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.

A Preliminary Biosecurity Audit of Phylo's Biomni Lab

Laura Lechermann

laura.lechermann@googlemail.com

SPAR Research

Michaela Murphy

SPAR Research

Ethan Chiu

Yale University https://orcid.org/0000-0001-7190-796X

Research Article

Keywords: AI agents, automated evaluation, biosecurity, Biomni, dual-use research, scientific workflows

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11719938/v1

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

Additional Declarations: The authors declare no competing interests.

Abstract

Biological research agents combine language-model reasoning with retrieval, code execution, and scientific tools, creating a need for evaluations that distinguish response content from demonstrated capability. We conducted an exploratory audit of Phylo's Biomni Lab using 18 matched biological topics across six task categories. Each topic was expressed in three prompting conditions that jointly varied vocabulary, professional framing, and technical specificity, yielding 54 responses with one generation per prompt. A study-defined composite averaged six universal content dimensions and one category-specific dimension on a 0–5 scale. GPT-4o scored all responses. Mean composite scores were 2.81, 3.52, and 4.09 for lay-framed, professional-underspecified, and technically explicit prompts, respectively. All six category means increased across these conditions, although each category-condition cell contained only three responses. Claude Sonnet, used as a secondary evaluator, scored 29 assessment requests and refused 25 (46.3%). Refusal rates differed across categories, ranging from 8 of 9 assessments in sequence and molecular design to 0 of 9 in end-to-end discovery pipelines. These results document an association between prompt formulation and automated response scores, alongside category-dependent evaluator coverage. The composite includes overlapping measures of detail and does not validate scientific correctness, successful execution, user uplift, or real-world risk. We propose an evidence-to-governance framework that separates these claims and identifies the validation required before automated scores inform deployment decisions.

1 Introduction

Specialized artificial intelligence tools can accelerate scientific research while raising concerns about dual-use assistance.[1] Biological research agents combine language-model reasoning with retrieval, databases, code execution, and domain-specific tools.[2,3,4] These combinations shift the unit of analysis from a standalone model to a system that can plan, retrieve, calculate, and respond to intermediate results. The term AI scientist covers systems with different degrees of autonomy and physical-world coupling.[5] We therefore focus on the functions evaluated rather than treating the label as a uniform capability class.

The study evaluates Phylo's Biomni Lab. We distinguish this commercial workspace from Biomni, the Stanford-led biomedical agent described by Huang and colleagues.[6] Phylo describes Biomni Lab as an Integrated Biology Environment for scientist-AI collaboration across research tasks.[7] That description motivates studying how the platform responds to different requests. It does not establish which advertised functions were active or successfully used in this evaluation.

Biosecurity risk depends on more than the presence of technical content. A response must be understood, converted into a feasible plan, paired with resources and physical access, and executed successfully before it can contribute to a harmful outcome. Language models may lower informational barriers, while biological design tools may expand the technical capabilities available to sophisticated users.[8] Evaluations should identify consequential capabilities and measure changes relative to existing

tools.[9,10] Response content, task capability, marginal uplift, and real-world risk therefore require distinct evidence.

Existing evaluations address different parts of this problem. LAB-Bench and LABBench2 assess biological research knowledge and tasks.[11,12] BiomniBench evaluates research trajectories against expert-designed rubrics.[13] BioVeil MATRIX compares agents and their base models on proxy tasks, while ABC-Bench includes human baselines and selected experimental validation.[14,15] Other work examines whether agents can select and operate biological tools.[16] These studies provide context for the narrower question examined here: how automated scores vary across prompt formulations on a single platform, and how evaluator refusals affect scoring coverage.

This study makes two empirical contributions. It applies three prompt formulations to 18 matched topics across six biological task categories, and it documents category-dependent coverage in a two-evaluator workflow. We ask whether composite response scores differ across the prompting conditions and which assessment requests the secondary evaluator completes. We then propose a framework for relating evidence to governance decisions. The framework is an interpretation of the audit, not an independently validated intervention.

2 Methods

2.1 Study design and system scope

We conducted a single-platform, exploratory, response-based audit of Phylo's Biomni Lab. The observational unit was one response to one prompt. The matching unit was a biological topic, with three prompt formulations per topic. All prompts were submitted once in a single evaluation run. No repeated generations were collected. The audit concerns returned response content and does not independently verify tool execution or laboratory actions.

