

Science Edge Evaluation: SEE the Missing Step Toward Real Scientific Discovery

Taolin Han^{1,3*}, Yuchen Zhang^{1,4*}, Jinghang Wang^{1*}, Yun Wu¹, Wai Yuet Chiu¹,
Zhaohai Li², Yifei Zhang^{1,5}, Jinxin Wang^{1,4}, Yuhao Zhou^{1,6}, Chen Zhao^{1,4}, Jiajia Li¹,
Jiaxin Li¹, Qile Jin¹, Kewei Sun¹, Shuang Wu¹, Weiqi Zhai¹, Renquan Lv^{1,6},
Junchao Li¹, Ruodan Chen¹, Qingteng Chen¹, Zhibo Yang², Hu Wei¹, Lin Qu¹,
Shuai Bai^{2,†}, Bing Zhao^{1,†}

¹Alibaba Group, ²Qwen Team, Alibaba Group, ³University of Chinese Academy of Sciences,
⁴Tsinghua University, ⁵University of Alberta, ⁶Zhejiang University

*Equal Contribution, †Correspondence

wangjinghang.wjh@alibaba-inc.com

 Code

Abstract

Large language models (LLMs) are increasingly involved in scientific discovery, yet it remains unclear whether they can support complex real laboratory science. Here we introduce **Science Edge Evaluation (SEE)**, a multimodal benchmark of expert-curated questions grounded in peer-reviewed literature and experimental practice in chemistry, biology, and materials science. Evaluation of 19 multimodal large language models (MLLMs) shows that even the best-performing model reaches only 48.7% accuracy. Moreover, general-purpose models outperform science-specialized models on average. In the visual-agent evaluation, the use of tools increases the best accuracy to 52.7%. Tool use can expand the information available to models, but more information does not necessarily lead to reliable scientific reasoning. The key challenge is whether models can manage tool-derived information within the boundaries of the original experimental evidence. Together, these findings reveal that current MLLMs still cannot reliably make justified and evidence-bounded inferences from experimental results, which is an essential capability in real scientific discovery. Bridging this gap requires MLLMs to transition from explaining established scientific concepts to deriving novel and evidence-based insights from experimental data.

1 Introduction

Large language models (LLMs) are transforming the scientific discovery process by accelerating some of its main stages, including literature screening, computational simulation, code synthesis, and, more recently, autonomous experimentation. As these models become more capable and widely adopted, they are beginning to shape the way scientific knowledge is produced, evaluated, and applied across disciplines. This emerging role positions LLMs not merely as tools for information retrieval or text generation, but as active contributors to the broader research workflow.

Current evaluation paradigms often prioritize breadth over depth and rely heavily on sanitized, exam-style problems that reward pattern matching and memorization. Such settings fail to capture the dynamic, multi-stage, and evidence-driven nature of real scientific discovery. As a result, they provide only a limited picture of whether a model can support realistic research workflows. Benchmarks grounded in authentic scientific tasks are therefore essential for measuring true analytical ability while reducing the risk of contamination from standard educational data.

Multimodality is equally indispensable. Real scientific reasoning rarely depends on text alone. Instead, it emerges from the joint interpretation of figures, spectra, microscopy images, tables, numerical measurements, and written context. Without evaluating this convergence of visual, numerical, and textual evidence, benchmark scores remain detached from the actual demands of laboratory science. A scientifically meaningful benchmark must therefore assess whether models can reason from heterogeneous evidence rather than from linguistic priors alone.

Scientific benchmarking must also evolve alongside the development of natural science itself. Historically, subjects such as physics, chemistry, biology, and medicine were often treated as separate domains. However, in modern research, critical problems increasingly arise at their intersections, where concepts, methods, and data types are deeply entangled. Thus, evaluating models only within isolated subjects misses a core requirement of real scientific reasoning, which is the ability to integrate knowledge across fields. Interdisciplinary questions are also critical for exposing brittle reasoning, hallucination triggers, and transfer failures at the boundaries of a model’s knowledge.

Taken together, these considerations reveal a central gap in current scientific evaluation. Existing benchmarks often test whether models know scientific facts or solve simplified problems, whereas real laboratory science requires models to reason from incomplete, heterogeneous, and cross-disciplinary experimental evidence. This distinction is increasingly important as LLMs enter scientific workflows and agentic research systems. Scientific evaluation should therefore move beyond knowledge recall toward evidence-grounded, multimodal, interdisciplinary, and uncertainty-aware reasoning.

To address this gap, we introduce **Science Edge Evaluation (SEE)**, a multimodal benchmark built from real scientific tasks in chemistry, biology, and materials science. Our evaluation shows that current MLLMs remain unreliable in realistic experimental settings, and that their limitations cannot be explained by knowledge access alone. Instead, **SEE** exposes a fundamental limitation in

evidence-based scientific reasoning. Our tool-augmented visual-agent analysis further shows that tool access can expand evidence acquisition, but does not eliminate the need for reliable evidence management across the interaction trajectory. Current models struggle to "SEE" multimodal observations across disciplines while maintaining rigorous adherence to the available evidence. In this sense, **SEE** evaluates not only whether models know scientific content, but whether they can derive justified, evidence-bounded insights from multimodal experimental data.

2 Related Work

General Academic Benchmarks. Academic benchmarks are essential for evaluating LLM and MLLM capabilities. General academic benchmarks such as MMLU [1], MMLU-Pro [2], and GPQA [3] evaluate models across broad academic domains with different levels of difficulty. The tested capabilities include general academic reasoning, scientific question answering, mathematical reasoning, and code generation. Region-specific extensions such as CMMLU [4] further broaden the evaluation surface. To challenge models at the frontier of expert-level reasoning, benchmarks such as HLE [5] have been specifically designed to incorporate closed-ended questions of exceptional difficulty.

Multimodal Benchmarks. Multimodal benchmarks such as ScienceQA [6] and MMMU [7] represent meaningful advances by extending evaluations beyond text-only paradigms through the incorporation of visual inputs covering scientific and academic subjects, with MMMU-Pro [8] further removing text-solvable questions to enforce true multimodal reasoning. Complementary efforts including MM-Vet [9], MathVista [10], and SciFIBench [11] target integrated multimodal capabilities, mathematical reasoning in visual contexts, and scientific figure interpretation, respectively. Olympiad- and contest-grade benchmarks such as OlympiadBench [12], the Chinese-oriented MMSciBench [13], the multilingual MME-SCI [14], and USNCO [15] built from chemistry-olympiad exams push difficulty further with bilingual or multilingual multimodal scientific problems. While current benchmarks are rooted in general academic and exam-style contexts, they fail to capture the experimental data and workflow-driven reasoning fundamental to authentic scientific inquiry.

Scientific Benchmarks. Some recent benchmarks are designed to evaluate models in more specialized scientific settings, including chemistry, biology, and materials science. These efforts focus on domain-specific reasoning rather than broad academic knowledge, exemplified by ChemBench [16], SciBench [17], SciEval [18], and LAB-Bench [19]. A few of them incorporate multimodal inputs such as figures, molecular structures, spectra, and tables, together with experimental results. For example, MaCBench [20] is a multimodal benchmark for chemistry and materials science that is closer to real research scenarios, while Matbench [21] targets materials property prediction. More recent efforts further probe modality-specific limitations, for example, ChemVTS-Bench [22] disentangles visual, textual, and symbolic chemical reasoning. Existing scientific benchmarks lack the disciplinary breadth, experimental realism, and diagnostic depth required to evaluate MLLMs against the complex, interdisciplinary challenges of authentic scientific research.

Scientific LLMs and Agents. Scientific LLMs have been developed through continued pre-training, instruction tuning, and domain adaptation on scientific corpora. Representative examples include general scientific and biomedical models such as Galactica [23], BioGPT [24],

BioMedLM [25], GatorTronGPT [26], Med-PaLM [27], and Meditron [28]; chemistry-, molecule-, and materials-oriented models or instruction resources such as ChemLLM [29], ChemDFM [30], Mol-Instructions [31], LLaSMol [32], HoneyBee [33], and MatterChat [34]; and multimodal biomedical or scientific models such as LLaVA-Med [35], Med-PaLM M [36], BioMedGPT [37], S1-VL [38], Intern-S2 Preview [39], Intern-S2 Preview-397B [40], and Intern-S1 Pro [41]. Recent scientific agents further extend LLMs from static question answering to tool-augmented and workflow-level scientific assistance, including ChemCrow [42], Coscientist [43], The AI Scientist [44], Agent Laboratory [45], and Co-Scientist [46]. While these efforts demonstrate the growing role of LLMs in scientific workflows, existing evaluations often focus on knowledge recall, domain-specific task performance, tool-use success, or final-answer accuracy. In contrast, **SEE** evaluates whether MLLMs can ground conclusions in multimodal experimental evidence, integrate concepts across disciplines, and reason within the boundaries of available evidence.

3 Method

We organize **SEE** through a staged construction and validation workflow, from real-world data sourcing and expert question design to AI-assisted quality control, expert review, and final benchmark acceptance (Figure 1).

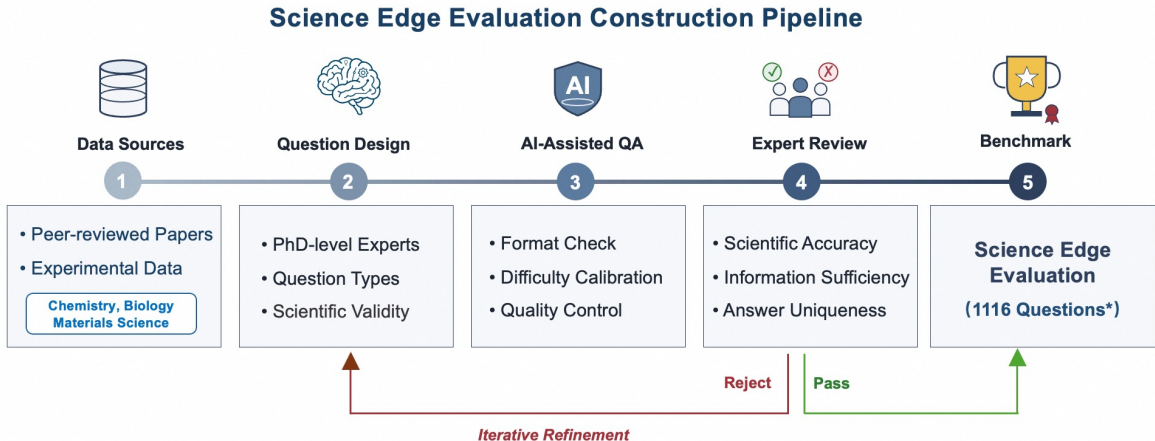


Figure 1. Benchmark construction pipeline of **SEE**. Questions are derived from peer-reviewed literature and expert experimental scenarios, standardized and calibrated with AI-assisted checks, reviewed by domain experts, and accepted only after passing scientific accuracy, information sufficiency, answer uniqueness, and evaluation-readiness checks. Of the 1,116 questions, 1,049 are publicly released; the remaining 67 are withheld because they involve unpublished experimental data from contributing experts.

3.1 Data Collection

SEE is a collaborative effort. The questions are contributed by active frontline experts who hold master’s or PhD degrees or are PhD candidates in chemistry, biology, materials science, or related interdisciplinary areas.

Sources. The questions are drawn from real scientific research settings. Sources include figures and data from peer-reviewed papers, recent research literature, and first-hand experimental

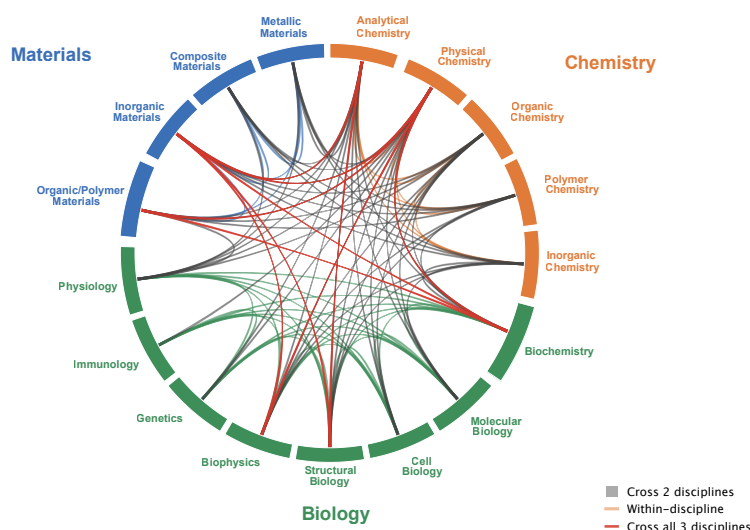


Figure 2. SEE consists of 1,116 questions spanning 3 disciplines and 17 reported sub-fields. Interdisciplinary questions are represented by lines connecting two or more sub-fields. Line thickness indicates the number of co-occurring questions across reported sub-fields. Gray lines represent cross-discipline overlap (spanning two disciplines).

data contributed by experts. Visual evidence covers common experimental data structures in laboratory research, including bioactivity measurements, Cryo-EM structures, Western blot and gel electrophoresis images, spectra such as infrared spectroscopy, nuclear magnetic resonance (NMR) and mass spectrometry (MS), microscopy images such as SEM, TEM and AFM, X-ray diffraction patterns, and thermal analysis curves.

