
BenchBench-Protocol: Evaluating Real-World Wet-Lab Protocol Reasoning and Modification

Aditya Sivakumar*
Benchling

Ashu Singhal
Benchling

Nicholas Larus-Stone
Benchling

Nithin Parsan*,†
Benchling

Abstract

We introduce BenchBench-Protocol, a benchmark for large language models of 149 protocol-modification tasks recovered from modifications that scientists made to published protocols during real experimental work. Adapting a published protocol to a new experiment is a routine task for a wet-lab scientist, and a correct modification requires accounting for prior choices and downstream steps. Recent life-science benchmarks have moved toward open-ended, rubric-graded tasks, but tasks are typically elicited from experts rather than reconstructed from real-world modifications. BenchBench-Protocol tasks are derived from differences between a published protocol and a version a scientist modified, which provides the basis for the query and the weighted rubric elements for a correct response. The benchmark draws from 96 source protocols across nine domains of wet-lab biology and only includes tasks rated highly after review by domain experts. We evaluate nine closed and open models; Claude Opus 5 scores highest at 59.2% normalized rubric score, with other models between 34.1% and 47.1%, and the benchmark remains unsaturated when taking the best of ten attempts. As models are increasingly helpful in life-sciences research, evaluating them on routine wet-lab tasks becomes correspondingly important. We present BenchBench-Protocol as both a grounded assessment of wet-lab reasoning and evidence for the utility of real-world experiments to construct benchmark tasks.

1 Introduction

Adapting an existing laboratory protocol for a new experiment is a common task for wet-lab scientists. Correct decisions require reasoning over previous choices and intended next steps, not just recalling facts or retrieving existing knowledge. Therefore, an important practical skill is being able to modify or extend an existing protocol while anticipating its downstream effects.

Recent frontier benchmarks of language models for life-science research have moved toward more open-ended, rubric-graded tasks that resemble research requests. These tasks are often constructed by asking experts to write tasks representative of their work. While this process results in high-quality tasks that are scientifically correct and plausible, they may diverge from the distribution of modifications used in real-world workflows, which are constrained by available reagents, instrument access, and prior results.

*Equal contribution.

†Corresponding author and team lead: nithin.parsan@benchling.com.

Benchmark Overview



Figure 1: Benchmark construction overview. Published protocols and scientist-adapted versions on Benchling from contributors are compared to produce tasks and weighted rubrics, which are retained after expert quality control.

We introduce BenchBench-Protocol, a benchmark of protocol-modification tasks. Rather than creating tasks from scratch, we recover them from real modifications scientists have made to published protocols. Tasks are realistic, open-ended, and free-response. Each task and rubric was then independently reviewed by multiple scientists with a PhD in a relevant field and at least three years of bench experience. Figure 1 summarizes this construction process.

We contribute:

- **BenchBench-Protocol:** a benchmark of 149 expert-reviewed protocol-modification tasks spanning nine domains of wet-lab biology, phrased as a scientist would pose the question to an LLM rather than as a formal problem statement.
- **Weighted rubrics:** expert-verified rubrics grounded in the modifications scientists made, used to score open-ended responses against the specific downstream dependencies a correct answer must address.

2 Related Work

Recent benchmarks have expanded beyond simple biological fact recall and increasingly focus on the diverse tasks a scientist may perform. Several benchmarks assess the research-assistant capabilities of a model: LAB-Bench has an extensive suite of biology research tasks (Laurent et al., 2024), with LABBench2 expanding these tests with free-response questions (ProtocolQA2, LitQA3, FigQA2, and others) that show whether models can troubleshoot protocols and understand and evaluate figures, tables, and data from research papers and databases (Laurent et al., 2026). BioProBench assesses model capabilities in error correction and protocol generation (Liu et al., 2025), while TroubleshootingBench evaluates the ability to identify and correct wet-lab error scenarios created by experts (OpenAI, 2025). BioLP-bench, like ProtocolQA and TroubleshootingBench, introduces a critical mistake into a protocol and measures whether a model can identify the responsible step (Ivanov, 2024).

Several benchmarks also address agentic and long-horizon capabilities for end-to-end research tasks. BixBench tests agents’ ability to execute code and reason over bioinformatics results (Mitchener et al., 2025). ScienceAgentBench and AstaBench extend agentic evaluations across scientific workflows more broadly (Chen et al., 2025; Bragg et al., 2025). GeneBench and GeneBench-Pro are frontier evaluations assessing multi-stage quantitative biology and carefully construct tasks to constrain the number of valid solution paths and permit exact verification (Li & Ho, 2026a,b). While

Table 1: Benchmark coverage by protocol domain.
Task and source protocol counts are shown for each of the nine domains.

Protocol domain	Tasks	Share	Protocols
Protein biochemistry and purification	32	21.5%	21
Cell culture and cell-based assays	27	18.1%	16
Molecular biology and cloning	26	17.4%	17
Staining, imaging, and cytometry	24	16.1%	17
Genomics, transcriptomics, and sequencing	19	12.8%	13
Microbiology and microbial methods	14	9.4%	8
Media, reagent prep, and recipe	4	2.7%	2
Model organism, in vivo, and primary tissue	2	1.3%	1
Chemical synthesis and conjugation	1	0.7%	1
Total	149	100.0%	96

these benchmark tasks require strong biological reasoning, most are limited to computational biology.

LifeSciBench is the closest analogous benchmark, containing expert-authored tasks with a rubric and a single-turn free-response format (Liu et al., 2026). One of its seven workflows explicitly covers the design and optimization of assays, constructs, and protocols. LifeSciBench is constructed by subject-matter experts with strong quality audits, but, similar to LABBench2 ProtocolQA2, the plausible modifications an expert may suggest to protocols may differ from the distribution of modifications in real-world workflows. BenchBench-Protocol complements and extends the suite of life-sciences capability benchmarks by deriving tasks from modifications scientists made to published protocols during real experimentation.

3 Benchmark Design and Construction

3.1 Task Formulation

In each task, the model receives a first-person, naturally phrased scientist query and a published protocol to extend or modify, and returns a free-text answer with a rationale. During evaluation, model tool use is unrestricted within the assigned search configuration described in Section 4.2 and Appendix B. Responses are graded by an LLM-as-a-judge that receives the system prompt, task, rubric, and response under evaluation. All evaluation attempts were graded with GPT-5.6 Terra at medium reasoning to standardize grading and minimize variation between model families. No significant scoring differences were observed across two representative grader models during testing (Appendix C). Each rubric criterion is presented to the grader separately so that its score reflects whether the response specifically satisfies that criterion.

3.2 Benchmark Coverage

BenchBench-Protocol contains 149 evaluation tasks drawn from 96 published protocols across nine domains of wet-lab biology, listed in Table 1. The published protocols are compared with protocols created by contributors on Benchling, a life-sciences platform used for wet-lab work across the major life-science domains.

3.3 Benchmark Construction

Tasks are constructed by comparing a published protocol against a version of the protocol that a scientist modified on Benchling. The differences between versions provide a record of what a scientist did to adapt the protocol for a new experimental goal. These differences supply the ground truth for the task, defining both the relevant modification and the downstream consequences that must be addressed. From each difference, a question and accompanying rubric are drafted, and a quality-control process removes questions that are unanswerable, malformed, or ambiguous. Review additionally validates that the differences are scientifically meaningful and relevant.

Table 2: Distribution of primary task intents. Task counts and shares are shown for each of the seven intents.

