VIALS: A BENCHMARK FOR VISUAL INTERPRETATION OF ARTIFACTS IN THE LIFE SCIENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

In professional life sciences workflows, scientists routinely interpret visual artifacts (gel blots, microscopy images, plasmid maps, flow cytometry plots, molecular structures, ...) to inform research decisions. We introduce VIALS, a visual question-answering benchmark with 161 such interpretation tasks, spanning the types of artifacts examined throughout experimental workflows in the biotech industry (rather than polished figures from publications and textbooks). While frontier vision-language models can now fluently describe natural images, we find that they are unable to accurately interpret these scientific images, reflecting limitations in domain knowledge and domain-specific visual reasoning capabilities. In contrast, scientists with relevant domain expertise find these visual interpretation tasks straightforward. AI that cannot similarly interpret such images will have limited utility in professional life sciences workflows, where such artifacts are central to how scientists reason, communicate, and make decisions.

1 INTRODUCTION

In the life sciences domain, information critical to research workflows is often encoded in visual artifacts such as gel blots, microscopy images, plasmid maps, phylogenetic trees, flow cytometry plots, docking views, and molecular structures. Current vision-language models (VLM) can generate reasonable descriptions of images, yet such descriptions do not guarantee an accurate expert interpretation of the critical information that enables scientific progress (Fu et al., 2024; Yue et al., 2025; Wu et al., 2026). Interpreting these scientific visual artifacts requires models to localize the relevant evidence, extract the correct labels or values, distinguish signal from noise, compare visual elements, and reason about the information—all of which require proper application of the underlying scientific concepts.

A model can fail at any of these steps even when it recognizes the artifact and understands the underlying biology. It may identify the wrong band in a gel lane, misread a flow cytometry gate, or overlook a relevant structural feature in a complex protein-ligand interaction. These failures point to capability shortcomings that go beyond text recognition: they involve localization, comparison, and interpretation of spatial or structural relationships. A further concern is that the final answer may still sound fluent and scientifically plausible, making errors difficult to detect. Previous biomedical evaluations have identified visual perception and grounding as sources of failure (Burgess et al., 2025; Liu et al., 2024). To progress toward more economically valuable AI for the biotech industry, it is thus important to assess: *How accurately can vision-language models interpret the scientific images routinely encountered in professional life sciences workflows?*

Existing benchmarks poorly cover this area, focusing on general academic knowledge (Center for AI Safety et al., 2026; Yue et al., 2024; 2025; Borisova et al., 2025) or solely on specialized biomedical modalities such as microscopy, pathology, or clinical imaging (He et al., 2020; Burgess et al., 2025; D'Cunha et al., 2026; Chen et al., 2024b). We introduce VIALS, a benchmark for **v**isual **i**nterpretation of **a**rtifacts in professional **l**ife **s**ciences. VIALS contains 161 visual question-answering (VQA) tasks spanning many industry-relevant scientific domains and artifact types. Each task pairs a scientific image with a question requiring the extraction and interpretation of visual evidence central to professional research workflows. The tasks are created and reviewed by PhD-level professional scientists with relevant domain and work expertise, ensuring they faithfully reflect real research workflows and have correct answers (Section 3). Figure 1 shows representative benchmark examples.

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.

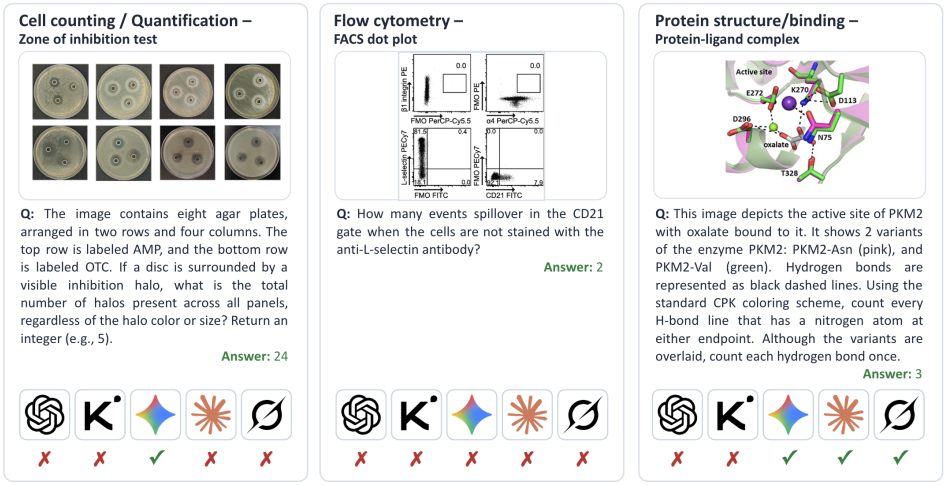


Figure 1: Three tasks from the VIALS benchmark. (*Left*) The **Zone of inhibition** artifacts depict agar plates containing growing microorganisms, and this task requires identifying the clear zones around the disks that reveal antibiotic activity—a common assessment in antibiotic discovery and clinical microbiology. (*Middle*) The **FACS dot plot** artifacts visualize the results of a flow cytometry assay to characterize cell populations, and this task requires identifying cell detection events for a particular biomarker—a key step in immunology research and cell therapy development. (*Right*) The **Protein-ligand complex** artifact is a 3D visualization from structural biology software, showing how a molecule fits into a protein, and this task requires identifying the molecular contacts that enable binding—a key step in structure-based drug discovery. For each task, the correctness of answers from various models is shown along the bottom (GPT-5.6 Sol, Kimi K3, Gemini 3.1 Pro, Claude Opus 5, and Grok 4.6, from left to right).

For scientists trained in the relevant domains, many VIALS tasks are everyday parts of scientific practice, and performing them accurately is critical for research progress. A model that cannot reliably perform these tasks cannot practically be trusted in high-stakes life sciences workflows. We evaluate all of today's best available multimodal models, observing that the top-performing models, GPT-5.6 Sol and Gemini 3.7 Flash, both achieve only 26.5% accuracy. Around 90% of unsuccessful tasks involve errors in reading the artifact correctly, including miscounting, misreading measurements, or missing spatial and structural relationships (Section 5.2). Our results show that reliable interpretation of scientific artifacts remains a major bottleneck for current multimodal models, which struggle to properly apply scientific domain knowledge during visual processing and reasoning.

2 Related Work

Existing benchmarks cover important aspects of scientific and multimodal visual reasoning, but largely target academic settings. MMMU and MMMU-Pro evaluate multimodal reasoning across many disciplines using academic materials such as exams and textbooks (Yue et al., 2024; 2025). SciVQR (Guo et al., 2026) and SciVQA Borisova et al. (2025) source conceptual questions from scientific figures in textbooks, exams, academic competitions, and publications. More specialized benchmarks target particular biomedical visual settings: GMAI-MMBench evaluates VQA across medical imaging modalities and clinical tasks (Chen et al., 2024b), while MicroVQA focuses on microscopy images (Burgess et al., 2025). These benchmarks provide important evaluations of scientific VQA, but do not target the visual artifacts and tasks encountered across professional life sciences work.

Humanity's Last Exam (HLE) evaluates expert-level questions across many academic disciplines, including VQA tasks in Biology/Medicine and Chemistry (Center for AI Safety et al., 2026). However, an independent audit estimates that 29% of HLE's biology and chemistry tasks have untrustworthy reference answers (Skarlinski et al., 2025). In contrast, every reference answer in VIALS is rigorously vetted via careful reviews of multiple independent attempts at the task by different experts (Section 3.5). Moreover, Figures 2 and 3 reveal that VIALS tasks are significantly more relevant to professional life sciences workflows than the tasks in HLE, which tend to be more academic/exam-like. See more comparisons in Appendix B.

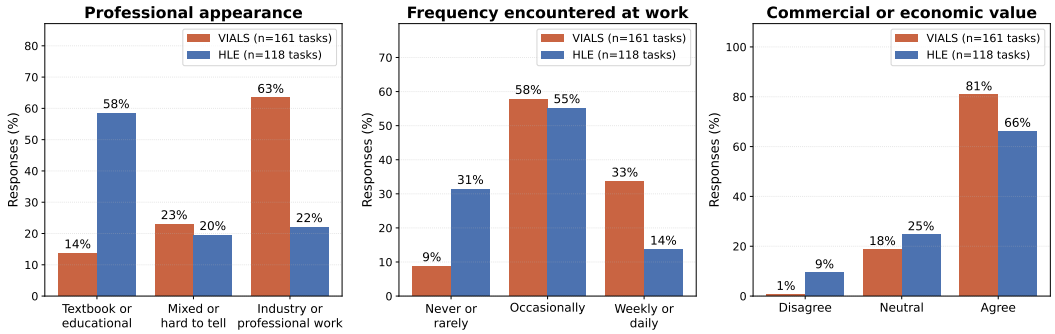
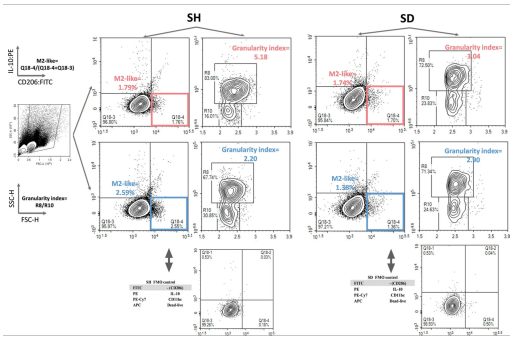
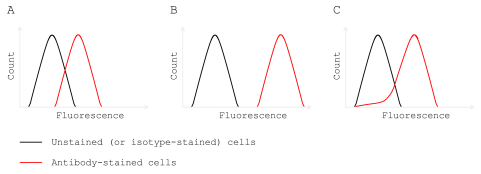


Figure 2: Results from an expert survey (detailed in Section 3.6) in which scientists with relevant domain expertise rated the professional relevance of individual benchmark tasks from VIALS and HLE's Chemistry and Biology/Medicine VQA domains. Professional scientists assess whether each artifact appears to stem from professional work rather than educational material, how often they encounter this type of artifact in their work, and whether this interpretation task has commercial or economic value.