Style. The questions in SEE include choice-based, short-answer, numerical, measurement, and image-processing tasks. Some questions contain explicit option lists, while others do not use a separate option field and are evaluated through canonical short answers or numerical answers. The dataset also records task-type metadata, including reasoning, measurement, and image-processing categories. Each entry contains text and associated visual files. These associated images may include both visual evidence presented with the question and images used for expert verification. As a result, the reported image counts are not the exact number of figures displayed in the question itself. Each entry is reported with a standard answer to support automatic evaluation and expert verification.

Labeling. Each question in SEE is assigned discipline labels spanning biological sciences, chemistry, and materials science. The reported analysis uses a final taxonomy of 17 fine-grained labels covering all 1,116 questions. This labeling scheme captures cross-disciplinary questions clearly. Detailed labeling methods and label distributions are included in the Supplementary Information.

3.2 Data Processing

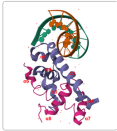
A multi-stage review process is used to ensure data quality, difficulty, and reliability. After collection and label assignment, the questions that do not meet our criteria are filtered out.

Standardization. The raw questions collected are standardized in terms of wording, terminology, answer structures, option numbering, numerical precision, tolerance records, and image organization. The images are normalized in format and resolution.

General Quality Control. After standardization, questions are checked for adherence to the required format, clear and objective answers, sufficient visual and textual evidence, and correct labels. We remove semantic duplicates and questions that do not fulfill the requirements. Each question is then cross-checked by at least two experts in the relevant domain.

Question Examples

(a) Biology
Molecular Biology | Structural Biology | Biochemistry
Western Blot | Protein-DNA
Question:

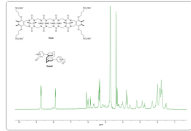


This is a schematic diagram of the local interaction between a transcription factor and DNA. Which α -helix in the figure is primarily responsible for nucleic acid recognition and interaction?

Answer: a2

Source: Experimental Data

(b) Chemistry
Analytical Chemistry | Organic Chemistry
¹H NMR
Question:



The ¹H NMR spectrum shown in the figure is obtained from a host-guest titration process. The specific experimental procedure involves gradually adding a concentrated solution of the host to a D₂O solution containing the guest and performing NMR measurements. Based on the provided host-guest structures, what is the guest-to-host ratio at this moment?

A. 1:0.3
B. 1:0.5
C. 1:0.9
D. 1:1.1

Answer: C

Source: Experimental Data

(c) Chemistry x Materials
Metallic Materials | Analytical Chemistry
Synchrotron XRD
Question:

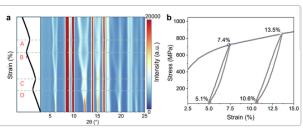


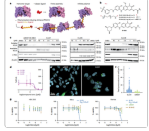
Figure (a) shows the in-situ SXRD diffraction peak evolution of a material during two loading-unloading-reloading (LUR) cycles, and Figure (b) shows the corresponding stress-strain curve. Four characteristic state points, A, B, C, and D (corresponding to different strain levels), are marked in Figure (a). Determine the strain values corresponding to A, B, C, and D in Figure (a).

A. 10.6%, 13.5%, 5.1%, 7.4%
B. 13.5%, 10.6%, 5.1%, 7.4%
C. 7.4%, 5.1%, 10.6%, 13.5%
D. 5.1%, 7.4%, 13.5%, 10.6%

Answer: A

Source: Experimental Data

(d) Biology
Molecular Biology | Cell Biology | Biochemistry
Western Blot | IF
Question:



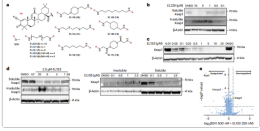
Based on the data in the figure below, which of the following conclusions is/are incorrect?

A. Bi-3802 only partially degrades BCL6, and the remaining intracellular BCL6 can still maintain its transcriptional function.
B. The downregulation of α -tubulin is caused by non-specific protein aggregation or cellular stress induced by EL221.
C. EL221 exerts killing effects only in BCL6-positive Mino and Ramos cells and is inactive in BCL6-negative HEK293 cells, indicating that BCL6 expression is a prerequisite for its efficacy.
D. The minimum effective concentration of EL221 for inducing soluble BCL6 downregulation is 5 μ M, and Figure 9 shows its cytotoxicity IC_{50} is approximately 5 μ M. These two values match perfectly, and the concentration threshold is a prerequisite for the effect.
E. α -tubulin can induce BCL6 to form a large number of puncta aggregates, whereas BCL6 is diffusely distributed after EL221 treatment.

Answer: BDE

Source: <https://doi.org/10.1038/s41598-025-02141-0>

(e) Biology x Chemistry
Organic Chemistry | Molecular Biology | Cell Biology | Biochemistry
Western Blot | IF
Question:



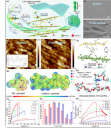
Based on the data in the figure below, which of the following conclusions is/are incorrect?

A. EL133 connects two BDX molecules via a linker consisting of 8 ethylene glycol units, whereas EL229 connects two BDX molecules via a linker consisting of 5 ethylene glycol units.
B. EL133 downregulates soluble Keap1, where low concentrations (0.01–0.1 μ M) show no obvious effect, and once the threshold (≥ 0.25 μ M) is reached, the effect is significantly enhanced as the concentration increases.
C. EL133 achieves significant downregulation of soluble Keap1 at 250 nM, which is far superior to the activity of EL229.
D. EL133 begins to show substantial accumulation of insoluble Keap1 at 6 h of treatment, indicating that EL133-induced Keap1 polymerization is a time-dependent gradual process rather than a transient effect.
E. Comparing the DMSO control group with the experimental group treated with 250 nM EL133 for 20 h, only Keap1 protein in the whole proteome shows significant downregulation.

Answer: ADE

Source: <https://doi.org/10.1038/s41598-025-02141-0>

(f) Chemistry x Materials
Physical Chemistry | Polymer Chemistry | Polymer Materials | Composite Materials
SEM | AFM | DFT
Question:



A research team developed a cellulose liquid crystal film (CLCF) based on ionic condensed-state engineering, as shown in the figure. Which of the following inferences regarding the structural assembly mechanism and mechanical properties of this film is/are correct?

A. Based on the electrostatic potential energy distribution map in panel (g) and the DFT calculation data in panel (h), there is a pronounced electrostatic attraction between the positively charged center of the imidazolium ring on the PIL chain (red region) and the negatively charged oxygen atoms on the CNC (blue region). This stable noncovalent interaction serves as one of the core driving forces for membrane self-assembly.
B. The SEM and AFM images in panels (b) through (e) demonstrate that, despite the incorporation of PIL and IL molecules, the CNC retains its characteristic chiral helical arrangement, indicating that the spatial structure of the membrane does not become disordered upon the addition of these components.
C. Comparing panels (i) and (j), it can be observed that as the PIL content increases, the mechanical behavior of the film transitions from "rigid and brittle" to "tough and soft"; when the ratio reaches 1:2, the tensile strength drops to its minimum because the polymer-dominated system leads to weakened effective stress transfer.
D. The curves in panel (k) clearly show that tensile strength decreases monotonically with increasing IL content. The IL primarily functions as a plasticizer and provides ionic transport channels.

Answer: ABD

Source: <https://doi.org/10.1038/s41598-025-02141-0>

Figure 3. Representative multimodal scientific questions in SEE, showing the associated figure, question stem, answer options, correct answer, discipline/technique tags, and data source.

Expert Quality Control. After AI-assisted checks, experts review the questions in terms of scientific accuracy. Questions with problems are returned to their contributors for revision. The revised questions then undergo the entire standardization and quality control process before inclusion.

Image-Ablation Checking. Finally, questions with images are checked to verify whether images are indispensable for solving the task. Questions that can be answered from text alone are removed or revised. This verification is consistent with previous multimodal benchmark practices [7, 11].

Question Evaluation. The evaluation of questions follows the standard scoring protocols established in large-scale evaluation frameworks [47]. For multiple choice questions, MLLM answers must *exactly* match the ground-truth answers (including single- and multiple-response questions). Numerical fill-in-the-blank questions are evaluated with tolerance level provided by experts when applicable. Non-numerical fill-in-the-blank questions are evaluated against standardized canonical answers.

Representative examples of the questions collected are shown in Figure 3, illustrating within-discipline and cross-disciplinary multimodal scientific questions.

3.3 Model Evaluation Environment

We evaluate 19 representative MLLMs with **SEE**, including 15 general-purpose models and four science-specialized models. The general-purpose models include Gemini 3.1 Pro [48], GPT-5.6-Sol (Max) [49], Claude Opus 5 (Max) [50], Qwen3.8-Max [51], GPT-5.5 (xhigh) [52], Kimi K3 [53], Gemini 3.5 Flash [54], Claude Opus 4.8 (Max) [55], Seed2.1 Pro [56], Qwen3.7-Plus [57], Seed2.0 Pro [58], MiniMax-M3 [59], Kimi K2.6 [60], GLM-5V-Turbo [61], and MiMo-V2.5 [62]. The science-specialized models include S1-VL [38], Intern-S2 Preview [39], Intern-S2 Preview-397B [40], and Intern-S1 Pro [41]. Only four science-specialized models are tested because few models in this category possess the multimodal capabilities required for our evaluation.

Unless otherwise specified, models are evaluated with default inference configurations. The only non-default settings are the fixed extended-thinking configurations used for GPT-5.5 (xhigh) and Claude Opus 4.8 (Max), as detailed in Supplementary Information B.1. For image-containing questions, we apply the preprocessing required by each model interface, including format conversion, resizing, and input organization. We introduce no manual prompt correction, per-question parameter tuning, or result filtering. Claude Opus 4.8 (Max) and Claude Opus 5 (Max) have safety-filtered no-response cases on biochemistry questions; under the strict binary scoring protocol, these instances are counted as incorrect.

Model outputs are evaluated with a strict binary LLM-as-a-judge pipeline [63]. Each response is extracted and marked as correct or incorrect using Gemini 3.1 Pro as the primary judge model for downstream accuracy computation and breakdown analysis. We also test GPT-5.5 as an alternative judge model and obtain similar results, indicating that the evaluation is robust to the choice of judge model (Cohen’s $\kappa \geq 0.99$; Supplementary Information B.4). Refusals, non-answers, and responses from which no final answer can be identified are counted as incorrect. Evaluation and judging prompts are detailed in Supplementary Information B.2, and subject-label coverage and cross-discipline composition of **SEE** are reported in Supplementary Information B.9.

4 Results and Discussion

4.1 Overall Performance on SEE

Following the evaluation protocol described in Section 3.3, we first report the overall accuracy of 19 representative MLLMs on **SEE**. All evaluated models achieve low accuracy on **SEE** in the standard evaluation setting (Figure 4). The best model, GPT-5.6-Sol (Max), reaches 48.7% accuracy, and no model exceeds 50%. Across the 19 models, accuracy ranges from 15.9% to 48.7%.

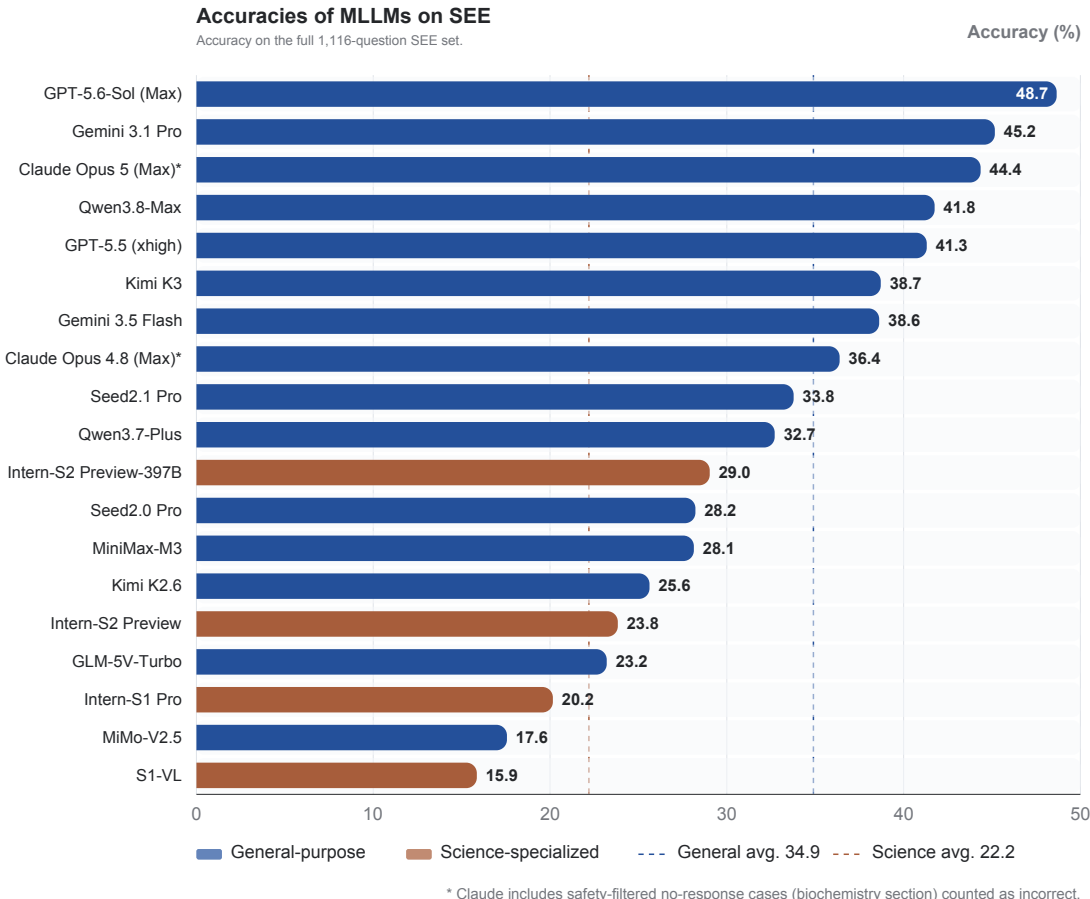


Figure 4. Accuracies of different models on **SEE**. General-purpose models are shown as blue bars, while science-specialized models are shown in brown. The dashed vertical lines indicate average accuracies in the two categories. Claude Opus 4.8 (Max) and Claude Opus 5 (Max) sometimes give no response to questions involving viral biology and pathogen structural characterization due to safety reasons (marked with *). These answers are counted as incorrect.