Primary task intent	Tasks	Share
Plan audit	38	25.5%
Measurement and result interpretation	35	23.5%
Troubleshooting	29	19.5%
Recordkeeping	23	15.4%
Procedure specification and adaptation	14	9.4%
Quantitative formulation	6	4.0%
Follow-up experiment	4	2.7%
Total	149	100.0%

Expert review

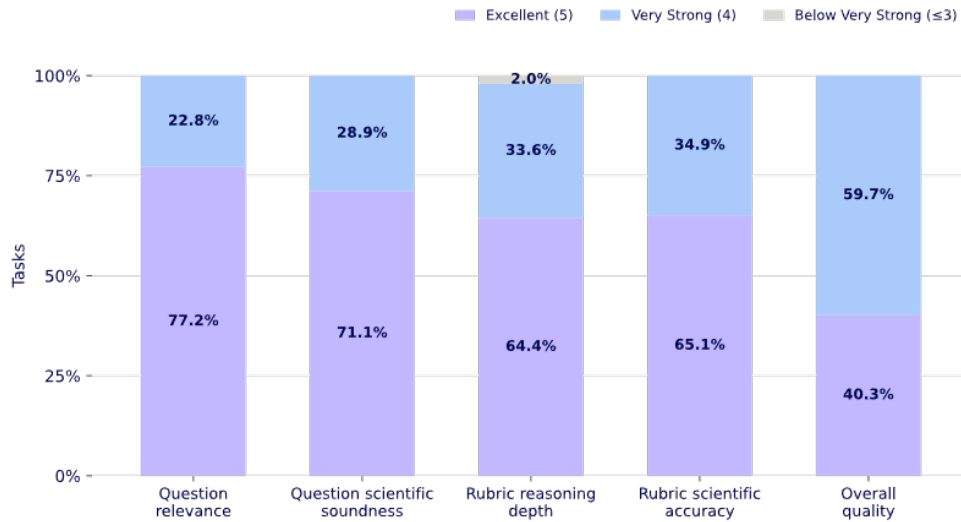


Figure 2: Expert review ratings by assessment dimension. Share of retained tasks rated Excellent, Very Strong, or below Very Strong on each of the five dimensions.

Tasks may carry multiple intent labels, which are not mutually exclusive, and each task is additionally assigned one primary intent. These labels allow performance to be broken down by the kind of judgment a task demands, rather than only by biological subject matter. Definitions of each source-protocol domain and task intent are provided in Appendix A. Table 2 reports primary intents.

3.4 Expert Review

An initial cohort of tasks was reviewed by domain experts and refined or corrected to yield a final set of 149 questions and associated rubrics. Reviewers rated each task on five dimensions: question relevance, question scientific soundness, rubric reasoning depth, rubric scientific accuracy, and overall question quality. Only tasks rated “Very Strong” or “Excellent” for question relevance, question scientific soundness, rubric scientific accuracy, and overall quality were retained in the final dataset.

Each task was reviewed by at least two independent reviewers, each of whom holds a PhD in a relevant field and has worked at least three years as a scientist in academia or industry. Figure 2 shows overall metrics across each measured dimension.

Additional sample tasks are shown in Appendix F.

Example task 1

Genomics Transcriptomics and Sequencing

Recovering RNA without counting the pellet as supernatant

Question

In a scaled RNeasy RT-style total RNA prep I used 1.0 mL reagent and 0.4 mL water. After the contaminant spin, the semisolid pellet looks compact, and I want to recover as much RNA-containing supernatant as the worksheet's 85% cap permits without scraping the pellet. If the contaminant pellet is roughly 10% of the 1.4 mL homogenate-water mixture, what transfer volume should I set as the practical upper target, and how should I set the isopropanol volume for total-RNA precipitation?

Protocol context

NucleoZOL Instruction Manual · MACHEREY-NAGEL
<https://www.mn-net.com/media/pdf/d3/a8/1b/Instruction-NucleoZOL.pdf>

Rubric

- | | | |
|---|---|--|
| <p>1 Calculates the transferable volume as about 1.07 mL, or practically 1.0-1.1 mL, by subtracting the roughly 0.14 mL pellet from the 1.4 mL mixture before applying the 85% cap.</p> | <p>2 Uses the correct denominator for the cap: the 85% cap is applied to the post-pellet RNA-containing supernatant, not to the original 1.4 mL homogenate-water mixture; answers that set the transfer around 1.19 mL fail this criterion.</p> | <p>3 Explains the practical consequence and precipitation scaling: avoid chasing the last liquid because disturbing the DNA/protein/polysaccharide pellet risks carryover, and add 100% isopropanol as one volume matched to the actual transferred supernatant.</p> |
| 10 pts | 10 pts | 5 pts |

Example 1: Sample task from genomics, transcriptomics, and sequencing. The scientist query, source protocol, and weighted rubric criteria are shown.

4 Experimental Setup

4.1 Models Evaluated

Nine models were evaluated: Claude Opus 5, GPT-5.6 Sol, GPT-5.6 Terra, GPT-5.6 Luna, Gemini 3.6 Flash, Grok 4.5, Kimi K3, GLM 5.2, and Inkling. All models were run at their highest available reasoning settings. Open models were run on Baseten, with exact settings detailed in Appendix B. Results are aggregated over ten attempts per task. Attempts that did not produce a valid response because of misformatted responses or refusals were assigned a score of zero; invalid-response rates by model are reported in Appendix I.

4.2 Evaluation Method

Each model receives the query, a published protocol, and internet access, and returns a single free-text response as described in Section 3.1. Claude Opus, GPT-series, Gemini, and Grok models had access to their providers' first-party search tools, while open models were given internet access through Exa configured at "auto" effort. Tool use was unrestricted within those configurations. LLM graders were not given web search or tools. Appendix C reports representative agreement results between two grader models.

4.3 Metrics

Model performance is reported using a normalized rubric score between 0 and 1. Each rubric criterion is scored as 0, 1, or 2, where 2 is full credit and 1 is partial credit. Each criterion additionally has a weight reflecting its relative importance to the overall task. A coarse three-point scale was used instead of a wider numeric range to improve LLM-as-a-judge stability. A response's normalized score is the sum of weighted criterion scores divided by the total weighted points available, and the average across ten attempts per task is reported. Confidence intervals are calculated by bootstrap resampling at the source-protocol level because tasks derived from the same protocol are not independent. The expected best-of- k score is computed by sampling k of the ten attempts per question without replacement, taking the maximum normalized score, and averaging across resamples.

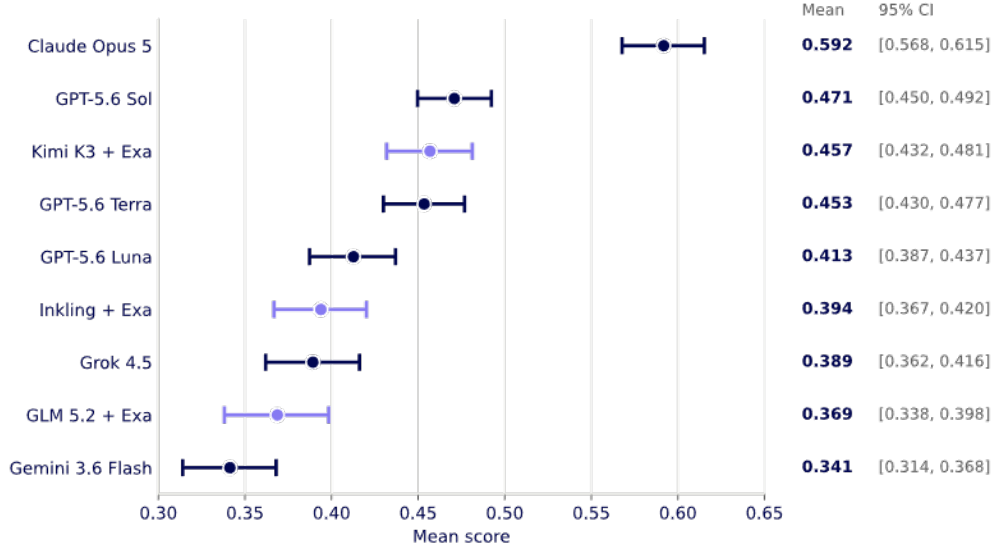


Figure 3: Mean normalized rubric score by model. Means across 149 tasks are shown with 95% confidence intervals.