(a) **Flow cytometry task in VIALS.** *Q.* What is the largest minus the smallest background-corrected high-granularity M2-like yield per 100,000 parent events? Round to the nearest whole cell. *A.* 989 cells per 100,000 parent events



(b) **Flow cytometry task in HLE.** *Q.* Three flow cytometry histograms (A, B, C) show cells stained with three different fluorescently-labelled antibodies. Which antibody stained *all* cells in the sample? *A.* **A**

Figure 3: Example tasks from VIALS (*left*) and HLE (*right*) in the flow cytometry domain, contrasting a professional gating workflow with a more academic textbook/exam-like conceptual question. In immunology or cell therapy research, interpretation tasks like the one on the left help a scientist determine whether a treatment changed immune cell composition. This is direct interpretation of experimental output rather than a conceptual test of principles like the right-hand task.

3 The VIALS Benchmark

VIALS evaluates whether VLMs can interpret professional life science artifacts and use the visual evidence they contain to answer core scientific questions. Each task in the benchmark consists of one or more images, a question, and a short ground-truth answer (held-out and solely used for grading).

3.1 Scoring model outputs

VIALS is intended to evaluate VLMs with reasoning capabilities (Xu et al., 2025). Our benchmark solely evaluates a model's generated final answer, not its intermediate reasoning process. To better reflect how surveyed scientists reported wanting to incorporate AI in their workflows, we design tasks to have open-domain text answers rather than multiple-choice. Our benchmark uses semantic grading rather than exact-match evaluation because correct responses may differ from the reference answer in formatting, units, numerical representation, or wording while remaining scientifically equivalent (although questions are authored to minimize ambiguity where applicable).

As the ground-truth answers are short-form, we can effectively employ a simple LLM-as-a-judge approach to determine whether each model-generated response is correct (Center for AI Safety et al., 2026; Wei et al., 2024; 2025). Using the grading prompt in Appendix C.1, our LLM judge takes in the question, the model response, and the reference ground-truth answer. For certain tasks that experts determine have multiple acceptable answers, the LLM judge additionally receives this (expert-determined) acceptable range. Because the judging task is straightforward when given the short-

Domain	Representative image types	Scientific fields	Share
Phylogenetics	Rectangular and circular phylogenetic trees; branch-length and clade diagrams; taxon and clade labels	Evolutionary biology, microbiology, genomics	14.3%
Flow cytometry	Gating hierarchies, density/contour plots, FMO overlays, histograms, and multi-parameter panels	Immunology, biomedical research, cell biology	14.3%
Blotting	Western blots, SDS-PAGE, and 2D gels; lane- and band-level intensity comparisons	Protein biochemistry, molecular biology	14.3%
Cell counting/Quantification	Fluorescence images, colony and plaque images, blood smears, and confluence estimation	Molecular biology, microbiology, cell biology	14.3%
Plasmid maps	Circular and linear plasmid maps; restriction maps; feature annotations and positions	Molecular biology, synthetic biology, virology	13.0%
Protein structure	Protein–ligand binding diagrams, active-site views, 3D structure and residue contact maps	Structural biology, biochemistry, drug discovery	13.0%
Molecular structure	Small-molecule skeletal structures; substituent identification and reaction transformations	Medicinal chemistry, synthetic chemistry	12.4%
Other	Pedigrees, Lateral flow assays, etc.	Genetics, systems biology, and related fields	4.3%

Table 1: Artifact domains covered by VIALS, their representative image types, the scientific fields in which those artifacts are commonly found, and each domain's prevalence in the final 161-task benchmark.

form reference answer, we employ a fast, cost-effective model as the judge: OpenAI's GPT-5-mini. We validate the LLM judge against independent human expert grading of 4,685 model-responses across all 161 tasks, where it agrees with human judgments on 99.9% of cases. When testing various judge models for VIALS tasks, we observed little evidence of self-preference, verbosity, or style bias in the LLM-judge (Ye et al., 2025).

3.2 Domains covered in the benchmark

VIALS covers a range of visual artifacts used across professional life sciences research. Artifact domains were selected according to their criticality in high-value research workflows across the life sciences industry, particularly in the biotechnology and pharmaceutical sectors. For instance, many activities in pharmaceutical research depend on interpreting such artifacts, such as experimental characterization, assay development/screening, and drug discovery/development (Chen et al., 2024a; Busby et al., 2020; Sinha & Vohora, 2018).

Table 1 summarizes the artifact categories, representative image types, and the scientific fields in which they are commonly used. These include phylogenetic trees, flow cytometry plots, blots and gels, plasmid maps, cell-counting and quantification images, protein-structure and binding views, and small-molecule structures. Each task is assigned a primary artifact domain and one or more secondary tags describing the specific artifact subtype, such as **Western blot** versus **SDS-PAGE** within the **blotting** domain, or **linear** versus **circular trees** within the **phylogenetics** domain.

3.3 Artifact sourcing and task construction

VIALS includes artifacts that are unpublished real experimental data, artifacts created through procedural generation with scientific software, and artifacts sourced from open-access publications with a CC-BY license. The benchmark emphasizes messy data artifacts that are typical in real-world research. Procedural generation using scientific software was used for select artifact types: phylogenetic trees, plasmid maps, and blots (all of which were expert-verified as realistic).

To create tasks for the benchmark, domain experts provide one or more images and a question that reflect their actual work experience, and then provide the ground truth answer for the question. Beyond representing core tasks from their work, the questions are designed to be answerable from the

provided artifact while requiring both the extraction of relevant visual evidence and the application of scientific knowledge or reasoning.

3.4 TASK CONTRIBUTORS

VIALS was developed by 31 contributors[1] with deep professional expertise in at least one of the category domains. Contributors trained at institutions including Harvard Medical School, MIT, Stanford, UC Berkeley, UCLA, NYU, Yale, and Oxford, and had prior work experience at organizations including Johnson & Johnson, Amgen, Bristol Myers Squibb, Abbott Laboratories, Mayo Clinic, the National Institutes of Health, and the U.S. Food and Drug Administration. Overall, 87% had doctoral-level training, and 74% had more than six years of professional work experience as a researcher in the life sciences industry. All contributors passed a rigorous qualification assessment, demonstrating scientific expertise in at least one of the following areas: immunology, microbiology, molecular biology, genetics and genomics, biochemistry, oncology, virology, neuroscience, bioinformatics, or drug discovery. Contributors created and reviewed benchmark tasks only within their areas of expertise, ensuring that the tasks reflect critical activities from their own professional experience.

3.5 QUALITY CONTROL PIPELINE

Each task passes through multiple stages of human review to ensure it is realistic and correctly specified (Skarlinski et al., 2025). A domain expert constructs the task by selecting or generating the relevant artifact, writing the question, and providing a reference answer. A second (and often third) expert with relevant domain expertise then answers the same task independently, without seeing the original answer. These independent task attempts (along with writing detailed justifications for their answers) took experts 16 minutes on average. Another expert reviewer checks that the images are understandable and realistic, and the question represents a meaningful life sciences task that is well-specified and answerable from the provided information. In addition, reviewers grade the correctness of model-generated answers and the independent expert attempts, to finalize the ground-truth answer and whether a range of values should be considered acceptable. A task is retained only if the independent experts' answers agree with the original reference answer or experts reach a consensus in follow-up reviews.

During benchmark construction, we additionally ensured task diversity by checking whether incoming tasks are similar to others in the benchmark. We also ensure tasks are not already found online (i.e. in frontier models' pretraining data). Whenever the prevalence of tasks from certain domains fell below their target representation (estimated via expert surveys of common life sciences workflows), we encouraged the relevant experts to create more tasks in these underrepresented domains until the benchmark reflected the target domain distribution in Table 1.

3.6 EXPERT ASSESSMENT OF PROFESSIONAL RELEVANCE

Beyond validating task correctness, we also asked expert reviewers to rate the professional relevance of VIALS tasks within their domain of expertise (rating tasks created by other experts, which they were not familiar with). We asked these same experts to similarly assess 118 VQA tasks from HLE's Biology/Medicine and Chemistry subset[2]. Reviews were conducted blind, such that reviewers were unaware of the source of a task while rating it.

Figure 2 and Appendix C.4 detail the survey questions and results, revealing that experts are more likely to classify VIALS artifacts as resembling professional work than those in HLE (63% versus 22%). Experts also reported encountering VIALS artifact types more frequently in their work, with 33% selecting weekly or daily compared with 14% for HLE. For 81% of VIALS tasks, experts agreed that interpreting the artifact has commercial or economic value, compared with 66% for HLE. These assessments indicate that visual tasks in VIALS are not only scientifically meaningful, but also closely connected to professional life science practice.

4 EVALUATION SETUP

We evaluate several frontier VLMs from different providers. All multimodal models receive the same prompt, which includes only the image(s) and the question. Appendix C.1 provides the full

[1] Benchmark contributors were sourced from a professional talent network.

[2] No reviewer felt familiar with the 119th such HLE task, so our expert assessment of professional relevance omitted it.