The result at the discipline level shows similar aggregate precision in all three main disciplines (Table 8 in Supplementary Information B.8). Across all evaluated models, average accuracy is 32.5% for chemistry, 31.2% for biology, and 29.0% for materials science.

Performance differences become clearer at the sub-discipline level (Figure 5; full per-model breakdown in Table 8). The lowest mean accuracies occur in organic and polymer materials (25.6%), polymer chemistry and physics (28.7%), composite materials (28.7%), and inorganic chemistry (28.3%). The highest reported sub-discipline accuracy is organic chemistry (37.9%), followed by metallic materials (34.8%), analytical chemistry (34.5%), and immunology (34.0%). Sub-disciplines with small sample sizes should be interpreted with care.

The radar plot shows that no single model demonstrates universal dominance throughout scientific spectrum. However, proficiency remains highly domain-contingent. Additionally, we penalized Claude Opus 4.8 and Claude Opus 5 for a subset of safety-related refusals (marked with asterisks in the related figures and tables), which are triggered exclusively in the biochemistry section by questions involving virus–host interactions, pathogen structural biology, and related experimental techniques (e.g., cryo-EM, immunoblotting, viral neutralization assays), to ensure a fair comparison across all models.

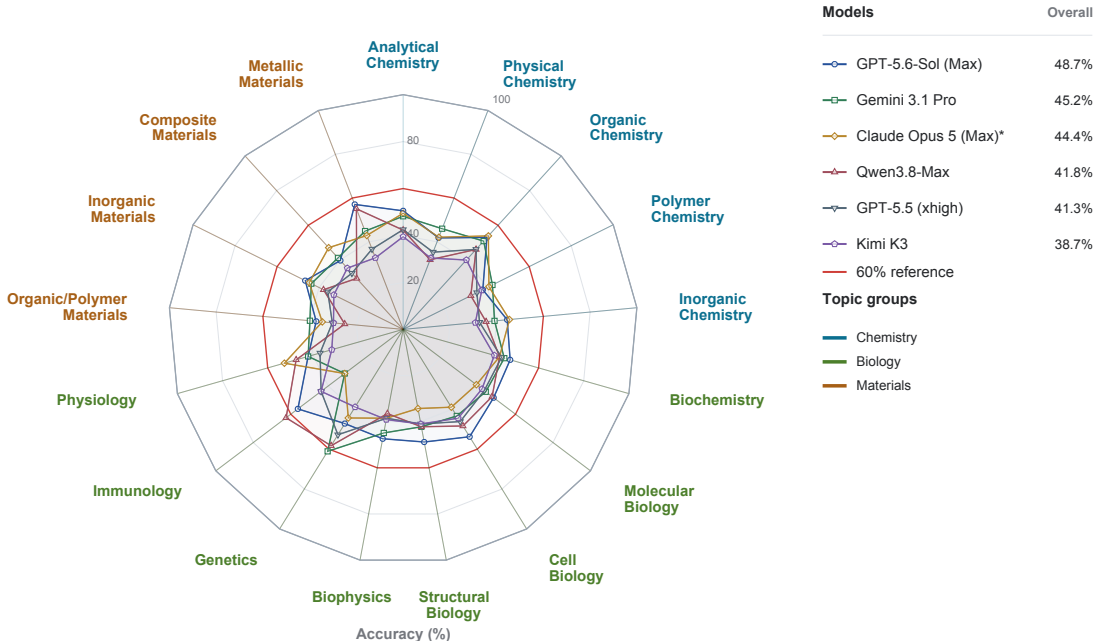


Figure 5. Results by discipline of the 6 overall strongest models are summarized in the radar plot. Each axis reports accuracy on one reported sub-discipline label. Axis labels and spokes are colored by broad discipline. Concentric rings mark 20%, 40%, 60%, 80%, and 100% accuracy respectively. The red ring highlights the 60% reference level. The asterisk on Claude Opus 5 (Max) indicates safety-filtered no-response cases in biochemistry, which are counted as incorrect.

4.2 Test on Science-Specialized Models

We next compare general-purpose and science-specialized MLLMs to examine whether domain adaptation improves performance on **SEE**. Within the tested model set, the four science-specialized models remain below the general-purpose average on **SEE** (Figure 4), with an average accuracy of 22.2% compared to 34.9% for the general-purpose models. However, Intern-S2 Preview-397B reaches 29.0%, outperforming several mid-tier general-purpose models and narrowing the gap relative to earlier science-specialized models.

We observe that the phenomenon of leading general-purpose models outperforming domain-specific models is not unique to our setting, but also appears in other fields such as clinical medicine [64]. This suggests that simply specializing a model through post-training or augmenting it with retrieval-augmented generation (RAG) does not necessarily lead to improved performance.

These results suggest that success on **SEE** requires more than domain-specific factual familiarity.

Science-specialized training may improve exposure to scientific terminology, concepts, and task formats, but **SEE** requires models to combine such knowledge with visual experimental evidence and specific experimental context. The gap between science-specialized and leading general-purpose models therefore indicates that robust scientific reasoning in real experimental settings depends not only on domain adaptation, but also on broader multimodal understanding and flexible evidence-grounded reasoning.

4.3 Interdisciplinary Analysis

As noted above, greater exposure to a specific scientific domain does not necessarily translate into reliable reasoning across broader scientific contexts. In practice, scientific discovery often crosses disciplinary boundaries and requires integrating knowledge, assumptions, and observations from multiple fields. To reflect this reality, **SEE** explicitly includes interdisciplinary questions spanning biology, chemistry, and materials science. This design is central to the benchmark, whose goal is not only to test whether models know isolated scientific facts, but also to evaluate whether they can integrate heterogeneous evidence across disciplinary contexts.

We first examine the performance gap between single-discipline and cross-discipline questions. Across all models, the average accuracy is 35.1% on questions within a single main discipline, but drops to 28.7% on questions that span multiple main disciplines. This 6.4% gap suggests that cross-disciplinary settings impose additional coordination demands beyond those captured by single-discipline evaluation.

We next examine cross-subdisciplinary complexity. Averaged across all models, performance remains similar for questions annotated with one or two subdisciplinary labels, with accuracies of 36.2% and 35.6%, respectively. However, the accuracy drops to 29.5% for questions with three labels and further to 28.2% for questions with four labels. This pattern suggests that performance degrades as a task requires coordination across a larger number of scientific concepts, experimental techniques, or forms of evidence.

Together, these results show that interdisciplinary tasks provide a stringent stress test of model robustness. Such tasks require models to integrate scientific terminology, experimental assumptions, visual evidence, and multistep reasoning within a single response. The interdisciplinary design of **SEE** therefore assesses whether current MLLMs can coordinate knowledge and evidence across the boundaries commonly encountered in real-world research. To be considered robust in a research setting, a model must maintain coherent reasoning even when a problem falls outside the distributions most frequently represented in training.

4.4 Modality Ablation Study

To evaluate the level of hallucination of MLLMs by investigating their ability to recognize missing information, we performed a blind-modality diagnostic by removing image inputs while preserving the original textual queries across 18 models. This experiment serves as a critical test for grounding. A truly intelligent system should identify when a scientific conclusion is impossible without visual evidence. Across 20,088 text-only instances, models identified missing information in only 921

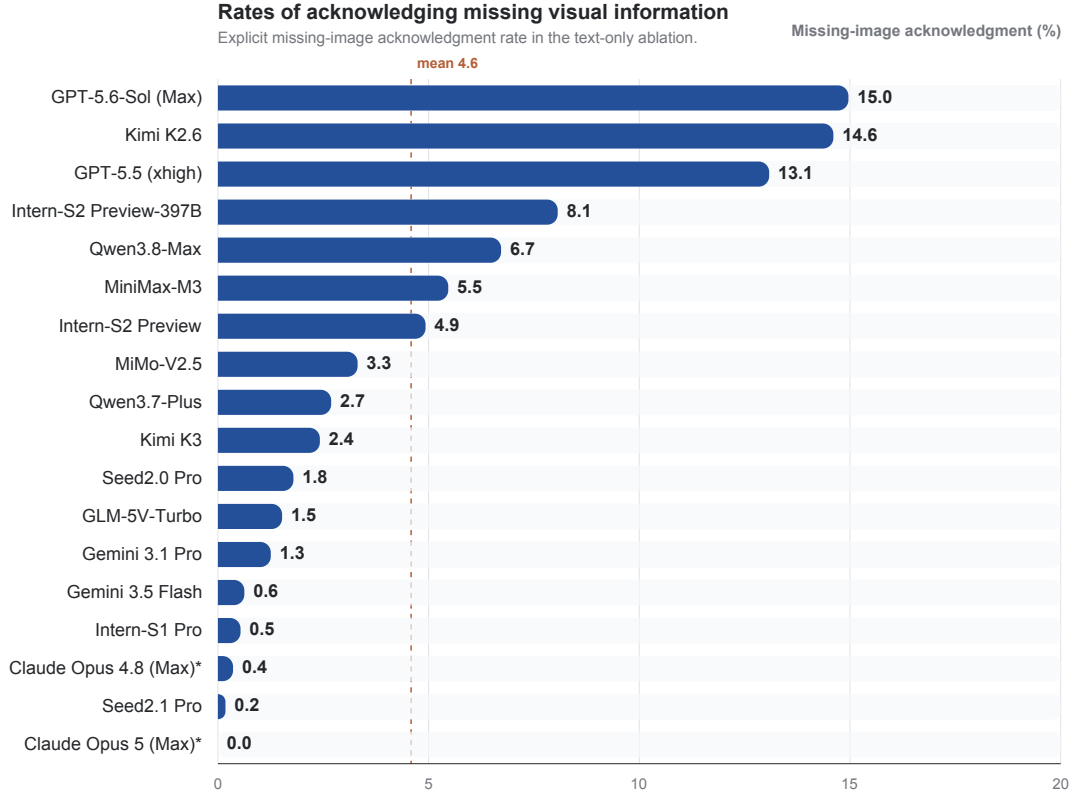


Figure 6. Explicit missing-image acknowledgment rates in the text-only ablation test. The denominator is the number of text-only instances for each model. The numerator is the number of responses that explicitly recognize missing image or visual evidence. Empty outputs, generation errors, generic refusals, and other non-answer cases are not counted. The dashed vertical line marks the average rate across the models. Asterisks denote safety-filtered no-response cases counted under the strict evaluation protocol.

cases (4.6%, Figure 6). This remarkably low rate suggests that current MLLMs are prone to blind hallucination, attempting to derive scientific answers from linguistic priors rather than admitting a lack of empirical evidence. Removing the image also lowers accuracy for every model, by 12.2 percentage points on average; this and further text-only diagnostics are reported in Supplementary Information B.7 (Figure 9).

This result reveals limited awareness of evidential boundaries in current MLLMs. When visual inputs are removed, models often continue to answer based on textual cues, prior knowledge, or common scientific patterns, rather than recognizing that the available evidence is insufficient. This behavior is especially concerning in real-world scientific settings, where decisions are often made under incomplete or ambiguous information and where unsupported but confident answers may pose greater risks than appropriate abstention. Thus, this ablation analysis provides a diagnostic signal that current MLLMs do not yet consistently reason within the limits of experimental evidence.