5 Results

Performance on BenchBench-Protocol varies substantially across models, task intents, and source-protocol domains. Claude Opus 5 is the strongest overall model, with a mean normalized score of 59.2% (95% CI: 56.8–61.5%). Scores for the other evaluated models range from 34.1% to 47.1%. Scores for all evaluated models are shown in Figure 3. BenchBench-Protocol has at least 40% headroom of unearned weighted rubric credit across model families, allowing future frontier-model assessment.

5.1 Performance Under Repeated Sampling

Because a scientist can compare several candidate answers before acting on one, expected best-of- k performance is reported as an upper bound on single-model performance.

The ranking of model performance is not stable in k . GPT-5.6 Sol and Kimi K3 are statistically indistinguishable at $k = 1$, but separate cleanly at $k = 10$, with Kimi K3 performing better. Because absolute gains are bounded by the unearned credit at $k = 1$, gains are reported as a fraction of headroom captured. Complete results are in Appendix E; Appendix D shows the distribution of model performance across tasks.

At ten attempts, Claude Opus 5 captures 39.5% of available headroom and Kimi K3 captures 31.5%, while all three GPT-5.6 models capture between 19.6% and 22.5%. Models that capture less headroom over multiple attempts tend to be more predictable. Under oracle selection, the strongest model still leaves roughly one quarter of rubric credit unearned, indicating that the benchmark is not saturated at any k .

5.2 Performance by Task Intent

Figure 5 demonstrates strong differences in performance across task intents.

Analyzing failure cases heuristically, models perform comparatively well on tasks that require reasoning from experimental evidence to the conclusions it supports. Models perform comparatively worse on tasks that require tacit knowledge or intuition about what an expert would consider the “most likely” or “best” approach, which may account for the discrepancy in troubleshooting and adaptation.

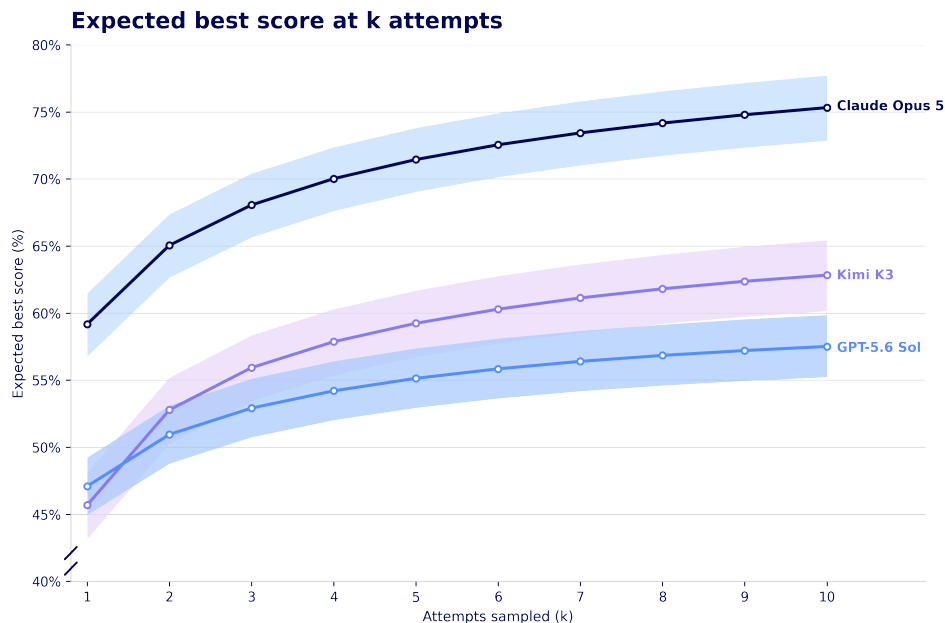


Figure 4: Expected best-of- k score for three representative models. Expected maximum score is shown for k of ten attempts per task with 95% confidence bands. Complete nine-model results are in Table E2.

5.3 Performance by Source-Protocol Domain

Claude Opus and GPT-series models perform strongly on staining and imaging tasks relative to open-source models, which may be due to general improvements in multimodal reasoning in these models versus open series. Figure 6 summarizes these results; full values are reported in Table E1.

5.4 Model Token Usage and Cost

Figure 7 visualizes average input and output token consumption across the benchmark task suite. Grok 4.5 consumes the most input tokens, likely because of a high volume of web searches and ingested content. At maximum reasoning settings, models perform frequent searches, leading to high volumes of input and output tokens. Gemini 3.6 Flash is excluded because search input tokens from Gemini search are billed separately and not precisely delineated in Vertex API calls, making total-input-token comparisons unfair. Average cost per answer attempt across model families, excluding web search and hosted-tool charges, is included in Appendix G. The performance–cost Pareto frontier at maximum reasoning is included in Appendix H.

6 Limitations and Future Work

6.1 Limitations

BenchBench-Protocol measures reasoning about realistic modifications to wet-lab protocols, but it does not directly measure whether models improve experimental outcomes at the bench. The benchmark contains 149 tasks across nine domains of wet-lab biology, and while scientists perform the bulk of their protocol work in these areas, no single benchmark can span the full range of modifications that arise in this setting. Tasks were also selected to be difficult, so scores may not reflect performance on routine adjustments that make up the majority of day-to-day laboratory work.

The evaluation was conducted in a single-turn setting, with the model’s output as the final response. Real use is almost always multi-turn, with scientists clarifying limitations in reagents, instruments, and time. Model performance may improve if the model is allowed to ask clarifying questions.



Figure 5: Mean normalized rubric score by primary task intent. Cells give mean scores in percent, with task counts per intent.