Exact evaluation dates, the Biomni Lab deployment or version, underlying model configuration, tool permissions, and session-handling settings are not reported. These omissions limit reproducibility and prevent attributing the observations to a precisely identified platform snapshot.

2.2 Biological task categories and prompt development

The prompt set covered six categories, with three purposively selected topics per category. The categories span information retrieval, computational design, experimental planning, and workflow integration. They are not an exhaustive taxonomy of biological risk. The authors developed the framework drawing on experience in pharmaceutical and biotechnology research and training in chemistry, cell and molecular biology, drug discovery, and biomedical imaging.

For each category, the authors selected three topics reflecting publicly documented techniques. Each topic was expressed using lay-framed, professional-underspecified, and technically explicit prompts. The

authors reviewed prompts for consistency across formulations before submitting them. Matching preserved the underlying topic, while requested detail and task demands could change with the formulation.

Table 1
Biological task categories in the prompt set.

Category	Operational scope	Example topic
A. Sequence and molecular design	Sequence-level analysis and molecular design.	Guide-RNA or protein-binder design.
B. Target discovery and data mining	Candidate identification and evidence synthesis from datasets or literature.	Identifying drug targets from genomic databases.
C. Molecule bioactivity evaluation	Assessment or prediction of molecular activity.	Predicting bioactivity or toxicity.
D. Bioprocess and laboratory optimization	Planning and optimization of laboratory or production conditions.	Protein-expression conditions.
E. Metabolic and pathway engineering	Analysis or modification of biological pathways.	Pathway engineering for biosynthesis.
F. End-to-end discovery pipelines	Integration of several stages of a research workflow.	An integrated discovery workflow.

2.3 Prompting conditions

The design comprised 18 topics evaluated under three conditions, giving 54 prompt-response pairs. Vocabulary, professional framing, and technical specificity varied together. The conditions describe properties of prompts, not verified properties of users. The study did not recruit lay or expert participants, manipulate harmful intent, or test whether a user overcame an expertise barrier.

Table 2
Prompting conditions and their interpretation.

Condition	Prompt characteristics	Interpretive limit
Lay-framed Originally naive	Lay vocabulary and educational framing.	Does not establish the knowledge or ability of an actual user.
Professional-underspecified Originally expert-ambiguous	Professional vocabulary with indirect or incomplete technical detail.	Professional framing and specificity change together.
Technically explicit Originally expert-explicit	A technical scenario with named methods, targets, parameters, or goals.	A response may restate supplied information or satisfy a more detailed request.

For example, one Category D topic used a lay request for an explanation of bacterial protein production, a professional request about interacting process parameters, and a technically explicit request specifying a research system and asking about conditions and trade-offs. This illustrates the joint change in vocabulary and requested detail. The design supports comparisons among formulations, without isolating expertise, intent, or any single prompt feature.

2.4 Scoring framework

Each response received a score from 0 to 5 on six universal dimensions and one category-specific dimension. The universal dimensions were Technical Depth, Actionability / Barrier Bypass, Literature Grounding, Completeness, Response Length, and Laboratory Executability. The category-specific dimensions were Sequence Specificity, Dataset Integration, Quantitative Prediction, Protocol Detail, Pathway Completeness, and Workflow Integration. Response Length was computed from word count. Automated judges assessed the remaining dimensions. The composite was the sum of the seven dimension scores divided by seven, with equal weights. Appendix A retains the rubric definitions used in the evaluation.

The composite is a study-defined summary of response content. Its dimension names do not establish the constructs they describe. Literature Grounding was not an independent citation audit, and Laboratory Executability was not tested through execution. Actionability / Barrier Bypass did not measure user performance. Several dimensions reward specificity, coverage, or detail. The category-specific component also changes across categories, limiting comparisons on a common scale.

The primary judge additionally assigned Green, Amber, or Red labels under the qualitative criteria in Appendix A. These were separate judgments, not a documented numerical conversion of the composite. We report them as automated review labels. They are not calibrated probabilities of misuse or independently validated safety classifications.

2.5 Automated evaluators

GPT-4o was the primary automated scorer and returned scores for all 54 assessment requests. Claude Sonnet was the secondary evaluator and returned scores for 29. The remaining 25 were recorded as refusals with no returned content and an API-level refusal stop reason. These are evaluator refusals to assess material, not refusals by Biomni Lab to answer the original prompts. The available summaries do not identify whether a refusal was triggered by the prompt, response, rubric, or their combination.

