Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Jinghang Wang

wjh2018@mail.ustc.edu.cn

Alibaba Group

Taolin Han

Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences

Yuchen Zhang

School of Basic Medical Sciences, Tsinghua University

Yun Wu

Alibaba Group

Wai Yuet Chiu

Alibaba Group

Zhaohai Li

Qwen Team, Alibaba Group

Yifei Zhang

Department of Chemistry, University of Alberta

Jinxin Wang

Department of Chemistry, Tsinghua University

Yuhao Zhou

College of Animal Sciences, Zhejiang University

Chen Zhao

Department of Chemistry, Tsinghua University

Jiajia Li

Alibaba Group

Jiaxin Li

Alibaba Group

Qile Jin

Alibaba Group

Kewei Sun

Alibaba Group

Shuang Wu

Alibaba Group

Weiqi Zhai

Alibaba Group

Renquan Lv

School of Chemical Engineering and Bioengineering, Zhejiang University

Junchao Li

Alibaba Group

Ruodan Chen

Alibaba Group

Qingteng Chen

Alibaba Group

Zhibo Yang

Qwen Team, Alibaba Group

Hu Wei

Alibaba Group

Lin Qu

Alibaba Group

Shuai Bai

Qwen Team, Alibaba Group

Bing Zhao

Alibaba Group

Article

Keywords:

Posted Date: October 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10734118/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

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

Taolin Han[1,3*], Yuchen Zhang[1,4*], Jinghang Wang[1*], Yun Wu[1*], Wai Yuet Chiu[1], Zhaohai Li[2], Yifei Zhang[1,5], Jinxin Wang[1,6], Yuhao Zhou[1,7], Chen Zhao[1,6], Jiajia Li[1], Jiaxin Li[1], Qile Jin[1], Kewei Sun[1], Shuang Wu[1], Weiqi Zhai[1], Renquan Lv[1,8], Junchao Li[1], Ruodan Chen[1], Qingteng Chen[1], Zhibo Yang[2], Hu Wei[1], Lin Qu[1], Shuai Bai[2,†], Bing Zhao[1,†]

[1]Alibaba Group; [2]Qwen Team, Alibaba Group; [3]Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences; [4]School of Basic Medical Sciences, Tsinghua University; [5]Department of Chemistry, University of Alberta; [6]Department of Chemistry, Tsinghua University; [7]College of Animal Sciences, Zhejiang University; [8]School of Chemical Engineering and Bioengineering, Zhejiang University.

[*]These authors contributed equally. [†]Corresponding authors

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 18 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.

Large language models (LLMs) and their agentic extensions are increasingly used as assistants for scientific discovery[1–3]. Current evaluations frame science capability along either disciplinary knowledge or procedural research skill. The former uses text-only suites to score knowledge breadth over curated closed-ended items[4–7]. Expert-level and domain-specific tests raise the difficulty of those items with frontier academic examinations and problems drawn from literature[8–11], while Olympiad and exam-grade collections push the difficulty even further[12–15]. The latter requires agentic systems to assist or execute research workflows, from tool-augmented chemical experimentation and autonomous laboratories to end-to-end paper production[16–21]. Their evaluations reward procedures, including whether a plan was executed, a tool call was successful, or a workflow ended with an answer[22]. Even benchmarks that claimed to be built closer to laboratory practice mainly ask whether a domain-specific answer is correct[23–26]. High

scores in these protocols certify knowledge coverage and procedural fluency[27], but do not certify the reasoning process that research depends on[28].

The distinction at issue is not one of difficulty but of what the task demands. In the laboratory, conclusions are assembled from imperfect observations, including spectra, gel images, or diffraction patterns, together with tabular measurements and textual context. Interpreting such evidence correctly is an expert skill[29]. A critical judgment, moreover, lies in determining which conclusions the observations can support and where the limits of that support lie. Existing protocols bypass this judgment. When an item is scored only against a single verified answer, a correct and evidence-bounded conclusion is indistinguishable from one that is correct despite overreaching reasoning. LLMs assert unsupported content fluently with poor calibration[30–32]. What remains unmeasured, in both static and agentic settings, is the evidentiary boundary, i.e. whether a model reasons within the limits of the data it was given.

To close this gap, we introduce Science Edge Evaluation (SEE), a multimodal benchmark constructed from authentic scientific tasks in chemistry, biology, and materials science. Its questions are adapted from peer-reviewed literature and experimental practice such that every answer must be licensed by the observations provided; SEE thus scores not only whether a model reaches the correct conclusion, but also whether that conclusion remains within the bounds of the available evidence. Our evaluation reveals that current multimodal LLMs (MLLMs) remain unreliable in realistic experimental settings and that their failures cannot be attributed to limited knowledge access alone. Rather, SEE exposes a fundamental deficiency in evidence-based scientific reasoning. A complementary analysis of tool-augmented visual agents further shows that tool access expands evidence acquisition but does not obviate the need for reliable evidence management across the interaction trajectory. In short, current models struggle to "SEE" multimodal observations across disciplines while maintaining rigorous adherence to the evidence at hand. By aligning evaluation with laboratory practice, SEE measures not only whether models possess scientific knowledge, but whether they can derive justified evidence-bounded conclusions from multimodal experimental data.

Results

Current MLLMs show limited reliability on SEE

SEE contains 1,116 expert-curated questions derived from real scientific tasks in chemistry, biology and materials science, with 1,049 questions publicly released (Figure 1a). The benchmark covers 17 reported sub-fields across three disciplines (Figure 1b) and includes both single-discipline and cross-disciplinary tasks (Figure 1c). Each question pairs textual context with multimodal experimental evidence, including figures, spectra, microscopy images, tables, measurement results and other laboratory data (Figure 1d); representative examples are shown in Extended Data Fig. 1. This design allows us to evaluate whether MLLMs can reason from experimental evidence rather than rely only on scientific knowledge or text-solvable cues.