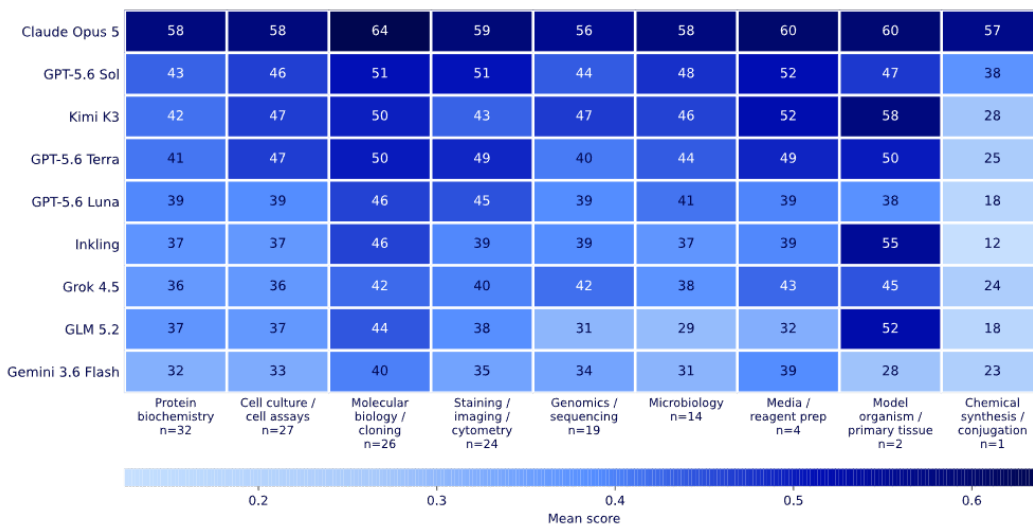


Figure 6: Mean normalized rubric score by source-protocol domain. Cells give mean scores in percent, with task counts per domain.

Model evaluation used the maximum reasoning settings offered by each model family to standardize the greatest available performance. This results in substantially higher token usage and cost than default or recommended settings, which affects the performance–cost frontier. Models may also perform differently in other harnesses or multi-agent setups.

Human expert baselines were not directly assessed on the benchmark tasks. The expert quality-control process places strong emphasis on answerability, which is a proxy adopted in place of a direct expert-performance baseline.

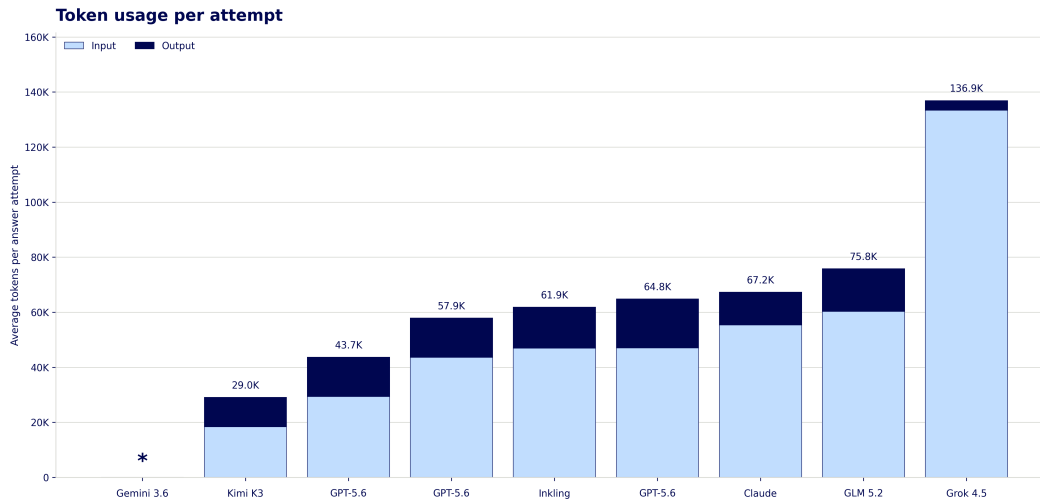


Figure 7: Token usage per answer attempt. Average input and output tokens are shown per model; Gemini 3.6 Flash is excluded (asterisk).

Responses were graded by GPT-5.6 Terra, which is also one of the nine evaluated models. Model grading of open-ended responses is imperfect in ways a coarse rubric scale reduces but does not remove, and the three-point scale itself limits how finely a response can be distinguished from a stronger one. Search tooling also differs between conditions, with closed models using first-party search and open models using Exa at “auto” effort.

6.2 Future Work

Future work can focus further on multi-turn settings that mirror how a wet-lab scientist may collaborate with a model. These can include tasks that rely on previous data or experiments to plan follow-up work, assess models’ abilities to troubleshoot or debug real experimental failures, or correctly elicit user intent when underspecified. Further analysis of the existing benchmark can be conducted to understand failure modes of current models and assess performance in consensus or multi-model setups.

7 Conclusion

In this paper, we make the following contributions:

- **Reviewed benchmark task corpus:** We introduce BenchBench-Protocol, 149 protocol-modification tasks spanning nine domains of wet-lab biology, recovered from modifications scientists made to published protocols during real experimental work. Tasks and rubrics were independently assessed by multiple scientists holding a PhD in a relevant domain and three or more years of bench experience.
- **Rubrics:** We provide weighted rubrics grounded in the modification that was actually performed, scoring responses against the specific downstream dependencies a correct answer must address rather than against general plausibility.
- **Model benchmarking:** We evaluate nine models across all 149 tasks, under repeated sampling as well as single attempts, and characterize their failures.

BenchBench-Protocol measures whether models can adapt an existing protocol while tracking the downstream steps a change affects. Claude Opus 5 scores highest at 59.2%, with other models’ scores ranging from 34.1% to 47.1%, and repeated sampling captures between one fifth and two fifths of unearned credit.

Kimi K3 approaches and then exceeds GPT-5.6 Sol in expected score@ k at ten attempts. Heuristic analysis of traces and responses shows that GPT-5.6 Sol had more consistent responses and more

closely followed exact protocol constraints, while Kimi K3 had greater variation, producing both weaker and more complete responses. The heavier upper tail of scores gave Kimi K3 a higher expected best-of- k score.

Models differ in overall level rather than in profile, so current systems are not differentially suited to particular areas of wet-lab reasoning. Protocol modification is routine work where errors propagate into downstream results, and grounding evaluation in modifications scientists actually made offers a more direct way to measure progress on it.

Acknowledgments

We thank Lauren DeVos for her support.

References

- Jonathan Bragg et al. AstaBench: Rigorous benchmarking of AI agents with a scientific research suite, 2025.
- Ziru Chen, Shijie Chen, Yuting Ning, Qianqian Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zhe Lu, Vikram Dey, Meng Xue, Francis N. Baker, Brian Burns, Daniel Adu-Ampratwum, Xiaoning Huang, Xuefei Ning, Song Gao, Yu Su, and Huan Sun. ScienceAgent-Bench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, 2025.
- Ivan Ivanov. BioLP-bench: Measuring understanding of biological lab protocols by large language models. *bioRxiv*, 2024. doi: 10.1101/2024.08.21.608694.
- Jon M. Laurent, Joseph D. Janizek, Miguel Ruzo, Madeline M. Hinks, Michael J. Hammerling, Siddharth Narayanan, Manvitha Ponnampati, Andrew D. White, and Samuel G. Rodriques. LAB-Bench: Measuring capabilities of language models for biology research, 2024.
- Jon M. Laurent, Albert Bou, Marlon Pieler, Conor Igoe, Alex Andonian, Siddharth Narayanan, Jonathan Braza, Alejandro Sanchez Vassopoulos, Jacob L. Steenwyk, Benjamin Lash, Andrew D. White, and Samuel G. Rodriques. LABBench2: An improved benchmark for AI systems performing biology research, 2026.
- James H. Li and Andrew J. Ho. GeneBench: Assessing AI agents for multi-stage inference problems in genomics and quantitative biology. *bioRxiv*, 2026a. doi: 10.64898/2026.04.22.720113.
- James H. Li and Andrew J. Ho. GeneBench-Pro: Evaluating multistage statistical reasoning in genomics, quantitative biology, and translational biomedicine. *bioRxiv*, 2026b. doi: 10.64898/2026.06.29.735386.
- Anthony Liu, Andrew Ho, Alexander M. Droste, David Martin, Eric Wong, Eric Zhou, Ian Zhou, Jeffrey Park, Jialu Jiao, Kai-Rou Skelly, Kevin Kim, Karthik Rao, Masatoshi Uehara, Madison Marion, Nicholas Fitzgerald, Ricardo Dias, Suyash Shringarpure, Yiqing Yuan, and Yana Wang. LifeSciBench: Evaluating language models on realistic, expert-level tasks in the life sciences, 2026. Preprint.
- Yuyang Liu, Lijun Lv, Xue Zhang, Li Yuan, and Yonghong Tian. BioProBench: A corpus and benchmark for biological protocol reasoning in autonomous science, 2025.
- Lindsay Mitchener, Jon M. Laurent, Benjamin Tenmann, Siddharth Narayanan, Geemi P. Wellawatte, Andrew D. White, Lorenzo Sani, and Samuel G. Rodriques. BixBench: A comprehensive benchmark for LLM-based agents in computational biology, 2025.
- OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. TroubleshootingBench, Section 5.2.1.5.