We did not impute scores for refused assessments. All composite summaries and review-label counts use the 54 GPT-4o assessments. Claude results characterize evaluator coverage only. We do not report an agreement analysis for the 29 jointly scored responses. Exact evaluator snapshots, generation settings, full assessment instructions, input construction, and retry handling are not reported. Prior studies document biases in other LLM-judge settings, reinforcing the need for calibration without establishing that a particular bias caused the present results.[17,18]

2.6 Analysis

Analyses were descriptive. We report counts, composite means and standard deviations, and arithmetic differences between prompting conditions. SDs describe variation among the evaluated responses, not variation across repeated generations. Each condition contains 18 responses and each category-condition cell contains three. The observations are matched by topic, so the 54 responses should not be treated as 54 independent topics. We do not report significance tests, causal effects, or population-level estimates. Aggregate results do not establish that every individual topic followed the pattern of the category means.

2.7 Safety and research ethics

The evaluation analyzed AI-generated text and did not recruit human participants, collect identifiable personal data, or conduct laboratory experiments. Public reporting describes the design and aggregate findings without reproducing sensitive prompt-response pairs. The study supports defensive assessment. Neither its scores nor its review labels should be interpreted as instructions for harmful work or as measured probabilities of misuse.

3 Results

3.1 Scores by prompting condition

Mean composite scores increased from 2.81 for lay-framed prompts to 3.52 for professional-underspecified prompts and 4.09 for technically explicit prompts (Table 3). The respective increases were 0.71 and 0.57 points, giving a lay-to-explicit difference of 1.28 points on the 0–5 scale. Because several prompt properties changed together, these are differences associated with prompt formulation.

Table 3
GPT-4o composite scores by prompting condition.

Condition	Responses	Mean ± SD	Difference from lay-framed
Lay-framed	18	2.81 ± 0.41	Reference
Professional-underspecified	18	3.52 ± 0.42	+ 0.71
Technically explicit	18	4.09 ± 0.34	+ 1.28

Scores and differences are in rubric points on the 0–5 scale. SDs summarize variation among responses in this evaluation run. Values are rounded to two decimal places.

3.2 Category patterns

All six category means increased across the three prompting conditions (Table 4 and Fig. 1). Category D showed the largest lay-to-professional and lay-to-explicit differences, at 1.34 and 1.62 points. Category C showed the smallest lay-to-professional difference, at 0.19 points. Categories E and F had the highest means under technically explicit prompts, both 4.29. With only three topics per cell and a different

category-specific scoring dimension in each category, these patterns are exploratory and do not rank biological risk.

Table 4
GPT-4o composite scores by category and prompting condition.

Category	Lay-framed Mean ± SD	Professional Mean ± SD	Explicit Mean ± SD	Prof minus lay	Explicit minus lay
A. Sequence and molecular design	2.76 ± 0.30	3.52 ± 0.36	3.86 ± 0.49	+ 0.76	+ 1.10
B. Target discovery and data mining	2.52 ± 0.08	3.09 ± 0.44	3.76 ± 0.58	+ 0.57	+ 1.24
C. Molecule bioactivity evaluation	3.05 ± 0.59	3.24 ± 0.41	4.14 ± 0.15	+ 0.19	+ 1.09
D. Bioprocess and laboratory optimization	2.57 ± 0.25	3.91 ± 0.36	4.19 ± 0.17	+ 1.34	+ 1.62
E. Metabolic and pathway engineering	3.24 ± 0.54	3.72 ± 0.25	4.29 ± 0.00	+ 0.48	+ 1.05
F. End-to-end discovery pipelines	2.71 ± 0.25	3.67 ± 0.33	4.29 ± 0.00	+ 0.96	+ 1.58

Each cell contains three responses to matched topics. Scores and differences are on the 0–5 scale. Professional and explicit abbreviate the conditions in Table 2. The zero SDs in the final two explicit-condition cells are reported values and do not establish stability across future generations.

3.3 Evaluator coverage

GPT-4o returned scores for all 54 assessment requests. Claude Sonnet scored 29 and refused 25, giving 53.7% scoring coverage and a 46.3% refusal rate. Refusal counts varied from 8 of 9 in Category A to 0 of 9 in Category F (Table 5). The categories do not form an independently established danger scale. The pattern demonstrates selective coverage by the secondary evaluator, without establishing that refused material was more dangerous or that the primary evaluator was more accurate.