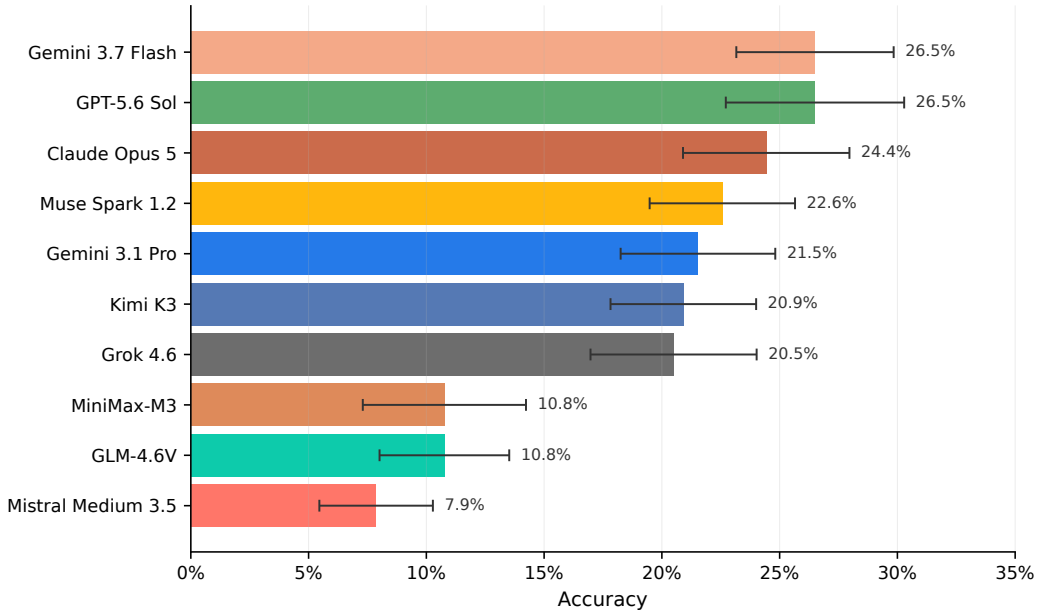


Figure 4: Performance of various vision-language models on VIALS (averaged across 3 independent rollouts per task). Error bars show 95% confidence intervals computed by bootstrap resampling across the three rollouts for each task (reflecting variability in *accuracy* under re-evaluation with one attempt per task). Claude Fable 5 could not be evaluated as its safety guardrails blocked most of the requests from this benchmark.

prompt and configurations used. We independently run each model three times on each task, and all results report its average accuracy across the three rollouts.

5 Results

5.1 Overall model performance

Figure 4 shows overall accuracy of each vision-language model on VIALS. The top-performing models, GPT-5.6 Sol and Gemini 3.7 Flash, achieve only 26.5% accuracy, with Gemini 3.7 Flash incurring lower costs (Figure A2). The Pass^3 results in Figure A1 show that these models are only able to handle under 17% of VIALS tasks if we require that all 3 rollouts from the model are correct, a basic requirement for trusting a model in critical life sciences research.

Performance differs substantially across domains (Table 2). Phylogenetics is the strongest domain for the leading models, with GPT-5.6 Sol and Muse Spark 1.2 reaching 43.5% and 34.8% accuracy, respectively. Results are lower in most other domains. The best scores are 38.1% on plasmid maps (Muse Spark 1.2), 33.3% on both protein structure (Muse Spark 1.2) and flow cytometry (Kimi K3 and Grok 4.6, tied), 31.7% on molecular structure (GPT-5.6 Sol), 30.4% on blotting (GPT-5.6 Sol and Claude Opus 5), and 21.7% on cell counting/quantification (Gemini 3.7 Flash). Current VLMs generally struggle the most in cell counting/quantification tasks and protein structure tasks.

5.2 Failure mode analysis

Figure 5 shows that failures are not limited to extracting information from the artifact. Models may identify salient visual evidence yet still apply the wrong representational convention, substitute prior expectations for evidence in the artifact, or fail to use the extracted information consistently in the final inference. These examples highlight a distinct challenge in scientific visual reasoning: correctly interpreting visual evidence in the context of the underlying scientific representation.

To understand where models fail, we assign a primary failure mode for each model on each task with at least one incorrect rollout via axial coding with an LLM classifier (see Appendix C.2 for details). We consider seven failure categories: *Visual quantification error*, for errors in counting objects such as cells, colonies, or bands; *Incorrect value or feature selection*, for reading the wrong value, label, lane, gate, or feature; *Overlooked evidence*, for missing relevant information visible in the artifact; *Representation misinterpretation*, for incorrectly reading the structure or conventions of a plot, tree,

Model	Blotting	Cell Count. / Quant.	Flow Cytometry	Molecular Structure	Phylogenetics	Plasmid Maps	Protein Structure	Other
GPT-5.6 Sol	**30.4%**	15.9%	29.0%	**31.7%**	**43.5%**	22.2%	15.9%	**14.3%**
Gemini 3.7 Flash	29.0%	**21.7%**	31.9%	30.0%	29.0%	23.8%	23.8%	**14.3%**
Claude Opus 5	**30.4%**	20.3%	27.5%	18.3%	31.9%	25.4%	19.0%	**14.3%**
Muse Spark 1.2	15.9%	2.9%	21.7%	15.0%	34.8%	**38.1%**	**33.3%**	**14.3%**
Gemini 3.1 Pro	24.6%	18.8%	14.5%	20.0%	31.9%	25.4%	17.5%	**14.3%**
Kimi K3	21.7%	10.1%	**33.3%**	23.3%	27.5%	19.0%	12.7%	**14.3%**
Grok 4.6	14.5%	15.9%	**33.3%**	13.3%	30.4%	19.0%	17.5%	**14.3%**
GLM-4.6V	18.8%	11.6%	10.1%	3.3%	13.0%	12.7%	6.3%	4.8%
MiniMax-M3	8.7%	8.7%	13.0%	5.0%	17.4%	15.9%	7.9%	4.8%
Mistral Medium 3.5	15.9%	4.3%	8.7%	8.3%	7.2%	7.9%	4.8%	0.0%

Table 2: Per-domain accuracy on VIALS achieved by each vision-language model. Bold indicates the best performance in each domain.

Model	Visual quantification error	Incorrect value or feature selection	Overlooked evidence	Representation misinterpretation	Quantitative reasoning error	Misapplied principle	Fabricated finding
GPT-5.6 Sol	41.8%	14.2%	15.6%	14.2%	6.4%	3.5%	4.3%
Gemini 3.7 Flash	35.3%	18.0%	18.0%	16.5%	8.3%	3.0%	0.8%
Claude Opus 5	36.8%	16.9%	17.6%	14.0%	10.3%	4.4%	0.0%
Muse Spark 1.2	38.3%	26.3%	12.0%	16.5%	3.8%	2.3%	0.8%
Gemini 3.1 Pro	34.3%	12.9%	20.7%	15.0%	10.7%	5.0%	1.4%
Kimi K3	33.1%	20.9%	15.8%	16.5%	7.9%	4.3%	1.4%
Grok 4.6	33.1%	17.2%	19.3%	17.9%	7.6%	4.8%	0.0%
GLM-4.6V	28.3%	19.7%	17.8%	17.8%	9.9%	5.9%	0.7%
MiniMax-M3	32.1%	18.6%	19.2%	20.5%	5.8%	3.2%	0.6%
Mistral Medium 3.5	28.2%	19.9%	15.4%	21.8%	7.7%	6.4%	0.6%

Table 3: Distribution of failure modes by model.

plasmid map, or other scientific representation; *Quantitative reasoning error*, for arithmetic, bound, or unit errors after the relevant values have been identified; *Misapplied scientific principle*, for applying an incorrect scientific rule or technique; and *Fabricated finding*, for introducing an element or measurement not supported by the artifact.

Most failures occur while models are reading or interpreting the artifact. Across models, 83–93% of errors fall into the first four categories in Table 3. Visual quantification is the most common failure mode for every model, accounting for 28.2–41.8% of errors. Incorrect value or feature selection, overlooked evidence, and representation misinterpretation are also common across models.

Quantitative reasoning errors account for 3.8–10.7% of failures, while misapplied scientific principles account for 2.3–6.4%. Fabricated findings are rare, accounting for at most 4.3% of errors. Overall, the failure analysis suggests that the main difficulty is often identifying and correctly interpreting the information encoded in the artifact, rather than carrying out the subsequent calculation or applying a scientific principle.

5.3 Evaluating tool-assisted agents

Our VQA benchmark focuses on evaluating VLMs, with tasks being completed via single-turn inference calls to multimodal reasoning models. Human scientists can handle these tasks relatively quickly by simply viewing the image without relying on software or external tools. Competent scientific AI models should be able to do the same. Modern life sciences research is starting to generate visual artifacts at scale: high-content screening campaigns such as Cell Painting routinely produce images across hundreds of thousands of chemical and genetic perturbations (Bray et al., 2016; Chandrasekaran et al., 2023), and interpretation of assay readouts is a recurring, throughput-limiting step in drug discovery pipelines (Busby et al., 2020). A model that interprets these artifacts with expert-level accuracy and machine cost/speed can remove a bottleneck that is currently dependent on human experts. A less efficient AI system will yield less acceleration of the experimental loop.