4.5 Tool-Augmented Visual-Agent Evaluation

Finally, we test whether tool access is sufficient to narrow the capability gap exposed by **SEE**. We evaluate the six top-performing MLLMs that support both web search and code-interpreter tools

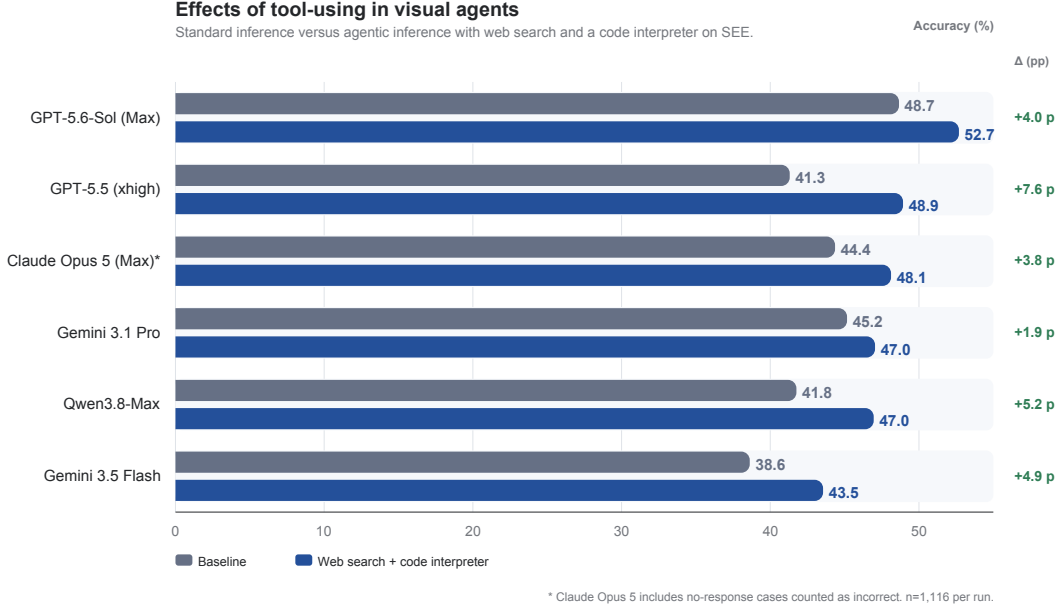


Figure 7. Comparison between standard multimodal inference and the tool-augmented visual-agent setting, in which models can iteratively use web search and a code interpreter before returning a final answer. Paired bars show the accuracy of each model under the two settings, and numbers on the right report the corresponding difference in percentage points. The asterisk on Claude Opus 5 (Max) denotes no-response cases, which are counted as incorrect.

in a tool-augmented visual-agent setting (Figure 7). Models receive the same original multimodal inputs as in the standard baseline, but can invoke a web search and a code interpreter before producing a final answer. The code interpreter supports programmatic inspection, measurement, and computation over the input image, whereas the web search provides external scientific background information. In this setting, all six models improve, with gains ranging from 1.9% to 7.6% and a mean improvement of 4.6%. GPT-5.5 (xhigh) shows the largest improvement, increasing from 41.3% to 48.9%. GPT-5.6-Sol (Max) achieves the highest tool-augmented accuracy of 52.7%. Nevertheless, significant errors remain, indicating that tool access alone does not make MLLMs reliable on **SEE**. Detailed configuration is provided in the Supplementary Information B.5.

Trajectory analysis shows that tool use has bidirectional effects on model outputs (Figure 8). Across the six models, 411 corrections are substantively attributed to tool use. Fact retrieval accounts for 47.7% of these cases, code-assisted image analysis for 41.6%, information cross-checking for 7.1%, and code-assisted computation for 3.6%. These results show that tools can expand the models’ ability to inspect visual evidence and access background information, partially compensating for the limitations of static multimodal inference.

Tools can also introduce new errors. Among 145 tool-attributed errors, failure modes gather in evidence integration failures (40.0%), action selection failures (26.9%), observation interpretation failures (21.4%), and endless overthinking (11.7%). Both beneficial and detrimental effects of tools occur in the tool-use trajectory (Figure 8). This indicates that tool augmentation does not simply provide models with additional information, but turns static multimodal reasoning into an evidence-management problem along the tool-use trajectory. Models must decide which actions

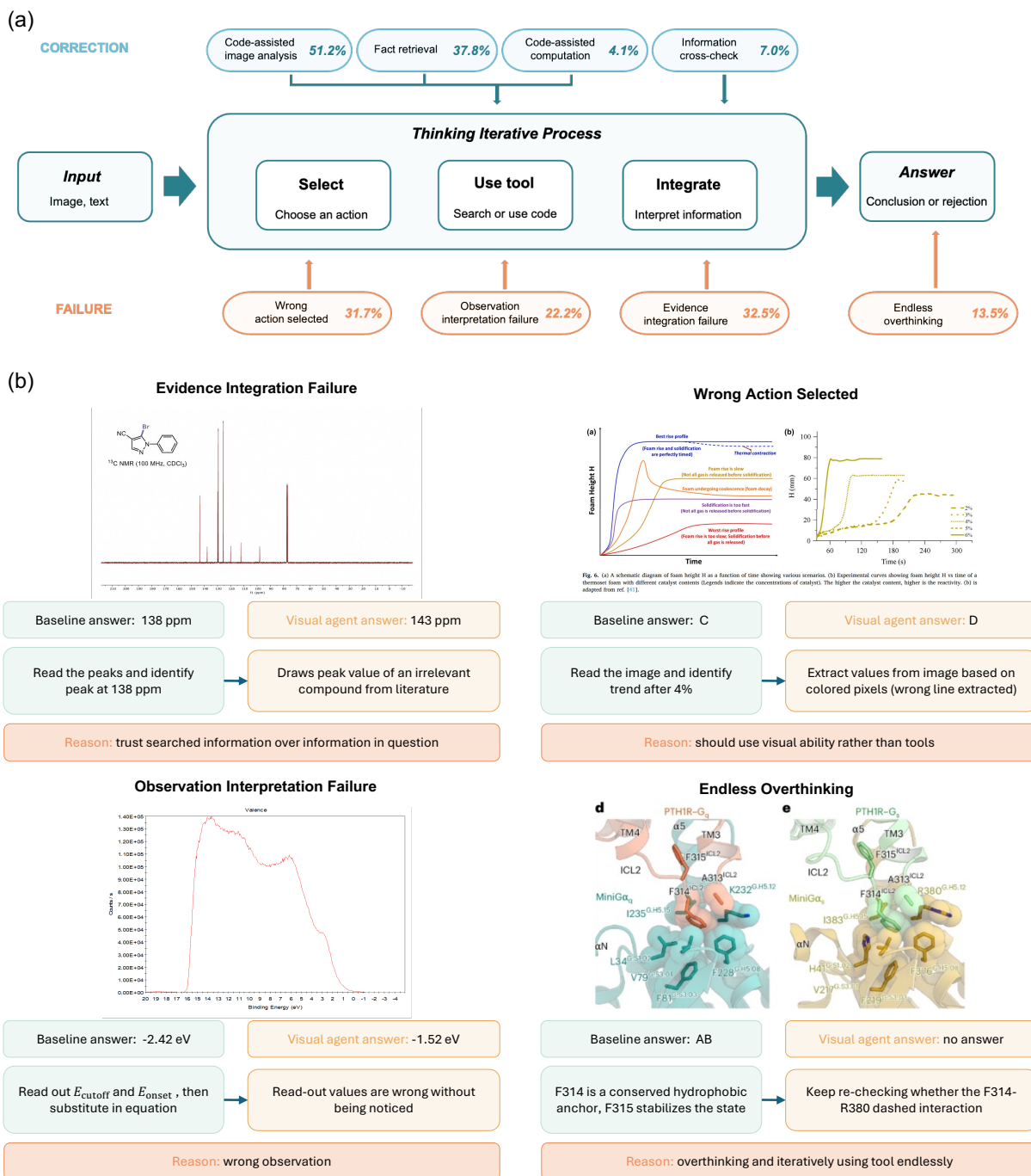


Figure 8. Effects of using tools in tool-augmented visual agents. (a) Illustration of tool effects on stages of tool-mediated visual-agent reasoning. Corrections are made when models select relevant actions, use tools to inspect or quantify visual evidence, retrieve key external facts, or verify intermediate interpretations. Failures can arise from wrong action selection, misinterpretation of observations, inappropriate integration of retrieved or computed information, or failure to control and terminate the interaction. The dashed loop indicates that unresolved uncertainty may trigger further tool use. (b) Cases for every failed scenario.

are relevant to the current problem, whether tool outputs are reliable, how retrieved or computed information should be integrated with the original experimental observation, and when to stop further interaction.

5 Model Failure Analysis

5.1 Weakness in Perception

Limited Visual Quantitative Precision. A substantial proportion of model failures are due to inadequate visual reading precision (tasks requiring the extraction of exact numerical values from spectra, standard curves, or instrument readouts). These failures expose that the visual encoders of current MLLMs are pattern recognizers rather than measurement instruments. They excel at categorical judgments, but perform poorly at fine-grained quantitative extraction. This perceptual inaccuracy can propagate downstream, causing the model to arrive at incorrect conclusions even when it possesses sound domain knowledge and applies valid reasoning logic. For scientific applications where quantitative precision is non-negotiable, this limitation constitutes a fundamental bottleneck that cannot be resolved by improvements to reasoning alone.

Prior Knowledge Suppressing Visual Input. When visual evidence presented in an image conflicts with patterns prevalent in the training data, models systematically defer to prior knowledge rather than faithfully attending to the actual visual input. Rather than genuinely "reading" an image, models tend to infer its content by reverse-mapping from the most frequently encountered concepts in training. This pattern is particularly noticeable in tasks involving molecular structure interpretation typical in biology: models directly match surface-level visual features to high-frequency terms encountered during training instead of decomposing the structure, identifying functional fragments, and progressively deriving the corresponding name or property as a human expert would. This pattern matching shortcut fails when confronted with atypical or non-canonical structures. Current alignment methods produce models that match patterns rather than observe structures. They excel at associating frequent labels with common visuals but lack the bottom-up logic necessary to parse the unknown.

5.2 Weakness in Inference

Beyond perceptual limitations, our analysis identifies three distinct inferential failure modes that reflect deeper architectural constraints in the way current MLLMs construct and evaluate logical representations.

Neglect of Global Logical Coherence. A recurring failure mode is the model's inability to maintain a coherent logical framework, often processing question components as disparate units rather than an integrated system. There are two typical patterns. First, models frequently overlook intra-option contradictions. If an option pairs two factually correct but logically incompatible premises, the model tends to evaluate them independently, missing the overall information. Second, there is a clear deficiency in multi-modal integration. When a problem relies on the interplay between several subfigures, models often perform a localized search within one subfigure while neglecting the broader logical constraints established by the remaining figures. This fragmented

processing prevents the model from forming a comprehensive relational structure, leading to conclusions that are locally plausible but globally invalid.

Overextrapolation of Experimental Conclusions. Another critical failure mode is the tendency to over-infer from partial experimental data. For example, an inquiry requires the synthesis of several independent experiments to reach a conclusion, the current models often treat preliminary or isolated results as conclusive evidence. Models frequently interpolate "missing" experimental steps and invent evidence to justify their final claims. This tendency to overextrapolate from individual observations undermines the rigorous evidentiary standards essential to valid scientific inference.

Multi-Capability Coordination Bottleneck. Many scientific challenges require the coordinated operation of multiple cognitive skills, and failure rates escalate sharply when faced with these demands. NMR spectral interpretation, for instance, requires a model to utilize precise visual perception, deep domain knowledge, and complex logical reasoning simultaneously. Because current MLLMs possess individual deficiencies in each area, the requirement for their serial execution leads to a multiplicative compounding of error. This explains why NMR-related tasks show disproportionately high failure rates across all architectures. In summary, failures on these tasks do not arise from a single weakness, but from the model’s inability to reliably coordinate multiple interdependent capabilities.

Collectively, these failure modes underscore a fundamental contradiction: while the essence of scientific inquiry lies in the ability to make rigorous judgments regarding out-of-distribution (OOD) phenomena, current models systematically regress to high-frequency patterns when confronted with unfamiliar evidence. Whether by prioritizing prior knowledge over visual observation or substituting "textbook" conclusions for incomplete evidence chains, models gravitate toward the path of least statistical resistance. This suggests that merely expanding the training database is insufficient, as the frontier of scientific discovery, by definition, resides at the boundary of available data.

6 Conclusion

We introduced **SEE** to evaluate if current MLLMs can support real laboratory science by reasoning from multimodal experimental evidence. Unlike benchmarks that primarily measure scientific knowledge recall or isolated domain competence, **SEE** is based on peer-reviewed literature and experimental practice in chemistry, biology, and materials science. Under the standard inference setting, the best-performing model among the 19 MLLMs evaluated reaches only 48.7% accuracy, and no model exceeds 50%. This low performance should not be interpreted simply as another hard benchmark result. Instead, it reveals a mismatch between the abilities captured by many existing scientific benchmarks and the abilities required for real reasoning process in experimental research.