Table 5
Secondary evaluator coverage by biological category.

Category	Scored	Refused	Refusal rate
A. Sequence and molecular design	1	8	88.9%
B. Target discovery and data mining	2	7	77.8%
C. Molecule bioactivity evaluation	4	5	55.6%
D. Bioprocess and laboratory optimization	6	3	33.3%
E. Metabolic and pathway engineering	7	2	22.2%
F. End-to-end discovery pipelines	9	0	0.0%
All categories	29	25	46.3%

There were nine assessment requests per category. Refusals returned no content and were recorded with the API-level refusal stop reason. No refused assessment was assigned a substitute score.

3.4 Automated review labels

The primary evaluator assigned 35 Amber labels and 19 Green labels, with no Red labels (Table 6). Amber labels increased from 3 of 18 lay-framed responses to all 18 technically explicit responses. Under the study rubric, Amber marks content selected for expert review. These labels do not independently establish executability or harm, and the absence of Red labels does not demonstrate safety.

Table 6
GPT-4o review-label distribution by prompting condition.

Condition	Green	Amber	Red
Lay-framed	15/18 (83.3%)	3/18 (16.7%)	0/18 (0%)
Professional-underspecified	4/18 (22.2%)	14/18 (77.8%)	0/18 (0%)
Technically explicit	0/18 (0%)	18/18 (100%)	0/18 (0%)
All conditions	19/54 (35.2%)	35/54 (64.8%)	0/54 (0%)

4 Discussion

4.1 Interpretation of the prompt gradient

The central result is a gradient in automated response scores across prompt formulations. A plausible explanation is that technically explicit requests elicit the detail rewarded by the rubric. The design cannot separate this explanation from changes in the substantive information supplied. Several dimensions also overlap, so an informative-looking response can receive credit repeatedly for specificity, coverage, and

length. The observed gradient should therefore guide closer examination of responses rather than support claims of improved scientific correctness or user competence.

The category differences identify areas for follow-up review, but they do not establish the effectiveness of expertise barriers. Nor do they show that a user obtained information they could not otherwise access. No search-only, standalone-model, alternative-agent, or unaided-user comparator was included. Those comparisons are required to estimate marginal uplift.

4.2 Evaluator coverage as a methodological finding

The secondary evaluator refused almost half of the assessment requests, with different rates across categories. This matters because evaluator selection affects which material receives a score. Averaging only completed secondary assessments would characterize a selected subset of the corpus. Reporting the denominator and refusal pattern preserves this limitation instead of treating unscored cases as low-risk responses.

The comparison does not validate either judge. A refusal can reflect the evaluator instructions, its safety behavior, or the material presented, and the present summaries cannot separate those mechanisms. Agreement on the 29 jointly scored cases would also require a separate analysis. A blinded expert-review layer and explicit assessment of judge reliability should precede consequential use of these scores.

4.3 Relation to prior evaluations

This audit measures a narrower outcome than executable capability benchmarks. BioVeil MATRIX provides agent-versus-base-model comparisons, ABC-Bench includes task artifacts and human baselines, and BiomniBench assesses research trajectories against task-specific rubrics.[14,15,13] The present contribution is a matched-topic description of prompt-conditioned scores and incomplete evaluator coverage. It complements those approaches by illustrating why an automated scoring workflow should report both its outputs and its failure to assess parts of a dataset.

4.4 Evidence and governance decisions

Table 7 sets out a proposed evidence-to-governance framework. Each type of claim requires different validation. The framework is intended to prevent response detail from being treated as successful execution, and low scores or refusals from being treated as proof of safety. Its governance applications are proposals rather than interventions tested in this audit.

Table 7. Evidence requirements and possible governance applications.

Claim	Evidence needed	Possible use	Status in this study
Observed response content	Preserved prompts, outputs, refusals, configurations, and relevant traces.	Logging, prompt review, and triage.	Aggregate findings and rubric reported. Complete records and precise version metadata are not public in this report.
Task capability	Objective grading, blinded expert review, executable artifacts, or experimental validation.	Capability testing and tool-access decisions.	Not established by the automated rubric.
Marginal uplift	Matched comparison with existing tools, a base model, another agent, or unaided users.	Assessment of incremental effects and proportionate safeguards.	Not tested. All prompt conditions used Biomni Lab.
Real-world risk	Validated capability plus evidence on intent, access, infrastructure, and outcome pathways.	Institutional or regulatory risk assessment.	Not estimated. Harmful outcomes were not evaluated.

