distinction between "Court" and "Avenue" despite the hint ruling it out. This is the linguistic shortcutting failure behind the $\sim$47% navigation scores in Table 2: the models read the map as text rather than as space.