A Protocol Domain and Task Intent Definitions

A.1 Protocol Domains

Each source protocol is assigned exactly one domain. Where a protocol spans two domains, we assign the domain corresponding to the step in which the task's modification falls.

Protein biochemistry and purification: Expression, extraction, chromatographic or affinity purification, buffer exchange, concentration, and biophysical or biochemical characterization of proteins. Includes enzymatic assays where the analyte is a purified protein. Excludes protein detection performed as a readout of a cell-based experiment, which is assigned to cell culture and cell-based assays.

Cell culture and cell-based assays: Maintenance, passaging, transfection, transduction, differentiation, and treatment of cultured cells, and assays whose readout is a property of the cell population. Includes viability, proliferation, reporter, and cytotoxicity assays. Excludes primary tissue and animal work, which is assigned to model organism, in vivo, and primary tissue.

Molecular biology and cloning: Nucleic acid manipulation: PCR, restriction and ligation, assembly cloning, mutagenesis, plasmid preparation, transformation, and sequence verification of constructs. Excludes library preparation intended for high-throughput sequencing, which is assigned to genomics, transcriptomics, and sequencing.

Staining, imaging, and cytometry: Sample fixation, permeabilization, labeling with antibodies or dyes, mounting, microscopy, image acquisition, flow cytometry, and analysis of the resulting images or cytometry data.

Genomics, transcriptomics, and sequencing: Nucleic acid extraction for sequencing, library preparation, quantification and quality control of libraries, sequencing runs, and associated sample handling. Excludes downstream computational analysis performed independently of sample handling.

Microbiology and microbial methods: Culture, isolation, enumeration, and characterization of bacteria, yeast, or other microorganisms, including growth curves, plating and colony counting, competent cell preparation, and antimicrobial susceptibility testing.

Media, reagent prep, and recipe: Preparation of media, buffers, stock solutions, and reagent mixtures, including formulation, sterilization, storage, and quality control, where the protocol's product is the reagent rather than an experimental result.

Model organism, in vivo, and primary tissue: Procedures on live animals or organisms, and harvest, dissociation, or handling of primary tissue, including dosing, husbandry-adjacent steps required by the protocol, and tissue processing before downstream assay.

Chemical synthesis and conjugation: Synthesis, modification, or conjugation of small molecules and bioconjugates, including labeling chemistry, purification of synthetic products, and characterization by chemical rather than biological means.

A.2 Task Intents

Each task is additionally assigned a single primary intent, defined as the intent corresponding to the rubric criteria carrying the largest share of total weight.

Plan audit: The task requires evaluating a proposed or existing experimental plan for correctness, completeness, or internal consistency, and identifying what is missing, misordered, or unsupported.

Measurement and result interpretation: The task requires interpreting a quantitative or qualitative result (e.g., a yield, curve, gel, plot, or control outcome) and stating what it supports or rules out.

Troubleshooting: The task presents a failed, degraded, or anomalous outcome and requires identifying probable causes and the corrective action or diagnostic step that discriminates among them.

Recordkeeping: The task requires specifying what must be documented, versioned, or tracked for a modification to be interpretable or reproducible later, including lot numbers, instrument settings, deviations, and sample provenance.

Procedure specification and adaptation: The task requires modifying or extending a procedure for a new goal, sample type, scale, or reagent, and specifying resulting steps concretely enough to execute.

Quantitative formulation: The task requires a calculation where the numerical answer and correctness of its derivation are the object of assessment.

Follow-up experiment: The task requires proposing a subsequent experiment, including what it would measure and what its outcome would establish.

A.3 Labeling Procedure

Labels were assigned by GPT-5.6 Sol at extra-high reasoning. Task classifications were reviewed to ensure consistency and accuracy.

B Model Inference Configurations

Closed systems used provider-hosted APIs; Kimi K3, GLM 5.2, and Inkling were served through Baseten. All systems used the highest reasoning tier exposed by their provider.

Table B1: Model inference configurations. Access route, reasoning tier, and search provider are listed per model.

Model	Access	Reasoning	Search provider
Claude Opus 5	Anthropic API	max	First party / native
GPT-5.6 Sol	OpenAI API	max	First party / native
GPT-5.6 Terra	OpenAI API	max	First party / native
GPT-5.6 Luna	OpenAI API	max	First party / native
Gemini 3.6 Flash	Vertex AI	HIGH	First party / native
Grok 4.5	xAI Responses	high	First party / native
Kimi K3	Baseten (moonshotai/Kimi-K3)	max	Exa "auto"
GLM 5.2	Baseten (zai-org/GLM-5.2)	max	Exa "auto"
Inkling	Baseten (thinkingmachines/inkling)	max	Exa "auto"

C Alternate Grader Agreement

We regraded 20 tasks with Claude Opus 5 under an identical system prompt. Exact agreement was 81.2%, and within-one-level agreement was 99.95%, with Claude Opus 5 being a slightly more permissive grader.

Table C1: Grader agreement on a 20-task subset. Mean scores under each grader are shown with paired shift and 95% confidence interval.

Model	Terra	Opus	Shift	95% CI
Claude Opus 5	0.550	0.656	+0.106	[0.084, 0.130]
GPT-5.6 Sol	0.476	0.549	+0.073	[0.046, 0.104]
Kimi K3	0.432	0.498	+0.066	[0.034, 0.101]
GPT-5.6 Terra	0.465	0.513	+0.048	[0.022, 0.076]
GPT-5.6 Luna	0.425	0.479	+0.054	[0.031, 0.081]
Inkling	0.383	0.442	+0.059	[0.032, 0.088]
Grok 4.5	0.391	0.432	+0.041	[0.012, 0.074]
GLM 5.2	0.305	0.330	+0.025	[0.006, 0.045]
Gemini 3.6 Flash	0.295	0.338	+0.043	[0.024, 0.065]
Pooled across systems	0.414	0.471	+0.057	[0.038, 0.078]