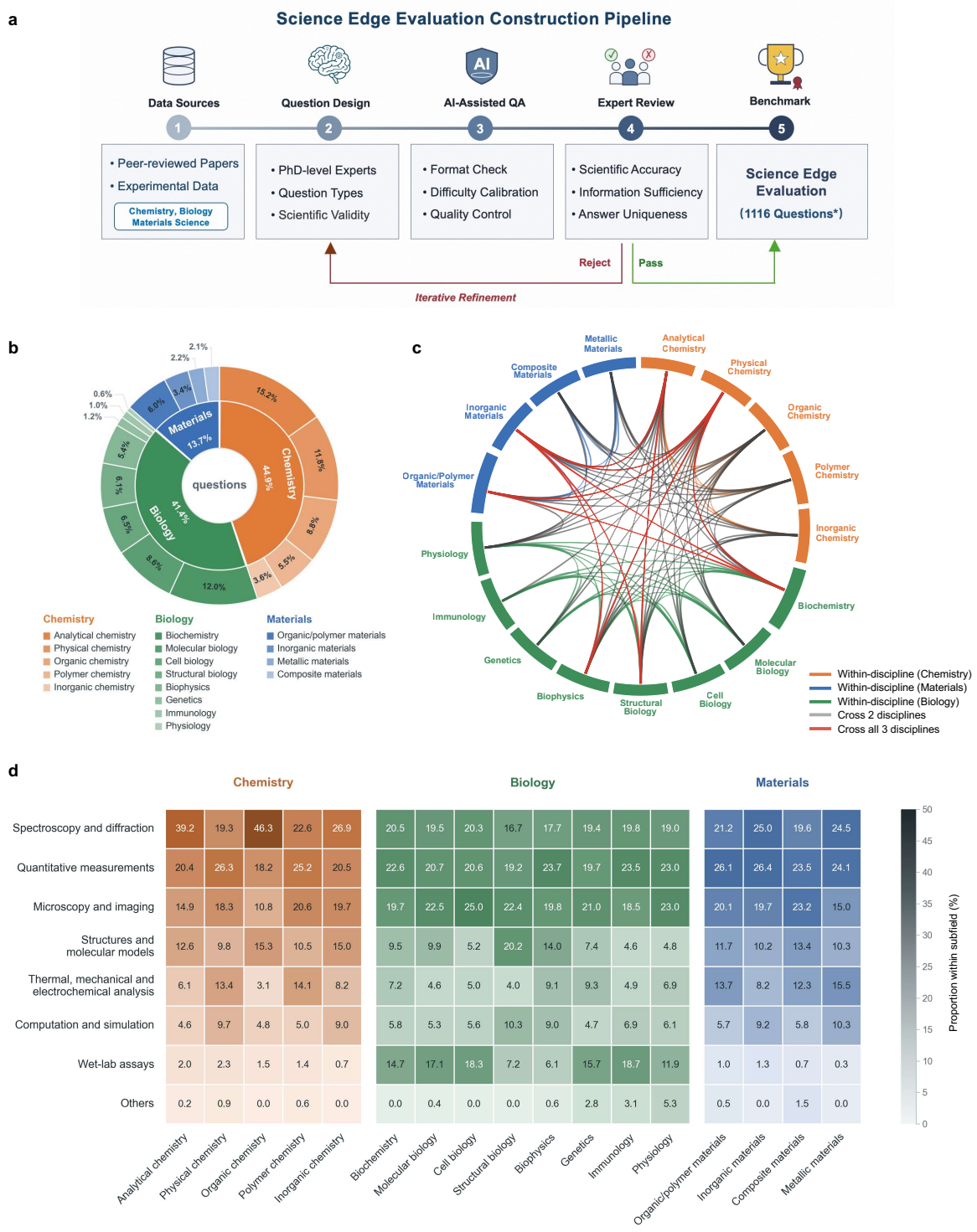


Figure 1. Overview of SEE construction and composition. **a**, Construction pipeline for the 1,116-question benchmark. Questions were generated from literature or real experimental scenarios, standardized with AI-assisted checks, reviewed by domain experts and collected after quality-control criteria were met. *1,049 questions are publicly released and 67 are withheld (contain unpublished data). **b**, Question distribution across three disciplines and 17 sub-fields. **c**, Co-occurrence of sub-field labels. Edges connect sub-field labels assigned to the same question, with thickness proportional to the number of shared questions and color indicating within- or cross-discipline overlap. **d**, Evidence types across sub-fields. Rows show eight real laboratory evidence types, and columns show sub-fields grouped by discipline.

Following the evaluation protocol described in the Methods, we first report the overall accuracy of 18 representative MLLMs on SEE. All evaluated models achieve low accuracy on SEE in the standard evaluation setting (Figure 2a). The best model, GPT-5.6-Sol (Max), reaches 48.7% accuracy, and no model exceeds 50%. Across the 18 models, accuracy ranges from 17.6% to 48.7%.

The result at the discipline-level shows similar aggregate precision in all three main disciplines (Table S8 in Supplementary Information S2.8). Across all evaluated models, average accuracy is 33.3% for chemistry, 32.3% for biology, and 29.6% for materials science.

Performance differences become clearer at the sub-discipline level (Figure 2b; full per-model breakdown in Table S8). The lowest mean accuracies occur in organic and polymer materials (26.0%), inorganic chemistry (29.0%), polymer chemistry and physics (29.3%), and composite materials (29.4%). The highest reported sub-discipline accuracy is organic chemistry (38.8%), followed by metallic materials (35.7%), analytical chemistry (35.3%), and immunology (35.2%). 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.