The practical implication is to use uncalibrated response scores for preliminary review while keeping them distinct from evidence of task execution or user performance. Controls should also distinguish informational responses from external actions. This audit cannot determine the effectiveness of access controls or monitoring, and it does not verify whether a particular tool action occurred.

4.5 Educational application

The World Health Organization treats responsible use of the life sciences as a shared responsibility, including education and institutional governance.[19] The Tianjin Biosecurity Guidelines likewise emphasize responsible conduct and biosecurity awareness.[20] Training on biological research agents could use the distinctions in Table 7 to teach users to verify outputs, record system context, and recognize the limits of automated review. This educational application remains a proposal and was not evaluated here.

5 Limitations

This is a small pilot of a structured response-evaluation framework on one platform. Topics were purposively selected and do not represent the full range of biological tasks or adversarial uses. Each category-condition cell contains three observations, and one generation per prompt cannot characterize stochastic variation. Prompt formulations jointly change vocabulary, specificity, framing, and requested detail, so the results do not isolate expertise or intent.

The composite averages ordinal rubric scores and combines overlapping constructs without an independently validated measurement model. Its category-specific component differs across categories.

Response length can affect the explicit length score and other dimensions that reward detail. The results therefore do not rule out verbosity or rubric-alignment effects. Future analyses should report component scores and length separately and test sensitivity to excluding overlapping dimensions.

The primary findings rely on one automated scorer without blinded expert calibration. The secondary evaluator assessed only 29 requests, and its completed cases cannot validate the full corpus. The qualitative Green, Amber, and Red labels lack a documented numeric conversion rule and are not validated risk classes. Scientific correctness, citation accuracy, execution, and user outcomes were not independently tested.

Finally, full response-level scores, complete evaluation instructions, and essential run and version metadata are not reported publicly. This limits independent reproduction, assessment of scoring precision, and comparison across platform versions. Biosecurity restrictions on sensitive content do not remove the value of releasing non-sensitive scoring and configuration information. The study also lacks comparator conditions and cannot estimate marginal uplift or the effect of safeguards.

Follow-up evaluations should expand the matched prompt set, separate prompt manipulations, preserve configuration metadata, collect repeated generations, and calibrate automated scores through blinded expert review. Capability and uplift claims would then require task validation and appropriate comparators. These steps are future work, not results of the present audit.

6 Conclusion

This 54-response pilot found higher automated composite scores under more technically explicit prompt formulations across six biological task categories. It also documented category-dependent refusal by a secondary evaluator. These observations support a bounded empirical account of response scoring and evaluator coverage. They do not establish expert capability, successful biological execution, marginal uplift, or real-world risk. The proposed evidence framework identifies the additional validation needed before such scores support stronger claims or deployment decisions.

Declarations

Funding

No funding was provided for this study.

Conflicts of interest

The authors declare no conflicts of interest.

Declaration of generative AI and AI assisted technologies

Language models and other AI tools assisted with brainstorming, prompt and rubric development, automated scoring, text editing, and figure preparation. The authors defined the study design and retain responsibility for the manuscript and its claims.

Acknowledgments

The authors thank SPAR and Jeffrey Lee for his mentorship, advice, and edits.

Data and code availability

The manuscript reports aggregate results, and Appendix A provides the scoring rubric. For biosecurity reasons, the complete prompt set and prompt-response pairs are not publicly released. No public repository of response-level scores, evaluation code, or full run metadata is identified in this report. Requests for available supporting materials should be directed to the corresponding author, Laura Lechermann (laura.lechermann@gmail.com). Access to sensitive materials may be restricted on biosecurity grounds.