The additional analyses clarify the nature of this mismatch. Science-specialized models do not close the performance gap. Meanwhile, interdisciplinary analysis shows that **SEE** requires models to coordinate concepts, methods, and different types of evidence across disciplinary boundaries.

Ablation analysis reveals that when visual evidence is removed, MLLMs rarely recognize that essential information is missing. Together, these findings suggest that the central limitation is not simply what the models know but whether they can extract and organize information from the evidence that is actually available in the problem.

The tool-augmented visual-agent experiment extends this conclusion from static multimodal inference to interactive scientific workflows. Tool access boosts accuracy to 52.7% (strongest model), showing that code interpreter and web search can partially compensate for the limitations of multimodal reasoning. However, trajectory analysis reveals that tool use is not always beneficial. The interaction process can help models inspect visual evidence through code-assisted image analysis, retrieve missing scientific facts, cross-check information, or perform code-assisted computation. Yet it can also introduce new errors through wrong action selection, misinterpretation of tool-derived observations, evidence integration failure, or endless overthinking. The diagnosis of tool-use trajectory provides an explicit view of tool-using in scientific reasoning rather than only the overall accuracy. The main challenge is therefore whether a model can manage tool-derived information by selecting appropriate actions, judging the reliability and relevance of tool outputs, integrating retrieved or computed result with the original information, and stopping when the available evidence is sufficient.

The failure analysis explains why this limitation matters for scientific reasoning. Current MLLMs often miss critical visual evidence, allow prior knowledge to override experimental observations, fail to maintain global logical coherence, and draw conclusions beyond what the evidence supports. These errors are not trivial. These deficiencies highlight a fundamental gap between pattern recognition and the rigorous, evidence-based logic required for authentic scientific inquiry.

These findings suggest a new stage for scientific evaluation where the focus shifts from single-response MLLMs toward visual agents and broader agentic scientific systems. Real discovery is rarely completed by answering one question from a fixed input. Instead, it requires an iterative process of proposing hypotheses, selecting tools, transforming or measuring observations, analyzing experimental outputs, and deciding which evidence should be collected next. Future benchmarks for scientific agents should therefore evaluate not only accuracy, but also the quality of the interaction trajectory. This includes assessing whether each action is justified by the available evidence, whether image operations, retrieval, and data analysis are appropriate, whether uncertainty is recognized before the next step, and whether conclusions are revised when new observations contradict prior assumptions. Therefore, agent-level evaluation should extend the evidence-based principle of **SEE** from static experimental reasoning to dynamic scientific decision-making.

Overall, **SEE** shows that the main barrier to scientifically useful AI is not simply more knowledge, stronger domain specialization, or broader tool access, but the ability to manage multimodal evidence and make justified evidence-bounded inferences from experimental results. The tool-augmented visual-agent analysis shows that this challenge persists in interactive workflows, where tools expand evidence access but also require reliable action selection, output evaluation, evidence integration, and termination control. The failure analysis reveals the same limitations. Current

MLLMs miss critical observations, let prior knowledge override evidence, lose global coherence, and overextend conclusions beyond the data. Future evaluations should therefore test not only what models know, but whether they can transform multimodal experimental observations into new, evidence-supported scientific understanding. This is the missing step toward real scientific discovery.

Data Availability

To facilitate the benchmarking and reproducibility of our work, the accompanying code and public datasets are available on [GitHub](#). The main benchmark contains 1,116 questions, of which 1,049 are publicly released. To support reproducibility, we also report headline results on the public subset in Supplementary Tables [5](#) and [6](#).

References

- [1] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. “Measuring Massive Multitask Language Understanding”. In: *International Conference on Learning Representations (ICLR)*. 2021. arXiv: 2009.03300 [cs.CY]. URL: <https://openreview.net/forum?id=d7KBjmI3GmQ>.
- [2] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, et al. “MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark”. In: *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*. 2024. arXiv: 2406.01574 [cs.CL]. URL: https://proceedings.neurips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets_and_Benchmarks_Track.html.
- [3] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. *GPQA: A Graduate-Level Google-Proof Q&A Benchmark*. 2023. arXiv: 2311.12022 [cs.AI]. URL: <https://arxiv.org/abs/2311.12022>.
- [4] Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. *CMMLU: Measuring Massive Multitask Language Understanding in Chinese*. 2023. arXiv: 2306.09212 [cs.CL]. URL: <https://arxiv.org/abs/2306.09212>.
- [5] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, et al. *Humanity’s Last Exam*. 2025. arXiv: 2501.14249 [cs.LG]. URL: <https://arxiv.org/abs/2501.14249>.
- [6] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. “Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022. arXiv: 2209.09513 [cs.CL]. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html.
- [7] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. “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)*. 2024, pp. 9556–9567. DOI: 10.1109/CVPR52733.2024.00913. arXiv: 2311.16502 [cs.CV]. URL: https://openaccess.thecvf.com/content/CVPR2024/html/Yue_MMMU_A_Massive_Multi-discipline_Multimodal_Understanding_and_Reasoning_Benchmark_for_CVPR_2024_paper.html.
- [8] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhua Chen, and Graham Neubig. *MMMU-Pro: A More Robust Multi-Discipline Multimodal Understanding Benchmark*. 2024. arXiv: 2409.02813 [cs.CL]. URL: <https://arxiv.org/abs/2409.02813>.
- [9] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. “MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities”. In: *International Conference on Machine Learning (ICML)*. Vol. 235. 2024,

- pp. 57730–57754. arXiv: [2308.02490 \[cs.CV\]](#). URL: <https://proceedings.mlr.press/v235/you24o.html>.
- [10] 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. arXiv: [2310.02255 \[cs.CV\]](#). URL: <https://openreview.net/forum?id=KUNzEQMWU7>.
 - [11] Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. *SciFIBench: Benchmarking Large Multimodal Models for Scientific Figure Interpretation*. 2024. arXiv: [2405.08807 \[cs.CV\]](#). URL: <https://arxiv.org/abs/2405.08807>.
 - [12] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. “OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. 2024, pp. 3828–3850. URL: <https://aclanthology.org/2024.acl-long.211/>.
 - [13] Xinwu Ye, Chengfan Li, Siming Chen, Wei Wei, and Robert Tang. “MMSciBench: Benchmarking Language Models on Chinese Multimodal Scientific Problems”. In: *Findings of the Association for Computational Linguistics: ACL 2025*. Vienna, Austria: Association for Computational Linguistics, 2025, pp. 14621–14663. DOI: [10.18653/v1/2025.findings-acl.755](#). URL: <https://aclanthology.org/2025.findings-acl.755/>.
 - [14] Jiacheng Ruan, Dan Jiang, Xian Gao, Ting Liu, Yuzhuo Fu, and Yangyang Kang. “MME-SCI: A Comprehensive and Challenging Science Benchmark for Multimodal Large Language Models”. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. Vol. 40. 11. 2026, pp. 8760–8768. DOI: [10.1609/aaai.v40i11.37829](#). arXiv: [2508.13938 \[cs.CL\]](#). URL: <https://ojs.aaai.org/index.php/AAAI/article/view/37829>.
 - [15] Yiming Cui, Xin Yao, Yuxuan Qin, Xin Li, Shijin Wang, and Guoping Hu. “Evaluating Large Language Models on Multimodal Chemistry Olympiad Exams”. In: *Communications Chemistry* 8 (2025). USNCO-V benchmark, p. 402. DOI: [10.1038/s42004-025-01782-x](#).
 - [16] Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Kevin Maik Jablonka, et al. “A Framework for Evaluating the Chemical Knowledge and Reasoning Abilities of Large Language Models against the Expertise of Chemists”. In: *Nature Chemistry* (2025). ChemBench. DOI: [10.1038/s41557-025-01815-x](#).
 - [17] Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. “SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models”. In: *International Conference on Machine Learning (ICML)*. 2024. URL: <https://proceedings.mlr.press/v235/wang24z.html>.
 - [18] Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. “SciEval: A Multi-Level Large Language Model Evaluation Benchmark for Scientific

- Research”. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 2024. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/29872>.
- [19] Jon M. Laurent, Joseph D. Janizek, Michael Ruzo, Michaela M. Hinks, Michael J. Hammerling, Siddharth Narayanan, Manvitha Ponnampati, Andrew D. White, and Samuel G. Rodrigues. *LAB-Bench: Measuring Capabilities of Language Models for Biology Research*. 2024. arXiv: [2407.10362](https://arxiv.org/abs/2407.10362) [cs.AI]. URL: <https://arxiv.org/abs/2407.10362>.
 - [20] Nawaf Alampara, Mara Schilling-Wilhelmi, Martı́ño Rı́os-Garcı́a, et al. “Probing the Limitations of Multimodal Language Models for Chemistry and Materials Research”. In: *Nature Computational Science* (2025). MaCBench. DOI: [10.1038/s43588-025-00836-3](https://doi.org/10.1038/s43588-025-00836-3).
 - [21] Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. “Benchmarking Materials Property Prediction Methods: The Matbench Test Set and Automatminer Reference Algorithm”. In: *npj Computational Materials* 6.1 (2020), p. 138. DOI: [10.1038/s41524-020-00406-3](https://doi.org/10.1038/s41524-020-00406-3).
 - [22] Zhiyuan Huang, Baichuan Yang, Zikun He, Yanhong Wu, Hongyu Fang, Zhenhe Liu, Dongsheng Lin, and Bing Su. *ChemVTS-Bench: Evaluating Visual–Textual–Symbolic Reasoning of Multimodal Large Language Models in Chemistry*. 2025. arXiv: [2511.17909](https://arxiv.org/abs/2511.17909) [cs.AI]. URL: <https://arxiv.org/abs/2511.17909>.
 - [23] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. *Galactica: A Large Language Model for Science*. 2022. arXiv: [2211.09085](https://arxiv.org/abs/2211.09085) [cs.CL]. URL: <https://arxiv.org/abs/2211.09085>.
 - [24] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. “BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining”. In: *Briefings in Bioinformatics* 23.6 (2022), bbac409. DOI: [10.1093/bib/bbac409](https://doi.org/10.1093/bib/bbac409).
 - [25] Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, and Jonathan Frankle. *BioMedLM: A 2.7B Parameter Language Model Trained on Biomedical Text*. 2024. arXiv: [2403.18421](https://arxiv.org/abs/2403.18421) [cs.CL]. URL: <https://arxiv.org/abs/2403.18421>.
 - [26] Cheng Peng, Xi Yang, Aokun Chen, Kaleb E. Smith, Nima PourNejatian, Anthony B. Costa, Cheryl Martin, Mona G. Flores, Ying Zhang, et al. “A Study of Generative Large Language Model for Medical Research and Healthcare”. In: *npj Digital Medicine* 6 (2023). GatorTronGPT, p. 210. DOI: [10.1038/s41746-023-00958-w](https://doi.org/10.1038/s41746-023-00958-w). URL: <https://www.nature.com/articles/s41746-023-00958-w>.
 - [27] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Amin Mohamed, Le Hou, Kevin Clark, Stephen R. Pfohl, et al. “Toward Expert-Level Medical Question Answering with Large Language Models”. In: *Nature Medicine* 31 (2025). Med-PaLM, pp. 943–950. DOI: [10.1038/s41591-024-03423-7](https://doi.org/10.1038/s41591-024-03423-7). URL: <https://www.nature.com/articles/s41591-024-03423-7>.
 - [28] Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, et al. *MEDITRON-70B: Scaling Medical Pretraining for Large Language Models*. 2023. arXiv: [2311.16079](https://arxiv.org/abs/2311.16079) [cs.CL]. URL: <https://arxiv.org/abs/2311.16079>.