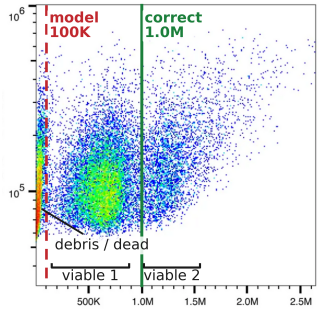
Nonetheless, our previous failure-mode analysis raises the question: how many errors made under direct visual evaluation can be recovered when models are given additional ways to iteratively inspect the same artifact? To answer this, we evaluate the same multimodal models in an *agentic setup with code execution*, where models can loop and iteratively crop, zoom, transform, measure, and otherwise programmatically inspect the supplied image before answering.

We evaluate several model and agent-harness configurations, including OpenAI's Codex, Claude Code, and OpenCode[3] (see Appendix C.3 for details). Table 4 shows that agentic tool access markedly improves performance for most models. However, these gains come with substantially greater token usage and costs: tool-assisted runs use 10–432× more tokens per attempt than direct evaluation, depending on the model–harness configuration. GLM-4.6V is an exception: its overall performance decreases with tool access.

[3] Agent-harness resources: Codex; Claude Code; OpenCode.

(a) Flow cytometry

"...dead and dying/apoptotic cells have not been removed. Identify the most optimum x-axis position of vertical boundary line between the two clusters of viable cells..."

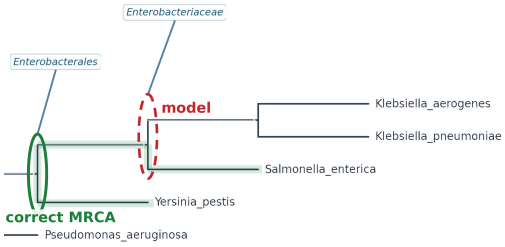


Model (claude-opus-5): 100K, axis-hugging cluster edge
Correct: 1.0M — minimum between viable populations

(a) **Wrong population.** The model selects the debris–viable boundary instead of the boundary between the two viable populations.

(b) Phylogenetics

"Reading the named internal nodes, what shared clade most specifically contains both Salmonella_enterica and Yersinia_pestis?"

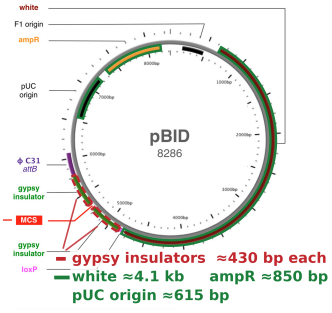


Model (gpt-5.6-sol): Enterobacteriaceae
Correct: Enterobacterales

(b) **Topology error.** The model relies on spatial proximity rather than the tree topology.

(c) Plasmid map

"Which feature(s), represented as colored rectangles in the plasmid map, are greater than or equal to 600bp in length?"

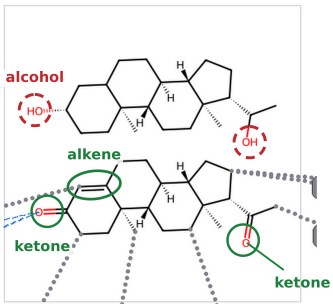


Model (glm-4.6v): gypsy insulators "typically ≥600 bp"
Correct: ampR, pUC origin, white — measured on the map

(c) **Prior over evidence.** The model uses a typical feature length rather than measurements from the plasmid map.

(d) Molecular structure

"What functional groups are present in the ligand?"



Model (grok-4.6): ketones, alkene — alcohols missed
Correct: alcohol, ketone, alkene

(d) **Feature not perceived.** The model correctly names the two ketone carbonyls and the alkene, but asserts that no alcohol groups are present on the ligand, missing the two hydroxyls on the steroid A-ring and side chain.

Figure 5: Representative model failures beyond visual extraction. Models can identify relevant features in scientific artifacts yet still use them incorrectly in downstream reasoning. Each panel shows the model's reading in red, and the correct reading supported by the artifact in green.

Model	Harness	Direct VLM	Tool-Assisted	Δ	Tok_D	Tok_T	Tok×
GPT-5.6 Sol	Codex	26.5	49.1	+22.6	4.1k	165k	40×
GPT-5.6 Sol	OpenCode	26.5	40.4	+13.9	4.1k	47k	11×
Claude Opus 5	Claude Code	24.4	64.2	+39.8	5.3k	959k	182×
Claude Opus 5	OpenCode	24.4	65.8	+41.4	5.3k	737k	139×
Gemini 3.1 Pro	OpenCode	21.5	35.2	+13.7	3.5k	1.0M	285×
Gemini 3.7 Flash	OpenCode	26.5	60.7	+34.2	3.7k	1.6M	432×
Grok 4.6	OpenCode	20.5	45.8	+25.3	10.4k	382k	37×
Kimi K3	OpenCode	20.9	64.2	+43.3	20.2k	557k	28×
GLM-4.6V	OpenCode	10.8	7.0	−3.7	5.5k	56k	10×

Table 4: Performance of tool-assisted agents on VIALS, across seven models run in different harnesses. **Direct VLM** vs. **Tool-Assisted** report accuracy (%) under the single/direct VLM call vs. agentic tool-assisted settings, respectively. Values are the average over $K=3$ rollouts. **Δ** reports the absolute change in accuracy in percentage points. $\textbf{Tok}_\text{D}$ and $\textbf{Tok}_\text{T}$ report mean tokens per attempt in the direct-VLM and tool-assisted settings, respectively; **Tok**× reports the ratio of tool-assisted to direct-VLM token usage.

These results suggest that many failures of the reasoning VLM are recoverable with additional inference-time resources and access to software tools. However, these gains incur a significant associated computational cost burden, despite the fact that scientists can quickly interpret the underlying artifacts in the relevant domains without such tools. For today's frontier VLMs, this indicates that there remains significant room for improvement in the base model's raw visual localization and reasoning capabilities, particularly as it pertains to such artifacts.

6 Conclusion

This paper introduces VIALS, a benchmark for evaluating how accurately VLMs can interpret visual artifacts from professional life sciences workflows that carry direct commercial and experimental consequences. Across leading available multimodal models, accuracy remains limited. Our error analysis indicates that many failures arise before higher-level reasoning: models often identify the general artifact type but misread labels, values, regions, or relationships needed to answer the question correctly. These results suggest that reliable interpretation of scientific images remains a substantial bottleneck for deploying VLMs in life sciences workflows.

Our agentic tool-assisted evaluation further decomposes the deficit of current frontier models into two gaps. First, a *perception gap*: when the same models are given code-execution tools to iteratively crop, zoom, and measure the artifact, accuracy improves by up to 43 points, indicating that much of the required scientific knowledge is present but cannot be applied through the neural network's direct visual inference (Tong et al., 2024). However, these gains come at 10–432× the token cost of direct inference, for interpretations that trained scientists can resolve in minutes. Second, a residual *interpretation gap*: even with unrestricted iterative inspection and programmatic tool use, the best agent solves only 65% of tasks, with failures concentrated in reading scientific representations and conventions along with proper application of other domain knowledge. Because every VIALS task is grounded in a real workflow with an expert-vetted answer, the benchmark measures both gaps against the standard that professional work requires.

VIALS has several limitations. The benchmark is limited in scale and does not capture the full range of visual artifacts or interpretation activities that encompass all life sciences research, focusing instead on high-frequency tasks in certain high-value workflows. The benchmark focuses on tasks which are relatively straightforward for scientists with the relevant Ph.D. to quickly accomplish, whereas professional research involves many other scientific artifact interpretations that are highly nontrivial due to ambiguity and experiment noise. VIALS also does not assess how well models can interpret scientific figures from publications nor academic textbooks/exams, which can require significantly more complex reasoning for humans to do accurately. Upon interpreting the visual artifacts from a life sciences workflow, human scientists must subsequently decide how to adjust their research plan—an important capability to evaluate models for that we leave to future benchmarks.

We hope VIALS advances the development of models that can interpret scientific artifacts accurately enough to be useful in professional life sciences research. To expand the benchmark, we are privately developing hundreds of additional tasks that will enable future benchmark refreshes and prevent overfitting to the current public benchmark.

AI USE STATEMENT

We used generative AI tools to assist with writing refinement and literature discovery. The authors drafted the manuscript sections and used AI assistants to improve clarity, grammar, and phrasing. We also used AI tools to help identify potentially relevant related work. The authors take responsibility for the final content of this work, including text, claims, and references produced or refined with AI assistance.

REPRODUCIBILITY STATEMENT

We provide the evaluation codebase and the VIALS dataset in the supplementary materials. The appendix includes evaluation prompts, model configurations, and methodological details to support replication of our experiments. We will also make the dataset publicly available on Hugging Face.

ETHICS STATEMENT

VIALS was developed with contributions and assessments from life sciences experts, who were compensated for their work. Task contributors were informed that they were creating tasks for a research benchmark and consented to participate.

The benchmark includes unpublished experimental artifacts, procedurally generated artifacts, and images from open-access publications. The dataset is released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. The accompanying attribution file documents per-image provenance, source information, original licenses, and redistribution licenses for all 161 tasks. Users are instructed to attribute the original sources when reusing individual images.

REFERENCES

Ekaterina Borisova, Nikolas Rauscher, and Georg Rehm. SciVQA 2025: Overview of the first scientific visual question answering shared task. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pp. 182–210. Association for Computational Linguistics, 2025.

Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M Gustafsdottir, Christopher C Gibson, and Anne E Carpenter. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, 11(9):1757–1774, 2016.

James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G. Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, Sarina M Hasan, Alexandra Johannesson, William D. Leineweber, Malvika G Nair, Ridhi Yarlagadda, Connor Zuraski, Wah Chiu, Sarah Cohen, Jan N. Hansen, Manuel D Leonetti, Chad Liu, Emma Lundberg, and Serena Yeung-Levy. MicroVQA: A multimodal reasoning benchmark for microscopy-based scientific research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19552–19564, 2025.