D Distribution of Average Performance by Model

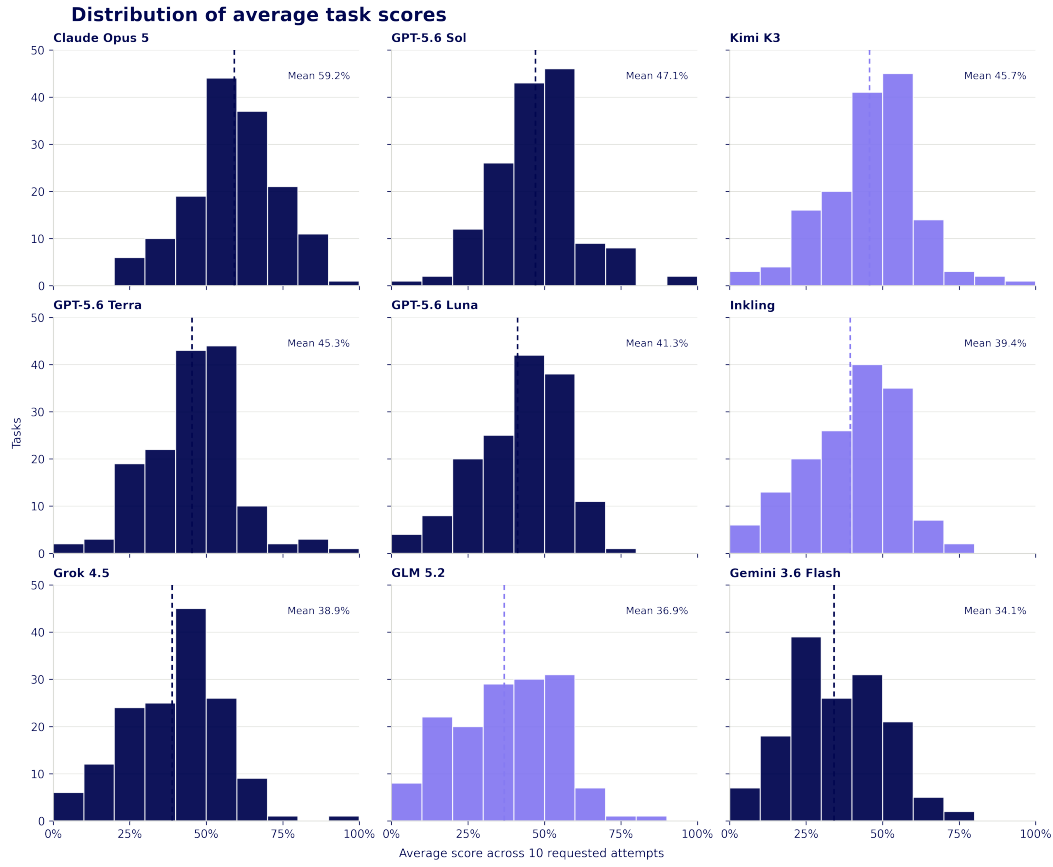


Figure D1: Distribution of average task scores by model. Per-task means across ten attempts are shown for each model, with the model mean marked.

E Performance by Source-Protocol Domain and Expected best-of- k Metrics

Table E1: Performance by source-protocol domain. Mean normalized rubric score is shown per domain with 95% confidence intervals.

Model	Protein	Cell	Molecular	Staining	Genomics	Microbiology	Media	Organism	Chemical
Claude Opus 5	.581 [.526,.634]	.584 [.531,.638]	.640 [.580,.692]	.589 [.534,.644]	.561 [.489,.626]	.583 [.535,.641]	.603 [.448,.758]	.596	.567
GPT-5.6 Sol	.430 [.382,.471]	.465 [.417,.515]	.510 [.452,.578]	.508 [.469,.552]	.438 [.371,.497]	.479 [.402,.540]	.515 [.508,.523]	.471	.375
Kimi K3	.416 [.361,.461]	.467 [.390,.537]	.496 [.446,.541]	.430 [.374,.489]	.470 [.407,.536]	.463 [.378,.545]	.520 [.500,.540]	.582	.275
GPT-5.6 Terra	.413 [.360,.459]	.466 [.403,.532]	.501 [.451,.566]	.487 [.438,.533]	.405 [.338,.469]	.439 [.372,.502]	.492 [.463,.521]	.500	.250
GPT-5.6 Luna	.394 [.339,.444]	.387 [.316,.455]	.461 [.400,.514]	.448 [.403,.490]	.387 [.316,.455]	.414 [.335,.479]	.395 [.394,.396]	.382	.183
Inkling	.374 [.326,.417]	.367 [.298,.434]	.461 [.401,.515]	.394 [.330,.457]	.388 [.317,.457]	.371 [.249,.471]	.393 [.360,.425]	.550	.125
Grok 4.5	.363 [.299,.420]	.365 [.291,.435]	.422 [.350,.498]	.397 [.338,.454]	.417 [.356,.476]	.376 [.299,.462]	.427 [.404,.450]	.446	.242
GLM 5.2	.372 [.305,.433]	.368 [.285,.449]	.441 [.367,.502]	.383 [.318,.446]	.306 [.259,.355]	.293 [.197,.393]	.322 [.256,.387]	.525	.183
Gemini 3.6 Flash	.318 [.257,.375]	.328 [.243,.411]	.399 [.341,.450]	.346 [.298,.390]	.342 [.277,.402]	.310 [.215,.400]	.387 [.344,.431]	.279	.233

Table E2: Expected best-of- k metrics. Scores at $k = 1, 2, 5$, and 10 are shown with the gain from $k = 1$ to $k = 10$.

System	@1	@2	@5	@10	Gain	95% CI @10
Claude Opus 5	.592	.651	.714	.753	+ .161	[.728,.777]
GPT-5.6 Sol	.471	.509	.551	.575	+ .104	[.552,.598]
Kimi K3	.457	.528	.592	.628	+ .171	[.601,.654]
GPT-5.6 Terra	.453	.497	.546	.577	+ .123	[.550,.603]
GPT-5.6 Luna	.413	.453	.498	.527	+ .115	[.501,.553]
Inkling	.394	.449	.506	.541	+ .147	[.512,.569]
Grok 4.5	.389	.445	.499	.533	+ .144	[.499,.565]
GLM 5.2	.369	.435	.503	.549	+ .181	[.521,.577]
Gemini 3.6 Flash	.341	.394	.446	.476	+ .134	[.448,.503]

F Additional Sample Tasks

Example task 2

Microbiology and Microbial Methods

A robot perfectly confounds plasmid identity with dispense order

Question

I'm automating 96 transformations with a different plasmid in every well. At time zero the robot makes one room-temperature reservoir containing the protocol's ssDNA/PEG/LiOAc/DTT mixture plus all of the yeast and gently vortexes it once. It dispenses the cell-containing mixture across the plate over 24 minutes and adds each well's plasmid immediately after that well is filled; the whole plate then starts the 30 C incubation together. Each plasmid appears once, in dispense order, and the first wells produced about twice as many colonies as the last wells. Because the protocol says plasmid may be added after yeast, a colleague wants to fit a smooth well-order correction and interpret the residuals as plasmid effects. Can this plate support that comparison, or is a rerun required? Give the smallest deck-order and control redesign that would make plasmid effects interpretable without changing the protocol's intended mixing states.