- [29] Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, et al. *ChemLLM: A Chemical Large Language Model*. 2024. arXiv: [2402.06852](https://arxiv.org/abs/2402.06852) [cs.CL]. URL: <https://arxiv.org/abs/2402.06852>.
- [30] Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Yi Xia, Bo Chen, Hongshen Xu, et al. “Developing ChemDFM as a Large Language Foundation Model for Chemistry”. In: *Cell Reports Physical Science* 6.4 (2025), p. 102523. DOI: [10.1016/j.xcrp.2025.102523](https://doi.org/10.1016/j.xcrp.2025.102523). URL: [https://www.cell.com/cell-reports-physical-science/fulltext/S2666-3864\(25\)00122-5](https://www.cell.com/cell-reports-physical-science/fulltext/S2666-3864(25)00122-5).
- [31] Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. “Mol-Instructions: A Large-Scale Biomolecular Instruction Dataset for Large Language Models”. In: *International Conference on Learning Representations (ICLR)*. 2024. arXiv: [2306.08018](https://arxiv.org/abs/2306.08018) [cs.CL]. URL: https://openreview.net/forum?id=Tl_sdsb619n.
- [32] Botao Yu, Frazier N. Baker, Ziqi Chen, Xia Ning, and Huan Sun. *LlaSMol: Advancing Large Language Models for Chemistry with a Large-Scale, Comprehensive, High-Quality Instruction Tuning Dataset*. 2024. arXiv: [2402.09391](https://arxiv.org/abs/2402.09391) [cs.CL]. URL: <https://arxiv.org/abs/2402.09391>.
- [33] Yu Song, Santiago Miret, Huan Zhang, and Bang Liu. “HoneyBee: Progressive Instruction Finetuning of Large Language Models for Materials Science”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, 2023, pp. 5724–5739. DOI: [10.18653/v1/2023.findings-emnlp.380](https://doi.org/10.18653/v1/2023.findings-emnlp.380). arXiv: [2310.08511](https://arxiv.org/abs/2310.08511) [cs.CL]. URL: <https://aclanthology.org/2023.findings-emnlp.380/>.
- [34] Yingheng Tang, Wenbin Xu, Jie Cao, Weilu Gao, Steve Farrell, Benjamin Erichson, Michael W. Mahoney, Andy Nonaka, and Zhi Yao. “A Multimodal Large Language Model for Materials Science”. In: *Nature Machine Intelligence* (2026). MatterChat. DOI: [10.1038/s42256-026-01214-y](https://doi.org/10.1038/s42256-026-01214-y). URL: <https://www.nature.com/articles/s42256-026-01214-y>.
- [35] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. “LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day”. In: *Advances in Neural Information Processing Systems*. Vol. 36. Curran Associates, Inc., 2023, pp. 28541–28564. arXiv: [2306.00890](https://arxiv.org/abs/2306.00890) [cs.CV]. URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/5abcdf8ecdcacba028c6662789194572-Abstract-Datasets_and_Benchmarks.html.
- [36] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaeckermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, et al. *Towards Generalist Biomedical AI*. 2023. arXiv: [2307.14334](https://arxiv.org/abs/2307.14334) [cs.AI]. URL: <https://arxiv.org/abs/2307.14334>.
- [37] Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. *BioMedGPT: Open Multimodal Generative Pre-trained Transformer for BioMedicine*. 2023. arXiv: [2308.09442](https://arxiv.org/abs/2308.09442) [cs.CV]. URL: <https://arxiv.org/abs/2308.09442>.
- [38] Qingxiao Li, Lifeng Xu, QingLi Wang, Yudong Bai, Mingwei Ou, Shu Hu, and Nan Xu. *S1-VL: Scientific Multimodal Reasoning Model with Thinking-with-Images*. We evaluate the S1-VL-32B-RL checkpoint. 2026. arXiv: [2604.21409](https://arxiv.org/abs/2604.21409) [cs.CV]. URL: <https://arxiv.org/abs/2604.21409>.

- [39] InternLM. *Intern-S2 Preview*. Official model page. 2026. URL: <https://huggingface.co/internlm/Intern-S2-Preview>.
- [40] InternLM. *Intern-S2 Preview-397B*. Official model page. 2026. URL: <https://huggingface.co/internlm/Intern-S2-Preview-397B>.
- [41] Shanghai AI Laboratory. *Intern-S1 Pro*. 2026. arXiv: 2603.25040 [cs.CL]. URL: <https://arxiv.org/abs/2603.25040>.
- [42] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. “Augmenting Large Language Models with Chemistry Tools”. In: *Nature Machine Intelligence* 6.5 (2024). ChemCrow, pp. 525–535. DOI: 10.1038/s42256-024-00832-8.
- [43] Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. “Autonomous Chemical Research with Large Language Models”. In: *Nature* 624.7992 (2023), pp. 570–578. DOI: 10.1038/s41586-023-06792-0.
- [44] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. *The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery*. 2024. arXiv: 2408.06292 [cs.AI]. URL: <https://arxiv.org/abs/2408.06292>.
- [45] Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. “Agent Laboratory: Using LLM Agents as Research Assistants”. In: *Findings of the Association for Computational Linguistics: EMNLP 2025*. Suzhou, China: Association for Computational Linguistics, 2025, pp. 5977–6043. DOI: 10.18653/v1/2025.findings-emnlp.320. arXiv: 2501.04227 [cs.HC]. URL: <https://aclanthology.org/2025.findings-emnlp.320/>.
- [46] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, et al. “Accelerating Scientific Discovery with Co-Scientist”. In: *Nature* (2026). DOI: 10.1038/s41586-026-10644-y. URL: <https://www.nature.com/articles/s41586-026-10644-y>.
- [47] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, et al. “Holistic Evaluation of Language Models”. In: *Transactions on Machine Learning Research (TMLR)* (2023). arXiv: 2211.09110 [cs.CL]. URL: <https://openreview.net/forum?id=i04LZibEqW>.
- [48] Google DeepMind. *Gemini 3.1 Pro*. Official model page. 2026. URL: <https://deepmind.google/models/gemini/pro/>.
- [49] OpenAI. *GPT-5.6-Sol*. Official announcement. 2026. URL: <https://openai.com/index/gpt-5-6/>.
- [50] Anthropic. *Claude Opus 5*. Official announcement. 2026. URL: <https://www.anthropic.com/news/claude-opus-5>.
- [51] Qwen Team, Alibaba Cloud. *Qwen3.8-Max*. Official announcement. 2026. URL: <https://qwen.ai/blog?id=qwen3.8>.
- [52] OpenAI. *GPT-5.5 (xhigh)*. Official announcement. 2026. URL: <https://openai.com/zh-Hans-CN/index/introducing-gpt-5-5/>.

- [53] Moonshot AI. *Kimi K3: Open Frontier Intelligence*. Technical report. 2026. URL: <https://arxiv.org/abs/2607.24653>.
- [54] Google DeepMind. *Gemini 3.5 Flash*. Official model page. 2026. URL: <https://deepmind.google/models/gemini/flash/>.
- [55] Anthropic. *Claude Opus 4.8 (Max)*. Official announcement. 2026. URL: <https://www.anthropic.com/news/claude-opus-4-8>.
- [56] ByteDance Seed. *Seed2.1 Pro*. Official product page. 2026. URL: https://seed.bytedance.com/en/seed2_1.
- [57] Qwen Team, Alibaba Cloud. *Qwen3.7-Plus: Multimodal Agent Intelligence*. Official announcement. 2026. URL: https://www.alibabacloud.com/blog/qwen3-7-plus-multimodal-agent-intelligence_603206.
- [58] ByteDance Seed. *Seed2.0 Pro*. Official product page. 2026. URL: <https://seed.bytedance.com/zh/seed2>.
- [59] MiniMax. *MiniMax-M3: Frontier Coding, 1M Context, Native Multimodality*. Official blog post. 2026. URL: <https://www.minimax.io/blog/minimax-m3>.
- [60] Moonshot AI. *Kimi K2.6: Advancing Open-Source Coding*. Official blog post. 2026. URL: <https://www.kimi.com/blog/kimi-k2-6>.
- [61] Zhipu AI (Z.ai). *GLM-5V-Turbo*. Official documentation. 2026. URL: <https://docs.z.ai/guides/vlm/glm-5v-turbo>.
- [62] LLM-Core Team, Xiaomi. *MiMo-V2.5*. Official model page. 2026. URL: <https://mimo.xiaomi.com/mimo-v2-5>.
- [63] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena”. In: *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*. 2023. arXiv: 2306.05685 [cs.CL]. URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.
- [64] Krithik Vishwanath, Anton Alyakin, Mrigayu Ghosh, Ali Hage, et al. “General-Purpose Large Language Models Outperform Specialized Clinical AI Tools on Medical Benchmarks”. In: *Nature Medicine* (2026). DOI: 10.1038/s41591-026-04431-5. URL: <https://www.nature.com/articles/s41591-026-04431-5>.

Supplementary Information

A Dataset Details

SEE contains 1,116 multimodal scientific questions. Each released entry includes the question text, associated visual files, a standard answer, source information, a UUID, discipline labels, and knowledge-point labels. The associated visual files are entry-level references: they may include images required by the question, but may also include images used in the solution rationale or expert verification, and therefore should not be read as a direct count of question-side figures.

The released question file includes both entries with explicit option lists and entries without a separate option field. The latter include short-answer, numerical, and visually grounded selection tasks whose answers are represented directly as canonical answers. The dataset metadata records three task-type categories: reasoning, measurement, and image processing. The source field is normalized into DOI-like literature sources, personal or experimental sources, and other source text or URL records.

B Evaluation

B.1 Evaluation Setup

We evaluate **SEE** on 19 representative MLLMs: Gemini 3.1 Pro, GPT-5.6-Sol (Max), Claude Opus 5 (Max), Qwen3.8-Max, GPT-5.5 (xhigh), Kimi K3, Gemini 3.5 Flash, Claude Opus 4.8 (Max), Seed2.1 Pro, Qwen3.7-Plus, Seed2.0 Pro, MiniMax-M3, Kimi K2.6, GLM-5V-Turbo, MiMo-V2.5, S1-VL, Intern-S2 Preview, Intern-S2 Preview-397B, and Intern-S1 Pro.

All models are evaluated using default inference settings, with two exceptions: GPT-5.5 is configured with extended thinking set to **xhigh**, and Claude Opus 4.8 is configured with extended thinking set to **max**. For S1-VL, we evaluate the S1-VL-32B-RL checkpoint. We do not introduce additional manual prompt optimization, response filtering, or post-hoc correction during evaluation. For questions containing image inputs, images are preprocessed according to the requirements of each model interface, including necessary format conversion, resizing, and input organization. Claude Opus 4.8 (Max) and Claude Opus 5 (Max) have safety-filtered no-response cases on biochemistry questions involving viral biology and pathogen structural characterization; these cases are counted as incorrect under the strict binary scoring protocol.

SEE contains 1,116 multimodal questions with both explicit-option and open-answer formats. The raw dataset records associated visual files at the entry level; in evaluation, models are given the image inputs retained in the evaluation payload for each question. Model performance is measured by accuracy, where a question is counted as correct only if the final model answer matches the ground-truth answer under the scoring rules described below.

B.2 Prompting Protocol

We use separate prompts for answer generation and answer judging. The generation prompt asks each evaluated model to analyze the question and return a JSON object containing both the

analysis and the final answer. The judging prompt is executed by Gemini 3.1 Pro; it compares the candidate response with the reference answer and returns a binary correctness label.

Listing 1. Default generation prompt used in the evaluation pipeline.

```
[System]
You are an intelligent assistant. Please read the question and images
carefully, and provide the correct answer.

[User]
Please answer the following question. If it is a choice question, it may
be either single-choice or multiple-choice.

[Question text]:
{question}

[Options]:
{options}

Please first provide the analysis process, and then provide the final
answer. Strictly output the following JSON format, and do not include
Markdown formatting:
{
  "analysis": "your analysis process",
  "answer": "final answer"
}
```

Listing 2. Default judging prompt used in the evaluation pipeline.

```
[System]
You are a strict examiner. Please judge whether the student's answer is
consistent with the reference answer.

[User]
Please judge whether the following answer is correct:

[Question]:
{question}

[Reference answer]:
{reference}

[Student answer] (it may be a short answer, or a full response containing
an analysis process):
{candidate}

[Task]:
1. If the student's answer is a full response containing an analysis
   process, first identify the final answer from it, and then compare it
   with the reference answer.
2. Judge whether the core meaning is consistent. For choice questions,
```

```
the letters must be identical. For fill-in-the-blank or short-answer
questions, the numerical value or key phrase must be identical. Format
differences such as "5" and "5.0" are allowed.
3. If the reference answer is a numerical range or contains an error
tolerance, such as "5-10", "[0.8, 1.2]", "greater than 100",
"5 +/- 0.5", "about 0.05", or "<3.2", the student's answer is correct
as long as the numerical value falls within the range. Range
boundaries are also treated as correct (closed interval). The
student's answer may also be a range; in this case, judge whether the
two ranges are substantially consistent, with similar centers and
widths.
4. If the student refuses to answer or no answer can be identified, mark
it as incorrect.

Strictly output the following JSON format, and do not include Markdown
formatting:
{
  "correct": true or false,
  "reason": "judgment reason within 50 characters"
}
```

For models that support multimodal inputs, images are provided together with the question text in the same evaluation instance. No additional tool use, retrieval, or external browsing is allowed in the standard evaluation setting.

B.3 Answer Extraction and Scoring

Candidate responses are scored by Gemini 3.1 Pro using the judging prompt in Listing 2. For entries with explicit option lists, including both single-choice and multiple-choice questions, the judge first extracts the final answer from the model response and then checks whether the predicted option letters exactly match the ground-truth option letters. Partial matches are not counted as correct.

For entries without a separate option field, answers are evaluated against canonical short answers or numerical answers. Numerical answers are treated as correct when they match the reference value or fall within the reference range or tolerance specified in the answer. Model outputs that refuse to answer, state that the question cannot be determined, fail to provide a final answer, or otherwise cannot be judged as matching the ground-truth answer are counted as incorrect. Thus, all final evaluation results follow a strict binary scoring protocol: each question is either correct or incorrect.

B.4 Judge Sensitivity Analysis

Because Gemini 3.1 Pro serves as both the default judge model and one of the evaluated models, we conduct a sensitivity analysis to verify that scores are not biased by same-model alignment. We re-judge all responses of a representative subset of five evaluated models using GPT-5.5 as an independent judge, covering both standard and text-only (no-image) evaluation settings (10 runs,

~11,000 judgments in total).