Scott A Busby, Seth Carbonneau, John Concannon, Christoph E Dumelin, YounKyoung Lee, Shin Numao, Nicole Renaud, Thomas M Smith, and Douglas S Auld. Advancements in assay technologies and strategies to enable drug discovery. *ACS Chemical Biology*, 15(10):2636–2648, 2020.

Center for AI Safety, Scale AI, and HLE Contributors Consortium. A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, 649:1139–1146, 2026.

Srinivas Niranj Chandrasekaran, Jeanelle Ackerman, Eric Alix, D. Michael Ando, John Arevalo, et al. JUMP Cell Painting dataset: Morphological impact of 136,000 chemical and genetic perturbations. *bioRxiv*, 2023.

Jinghong Chen, Anqi Lin, and Peng Luo. Advancing pharmaceutical research: A comprehensive review of cutting-edge tools and technologies. *Current Pharmaceutical Analysis*, 21(1):1–19, 2024a.

Jiuhai Chen and Jonas Mueller. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 5186–5200, 2024.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, Shaoting Zhang, Bin Fu, Jianfei Cai, Bohan Zhuang, Eric J. Seibel, Junjun He, and Yu Qiao. GMAI-MMBench: A comprehensive multimodal evaluation benchmark towards general medical AI. In *Advances in Neural Information Processing Systems*, volume 37, 2024b.

Ryan D'Cunha, Alejandro Lozano, Xiaoxiao Sun, Daniel Vela Jarquin, Min Woo Sun, Josiah Aklilu, James Burgess, Yuhui Zhang, Ryan Nayebi, Paola Avila, et al. MMBU: A massive multi-modal biomedical understanding benchmark to probe the perception capabilities of vision-language models. *arXiv preprint arXiv:2606.06696*, 2026.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal Large Language Models Can See but Not Perceive. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Longteng Guo, Xuanxu Lin, Dongze Hao, Tongtian Yue, Pengkang Huo, Jiatong Ma, Yuchen Liu, and Jing Liu. SciVQR: A multidisciplinary multimodal benchmark for advanced scientific reasoning evaluation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 577–601. Association for Computational Linguistics, 2026.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. PathVQA: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Yunyi Liu, Yingshu Li, Zhanyu Wang, Xinyu Liang, Lingqiao Liu, Lei Wang, Leyang Cui, Zhaopeng Tu, Longyue Wang, and Luping Zhou. A systematic evaluation of GPT-4V's multimodal capability for chest x-ray image analysis. *Meta-Radiology*, 2(4):100099, 2024.

Sandeep Sinha and Divya Vohora. Chapter 2 - drug discovery and development: An overview. In Divya Vohora and Gursharan Singh (eds.), *Pharmaceutical Medicine and Translational Clinical Research*, pp. 19–32. Academic Press, Boston, 2018.

Michael Skarlinski, Jon Laurent, Albert Bou, and Andrew White. About 30% of humanity's last exam chemistry/biology answers are likely wrong. FutureHouse Research, 2025. `https://www.futurehouse.org/research/hle-exam`.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, 2023.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. BrowseComp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.

Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33577–33587, 2026.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. LLaVA-CoT: Let vision language models reason step-by-step. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2087–2098. IEEE, 2025.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. Justice or prejudice? quantifying biases in llm-as-a-judge. In *International Conference on Learning Representations*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, 2024.

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. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 15134–15186. Association for Computational Linguistics, 2025.

A Additional Results

A.1 Pass@3 and Pass^3 performance, costs, and token efficiency

The main leaderboard reports mean accuracy across three rollouts. We additionally report Pass@3, the fraction of tasks answered correctly in at least one of the three attempts, and Pass^3, the fraction answered correctly in all three attempts.

Figure A1 compares mean accuracy, Pass@3, and Pass^3. Across models, Pass@3 tops out around 40%, while Pass^3 remains below 17%. The large gap between these metrics shows that models often fail to reproduce correct answers across repeated attempts. This matters in professional workflows, where a model that is correct once but unreliable on the next run is difficult to depend on.

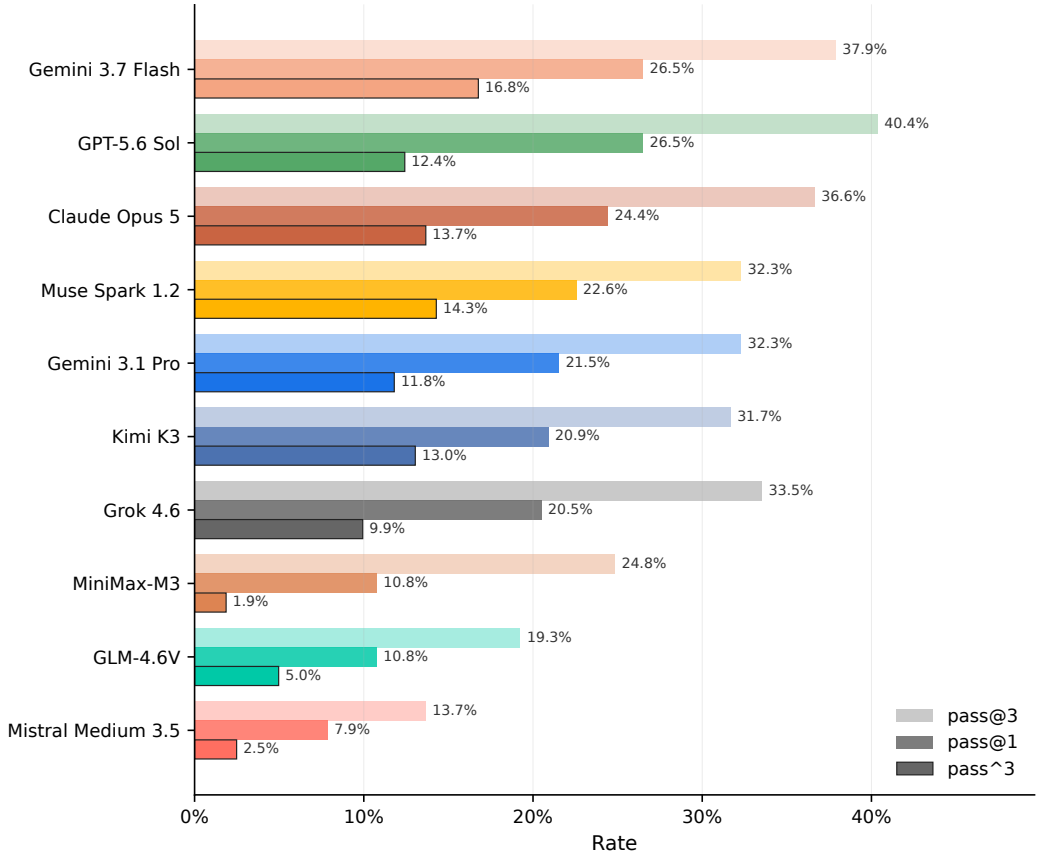


Figure A1: Mean accuracy (Pass@1), Pass@3, and Pass^3 achieved by different vision-language models. Pass@3 measures whether a task is solved correctly in at least one of three rollouts, while Pass^3 measures whether it is consistently solved across all three.

Figure A2 compares model accuracy with inference cost and output-token usage. We observe substantial variation in efficiency across models, with higher cost or longer outputs not consistently corresponding to higher accuracy.

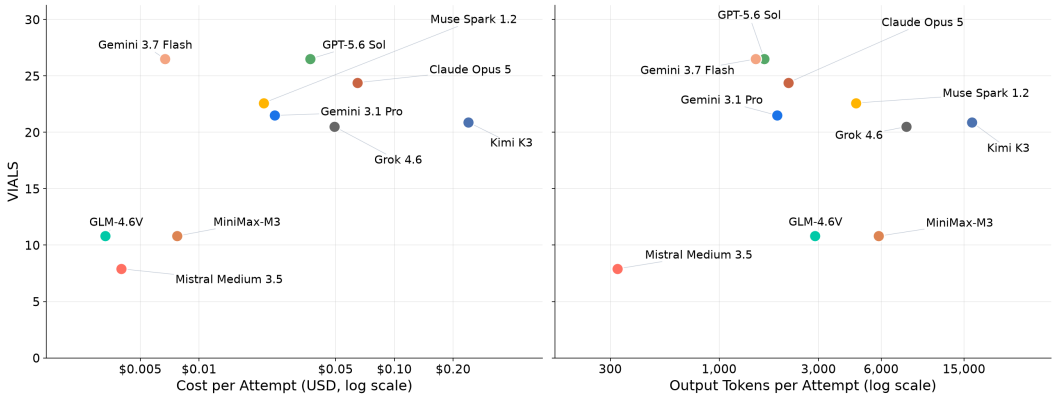


Figure A2: **Cost and output-token efficiency on VIALS.** Each point represents one model; accuracy is plotted against inference cost per attempt (**left**) and output tokens per attempt (**right**).

A.2 Self-reported model confidence

In high-stakes AI applications like life sciences research, it is valuable to have a calibrated measure of uncertainty in model responses to know when they can be trusted (Chen & Mueller, 2024). We evaluate whether the confidence reported by the model is predictive of the correctness of the response in VIALS (Tian et al., 2023). For each task (and for all models), we augment the prompt in Appendix C.1 by additionally asking the model to verbalize a confidence score for its answer, expressed on a five-point Likert scale.