General-purpose models outperform 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 three science-specialized models remain below the general-purpose average on SEE (Figure 2a), with an average accuracy of 24.3% 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[33]. 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

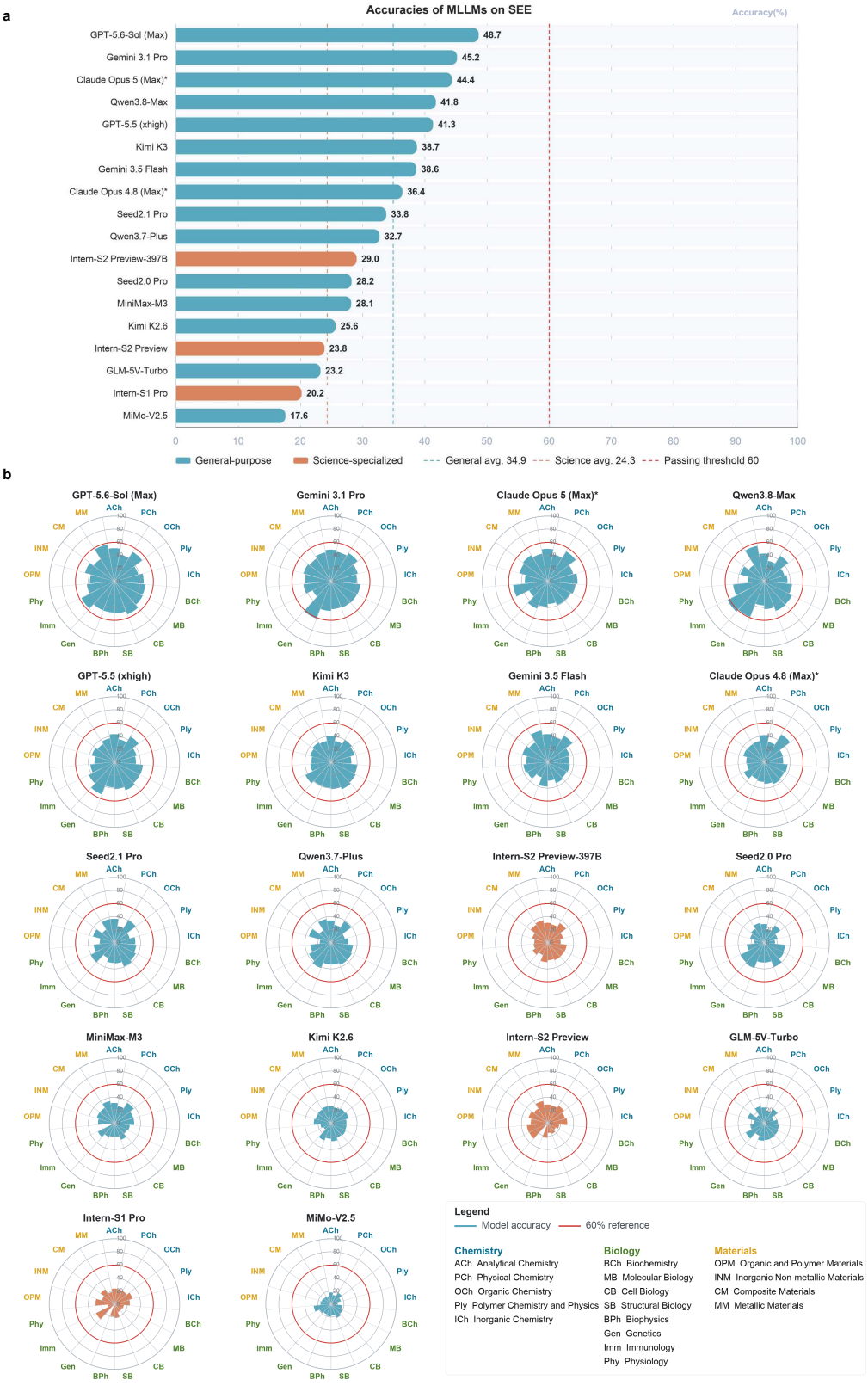


Figure 2. Model performance on SEE across overall and discipline-level evaluations. **a**, Overall accuracies of the evaluated MLLMs on SEE. General-purpose and science-specialized models are shown in blue and brown, respectively. Dashed vertical lines indicate the mean accuracy within each model category, with colors matching the corresponding bars. **b**, Discipline-level results for all 18 evaluated models, ordered by overall accuracy. Each axis reports accuracy on one reported sub-discipline label, with axis labels and spokes colored by broad discipline. Concentric rings mark 20%, 40%, 60%, 80% and 100% accuracy, and the red ring highlights 60%. Asterisks indicate safety-filtered no-response cases involving viral biology and pathogen structural characterization; these cases are counted as incorrect.

depends not only on domain adaptation, but also on broader multimodal understanding and flexible evidence-grounded reasoning.

Interdisciplinary complexity reduces model performance

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 36.1% on questions within a single main discipline, but drops to 29.5% on questions that span multiple main disciplines. This 6.6% 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 37.4% and 36.5%, respectively. However, the accuracy drops to 30.4% for questions with three labels and further to 29.0% 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.