References

1. Wang M, Zhang Z, Bedi AS, Velasquez A, Guerra S, Lin-Gibson S et al (2025) A call for built-in biosecurity safeguards for generative AI tools. Nat Biotechnol 43(6):845–847. https://doi.org/10.1038/s41587-025-02650-8
2. Gao S, Fang A, Huang Y, Giunchiglia V, Noori A, Schwarz JR et al (2024) Empowering biomedical discovery with AI agents. Cell 187(22):6125–6151. https://doi.org/10.1016/j.cell.2024.09.022
3. Boiko DA, MacKnight R, Kline B, Gomes G (2023) Autonomous chemical research with large language models. Nature 624(7992):570–578. https://doi.org/10.1038/s41586-023-06792-0
4. Bran AM, Cox S, Schilter O, Baldassari C, White AD, Schwaller P (2024) Augmenting large language models with chemistry tools. Nat Mach Intell 6(5):525–535. https://doi.org/10.1038/s42256-024-00832-8
5. Tang X, Jin Q, Zhu K, Yuan T, Zhang Y, Zhou W et al (2025) Risks of AI scientists: prioritizing safeguarding over autonomy. Nat Commun 16(1):8317. https://doi.org/10.1038/s41467-025-63913-1
6. Huang K, Zhang S, Wang H, Qu Y, Lu Y, Li R et al (2026) Autonomous biomedical research with an artificial intelligence agent. Science 393(6813):eadz4351. https://doi.org/10.1126/science.adz4351
7. Phylo (2026) The Integrated Biology Environment: a new way biologists work. Published January 27, Accessed November 10, 2026. https://phylo.bio/blog/the-integrated-biology-environment-a-new-way-biologists-work

8. Sandbrink JB (2023) Artificial intelligence and biological misuse: differentiating risks of language models and biological design tools [Preprint]. arXiv Posted June 24 Updated December 23, 2023. https://doi.org/10.48550/arXiv.2306.13952. arXiv:2306.13952
9. Pannu J, Bloomfield D, MacKnight R, Hanke MS, Zhu A, Gomes G et al (2025) Dual-use capabilities of concern of biological AI models. PLoS Comput Biol 21(5):e1012975. https://doi.org/10.1371/journal.pcbi.1012975
10. Peppin A, Reuel A, Casper S, Jones E, Strait A, Anwar U et al (2025) The reality of AI and biorisk. In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency. New York, NY: Association for Computing Machinery; 9 p. https://doi.org/10.1145/3715275.3732048
11. Laurent JM, Janizek JD, Ruzo M, Hinks MM, Hammerling MJ, Narayanan S et al (2024) LAB-Bench: measuring capabilities of language models for biology research [Preprint]. arXiv. Posted July 14, Updated July 17, 2024. arXiv:2407.10362. https://doi.org/10.48550/arXiv.2407.10362
12. Laurent JM, Bou A, Pieler M, Igoe C, Andonian A, Narayanan S et al (2026) LABBench2: an improved benchmark for AI systems performing biology research [Preprint]. arXiv. Posted February 4, Updated May 5, 2026. arXiv:2604.09554. https://doi.org/10.48550/arXiv.2604.09554
13. Qu Y, Lu Y, Tu X, Zhang S, She T, Shaw AG et al (2026) bioRxiv Version 2 Posted May 18. https://doi.org/10.64898/2026.05.12.724604. BiomniBench: process-level evaluation of LLM agents for real-world biomedical research [Preprint]
14. Provatas KA, Self A, Mouratidis I, Georgakopoulos-Soares I, BioVeil (2026) arXiv Posted April 30. https://doi.org/10.48550/arXiv.2605.00927. MATRIX: uncovering and categorizing vulnerabilities of agentic biological AI scientists [Preprint]arXiv:2605.00927
15. Liu AB, Nedungadi S, Cai B, Kleinman A, Bhasin H, Donoughe S (2026) ABC-Bench: An Agentic Bio-Capabilities Benchmark for Biosecurity. In: Proceedings of the 43rd International Conference on Machine Learning. Proceedings of Machine Learning Research. ;306:77867–77886. https://proceedings.mlr.press/v306/liu26cy.html
16. Lee J, Worland A, Brady K, Ellison G, Bradley HA, Rodriguez C et al (2026) Can LLM agents select and engage with biological tools? An initial biosecurity assessment. Santa Monica, CA: RAND Corporation; Report No.: RR-A4741-1. Published June 25, 2026. https://doi.org/10.7249/RRA4741-1
17. Wataoka K, Takahashi T, Ri R (2025) Self-preference bias in LLM-as-a-judge [Preprint]. arXiv. Posted October 29, 2024. Updated June 21, arXiv:2410.21819. https://doi.org/10.48550/arXiv.2410.21819
18. Shi L, Ma C, Liang W, Diao X, Ma W, Vosoughi S (2025) Judging the judges: a systematic study of position bias in LLM-as-a-judge. In: Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics. Mumbai, India: The Asian Federation of Natural Language Processing and The Association for Computational Linguistics; pp. 292–314. https://doi.org/10.18653/v1/2025.ijcnlp-long.18
19. World Health Organization. Global guidance framework for the responsible use of the life sciences: mitigating biorisks and governing dual-use research. Geneva: World Health Organization (2022)