Addition to task prompt for self-reporting confidence

After providing your final answer, report your confidence that the final answer is correct on the following scale:

1 = Very low: essentially guessing; substantial uncertainty
2 = Low: significant uncertainty; answer may well be incorrect
3 = Moderate: plausible answer, but meaningful ambiguity remains
4 = High: likely correct; only limited uncertainty remains
5 = Very high: clear interpretation; little meaningful uncertainty

Format your response exactly as:

```
REASONING: <your reasoning>
ANSWER: <final answer>
CONFIDENCE: <1--5>
```

Before grading, we strip the `CONFIDENCE:` line from the candidate answer and apply the same grading procedure as in the direct evaluation. We separately extract the integer following the final `CONFIDENCE:` marker for the confidence analysis. Responses with missing, non-integer, or out-of-range confidence values are excluded. We compare accuracy between low- or moderate-confidence responses ($c \leq 3$) and high-confidence responses ($c \geq 4$).

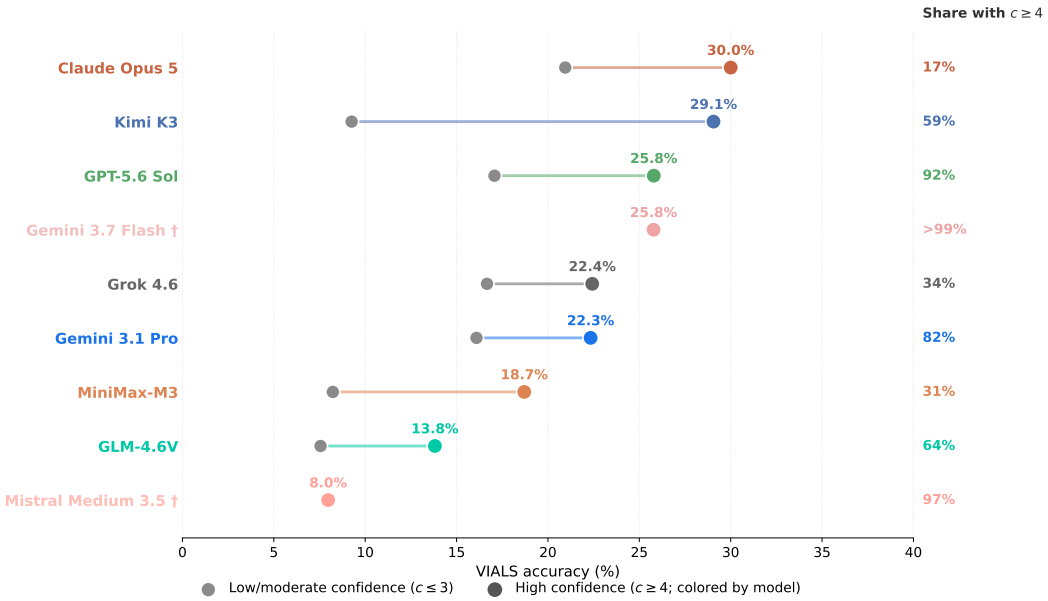


Figure A3: Accuracy of model responses amongst those with high vs. low self-reported confidence. Gray points show the accuracy of responses with low confidence ($c \leq 3/5$), while colored points show the accuracy of responses with high confidence ($c \geq 4/5$). The right column reports the share of model responses that were self-reported as high-confidence ($c \geq 4/5$). Models marked with † reported too few low-confidence responses for a reliable comparison.

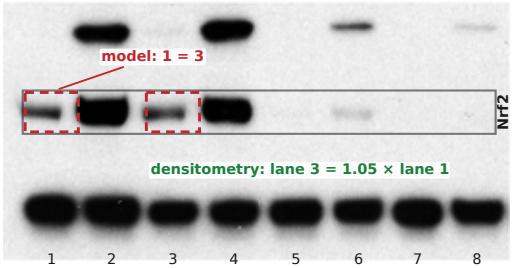
Of the nine models shown in Figure A3, seven produce enough lower-confidence responses for a meaningful comparison. For these models, high-confidence responses ($c \geq 4$) are about 6–20 percentage points more accurate than responses with $c \leq 3$. Gemini 3.7 Flash and Mistral Medium 3.5 are notable because both assign high confidence to more than 96% of their responses, leaving too few lower-confidence examples for a reliable comparison. Yet high confidence is still a poor indicator of correctness: even within the high-confidence bucket, accuracy reaches only 30% at best. Overall, these results show severe miscalibration in these VLMs' verbalized confidence estimates.

A.3 Additional failure examples

Here we display one representative example from each failure category to illustrate the range of model errors that occurred over the benchmark.

Incorrect value or feature selection

"Rank the conditions based on the expression of Nrf2 protein in descending order... Use the " = " sign in case of a tie"

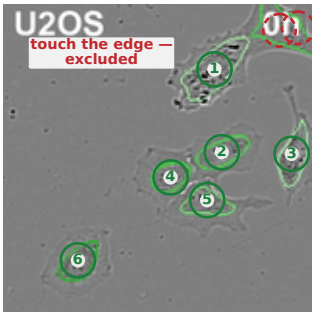


Model (4 of the 10 models): 2 = 4 > 1 = 3 > 6 > 5 = 7 = 8
Correct: 2 = 4 > 3 > 1 > 6 > 5 = 7 = 8

(a) **Tie called on a real difference.** Four of ten models rank lanes 1 and 3 as equal; densitometry puts lane 3 5 % above lane 1.

Visual quantification error

"In the 0h subpanel only, how many green outlined cell objects are fully contained within the image boundaries? Do not count green outlined objects that touch any edge"

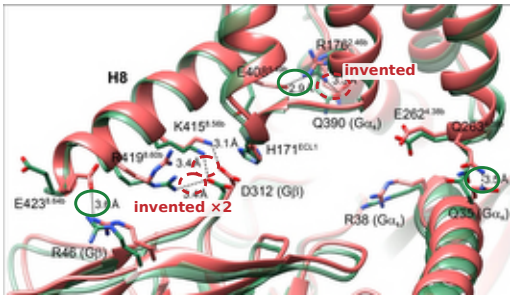


Model (gpt-5.6-sol): 5
Correct: 6 fully contained; 2 more touch the edge

(b) **Miscount under a stated rule.** The model applies the edge-exclusion rule correctly but enumerates five of the six objects.

Fabricated finding

"...which hydrogen bonds are uniquely formed in the bound structure and not in the apo structure?"

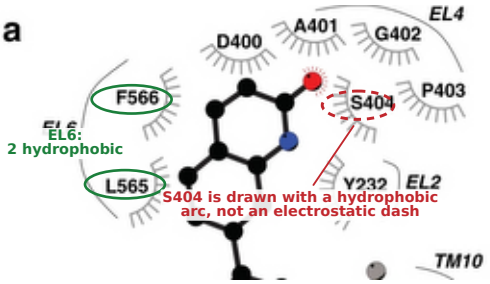


Model (gemini-3.1-pro): the 3 real bonds + 3 that do not exist
Correct: R46–E423, Q35–Q263, R176–E408

(c) **Plausible findings asserted.** The model reports all three real hydrogen bonds plus three that are not present.

Misapplied principle

"Report the ratio of the residues from the extracellular loop 6 that form a hydrophobic interaction... and the... loop 4 that form an electrostatic interaction"



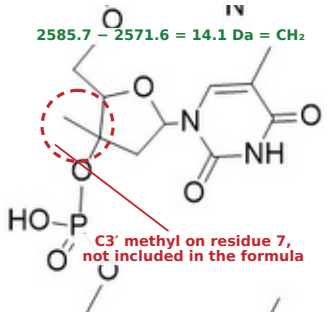
Model (mistral-medium-3-5): 2:1
Correct: 2:0 — EL4 has no electrostatic contact

(d) **Legend convention overridden.** S404 carries a hydrophobic arc; the model counts it as an electrostatic contact.

Figure A4: Representative VLM failure modes (1 of 2). Together with Figure A5, the seven panels provide one example from each error category in Table 3. Each panel shows the model's response in red and the interpretation supported by the artifact in green.

Quantitative reasoning error

"What is the molecular mass of the DNA sequence before the bioconjugation reaction? Provide the answer in Da and rounded to the first decimal place."

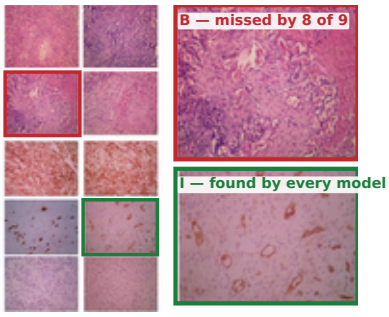


Model (claude-opus-5): 2571.6 Da
Correct: 2585.7 Da

(a) **One substituent short.** The arithmetic is internally correct; the C3′ methyl was never entered into the formula.

Overlooked evidence

"In the histopathology slides shown, which panel(s) show(s) clearly defined blood vessels?"

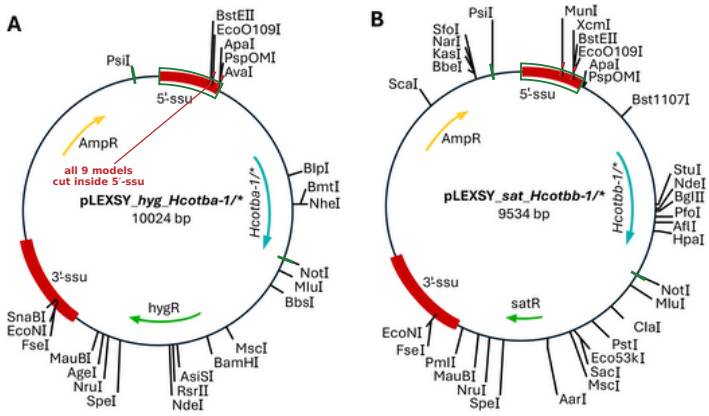


Model (5 of 9 answered exactly "I"): I
Correct: B, I

(b) **One of two instances found.** Every model finds the immunostained vessels in I; almost none of the models report B.