Table 1 reports inter-judge agreement. Across all runs, raw agreement exceeds 99.5% and Cohen’s κ exceeds 0.989, indicating near-perfect concordance. The maximum accuracy difference between the two judges is 0.46 percentage points. Notably, for Gemini 3.1 Pro’s own responses, the alternative judge assigns a marginally *higher* accuracy (+0.18 pp), ruling out self-favoring bias in the original judge.

Manual inspection of the disagreement cases reveals that Gemini 3.1 Pro’s judgments are more consistent with the intended scoring protocol. The original judge applies stricter matching criteria aligned with our evaluation rules (e.g., requiring identifiable final answers rather than accepting truncated reasoning traces, and enforcing format requirements specified in the judging prompt), while GPT-5.5 more readily accepts semantically plausible but formally non-conforming outputs. Based on this human verification, we retain Gemini 3.1 Pro as the default judge. Its conservative tendency works against, rather than in favor of, any hypothetical same-model bias.

Table 1. Inter-judge agreement between Gemini 3.1 Pro (default) and GPT-5.5 (independent). Agreement is the fraction of identically scored questions. κ is Cohen’s kappa. Δ Acc. is GPT-5.5 judge accuracy minus Gemini 3.1 Pro judge accuracy, in percentage points. The asterisk on Claude Opus 4.8 (Max) denotes safety-filtered no-response cases on biochemistry questions, counted as incorrect under the strict binary scoring protocol.

Evaluated Model	Agreement (%)	Cohen’s κ	Δ Acc. (pp)
Claude Opus 4.8 (Max)*	99.81	0.996	+0.46
Seed2.0 Pro	99.64	0.991	+0.36
Gemini 3.1 Pro	99.82	0.996	+0.18
GPT-5.5 (xhigh)	99.82	0.996	+0.14
Kimi K2.6	99.72	0.993	+0.09

B.5 Tool-Augmented Visual-Agent Evaluation Protocol

To test whether iterative, tool-augmented inference improves performance under more pragmatic research-assistance conditions, we evaluate six models in a visual-agent environment with both web search and a code interpreter: GPT-5.6-Sol (Max), GPT-5.5 (xhigh), Claude Opus 5 (Max), Gemini 3.1 Pro, Qwen3.8-Max, and Gemini 3.5 Flash. For each model, we use the tool capabilities officially provided in its evaluation environment rather than third-party or custom-built substitutes. Each tool-augmented visual-agent run uses the same 1,116-question evaluation set as its corresponding standard baseline and receives the same question text, options, and image inputs.

The tool-augmented visual-agent setting differs from the standard baseline by allowing each model to form a multi-step interaction trajectory at inference time. At each step, the model may invoke web search to gather external context or use the code interpreter to inspect and manipulate the visual input, perform computation, or verify an intermediate interpretation. The resulting text, numerical output, or processed visual observation is returned to the model and may inform its next action or final answer. No local database, curated retrieval corpus, or domain-specific scientific tool is provided. Tool use is optional rather than forced, and the model may terminate the trajectory and answer directly whenever it judges the available evidence sufficient. This design

distinguishes the tool-augmented visual-agent setting from the standard baseline, which requires a final answer from the original multimodal input without external retrieval or code execution.

Table 2. Configuration of the tool-augmented visual-agent setting. Each model receives the same benchmark inputs as in the standard multimodal baseline and uses the web-search and code-interpreter capabilities officially provided in its evaluation environment for iterative evidence acquisition and verification during inference. The asterisk on Claude Opus 5 (Max) denotes no-response cases, which are counted as incorrect under the strict binary scoring protocol.

Component	Configuration
Evaluated models	Six MLLMs with complete paired standard and tool-augmented visual-agent runs: GPT-5.6-Sol (Max), GPT-5.5 (xhigh), Claude Opus 5 (Max)*, Gemini 3.1 Pro, Qwen3.8-Max, and Gemini 3.5 Flash.
Input payload	Same question text, options, and evaluation image inputs as the standard baseline evaluation.
Added tools	The web-search and code-interpreter capabilities officially provided in each model’s evaluation environment; no third-party or custom-built substitutes are used. No local database, curated retrieval corpus, or specialized scientific tool is enabled.
Tool policy	The evaluated model decides whether to call available tools; tool use is optional rather than required for every question.
Scoring	Same binary judging protocol as the baseline; stage failures, refusals, unanswered outputs, and outputs without an identifiable final answer are counted as incorrect.

The tool-augmented visual-agent evaluation contains two stages. In the first stage, the evaluated model analyzes the multimodal question and may construct a variable-length visual-agent trajectory: it selects a tool action, observes the execution result, and either continues gathering or verifying evidence or terminates with the structured answer object described above. In the second stage, only the final answer produced by the trajectory is evaluated, using the same binary correctness protocol as in the baseline experiments. Intermediate tool outputs are retained for trajectory-level diagnostics but are not scored independently.

For multiple-choice questions, the final option letters must match the ground-truth answer exactly. For short-answer and numerical questions, the answer is compared against the standardized canonical answer and any expert-defined tolerance. Tool-augmented execution failures, refusals, unanswered outputs, and outputs from which no final answer can be identified are counted as incorrect, matching the conservative treatment used in the standard evaluation.

Human-in-the-loop trajectory attribution.. To diagnose how tool use changes an answer, we compare each paired standard and tool-augmented visual-agent trajectory for which correctness changes between the two settings. We use a human-in-the-loop procedure in which an attribution model first performs a structured preliminary analysis using the question, reference answer, standard response, tool-augmented visual-agent response, and available tool-call trace. It identifies the key step associated with the improvement or deterioration, determines whether and how web search or code execution plays a substantive role, and proposes a mechanism category with supporting evidence. Domain experts then inspect the original responses and tool trace, verify the proposed attribution against the task-specific experimental context, and revise the category or rationale when necessary. Improvement mechanisms include code-assisted figure analysis, retrieval of key scientific facts, search-based confirmation, and code-based computation. Tool-related

detrimental mechanisms are organized into four recurring failure categories: action-selection failure, observation-interpretation failure, evidence-integration failure, and endless overthinking. Their relationship to the fine-grained attribution types is summarized in Table 3. The resulting expert-validated attribution is used as a diagnostic analysis of the interaction trajectory rather than as an additional correctness score.

Table 3. Mapping between the recurring categories of tool-related failure and their fine-grained manifestations in the trajectory attribution analysis.

Failure category	Fine-grained manifestations
Action-selection failure	Inappropriate use of code, including an unsuitable image-processing method, data-processing assumption, calculation target, or retrieval direction.
Observation-interpretation failure	Code-logic or computation error; tool-mediated misreading of figures, crops, measurements, or computed outputs.
Evidence-integration failure	Over-reliance on retrieved information; noisy or misleading retrieval; inappropriate weighting of tool-derived evidence relative to the question-specific experimental evidence.
Endless overthinking	Excessive tool use and context degradation; unproductive repetition; failure to stop or produce a final answer.

Across the six models, 990 paired instances change correctness between the standard and tool-augmented settings: 648 change from incorrect to correct and 342 from correct to incorrect, yielding a net gain of 306 correct predictions. Trajectory attribution identifies substantive tool contributions in 411 improvements and 145 regressions. Among the remaining changes, some occur without tool calls, while others follow tool use but lack evidence that a specific tool output decisively changes the final answer; these cases are therefore classified primarily as variation in reasoning or visual interpretation.

Table 4. Paired correctness transitions and trajectory attribution by model. Wrong→Correct and Correct→Wrong report all answer changes between the standard and tool-augmented visual-agent settings. Tool-attributed improvements and regressions report the subsets for which web search or code execution is identified as playing a substantive role. Percentages are calculated relative to the corresponding number of improvements or regressions for each model.

Model	Wrong→Correct	Correct→Wrong	Net change	Tool-attributed improvements	Tool-attributed regressions
Claude Opus 5 (Max)	80	38	+42	24 (30.0%)	8 (21.1%)
Gemini 3.1 Pro	95	74	+21	24 (25.3%)	15 (20.3%)
Gemini 3.5 Flash	132	77	+55	75 (56.8%)	28 (36.4%)
GPT-5.5 (xhigh)	127	42	+85	103 (81.1%)	19 (45.2%)
GPT-5.6-Sol (Max)	113	68	+45	101 (89.4%)	50 (73.5%)
Qwen3.8-Max	101	43	+58	84 (83.2%)	25 (58.1%)
Overall	648	342	+306	411 (63.4%)	145 (42.4%)

B.6 Public-Subset Reproducibility

The main evaluation uses all 1,116 questions in **SEE**, including 67 withheld questions based on unpublished experimental data. To support reproducibility on the released benchmark, we recompute headline results on the 1,049 publicly released questions marked as public in the `open_source` field of the dataset metadata. Public-subset accuracies closely track the full-set

results, indicating that the reported conclusions are not driven by the withheld subset.

Table 5. Full-set and public-subset accuracy under the standard no-tool evaluation setting. The public subset contains the 1,049 released questions; the full set contains all 1,116 questions. Delta is public-subset accuracy minus full-set accuracy, in percentage points. Asterisks on Claude Opus 4.8 (Max) and Claude Opus 5 (Max) denote safety-filtered no-response cases on biochemistry questions, counted as incorrect under the strict binary scoring protocol.

Model	Full Correct	Full Acc. (%)	Public Correct	Public Acc. (%)	Delta (pp)
Gemini 3.1 Pro	504/1,116	45.2	478/1,049	45.6	+0.4
GPT-5.6-Sol (Max)	543/1,116	48.7	503/1,049	48.0	-0.7
Claude Opus 5 (Max)*	495/1,116	44.4	465/1,049	44.3	-0.0
Qwen3.8-Max	466/1,116	41.8	439/1,049	41.8	+0.1
GPT-5.5 (xhigh)	461/1,116	41.3	433/1,049	41.3	-0.0
Kimi K3	432/1,116	38.7	403/1,049	38.4	-0.3
Gemini 3.5 Flash	431/1,116	38.6	410/1,049	39.1	+0.5
Claude Opus 4.8 (Max)*	406/1,116	36.4	377/1,049	35.9	-0.4
Seed2.1 Pro	377/1,116	33.8	350/1,049	33.4	-0.4
Qwen3.7-Plus	365/1,116	32.7	346/1,049	33.0	+0.3
Intern-S2 Preview-397B	324/1,116	29.0	302/1,049	28.8	-0.2
Seed2.0 Pro	315/1,116	28.2	298/1,049	28.4	+0.2
MiniMax-M3	314/1,116	28.1	302/1,049	28.8	+0.7
Kimi K2.6	286/1,116	25.6	266/1,049	25.4	-0.3
Intern-S2 Preview	266/1,116	23.8	252/1,049	24.0	+0.2
GLM-5V-Turbo	259/1,116	23.2	245/1,049	23.4	+0.1
Intern-S1 Pro	225/1,116	20.2	215/1,049	20.5	+0.3
MiMo-V2.5	196/1,116	17.6	185/1,049	17.6	+0.1
S1-VL	177/1,116	15.9	169/1,049	16.1	+0.3

Table 6. Public-subset accuracy under the combined web-search-and-code-interpreter setting. Results are recomputed on the same 1,049 publicly released questions used in Table 5. The asterisk on Claude Opus 5 (Max) denotes no-response cases, which are counted as incorrect under the strict binary scoring protocol.

Model	Correct	Public Acc. (%)
GPT-5.6-Sol (Max)	547/1,049	52.1
GPT-5.5 (xhigh)	511/1,049	48.7
Claude Opus 5 (Max)*	502/1,049	47.9
Gemini 3.1 Pro	497/1,049	47.4
Qwen3.8-Max	488/1,049	46.5
Gemini 3.5 Flash	459/1,049	43.8

B.7 Text-Only Ablation

To examine whether models recognize missing visual evidence, we conduct a text-only ablation by removing all evaluation image inputs from the questions while retaining the textual question content. This setting is designed to test whether models can avoid unsupported conclusions when key information is missing.

Figure 9 reports the paired accuracies for each model; alongside the mean drop, the spread across models narrows from 31.1 to 14.0 points.