Visual ablation exposes weak evidence-boundary awareness in MLLM

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 cases (4.6%, Figure 3). This remarkably low rate suggests that current MLLMs are prone to blind hallucination, attempting to derive scientific answers from linguistic

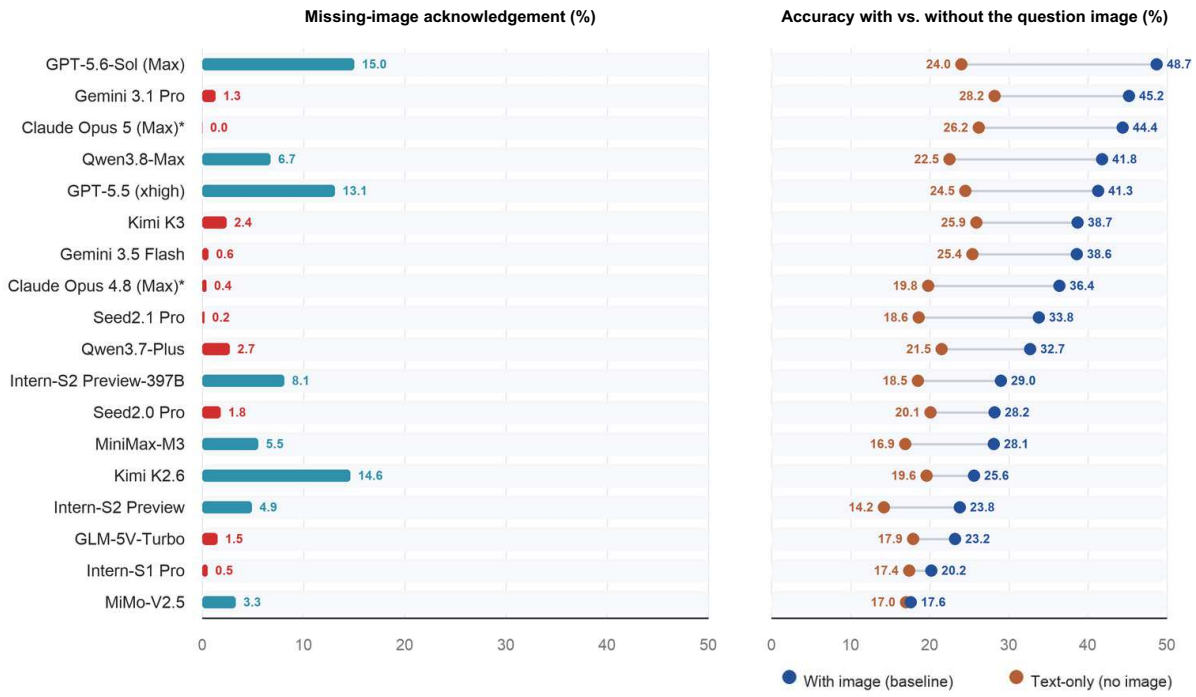


Figure 3. Text-only ablation results for missing-image acknowledgment and accuracy. Left panel, explicit missing-image acknowledgment rates in the text-only ablation test. The denominator is the number of text-only instances for each model. The numerator is the number of responses that explicitly recognize missing image or visual evidence. Empty outputs, generation errors, generic refusals, and other non-answer cases are not counted. Right panel, accuracy comparison between the original image-conditioned setting and the text-only setting after removing the question image. Blue points indicate the image-conditioned baseline, and brown points indicate the text-only setting. Model order is matched between panels. Asterisks denote safety-filtered no-response cases counted under the strict evaluation protocol.

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 S2.7 (Figure S1).

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.

Model failures reveal perceptual and inferential weaknesses

To understand why current MLLMs fail on SEE, we further inspect representative incorrect responses and identify five recurring failure modes (Figure 4). These errors fall into two categories: the first two are failures in perception, where models do not faithfully extract or use visual evidence, and the next two are failures in inference, where models cannot maintain evidence-

constrained reasoning over the available observations, measurements and context. Remaining failure modes that cannot be categorized are grouped as other errors. These errors are mainly due to inability to infer the expected number of correct options in multiple-choice questions.