Representation misinterpretation

"Cleavage of which two restriction sites would result in a fragment that includes the full length 5′-ssu sequence for both of the plasmids... shortest possible fragment length"



Model (all 9 models): PsiI + a site inside 5′-ssu
Correct: NotI, PsiI — NotI is the nearest shared site clear of it

(c) **Tick read without the feature.** Every model picks a site whose tick falls inside the 5′-ssu block, truncating what the fragment was meant to contain.

Figure A5: Representative failure modes (2 of 2), continuing Figure A4. Together, the two figures provide one example from each of the seven error categories in Table 3. Each panel shows the model's response in red and the interpretation supported by the artifact in green.

B FURTHER COMPARISON WITH HUMANITY'S LAST EXAM

Here we additionally compare VIALS against the Biology/Medicine and Chemistry VQA subset of Humanity's Last Exam (HLE) along two dimensions: semantic coverage and task construction.

Semantic coverage. We embed all VIALS questions and the corresponding HLE questions using OpenAI's `text-embedding-3-small` model. Figure A6 shows projections of these embeddings to two dimensions via UMAP using cosine distance, highlighting stark semantic differences between the questions from these two benchmarks.

Task construction. We also qualitatively examine HLE questions in scientific domains that overlap with VIALS. The examples in Figure A7 illustrate differences in how the two benchmarks use visual evidence. Some HLE questions heavily focus on testing knowledge, or use the image primarily to identify an entity rather than as the source of the scientific evidence needed to answer a meaningful research question. The depicted HLE examples clearly do not stem from professional life sciences work.

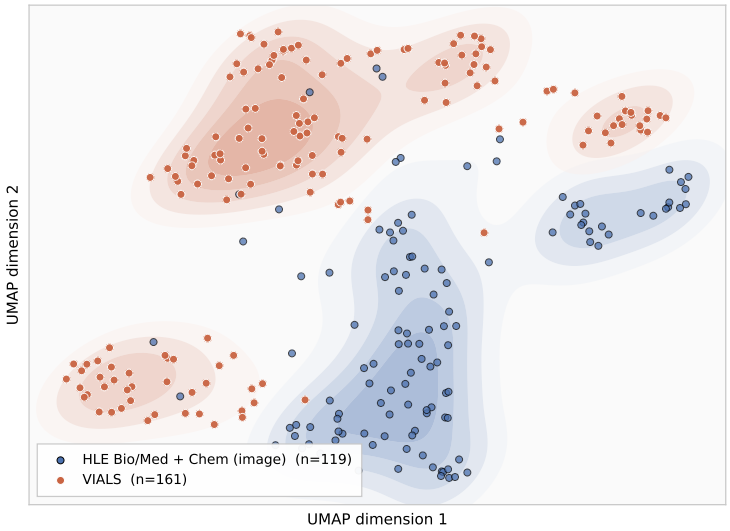


Figure A6: UMAP projection of embeddings for questions from VIALS and the Biology/Medicine and Chemistry VQA subset of HLE. Shaded regions indicate Gaussian kernel density estimate contours.



Q: "This image was taken in Acadia National Park. How many amino acids does the encoded XPH1 protein of this organism have?"
A: 110



Q: "An entomologist . . . spent June of 2024 on a collecting trip where they encountered the organism in the attached image. Based on the morphology of the observed specimen, what is the most likely collection locality?"
A: Luodong, Taiwan



Q: "X Y is pictured at an event on June 20, 2019. Another X Y has almost ceased to be used due to the toxicity of mercury salts. What is the X Y?"
A: Kucherov reaction

Figure A7: Examples of VQA tasks in HLE from scientific domains overlapping with VIALS.

C METHODOLOGICAL DETAILS

C.1 VLM EVALUATION

All evaluated models receive the same prompt. The input consists of one or more images together with the task question. The model is instructed to provide a brief reasoning process followed by a single final answer in the format below.

Task prompt for generating VLM responses

System prompt
You are an expert scientist. You will be shown a scientific image and a question. Study the image carefully, then respond with your reasoning and a single final answer.

User prompt

```
[image_1] ... [image_N]

{QUESTION}

Reason through the image and task, then commit to a single final
↪  answer. Keep the reasoning concise so your response ends with
↪  the ANSWER line. The answer must appear on its own line prefixed
↪  by 'ANSWER:' and contain only the final answer: no LaTeX
↪  brackets, no extra explanation, no units unless the question
↪  asks for them.

Format your response EXACTLY as:
REASONING: <your reasoning>
ANSWER: <final answer>
```

Answer extraction and grading. Model responses are graded using an LLM judge. We first extract the candidate answer as the substring following the final `ANSWER:` marker in the model's raw response. If no such marker is present, the candidate answer is treated as empty and the task is scored incorrect. Thus, compliance with the required answer format is part of the evaluation.

The judge receives the question, reference answer, and extracted candidate answer, and determines whether the candidate is semantically equivalent to the reference answer. The text-only grading prompt is shown below.

Grading prompt (text-only)

System prompt

You are grading a scientific answer. Given the question, the ground-truth final answer (GTFA), and a candidate answer, decide whether the candidate expresses the same intended answer as the GTFA in the context of the question. You do NOT see the image.

IMPORTANT: You are grading, not solving. Do NOT attempt to work out the correct answer to the question yourself and then compare the candidate to your own derivation. Treat the GTFA as the sole source of truth. Your only decision is whether the candidate expresses the same answer as the GTFA.

Apply the standard of a competent domain scientist. ACCEPT variations that an expert would recognize as immaterial to the answer:

- **Formatting:** punctuation, whitespace, dash / hyphen style, thousands separators, and equivalent scientific notation for the same underlying value.
- **Capitalization:** when the identity is otherwise unambiguous.
- **Units:** omitted when unambiguous from the question or GTFA, but only when the numeric value itself matches.
- **Lead bounds:** when an inclusive `LeadBoundLower`/`LeadBoundUpper` range is provided, a numeric candidate inside that range is equivalent even if it differs from the exact GTFA value.
- **Order and duplicates:** irrelevant when the answer is a set or unordered list.
- **Terminology:** synonymous scientific terms of the same specificity.

Numeric answers are STRICT. Do NOT invent tolerances, relative-error allowances, or "same order of magnitude" acceptance.

- The candidate must express the same numeric value as GTFA, after applying the formatting and unit rules above, unless LeadBounds apply.
- If the question asks for a stated reporting precision, any difference at that precision is NOT equivalent.
- Exact counts of discrete objects or events must match exactly; off-by-one is NEVER equivalent.
- Different coefficients at the same power of ten are NOT equivalent.

REJECT when the candidate materially disagrees with GTFA on any of:

- Missing or extra items in a list.
- The identity, count, direction, mechanism, or sign of a biological entity.
- Any numeric mismatch under the strict numeric rules above, unless inside LeadBounds.
- A broader or narrower term when the question requires a specific level of detail.
- Extra conflicting numeric values or alternate counts that disagree with GTFA.

Refusals, meta-answers ("the provided image," "cannot determine"), and empty responses are ALWAYS not equivalent, regardless of GTFA.

If evaluating equivalence would require seeing the image—for example, GTFA uses one labeling system (condition names, colour labels) and the candidate uses another (lane numbers, positional indices), and the mapping is not given in the question text—return `equivalent=false` with reasoning prefixed by `UNVERIFIABLE:`. Downstream tooling treats these cases as requiring a multimodal regrade or human review, rather than as confirmed model errors.

Reply using the provided JSON schema. Keep the reasoning to at most three sentences and name the specific principle applied.

User prompt template

Question:
{`QUESTION`}

Ground-truth final answer:
{`GTFA`}

Lead-assigned acceptable numeric range (inclusive):
lower={`LOWER`}, upper={`UPPER`}.
If the candidate is a numeric answer inside this range, mark it equivalent to the GTFA.
(this part of the prompt is omitted when the task has no lead bounds)

Candidate final answer:
{`MODEL_ANSWER`}

Are the candidate and ground-truth final answers equivalent?

Multimodal fallback. If the text-only judge returns `UNVERIFIABLE:`, we invoke the same judge model again with a multimodal request that has the task image attached. This fallback is used only when determining equivalence requires resolving information visible in the image, such as lane numbers, positional indices, or colour labels; the judge is not instructed to re-solve the task. In practice, the multimodal fallback is rarely needed. Across the main evaluation (10 models $\times$ 161 tasks $\times$ 3 rollouts), well under 1% of judge calls escalate to the multimodal judge. The typical trigger is a task where the reference answer names a sample condition while most models report the corresponding lane number, so determining equivalence requires reading the lane-to-sample mapping shown in the image.

The multimodal system prompt is identical to the text-only prompt except that the judge is informed that the task image is available and may be used only to resolve labeling ambiguities between the candidate answer and the ground-truth reference answer. The instruction to return `UNVERIFIABLE:` is removed, and the judge is instead instructed to decide equivalence directly using the image where necessary. The multimodal judge prompt is shown below.

Grading prompt (multimodal fallback)

User message

[task image attached]

Question:
{QUESTION}

Ground-truth final answer:
{GTFA}

Lead-assigned acceptable numeric range (inclusive):
lower={LOWER}, upper={UPPER}.
If the candidate is a numeric answer inside this range, mark it equivalent to the GTFA.
(omitted when the task has no lead bounds)

Candidate final answer:
{MODEL_ANSWER}

Using the image only to resolve labelling ambiguities, are the candidate and ground-truth final answers equivalent?