Protocol context

High-efficiency *S. cerevisiae* lithium acetate transformation · protocols.io
<https://www.protocols.io/view/high-efficiency-s-cerevisiae-lithium-acetate-trans-j4tcqwn.html>

Rubric

1	Explains why the optional post-yeast plasmid note does not validate this schedule and quantifies the unequal exposure: first versus last wells receive about 24 versus 0 minutes of pre-incubation plasmid contact, hence about 54 versus 30 total minutes including the common 30 C incubation; also identifies the out-of-order room-temperature PEG/LiOAc/DTT exposure.	2	Identifies the independent cell-dose/homogeneity risk from a 24-minute viscous, cell-containing PEG reservoir after only one gentle vortex, and explains why repeatedly vortexing strongly would not be a neutral fix once cells are present.	3	Rejects computational well-order correction as non-identifying because every plasmid is observed once and plasmid identity is perfectly confounded with dispense time, DNA-contact time, and reservoir drift; limits the existing plate to process-QC evidence rather than plasmid-effect inference.	4	Redesigns the run by vortexing a cell-free ssDNA/PEG/LiOAc/DTT master well on ice, predispersing master and plasmids, adding equal yeast last with only gentle mixing, and using synchronized short incubation blocks; additionally randomizes plasmid order and repeats a common control plasmid across early/middle/late positions to estimate residual order effects.
	10 pts		10 pts		10 pts		10 pts

Example F1: Sample task from microbiology and microbial methods. The scientist query, source protocol, and weighted rubric criteria are shown.

Example task 3

Microbiology and Microbial Methods

One plaque pick is not yet a clonal phage isolate

Question

I'm isolating lytic phage from soil against one host strain. My first double-layer overlay of the crude enrichment filtrate gave a lawn with three big clear plaques and, off to one side of the same plate, a few small turbid ones. I stabbed the largest of the clear plaques - nothing else within about a centimetre of it - with a sterile P200 tip, dropped the tip into 10 mL LB, pipetted up and down to release the phage, added 250 µL of mid-log host, shook it overnight at 37 °C, spun it, and pushed the supernatant through a 0.22 µm filter. That filtrate is about 2×10^9 PFU/mL, and when I plate it out every plaque looks the same: same size, sharp edges, completely clear. My plan now is to freeze this filtrate as the working stock for the project, run host range, efficiency of plating and TEM on it, and describe the isolate in the methods as a single phage. Is that plan sound as it stands? Tell me what I should actually do with this filtrate before any of that characterisation, and how I should be reading the plates as I go.

Protocol context

Isolation and characterization of bacteriophages from soil against food spoilage and foodborne pathogenic bacteria · Scientific Reports
<https://www.nature.com/articles/s41598-023-36591-6>

Rubric

1 Rejects the plan and supplies the mechanism: states that this picked-plaque-plus-overnight-broth-enrichment filtrate is not established as a single phage, because the overnight host enrichment amplifies whatever was co-picked instead of separating it (no single-particle bottleneck) and competition during that growth can let a minority co-picked phage dominate, so neither the $\sim 10^9$ PFU/mL titer nor the uniform plaque appearance is evidence of clonality. 10 pts	2 Specifies three cycles of single-plaque purification and the operations within each cycle: pick a well-isolated clear plaque, elute the tip into a small volume (about 1 mL) of SM buffer, prepare 1:10 serial dilutions of that eluate, and replat the higher dilutions with mid-log host by double-layer agar overlay, carrying the diluted eluate forward directly rather than amplifying it in broth between cycles. 10 pts	3 States that the existing filtrate must not be frozen or used as the working/characterisation stock, and that the stock must instead be re-derived from a single well-isolated plaque of the final purification cycle (eluted, propagated on fresh mid-log host, then clarified, 0.22 µm filtered and titered), so that host range, EOP and TEM are run on the purified stock rather than on the current filtrate. 10 pts	4 Gives the plate outcome that qualifies a cycle and its next pick: a countable plate with roughly 30-300 well-separated plaques, with no confluent or webbed lysis and no plaques touching their neighbours, and re-plating at a different dilution rather than picking when a plate falls outside that window. 5 pts
---	---	--	--

Example F2: Sample task from microbiology and microbial methods. The scientist query, source protocol, and weighted rubric criteria are shown.

Example task 4

Molecular Biology and Cloning

A plasmid-sized DNA band and an RNA doublet do not certify linearization

Question

I'm deciding whether to release an SP6-transcribed pCS2+ RNA batch or repeat the NotI linearization and IVT. The retained NotI-treated DNA had been run only beside a 1 kb DNA ladder: it showed a predominant band near the expected plasmid size and a faint faster band, but no undigested plasmid lane was included. The cleaned RNA is >1000 ng/µL by Qubit, while an ordinary native agarose gel shows a doublet, with one band near the predicted transcript length and one migrating more slowly. One scientist says the slower RNA band proves failed linearization and wants to discard the batch; another says the DNA band and Qubit concentration are enough to release it. What can these records actually certify, and what is the minimum ordered QC decision path that should determine whether re-linearization and a new IVT are required?

Protocol context

How to make Tol2 mRNA · protocols.io
<https://www.protocols.io/view/how-to-make-tol2-mrna-cqavvse6.html>

Rubric

1 Explains why the current evidence certifies neither complete linearization nor correct full-length RNA: a DNA ladder alone cannot resolve anomalously migrating circular plasmid topologies, Qubit measures RNA amount rather than size/integrity, and a native RNA doublet can result from secondary structure rather than two transcript lengths. 10 pts	2 Makes denaturing RNA electrophoresis with an RNA size ladder the first decision step and interprets both branches: collapse to one predicted-length band supports a native-structure/mobility artifact, whereas a persistent longer product warrants investigation of incomplete template linearization. 10 pts	3 In the persistent-long-product branch, requires comparing the retained NotI digest with undigested/circular plasmid, explains that a small circular fraction can yield abundant long RNA because phage polymerase is processive, and limits re-linearization/re-IVT to evidence of residual circular template rather than recommending it from the native gel alone. 10 pts
--	---	---

Example F3: Sample task from molecular biology and cloning. The scientist query, source protocol, and weighted rubric criteria are shown.

Example task 5

Protein Biochemistry and Purification

Bubbles in an anisotropy plate can bend the binding curve

Question

I am running a fluorescence-anisotropy titration of His6-MBP-MYC-N (1-154) against a 5' 6-FAM-labelled RNA oligo. Twelve half-log protein points per curve, 5 nM tracer, reactions assembled in a 96-well plate and then 18 uL transferred per well into a 384-well plate, top read at a single fixed Z-height with gains of 100/120/140 (excitation 470 +/- 20 nm, emission 518 +/- 20 nm), and every curve zeroed against wells containing protein dilution buffer alone. On most plates two or three scattered wells come back with unusually high raw counts and with polarization values that fall off the titration. There is no plate centrifuge in the room where I read, so the 1000 rpm, 1 min pre-read spin specified in the published rG4 FP method is not available to me at that point. How should I be handling these plates immediately before the read, and what does this failure mode mean for my analysis and for whether I could later adopt the published spin?

Protocol context

Fluorescence polarization assay · Bio-protocol Exchange
<https://bio-protocol.org/exchange/minidetail?id=7791167&type=30>

Rubric

1 States that the pre-read control here is manual/mechanical bubble removal after transfer into the 384-well plate - gently tapping the plate and confirming no air bubbles in any well, including the buffer-alone blank wells - rather than centrifugation.	2 Gives the FP-specific mechanism: because polarization is a ratio of parallel to perpendicular intensity a uniform brightness change would largely cancel, but a bubble/uneven meniscus in a low-volume fixed-Z-height top read refracts or back-reflects the polarized excitation and perturbs the two channels unequally (and can drive a channel toward saturation, e.g. at gain 140), so both raw counts and mP shift; concludes these wells are optical artifacts rather than binding signal and that raw channel intensities/gain series should be inspected rather than mP alone.	3 Identifies the propagation consequence of zeroing to protein-dilution-buffer-alone wells: an artifact in a blank or endpoint well shifts the offset/span applied to the entire curve and biases the fitted affinity rather than corrupting one data point, and prescribes replicated bubble-screened blanks plus flagging/excluding affected wells and refitting.	4 Sets conditions before adopting the 1000 rpm spin instead of treating it as interchangeable: apply it uniformly to all wells including blanks, re-verify or re-optimize Z-height and gain because spinning changes meniscus/fill geometry, preserve the equilibration temperature/dark-timing despite added handling, and confirm on a tapped-versus-spun comparison that blank baseline mP and the fitted Kd are unchanged.
5 pts	10 pts	10 pts	5 pts

Example F4: Sample task from protein biochemistry and purification. The scientist query, source protocol, and weighted rubric criteria are shown.