Across the models with available no-image runs, explicit missing-image acknowledgment is consistently rare (Table 7). Across 18 text-only runs, only 921 out of 20,088 instances (4.6%) are

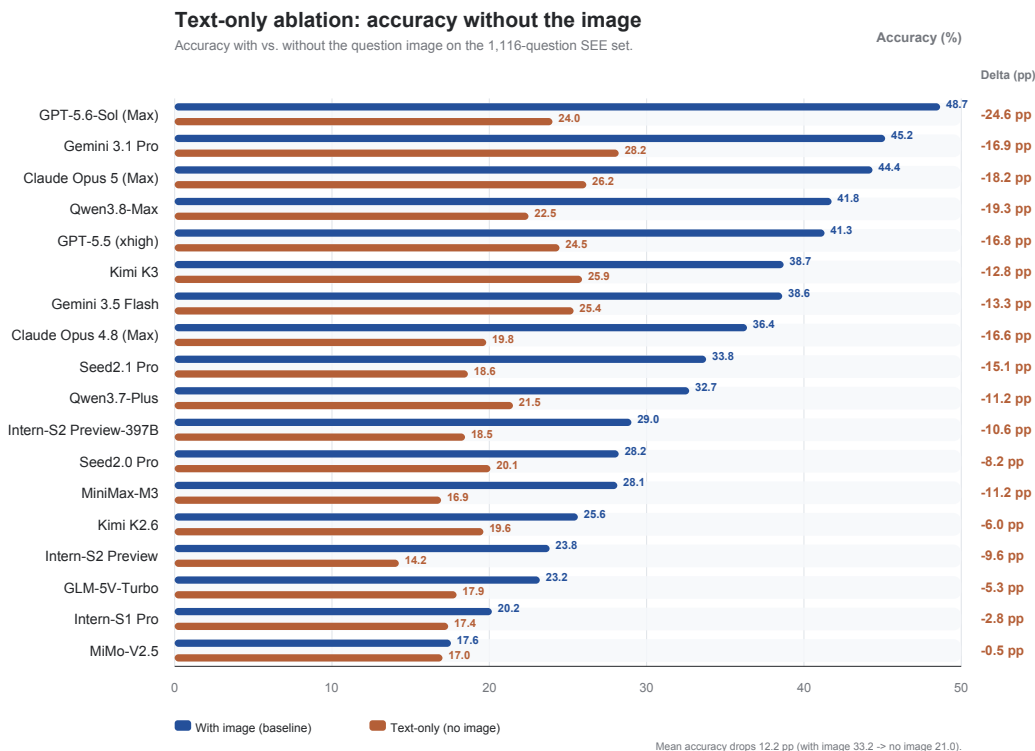


Figure 9. Accuracy with and without the question image for the 18 models with paired runs. Each model is scored twice on the same question set; the text-only run receives the identical question with the image removed. Accuracy follows the strict binary protocol, in which every instance stays in the denominator and refusals, non-answers, and generation errors count as incorrect. Delta is the text-only accuracy minus the with-image accuracy.

classified as explicit acknowledgments that necessary image evidence is missing. This pattern indicates that current MLLMs generally lack robust evidence-boundary awareness under visually under-specified conditions.

In this ablation setting, only explicit missing-image acknowledgments are treated as evidence-aware behavior for the main analysis. Empty outputs, generation errors, generic refusals, and other non-answer cases are not counted toward this metric because they do not directly show recognition that visual evidence is unavailable. Unsupported final answers are analyzed as evidence-insensitive attempts.

B.8 Per-Discipline Results

We report model performance across the three main disciplines (chemistry, biology, materials) and the final set of 17 fine-grained discipline labels used for reported sub-discipline analysis. Because **SEE** uses multi-label annotation, a single question may contribute to multiple labels; counts and accuracies follow label-membership semantics.

Table 8 reports per-sub-discipline accuracy for all 19 evaluated models, ordered by overall accuracy.

B.9 Subject Label Coverage

SEE exhibits pronounced multi-label and cross-disciplinary characteristics. Under the final

Table 7. Explicit missing-image acknowledgments in the text-only ablation. The denominator is all text-only instances for each model; the numerator includes only responses with `refusal_type=no_image`, i.e., responses explicitly classified as acknowledging missing image or visual evidence. Empty outputs, generation errors, generic refusals, and other non-answer cases are not counted.

Model	Total	Explicit Missing-Image Ack.	Rate (%)
GPT-5.6-Sol (Max)	1,116	167	15.0
Kimi K2.6	1,116	163	14.6
GPT-5.5 (xhigh)	1,116	146	13.1
Intern-S2 Preview-397B	1,116	90	8.1
Qwen3.8-Max	1,116	75	6.7
MiniMax-M3	1,116	61	5.5
Intern-S2 Preview	1,116	55	4.9
MiMo-V2.5	1,116	37	3.3
Qwen3.7-Plus	1,116	30	2.7
Kimi K3	1,116	27	2.4
Seed2.0 Pro	1,116	20	1.8
GLM-5V-Turbo	1,116	17	1.5
Gemini 3.1 Pro	1,116	14	1.3
Gemini 3.5 Flash	1,116	7	0.6
Intern-S1 Pro	1,116	6	0.5
Claude Opus 4.8 (Max)*	1,116	4	0.4
Seed2.1 Pro	1,116	2	0.2
Claude Opus 5 (Max)*	1,116	0	0.0
Total / Mean	20,088	921	4.6

*Safety-filtered no-response cases, counted as incorrect under the strict binary scoring protocol.

Table 8. Per-model accuracy by reported sub-discipline for all 19 evaluated models, ordered by overall accuracy. Values are percentages; bold indicates the best model for each label. Asterisks on Claude Opus 4.8 (Max) and Claude Opus 5 (Max) denote no-response cases triggered by safety filters on biochemistry questions involving viral biology and pathogen structural characterization; under the strict binary scoring protocol, these instances are counted as incorrect.

Main Discipline	Sub-Discipline	Model Accuracy (%)																		
		GPT-5.6-Sol (Max)	Gemini 3.1 Pro	Claude Opus 5 (Max)*	Qwen3.8-Max	GPT-5.5 (xhigh)	Kimi K3	Gemini 3.5 Flash	Claude Opus 4.8 (Max)*	Seed2.1 Pro	Qwen3.7-Plus	Intern-S2 Preview-397B	Seed2.0 Pro	MiniMax-M3	Kimi K2.6	Intern-S2 Preview	GLM-5V-Turbo	Intern-S1 Pro	MiMo V2.5	SL-12
Chemistry	Analytical Chemistry	50.5	48.2	49.5	42.3	42.5	39.5	42.5	40.9	36.4	33.9	30.2	28.4	31.1	25.9	28.6	24.1	23.4	17.5	19.3
	Physical Chemistry	41.8	46.0	42.1	31.8	35.3	32.6	39.2	31.2	25.8	27.6	28.5	21.1	27.0	24.6	27.9	19.6	23.1	13.9	16.6
	Organic Chemistry	53.0	50.9	53.9	46.1	46.1	40.0	47.4	49.6	40.9	38.7	35.2	30.9	37.0	27.0	31.7	25.7	23.5	20.4	22.2
	Polymer Chemistry and Physics	37.9	42.5	40.8	32.2	35.1	37.4	35.1	29.9	24.1	21.8	24.7	21.3	29.9	28.2	27.6	14.9	28.7	15.5	17.8
	Inorganic Chemistry	44.5	39.1	45.5	35.5	32.7	30.9	32.7	29.1	31.8	26.4	20.0	16.4	30.0	23.6	30.0	20.9	23.6	10.0	14.5
Biology	Biochemistry	47.4	44.8	42.7	42.7	43.6	40.4	34.0	35.8	33.4	33.7	29.1	32.3	23.5	23.8	17.7	22.7	14.2	16.9	11.0
	Molecular Biology	48.3	44.1	39.1	47.5	43.3	42.1	32.6	34.5	36.4	35.6	28.7	33.7	21.5	24.9	13.8	23.8	13.0	19.5	10.7
	Cell Biology	53.8	43.4	39.0	48.4	46.2	44.5	34.6	36.3	39.0	36.3	28.6	39.6	26.9	23.6	17.0	24.7	13.2	22.5	11.0
	Structural Biology	48.8	42.2	34.3	42.2	41.0	41.0	28.3	33.1	31.3	33.1	26.5	30.1	21.1	27.1	14.5	25.3	21.1	17.5	12.7
	Biophysics	47.4	44.9	38.5	36.5	38.5	39.1	37.2	30.1	30.8	32.7	29.5	30.1	20.5	23.1	21.8	26.3	18.6	18.6	16.7
	Genetics	47.2	61.1	44.4	58.3	52.8	38.9	25.0	27.8	25.0	41.7	19.4	41.7	22.2	27.8	2.8	19.4	2.8	16.7	2.8
	Immunology	56.2	31.2	31.2	62.5	43.8	43.8	34.4	34.4	40.6	37.5	25.0	40.6	28.1	12.5	31.2	31.2	31.2	18.8	12.5
	Physiology	42.1	42.1	52.6	47.4	36.8	31.6	36.8	26.3	31.6	31.6	21.1	26.3	0.0	15.8	31.6	21.1	10.5	26.3	15.8
Materials	Organic and Polymer Materials	37.2	39.9	34.6	25.0	30.3	29.8	32.4	27.1	22.3	17.0	20.2	14.9	25.0	26.1	25.5	15.4	28.7	16.0	18.6
	Inorganic Non-metallic Materials	46.6	43.7	44.7	37.9	35.9	33.0	39.8	33.0	34.0	35.0	24.3	22.3	28.2	24.3	30.1	26.2	19.4	14.6	16.5
	Composite Materials	39.7	41.2	47.1	29.4	32.4	35.3	35.3	32.4	23.5	22.1	32.4	25.0	27.9	25.0	26.5	14.7	26.5	13.2	16.2
	Metallic Materials	57.1	44.9	42.9	55.1	36.7	32.7	49.0	32.7	36.7	36.7	34.7	30.6	34.7	26.5	34.7	26.5	18.4	12.2	18.4

17-label reporting taxonomy, all 1,116 questions are covered by the reported fine-grained labels. The overall model evaluation is therefore based on the same 1,116-question set used for reported label coverage.

Cross-Discipline Composition. Under the 17-label reporting taxonomy, questions labeled with biology only total 370 (33.2%); questions jointly labeled with chemistry and materials total 346

Table 9. Mean model accuracy across 19 evaluated models for each reported fine-grained discipline label. Labels with smaller sample sizes (e.g., Immunology, Physiology, Genetics) should be interpreted with care due to their limited support.

Sub-discipline	Main Discipline	# Questions	Mean Acc. (%)
Organic Chemistry	Chemistry	230	37.9
Metallic Materials	Materials	49	34.8
Analytical Chemistry	Chemistry	440	34.5
Immunology	Biology	32	34.0
Cell Biology	Biology	182	33.1
Molecular Biology	Biology	261	31.2
Biochemistry	Biology	344	31.0
Inorganic Non-metallic Materials	Materials	103	31.0
Biophysics	Biology	156	30.6
Genetics	Biology	36	30.4
Structural Biology	Biology	166	30.1
Physical Chemistry	Chemistry	337	29.3
Physiology	Biology	19	28.8
Composite Materials	Materials	68	28.7
Polymer Chemistry and Physics	Chemistry	174	28.7
Inorganic Chemistry	Chemistry	110	28.3
Organic and Polymer Materials	Materials	188	25.6

Table 10. Fine-grained discipline-label coverage under the 17-label reporting taxonomy.

Fine-Grained Label	Main Discipline	# Questions
Analytical Chemistry	Chemistry	440
Biochemistry	Biology	344
Physical Chemistry	Chemistry	337
Molecular Biology	Biology	261
Organic Chemistry	Chemistry	230
Organic and Polymer Materials	Materials	188
Cell Biology	Biology	182
Polymer Chemistry and Physics	Chemistry	174
Structural Biology	Biology	166
Biophysics	Biology	156
Inorganic Chemistry	Chemistry	110
Inorganic Non-metallic Materials	Materials	103
Composite Materials	Materials	68
Metallic Materials	Materials	49
Genetics	Biology	36
Immunology	Biology	32
Physiology	Biology	19

Table 11. Distribution of reported subject-label cardinality among questions covered by the 17-label taxonomy. Most questions span multiple sub-disciplines, reflecting the cross-concept, cross-direction, and cross-discipline nature of real experimental scenarios rather than isolated single-topic knowledge.

Labels per Question	# Questions	Share (%)
1	96	8.6
2	416	37.3
3	458	41.0
4	137	12.3
5	9	0.8

Table 12. Main-discipline coverage by question count and by label annotation count. Chemistry and biology coverage is high; materials coverage is comparatively lower.

Main Discipline	# Questions	Coverage (%)	# Label Annotations	Share of Annotations (%)
Chemistry	732	65.6	1,291	44.6
Biology	538	48.2	1,196	41.3
Materials	363	32.5	408	14.1

(31.0%); chemistry-only questions total 219 (19.6%); and chemistry-with-biology questions total 164 (14.7%). Materials-only questions total just 13 (1.2%). A small number of questions combine all three main disciplines (3 questions, 0.3%) or biology with materials only (1 question, 0.1%). This distribution matches the strong overlap between materials science and chemistry in real experimental practice, and accordingly model accuracy on materials labels should be interpreted as reflecting joint chemistry–materials reasoning rather than isolated materials knowledge.

Long-Tail Fine-Grained Labels. Analytical chemistry (440 questions) is the most frequent reported fine-grained label, followed by biochemistry (344) and physical chemistry (337). At the tail, metallic materials, genetics, immunology, and physiology have only 49, 36, 32, and 19 questions respectively. Accuracy estimates on tail labels are more sensitive to a small number of questions and should be read with sample size in mind.