

Supplementary Information

S1 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 (Extended Data Fig. 1). 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.

S2 Evaluation

S2.1 Evaluation Setup

We evaluate SEE on 18 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, 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`. 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

576 measured by accuracy, where a question is counted as correct only if the final model answer
577 matches the ground-truth answer under the scoring rules described below.

578 S2.2 Prompting Protocol

579 We use separate prompts for answer generation and answer judging. The generation prompt asks
580 each evaluated model to analyze the question and return a JSON object containing both the
581 analysis and the final answer. The judging prompt is executed by Gemini 3.1 Pro; it compares
582 the candidate response with the reference answer and returns a binary correctness label.

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

```
583 [System]
584 You are an intelligent assistant. Please read the question and images
585 carefully, and provide the correct answer.
586
587 [User]
588 Please answer the following question. If it is a choice question, it may
589 be either single-choice or multiple-choice.
590
591 [Question text]:
592 {question}
593
594 [Options]:
595 {options}
596
597 Please first provide the analysis process, and then provide the final
598 answer. Strictly output the following JSON format, and do not include
599 Markdown formatting:
600 {
601   "analysis": "your analysis process",
602   "answer": "final answer"
603 }
604
605 }
```

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

```
606 [System]
607 You are a strict examiner. Please judge whether the student's answer is
608 consistent with the reference answer.
609
610 [User]
611 Please judge whether the following answer is correct:
612
613 [Question]:
614 {question}
615
616 [Reference answer]:
617 {reference}
618
619 [Student answer] (it may be a short answer, or a full response containing
```

```

621 an analysis process):
622 {candidate}
623
624 [Task]:
625 1. If the student's answer is a full response containing an analysis
626    process, first identify the final answer from it, and then compare it
627    with the reference answer.
628 2. Judge whether the core meaning is consistent. For choice questions,
629    the letters must be identical. For fill-in-the-blank or short-answer
630    questions, the numerical value or key phrase must be identical. Format
631    differences such as "5" and "5.0" are allowed.
632 3. If the reference answer is a numerical range or contains an error
633    tolerance, such as "5-10", "[0.8, 1.2]", "greater than 100",
634    "5 +/- 0.5", "about 0.05", or "<3.2", the student's answer is correct
635    as long as the numerical value falls within the range. Range
636    boundaries are also treated as correct (closed interval). The
637    student's answer may also be a range; in this case, judge whether the
638    two ranges are substantially consistent, with similar centers and
639    widths.
640 4. If the student refuses to answer or no answer can be identified, mark
641    it as incorrect.
642
643 Strictly output the following JSON format, and do not include Markdown
644 formatting:
645 {
646   "correct": true or false,
647   "reason": "judgment reason within 50 characters"
648 }

```

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

653 S2.3 Answer Extraction and Scoring

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

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

S2.4 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, $\sim 11,000$ judgments in total).

Table S1 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 S1. 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

S2.5 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 per-model tool configuration is summarised in Table S2. 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 S2. 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 execution-control failure. Their relationship to the fine-grained attribution types is summarized in Table S3. The resulting expert-validated attribution is used as a diagnostic analysis of the interaction trajectory rather than as an additional correctness score.

Table S3. 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.
Execution-control failure	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 (Table S4): 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.

S2.6 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 S5 reports the standard no-tool setting and Table S6 the tool-augmented setting.

Table S4. 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%)

Table S5. 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

Table S6. 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 S5. 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

S2.7 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 S1 reports the paired accuracies for each model; alongside the mean drop, the spread across models narrows from 31.1 to 14.0 points.