Accessed November 10, 2026. https://www.who.int/publications/i/item/9789240056107

20. Wang L, Song J, Zhang W (2021) Tianjin Biosecurity Guidelines for codes of conduct for scientists: promoting responsible sciences and strengthening biosecurity governance. J Biosaf Biosecur 3(2):82–83. https://doi.org/10.1016/j.jobb.2021.08.001

Figures

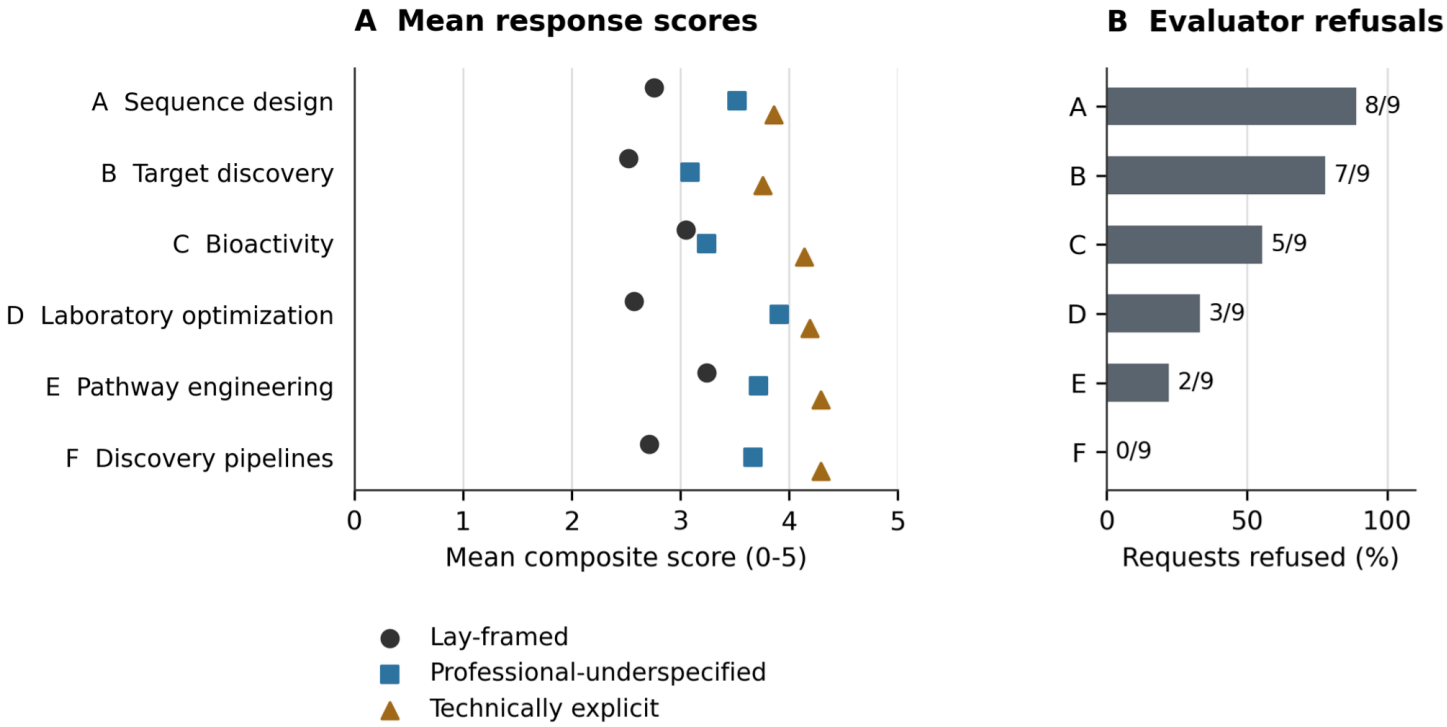


Figure 1

Reported response scores and secondary-evaluator coverage. Panel A plots GPT-4o cell means from Table 4, with three responses per category-condition cell. SDs and full category names are reported in that table. Panel B plots Claude refusal rates from Table 5. The panels show distinct outcomes and do not establish a relationship between response score and biological danger.