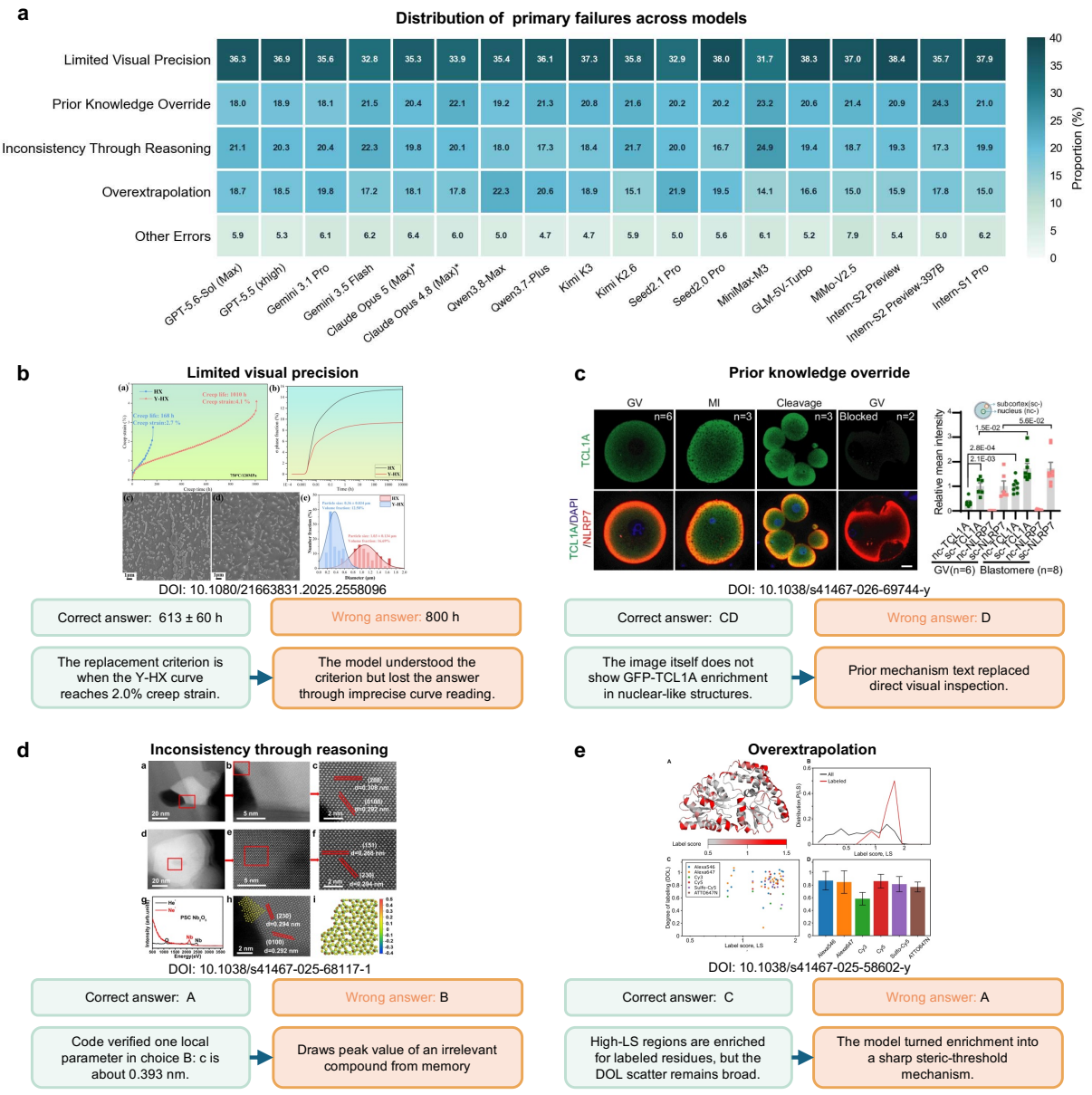


Figure 4. Failure-mode distribution and representative error cases on SEE. **a**, Heatmap showing the proportions of five primary failure modes across all 18 models. Each incorrect response is assigned to one primary failure mode, and percentages indicate the proportion of categorized errors for each model. Asterisks indicate safety-filtered no-response cases involving viral biology and pathogen structural characterization; these cases are counted as incorrect. **b**, Limited visual precision arises when a model identifies the relevant criterion but cannot accurately read quantitative information from the figure. **c**, Prior knowledge override occurs when a model substitutes a familiar scientific mechanism for direct inspection of the visual evidence. **d**, Inconsistency through reasoning occurs when a model fails to integrate local observations into a coherent global conclusion and instead selects an irrelevant memorized feature. **e**, Overextrapolation occurs when a model turns partial or broad experimental evidence into a stronger mechanistic conclusion than the data support. Green boxes indicate the evidence supporting the correct answer, and orange boxes indicate the model's wrong answer and failure pattern.

Limited Visual 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 Override 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.

Inconsistency Through Reasoning 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 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.

These failure modes indicate that current MLLMs fail in SEE not due to a single deficiency, but because perception and inference break down jointly. Visual imprecision, prior-driven interpretation, and weak evidence control often interact, so small errors in reading or selecting evidence cascade into incorrect scientific conclusions. Simply scaling data or model size is therefore unlikely to resolve this gap, because scientific discovery often lies precisely beyond the distribution of previously seen examples.

Tool-augmented visual agents shift the bottleneck to evidence management

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 in a tool-augmented visual-agent setting (Figure 5a). 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, substantial errors remain, indicating that tool access alone does not make MLLMs reliable on SEE. Detailed configuration is provided in the Supplementary Information S2.5.

Trajectory analysis shows that tool use has bidirectional effects on model outputs (Figure 5b). Across the six models, 411 corrections are substantively attributed to tool use. Code-assisted image analysis accounts for 51.2% of these cases, fact retrieval for 37.8%, information cross-checking for 7.0%, and code-assisted computation for 4.1%. 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 (32.5%), action selection failures (31.7%), observation interpretation failures (22.2%), and endless overthinking (13.5%). Both beneficial and detrimental effects of tools occur in the tool-use trajectory (Figure 5b). 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 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.