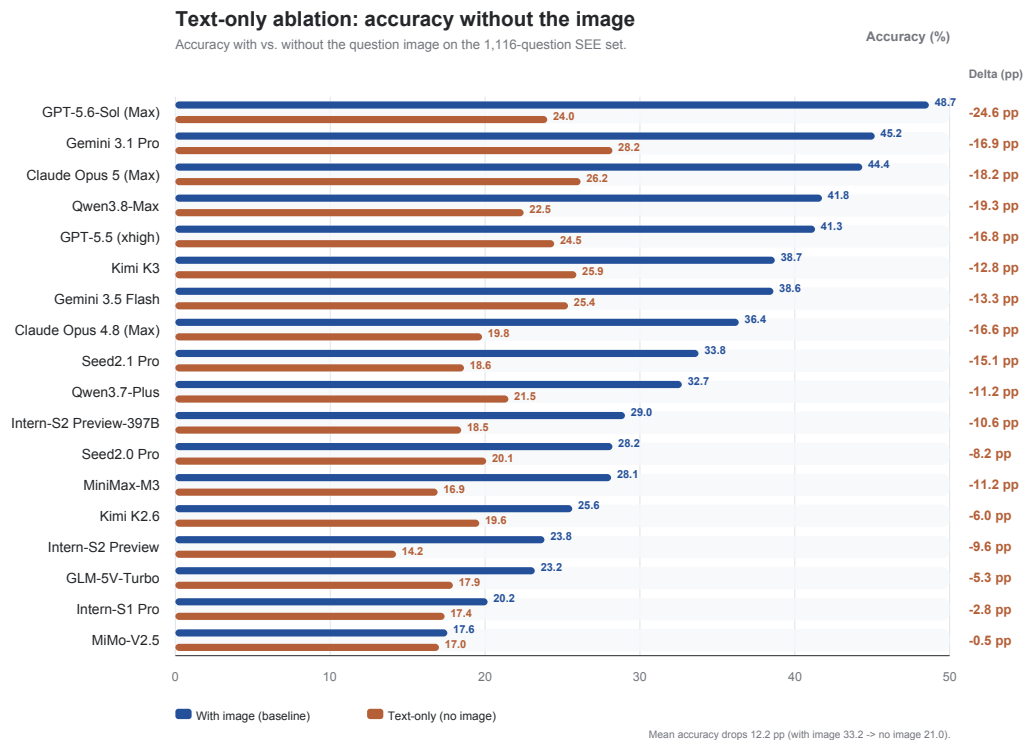


Figure S1. Accuracy with and without the question image for the 18 models with paired runs. Each model is scored twice on the same question set; the text-only run receives the identical question with the image removed. Accuracy follows the strict binary protocol, in which every instance stays in the denominator and refusals, non-answers, and generation errors count as incorrect. Delta is the text-only accuracy minus the with-image accuracy.

Across the models with available no-image runs, explicit missing-image acknowledgment is consistently rare (Table S7). Across 18 text-only runs, only 921 out of 20,088 instances (4.6%) are 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.

Table S7. 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.

S2.8 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 S8 reports per-sub-discipline accuracy for all 18 evaluated models, ordered by overall accuracy, and Table S9 reports the mean accuracy per label averaged over models.

S2.9 Subject Label Coverage

SEE exhibits pronounced multi-label and cross-disciplinary characteristics. Under the final 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. Table S10 lists the question count per fine-grained label, Table S11 the distribution of labels per question, and Table S12 main-discipline coverage by question and by annotation count.

Table S8. Per-model accuracy by reported sub-discipline for all 18 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-4o-S3aI (Max)	Gemini 1.5 Pro	Claude Opus 5 (Max)*	Qwen2.5-Max	GPT-4o (highlight)	Kimi K3	Gemini 1.5 Flash	Claude Opus 4.8 (Max)*	Sora2.1 Pro	Qwen2.7-Plus	Intern-S2 Preview-307B	Sora2.0 Pro	Mistral-M3	Kimi K2.6	Intern-S2 Preview	GLM-3V-Turbo	Intern-S1 Pro	Mistral-V2.5	
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	
	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	
	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	
	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	
	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	
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	
	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	
	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	
	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	
	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	
	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	
	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	
	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	
	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
		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
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	
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	

Table S9. Mean model accuracy across 18 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	38.8
Metallic Materials	Materials	49	35.7
Analytical Chemistry	Chemistry	440	35.3
Immunology	Biology	32	35.2
Cell Biology	Biology	182	34.3
Molecular Biology	Biology	261	32.4
Biochemistry	Biology	344	32.2
Genetics	Biology	36	31.9
Inorganic Non-metallic Materials	Materials	103	31.8
Biophysics	Biology	156	31.3
Structural Biology	Biology	166	31.0
Physical Chemistry	Chemistry	337	30.0
Physiology	Biology	19	29.5
Composite Materials	Materials	68	29.4
Polymer Chemistry and Physics	Chemistry	174	29.3
Inorganic Chemistry	Chemistry	110	29.0
Organic and Polymer Materials	Materials	188	26.0

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 (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.

Table S10. 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 S11. 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 S12. 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

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.