G Average Answer-Generation Cost per Attempt

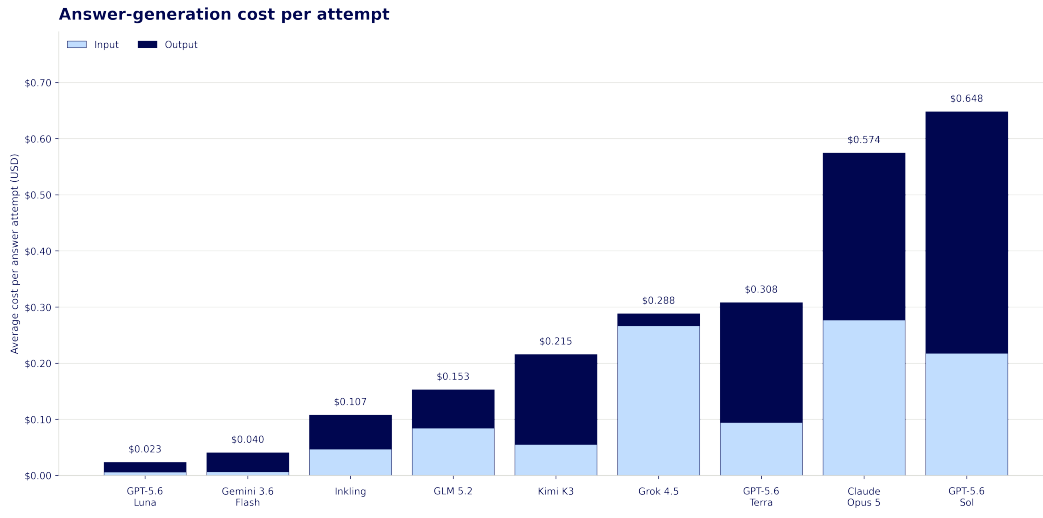


Figure G1: Answer-generation cost per attempt. Average input and output cost is shown per model, excluding web search and hosted-tool usage.

Average cost per attempt across model families was computed using input and output token counts and API pricing on August 13, 2026, for Gemini 3.6 Flash, GPT-series models, Claude Opus 5, and Grok 4.5. Pricing for GLM 5.2, Inkling, and Kimi K3 is representative pricing from Baseten's model inference API service, collected on August 13, 2026. Costs exclude web search or hosted-tool usage.

H Performance–Cost Frontier

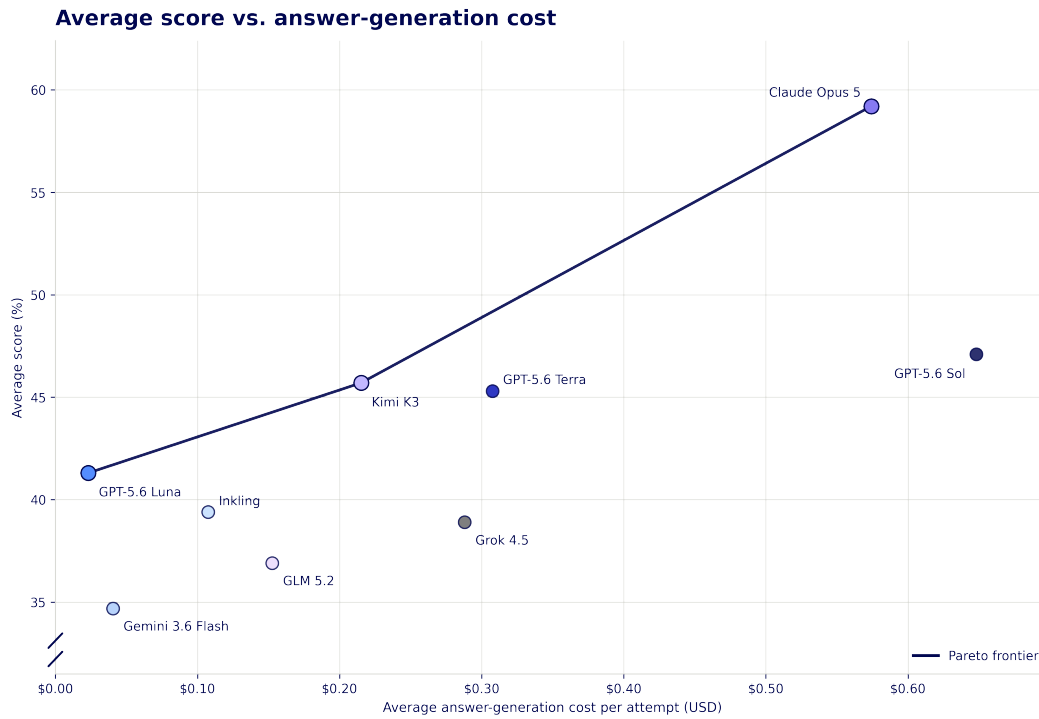


Figure H1: Average score against answer-generation cost per attempt. The Pareto frontier is drawn over GPT-5.6 Luna, Kimi K3, and Claude Opus 5; the score axis is truncated.

Figure H1 shows that GPT-5.6 Luna, Kimi K3, and Claude Opus 5 lie on the performance–cost Pareto frontier. Claude Opus 5 achieves the highest score (59.2%) while costing less than GPT-5.6 Sol, and Kimi K3 slightly outperforms GPT-5.6 Terra at lower cost. All models were assessed at their maximum reasoning settings, which can strongly affect costs and therefore the Pareto frontier.

I Model Misformatted Responses and Refusals

Table I1: Rate of refusals and misformatted outputs across model families.

Model	Misformatted	Refusals	Total
Claude Opus 5	0 / 1,490	0 / 1,490	0.0%
GPT-5.6 Sol	0 / 1,490	0 / 1,490	0.0%
Kimi K3	62 / 1,490	0 / 1,490	4.2%
GPT-5.6 Terra	0 / 1,490	0 / 1,490	0.0%
GPT-5.6 Luna	0 / 1,490	0 / 1,490	0.0%
Inkling	28 / 1,490	0 / 1,490	1.9%
Grok 4.5	10 / 1,490	0 / 1,490	0.7%
GLM 5.2	181 / 1,490	0 / 1,490	12.1%
Gemini 3.6 Flash	0 / 1,490	106 / 1,490	7.1%

GLM 5.2 produced misformatted responses on 181 of 1,490 attempts (12.1%), followed by Kimi K3 on 62 attempts (4.2%), Inkling on 28 attempts (1.9%), and Grok 4.5 on 10 attempts (0.7%). Gemini 3.6 Flash refused 106 of 1,490 requests (7.1%). Claude Opus 5 and all three GPT-5.6 models produced valid responses for every attempt. Unsuccessful attempts because of refusals or misformatting were assigned a score of zero.