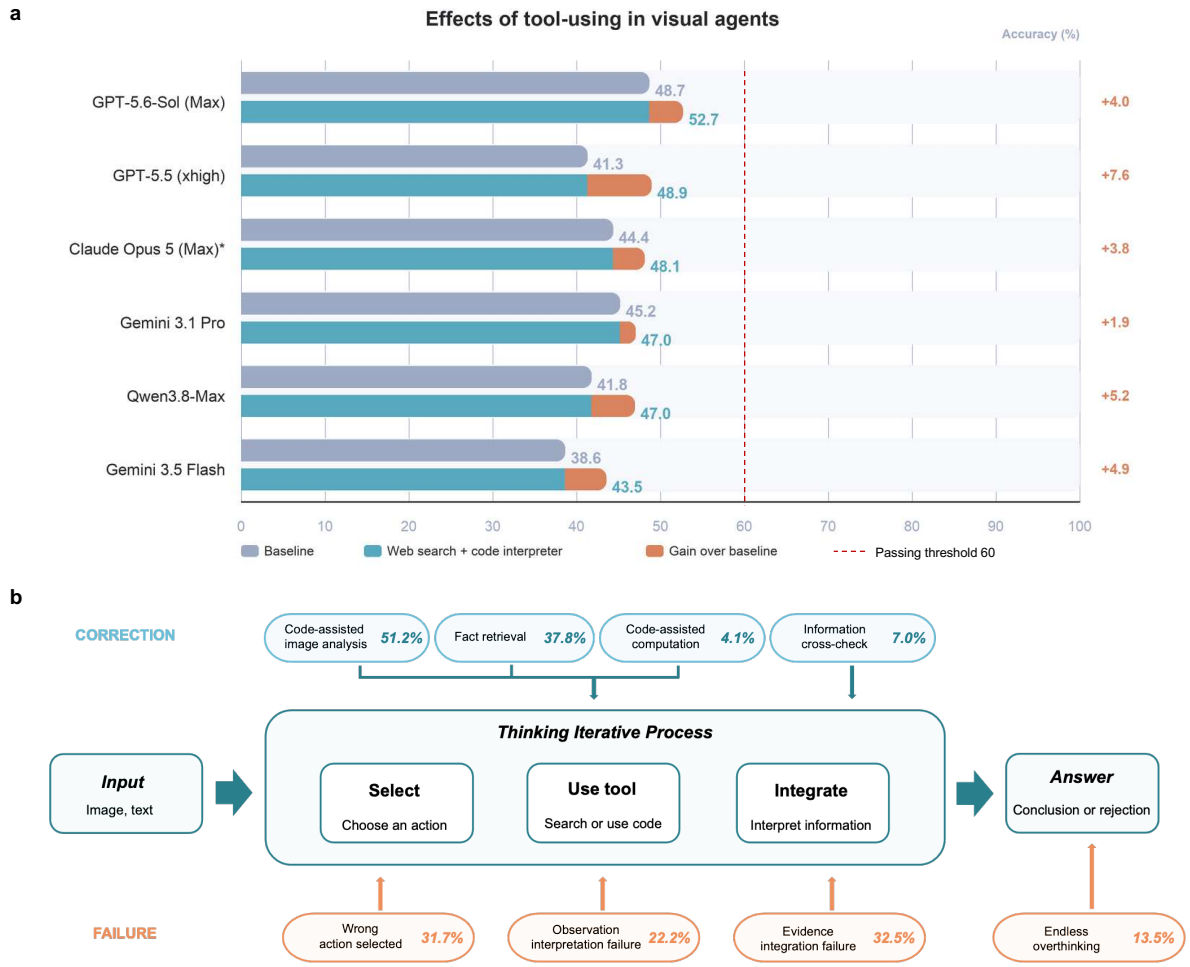


Figure 5. Tool-augmented visual-agent evaluation on SEE. **a**, Comparison between standard multimodal inference and the tool-augmented visual-agent setting, in which models can iteratively use web search and a code interpreter before returning a final answer. Paired bars show the accuracy of each model 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. **b**, Schematic of tool effects during tool-mediated visual-agent reasoning. Corrections arise 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.

Discussion

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. In the standard inference setting, the best-performing model among the 18 MLLMs evaluated reaches only 48.7% accuracy, and no model exceeds 50%. This low performance should not be interpreted simply as another difficult benchmark result. Instead, it reveals a mismatch between the abilities captured by many existing scientific benchmarks and the abilities required for the 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. The access of tools boosts accuracy to 52.7% (strongest model), showing that the 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 to 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.

Methods

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 1a).

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. The resulting benchmark spans three broad disciplines and 17 reported sub-fields, including both single-discipline and cross-disciplinary questions (Figure 1b).

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 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. Label distributions are reported in Supplementary Information S2.9.

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.

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[34,35].

Question Evaluation The evaluation of questions follows the standard scoring protocols established in large-scale evaluation frameworks[36]. 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 Extended Data Fig. 1, illustrating within-discipline and cross-disciplinary multimodal scientific questions.

Model Evaluation Environment

We evaluate 18 representative MLLMs with SEE, including 15 general-purpose models and three science-specialized models. The general-purpose models include Gemini 3.1 Pro[37], GPT-5.6-Sol (Max)[38], Claude Opus 5 (Max)[39], Qwen3.8-Max[40], GPT-5.5 (xhigh)[41], Kimi K3[42], Gemini 3.5 Flash[43], Claude Opus 4.8 (Max)[44], Seed2.1 Pro[45], Qwen3.7-Plus[46], Seed2.0 Pro[47], MiniMax-M3[48], Kimi K2.6[49], GLM-5V-Turbo[50], and MiMo-V2.5[51]. The science-specialized models include Intern-S2 Preview[52], Intern-S2 Preview-397B[53], and Intern-S1 Pro[54]. Only three 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 S2.1. 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[55]. 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 S2.4). 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 S2.2, and subject-label coverage and cross-discipline composition of SEE are reported in Supplementary Information S2.9.

Tool-augmented visual-agent evaluation

To evaluate whether tool access improves scientific reasoning beyond the standard multimodal setting, we further test six models that support both web search and code-interpreter tools in a tool-augmented visual-agent setting. Each model receives the same question text, options, and image inputs as in the standard evaluation, but may iteratively invoke web search or a code interpreter before producing a final answer. Tool use is optional rather than forced, and no curated retrieval corpus, local database, or domain-specific scientific tool is provided. Final answers are scored with the same binary judging protocol as the standard evaluation. We also retain the tool-use trajectories for diagnostic attribution analysis, in which tool-associated improvements and regressions are inspected with a human-in-the-loop procedure. Detailed configurations, trajectory attribution criteria, and tool-failure categories are provided in Supplementary Information S2.5.

Data availability

SEE contains 1,116 expert-curated questions, of which 1,049 are publicly released. The public benchmark subset, associated metadata, answer keys, evaluation code and public-subset results are available at GitHub. The remaining 67 questions are not publicly released because they contain unpublished experimental data contributed by domain experts. To support reproducibility and comparability between the full and public subsets, headline results on the public subset are reported in Supplementary Tables S5 and S6. Additional aggregate statistics and evaluation details are provided in the Supplementary Information.

Code availability

The code used for benchmark construction, model evaluation, answer judging, result analysis and figure generation is available at GitHub. The repository includes the public benchmark subset, evaluation scripts, judging prompts, analysis scripts and instructions required to reproduce the public-subset results reported in this study.

Acknowledgements

We thank the non-author domain experts and annotators who contributed candidate questions, experimental scenarios, expert validation and quality control for SEE. We also thank colleagues who provided discussion and feedback during the development of the benchmark and evaluation pipeline.