Evaluation configuration. For each (task, model) pair, we generate three rollouts ($K=3$) and score each independently. We sample all models at temperature 1.0 with a $32{,}768$-token completion limit. For each VLM provider, we use its default inference configuration. This includes OpenAI's default reasoning effort for GPT-5.6, adaptive thinking for Claude Opus and the Gemini models, native reasoning modes for Grok, Kimi K3, MiniMax-M3, and GLM-4.6V, and standard inference for Muse Spark and Mistral Medium.

Model requests that fail, are refused, or return an empty response, are retried three additional times. If still unsuccessful, these responses are counted as incorrect rather than omitted from the evaluation. Out of 483 total responses (3 rollouts over 161 tasks): Muse Spark 1.2 has 13 such responses (2.7%), Kimi K3 has 6 (1.2%), Grok 4.6 has 2 (0.4%), Mistral Medium 3.5 has 1 (0.2%), and the other models have none.

C.2 Failure classification methodology

For the failure analysis in Section 5.2, we assign each (task, model) pair with at least one incorrect rollout a single primary failure mode using an LLM classifier. Each pair contributes a single observation to our failure mode analysis, regardless of the number of incorrect rollouts. We use Claude Sonnet 4.5 as the LLM classifier and fallback to GPT-4o in rare cases where the primary classifier refuses or errors.

The failure mode classifier does not receive the task image. Its input consists of the question, the ground-truth final answer (GTFA), and the failing attempt's final answer and reasoning trace. The classifier is instructed to identify the primary reason the response is incorrect relative to the GTFA, rather than to re-solve the task. Each (task, model) pair is classified once using its first incorrect attempt, so multiple incorrect rollouts for the same pair contribute a single observation to the failure-mode distribution.

Failure-mode classification prompt

You classify why an AI model got a life-sciences image-interpretation task wrong.

You do NOT see the image. Base your classification on:

1. the question,
2. the ground-truth final answer (GTFA), and
3. the failing attempt's final answer and reasoning trace.

IMPORTANT: You are classifying, not solving. Do NOT attempt to work out the correct answer to the question yourself. Treat the ground truth (GTFA) as the sole source of truth and assume the LLM judge's per-attempt correct/wrong decisions are final. Your only job is to name why the failing attempt diverged from the GTFA.
Focus on the substantive scientific or artifact-interpretation error rather than generic labels such as raw perception, OCR, or "couldn't read the image" failure. If the model's reasoning identifies the right image features but reaches the wrong scientific conclusion, that's still a knowledge failure.

Assign exactly ONE primary failure mode:

- ASSAY_COUNTING_ERROR (Visual quantification error): failed to quantify a biologically-meaningful countable – cells, colonies, PCR bands, atoms in a scaffold, ligand contacts, residues in an interface. A domain expert knows WHICH objects count under the assay's rules (e.g. count only colonies above a size threshold; exclude edge artifacts). The failure is not knowing that scientific rule of counting.

- `MISINTERPRETED_DIAGNOSTIC_VALUE` (Incorrect value or feature selection): picked the wrong reading on a scale that a biologist would call diagnostic – the actionable band MW, the axis threshold, the gated %, the well/lane that carries the readout. The failure is not knowing which measurement is scientifically load-bearing.
- `DIAGRAM_CONVENTION_ERROR` (Representation misinterpretation): misread a domain-specific diagrammatic convention – phylogenetic branching order or branch-length semantics, pedigree inheritance arrows, plasmid feature order along the sequence axis, gel-lane layout. The failure is not knowing what the diagram encodes, not what shapes are on it.
- `DOMAIN_QUANTITATIVE_ERROR` (Quantitative reasoning error): got the raw readings roughly right but botched the quantitative-biology reasoning on top – dilution factor, MOI, Kd, unit conversion, bounds on a rate. Domain math, not generic arithmetic.
- `FABRICATED_FINDING` (Fabricated finding): asserted a scientific finding not supported by the figure – invented a band, residue, colony, cell state. The failure is generating a domain-plausible fiction instead of admitting uncertainty.
- `MISSED_LOAD_BEARING_EVIDENCE` (Overlooked evidence): missed a scientifically load-bearing feature that WAS present – an important band, residue contact, cell subset, control lane. The failure is not knowing what a competent biologist wouldn't overlook.
- `MISAPPLIED_DOMAIN_RULE` (Misapplied principle): applied the wrong scientific principle – inheritance mode, enzyme kinetics, restriction-digest logic, gating hierarchy, phylogenetic parsimony.

Always name a substantive cause (do NOT answer with a circular label like "chose the wrong option"): pick the specific piece of scientific knowledge or reasoning that broke.
Your `reason` field should be 2-3 complete sentences (roughly 30-90 words). It must be long enough that a domain expert can sanity-check your classification without opening the log: name the specific scientific concept, quote or paraphrase the model's wrong step, and briefly say what a competent expert would have done instead. Do not truncate mid-sentence. Do not exceed 4 sentences.
Reply with a JSON object matching the schema.

C.3 TOOL-ASSISTED AGENT EVALUATION

Here we detail the methodology behind the alternative agentic tool-assisted setting evaluated in Section 5.3. All discussions outside of this section pertain to the primary way we recommend running VIALS, where tasks are completed via single-turn inference calls to reasoning VLMs.

Execution environment. The tool-assisted evaluation uses the same VIALS tasks as the direct evaluation, but gives each model access to a sandboxed Linux environment with code execution. Task images are placed in the `/task/` directory, where agents may inspect, transform, and analyze them using the shell and common scientific Python packages, including `numpy`, `scipy`, `pillow`, `opencv-python-headless`, `scikit-image`, `matplotlib`, `pandas`, `pytesseract`, `imageio`, `networkx`, and `sympy`. Internet access and external databases are disabled.

Inference cost. Table A1 reports mean USD cost per attempt under the direct and tool-assisted settings for the same model–harness configurations as Table 4. Direct costs are calculated from logged input and output token usage using the corresponding provider API prices, while tool-assisted costs are taken from the per-trial costs reported by the Harbor agent runtime. Mean tool-assisted cost per attempt ranges from 0.011 for GLM-4.6V to 1.047 for Claude Opus 5 with Claude Code. For most models, tool-assisted evaluation costs roughly 3–16$\times$ as much per attempt as direct evaluation, with substantially larger ratios for models with very low direct inference costs (29$\times$ for Gemini 3.1 Pro and 56$\times$ for Gemini 3.7 Flash).

Model	Harness	$\textbf{Cost}_{\text{D}}$	$\textbf{Cost}_{\text{T}}$	Cost×
GPT-5.6 Sol	Codex	\$0.037	\$0.271	7×
GPT-5.6 Sol	OpenCode	\$0.037	\$0.155	4×
Claude Opus 5	Claude Code	\$0.065	\$1.047	16×
Claude Opus 5	OpenCode	\$0.065	\$0.917	14×
Gemini 3.1 Pro	OpenCode	\$0.024	\$0.716	29×
Gemini 3.7 Flash	OpenCode	\$0.007	\$0.374	56×
Grok 4.6	OpenCode	\$0.049	\$0.337	7×
Kimi K3	OpenCode	\$0.239	\$0.676	3×
GLM-4.6V	OpenCode	\$0.003	\$0.011	3×

Table A1: Per-attempt inference cost under direct and tool-assisted evaluation. $\textbf{Cost}_{\text{D}}$ is calculated from logged direct-evaluation token usage using the corresponding provider API prices. Cost_{T} is the mean per-attempt cost reported by the Harbor agent runtime. **Cost**× is the ratio of tool-assisted to direct cost.

Agent prompt. Each agent receives the following instruction, with a task's specific question substituted for the {`QUESTION`} placeholder.

Task prompt for agent

You are an expert scientist working inside a Linux container.

The task image is available at `/task/image.png`. You have access to a shell and may use any software available in the container to inspect and analyze the provided files. The following Python libraries are already installed:

- `numpy`, `scipy`, `sympy`
- `pillow`, `imageio`, `opencv-python-headless`, `scikit-image`
- `matplotlib`, `pandas`, `networkx`
- `pytesseract` (backed by the `tesseract-ocr` CLI)

Do not use the internet, external databases, or information outside the files provided for this task. You may write scratch files under `/task/` while you work. You may also inspect the original image or cropped regions as additional vision inputs.

Answer the question below. When finished, write your response to `/task/answer.txt`, with the final line in the following format:

```
ANSWER: <your final answer>
```

Question

{`QUESTION`}

The agent writes its final response to `/task/answer.txt`. The answer following the final `ANSWER:` marker is scored using the same semantic grading procedure as the direct evaluation (Appendix C.1).

C.4 Survey questions for expert professional relevance assessment

Here are the survey questions we asked experts to answer in the professional relevance assessment from Figure 2 and Section 3.6:

1. *Does this image look more like an educational figure, or like something you'd encounter in real industry / professional life sciences work?*
   Choose one of: {Real industry / professional work; Mixed / hard to tell; Textbook / Educational}
2. *How often do you come across this type of image in your work?*
   Choose one of: {Daily; Weekly; Occasionally; Rarely; Never}
3. *Interpreting this type of image is a task that has real commercial / economic value in life sciences — it's used in professional settings like drug discovery, biotech or pharma R&D, diagnostics, or QA, where getting it right matters.*
   Choose one of: {Strongly agree; Agree; Neutral; Disagree; Strongly disagree}