Author contributions

Taolin Han, Yuchen Zhang, Jinghang Wang and Yun Wu contributed equally to this work. Taolin Han, Yuchen Zhang, Jinghang Wang, Yun Wu, Wai Yuet Chiu, Zhaohai Li, Shuai Bai and Bing Zhao contributed to conceptualization, methodology, investigation, formal analysis and writing–original draft. Yifei Zhang, Jinxin Wang, Yuhao Zhou, Chen Zhao, Jiajia Li, Jiaxin Li, Qile Jin, Kewei Sun, Shuang Wu, Weiqi Zhai, Renquan Lv, Junchao Li, Ruodan Chen, Qingteng Chen, Zhibo Yang, Hu Wei and Lin Qu contributed to data curation, validation, investigation and writing–review and editing. Shuai Bai and Bing Zhao supervised this project. All authors discussed the results and approved the manuscript.

Competing interests

Several authors (Zhaohai Li, Zhibo Yang and Shuai Bai) are employees of Qwen Team, which develops Qwen models evaluated in this study. The authors declare no other competing interests.

Additional information

Supplementary information is available for this paper. Correspondence and requests for materials should be addressed to Shuai Bai or Bing Zhao.

References

1. Wang, H. et al. Scientific discovery in the age of artificial intelligence. *Nature* **620** (2023), 47–60. DOI: 10.1038/s41586-023-06221-2.
2. Ashyrmamatov, I. et al. A survey on large language models in biology and chemistry. *Exp. Mol. Med.* (2025). Advance online publication. DOI: 10.1038/s12276-025-01583-1.
3. Huang, K. et al. Autonomous biomedical research with an artificial intelligence agent. *Science* (2026), eadz4351. DOI: 10.1126/science.adz4351.
4. Hendrycks, D. et al. Measuring massive multitask language understanding. *Int. Conf. Learn. Represent. (ICLR)*. 2021.
5. Wang, Y. et al. MMLU-Pro: a more robust and challenging multi-task language understanding benchmark. *Adv. Neural Inf. Process. Syst. (NeurIPS) Datasets and Benchmarks Track*. 2024.
6. Li, H. et al. CMMLU: measuring massive multitask language understanding in Chinese. *Find. Assoc. Comput. Linguist.: ACL 2024*. Association for Computational Linguistics, 2024, 11260–11285. DOI: 10.18653/v1/2024.findings-acl.671.
7. Wang, X. et al. SciBench: evaluating college-level scientific problem-solving abilities of large language models. *Int. Conf. Mach. Learn. (ICML)*. 2024. URL: https://proceedings.mlr.press/v235/wang24z.html.
8. Rein, D. et al. GPQA: a graduate-level Google-proof Q&A benchmark. *Conf. Lang. Model. (COLM)*. 2024. URL: https://openreview.net/forum?id=Ti67584b98.
9. Center for AI Safety, Scale AI, and HLE Contributors Consortium. A benchmark of expert-level academic questions to assess AI capabilities. *Nature* **649** (2026), 1139–1146. DOI: 10.1038/s41586-025-09962-4.
10. Mirza, A. et al. A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists. *Nat. Chem.* **17** (2025), 1027–1034. DOI: 10.1038/s41557-025-01815-x.
11. Alampara, N. et al. Probing the limitations of multimodal language models for chemistry and materials research. *Nat. Comput. Sci.* **5** (2025), 952–961. DOI: 10.1038/s43588-025-00836-3.
12. He, C. et al. OlympiadBench: a challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. *Proc. Annu. Meet. Assoc. Comput. Linguist. (ACL)*. 2024, 3828–3850. URL: https://aclanthology.org/2024.acl-long.211/.

13. Ye, X., Li, C., Chen, S., Wei, W., and Tang, R. MMSciBench: benchmarking language models on Chinese multimodal scientific problems. *Find. Assoc. Comput. Linguist.: ACL 2025*. Vienna, Austria: Association for Computational Linguistics, 2025, 14621–14663. DOI: 10.18653/v1/2025.findings-acl.755.

14. Ruan, J. et al. MME-SCI: a comprehensive and challenging science benchmark for multimodal large language models. *Proc. AAAI Conf. Artif. Intell. (AAAI)*. Vol. 40. 2026, 8760–8768. DOI: 10.1609/aaai.v40i11.37829.

15. Cui, Y. et al. Evaluating large language models on multimodal chemistry olympiad exams. *Commun. Chem.* **8** (2025), 402. DOI: 10.1038/s42004-025-01782-x.

16. Bran, A. M. et al. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **6** (2024), 525–535. DOI: 10.1038/s42256-024-00832-8.

17. Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature* **624** (2023), 570–578. DOI: 10.1038/s41586-023-06792-0.

18. Abolhasani, M. and Kumacheva, E. The rise of self-driving labs in chemical and materials sciences. *Nat. Synth.* **2** (2023), 483–492. DOI: 10.1038/s44160-022-00231-0.

19. Gottweis, J. et al. Accelerating scientific discovery with Co-Scientist. *Nature* **655** (2026), 487–496. DOI: 10.1038/s41586-026-10644-y.

20. Lu, C. et al. The AI Scientist: towards fully automated open-ended scientific discovery. 2024. Preprint at https://arxiv.org/abs/2408.06292.

21. Schmidgall, S. et al. Agent Laboratory: using LLM agents as research assistants. *Find. Assoc. Comput. Linguist.: EMNLP 2025*. Suzhou, China: Association for Computational Linguistics, 2025, 5977–6043. DOI: 10.18653/v1/2025.findings-emnlp.320.

22. Liu, X. et al. AgentBench: evaluating LLMs as agents. *Int. Conf. Learn. Represent. (ICLR)*. 2024.

23. Laurent, J. M. et al. LAB-Bench: measuring capabilities of language models for biology research. 2024. Preprint at https://arxiv.org/abs/2407.10362.

24. Huang, Z. et al. ChemVTS-Bench: evaluating visual–textual–symbolic reasoning of multimodal large language models in chemistry. 2025. Preprint at https://arxiv.org/abs/2511.17909.

25. Dunn, A., Wang, Q., Ganose, A., Dopp, D., and Jain, A. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Comput. Mater.* **6** (2020), 138. DOI: 10.1038/s41524-020-00406-3.

26. Liu, W. et al. PolyReal: a benchmark for real-world polymer science workflows. 2026. Preprint at https://arxiv.org/abs/2604.02934.

27. Sainz, O. et al. NLP evaluation in trouble: on the need to measure LLM data contamination for each benchmark. *Findings Assoc. Comput. Linguist.: EMNLP 2023*. Association for Computational Linguistics, 2023, 10776–10787. DOI: 10.18653/v1/2023.findings-emnlp.722.

28. Lightman, H. et al. Let's verify step by step. *Int. Conf. Learn. Represent. (ICLR)*. 2024.

29. Bucci, E. M. Automatic detection of image manipulations in the biomedical literature. *Cell Death Dis.* **9** (2018), 400. DOI: 10.1038/s41419-018-0430-3.

30. Ji, Z. et al. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55** (2023), 1–38. DOI: 10.1145/3571730.

31. Lin, S., Hilton, J., and Evans, O. TruthfulQA: measuring how models mimic human falsehoods. *Proc. Annu. Meet. Assoc. Comput. Linguist. (ACL)*. Association for Computational Linguistics, 2022, 3214–3252. DOI: 10.18653/v1/2022.acl-long.229.

32. Kadavath, S. et al. Language models (mostly) know what they know. 2022. Preprint at https://arxiv.org/abs/2207.05221.

33. Vishwanath, K. et al. General-purpose large language models outperform specialized clinical AI tools on medical benchmarks. *Nat. Med.* **32** (2026), 2405–2409. DOI: 10.1038/s41591-026-04431-5.

34. Yue, X. et al. MMMU: a massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. 2024, 9556–9567. DOI: 10.1109/CVPR52733.2024.00913.

35. Roberts, J. et al. SciFIBench: benchmarking large multimodal models for scientific figure interpretation. *Adv. Neural Inf. Process. Syst. (NeurIPS)*. 2024.

36. Liang, P. et al. Holistic evaluation of language models. *Trans. Mach. Learn. Res.* (2023). ISSN 2835-8856. URL: https://openreview.net/forum?id=iO4LZibEqW.

37. Google DeepMind. Gemini 3.1 Pro. 2026. URL: https://deepmind.google/models/gemini/pro/.

38. OpenAI. GPT-5.6-Sol. 2026. URL: https://openai.com/index/gpt-5-6/.

39. Anthropic. Claude Opus 5. 2026. URL: https://www.anthropic.com/news/claude-opus-5.

40. Qwen Team, Alibaba Cloud. Qwen3.8-Max. 2026. URL: https://qwen.ai/blog?id=qwen3.8.

41. OpenAI. GPT-5.5 (xhigh). 2026. URL: https://openai.com/zh-Hans-CN/index/introducing-gpt-5-5/.

42. Moonshot AI. Kimi K3: open frontier intelligence. 2026. URL: https://arxiv.org/abs/2607.24653.

43. Google DeepMind. Gemini 3.5 Flash. 2026. URL: https://deepmind.google/models/gemini/flash/.

44. Anthropic. Claude Opus 4.8 (Max). 2026. URL: https://www.anthropic.com/news/claude-opus-4-8.

45. ByteDance Seed. Seed2.1 Pro. 2026. URL: https://seed.bytedance.com/en/seed2_1.

46. Qwen Team, Alibaba Cloud. Qwen3.7-Plus: multimodal agent intelligence. 2026. URL: https://www.alibabacloud.com/blog/qwen3-7-plus-multimodal-agent-intelligence_603206.

47. ByteDance Seed. Seed2.0 Pro. 2026. URL: https://seed.bytedance.com/zh/seed2.

48. MiniMax. MiniMax-M3: frontier coding, 1M context, native multimodality. 2026. URL: https://www.minimax.io/blog/minimax-m3.

49. Moonshot AI. Kimi K2.6: advancing open-source coding. 2026. URL: https://www.kimi.com/blog/kimi-k2-6.

50. Zhipu AI (Z.ai). GLM-5V-Turbo. 2026. URL: https://docs.z.ai/guides/vlm/glm-5v-turbo.

51. LLM-Core Team, Xiaomi. MiMo-V2.5. 2026. URL: https://mimo.xiaomi.com/mimo-v2-5.

52. InternLM. Intern-S2 Preview. 2026. URL: https://huggingface.co/internlm/Intern-S2-Preview.

53. InternLM. Intern-S2 Preview-397B. 2026. URL: https://huggingface.co/internlm/Intern-S2-Preview-397B.

54. Shanghai AI Laboratory. Intern-S1 Pro. 2026. Preprint at https://arxiv.org/abs/2603.25040.

55. Zheng, L. et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Adv. Neural Inf. Process. Syst. (NeurIPS) Datasets and Benchmarks Track*. 2023.

Extended Data



Extended Data Fig. 1. Representative questions in SEE. Four examples illustrating the range of disciplinary scopes, evidence types and answer formats in the benchmark. **a**, A single-discipline biology question. **b**, A single-discipline chemistry question. **c**, A cross-disciplinary biology–chemistry question combining chemical structures, Western blots and a volcano plot. **d**, A cross-disciplinary chemistry–materials question that assembles electrostatic potential maps, DFT results, SEM and AFM images and stress–strain curves. In every panel, labels with colors annotate sub-fields and outlined labels give the associated experimental technique or knowledge point. Answers are scored by exact match. B oth single-response (**b**) and multiple-response (**a**, **c**, **d**) formats occur. The source line gives the literature DOI, or "Private Data" for questions built on unpublished experimental data contributed by domain experts.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf