LifeSciBench: Evaluating Language Models on Realistic, Expert-Level Tasks in the Life Sciences

Amelia Liu[1], Andrew Ho[1], Anne Marie Droste[2], David Martin[1], Edmund Wong[1], Edward Zhou[1], Isabelle Zhou[1], Joshua Park[1], Joy Jiao[1], Katie-Rose Skelly[2], Kenny Kim[1], Kevin Rao[1], Masatoshi Uehara[1], Max Marion[2], Nicole Fitzgerald[2], Rachel Dias[1], Suyash Shringarpure[1], Yuan Yuan[1], and Yunyun Wang[1]

[1]OpenAI
[2]Tacit Labs

Abstract

We introduce LifeSciBench, a benchmark of 750 expert-authored tasks designed to evaluate whether language models can handle realistic life science research work. At present, the vast majority of biological benchmarks fail to capture the complexity of research-level work; questions are typically narrowly scoped and purely knowledge-based, while real-world work is often ambiguous and requires multiple judgment calls. Additionally, almost all existing benchmarks are scoped to at best a small collection of scientific domains. There is no existing benchmark in the life sciences with both the requisite breadth and depth required to convincingly measure proficiency in real-world professional settings. LifeSciBench addresses this gap by spanning seven scientific workflows and seven life science domains, with each task paired with an expert-written rubric. Across five frontier and domain-specialized models, GPT-Rosalind performs best with a problem-weighted normalized score of 0.576 and a 36.1% task pass rate, but the benchmark remains far from saturated: no model passes 171 tasks (22.8%), and 261 tasks (34.8%) have a best-model pass rate below 20%. LifeSciBench serves as a high-resolution evaluation of practical scientific reasoning and operational decision-making in biology.

1 Introduction

Recent advances in large language models (LLMs) have produced agentic systems that can reason, use software tools, and solve highly specialized domain-specific problems with increasing sophistication. These capabilities are especially relevant to the life sciences, where research progress relies upon a combination of reasoning from ambiguous yet complex evidence, careful experimental design, and precise decision-making in uncertain settings. To be useful to professionals or academics in the life sciences, models must be capable of handling tasks that reflect the structure and constraints of real scientific work. Existing life science evaluations measure important capabilities, but they generally focus on the precise retrieval of factual knowledge or are scoped purely within the domain of computational analysis. Such benchmarks, which involve isolated questions with clean reference answers, provide limited insight into whether models can support the broader range of reasoning-based capabilities and judgement calls required in realistic life science workflows.

Accordingly, we present LifeSciBench, a benchmark containing 750 expert-level problems designed to evaluate the ability of models to perform frontier-level work in the life sciences. The benchmark is designed around tasks that require a combination of domain-specific reasoning, data analysis, and

literature search. Additionally, the benchmark does not only evaluate the ability of the model to produce the correct analysis in response to a well-specified question; the questions are intentionally left somewhat open-ended, such that LifeSciBench also measures the model's understanding of the appropriate level of detail to supply in its responses.

LifeSciBench is designed to capture aspects of real-world research work that are often missing from existing evaluations:

1. **Complex artifacts:** LifeSciBench requires models to reason over artifacts such as images, documents, sequence files, molecular structures, and web references.
2. **Situational ambiguity:** In practice, most user interactions with LLMs are not written in a perfectly precise manner. LifeSciBench evaluates whether models can correctly interpret ambiguous context and make justified recommendations from incomplete evidence.
3. **Realism:** Tasks are organized around the actual functions that scientists perform in practice, from interpreting data and generating hypotheses to evaluating translational risk and communicating findings.
4. **Expert-authored and independently reviewed:** Each task and rubric was written by a domain specialist and reviewed through a multi-round QA process.

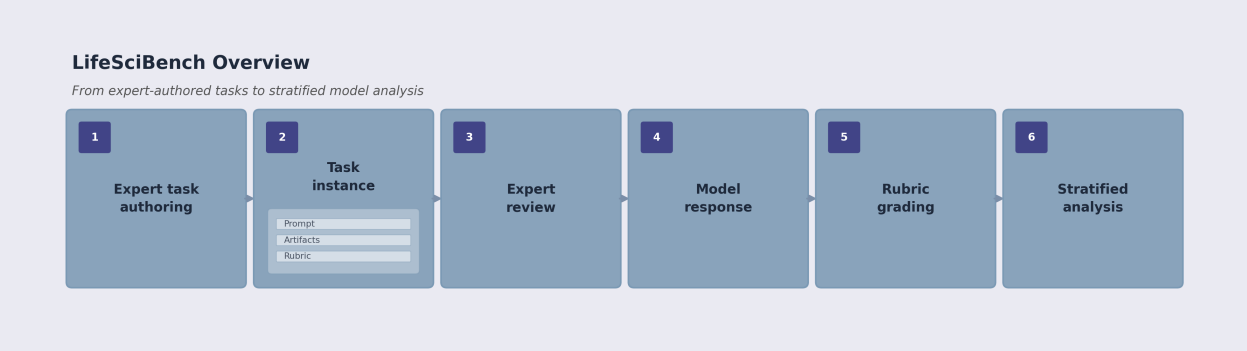


Figure 1: LifeSciBench overview. Expert scientists write tasks consisting of prompts, artifacts, and rubrics; tasks undergo expert review; models produce single-turn responses; responses are graded against task-specific rubrics and analyzed by workflow, domain, evidence type, and capability category.

2 Related Work

Recent life science benchmarks have expanded beyond factual biology question answering and now tend to differ in the particular type of research capability they assess. One line of work evaluates research-assistant behaviors such as literature retrieval, database lookup, figure and table interpretation, and reasoning about bench protocols. For instance, LAB-Bench introduced a broad suite of biology research tasks along these lines (Laurent et al., 2024). Recently, LABBench2 extended this direction with more realistic, open-response settings, including tasks involving patents, clinical trials, and messy data (Laurent et al., 2026).

These benchmarks are closest in spirit to LifeSciBench because they move toward practical biology research tasks, but they still fail to thoroughly evaluate expert scientific judgment; in practice, almost all research decisions require not only the identification or manipulation of biological information, but also the weighing of imperfect evidence, reasoning through experimental constraints, clear justification of recommended next steps, and the production of clear, actionable outputs.

A second line of work evaluates agentic capabilities in computational biology. For example, BixBench tests LLM agents on bioinformatics scenarios requiring code execution, dataset exploration, and multi-step analysis, but is mostly saturated by current frontier models (Mitchener et al., 2025). More recently, GeneBench evaluates model capabilities on extremely challenging, multi-stage quantitative biology tasks involving noisy data, quality control, statistical modeling, confounding, and downstream inference (Li et al., 2026). Uniquely, the reliance of GeneBench upon a combination of purely synthetic data and a comprehensive set of ablation studies ensures that there exists only a single unique path that models can take to arrive at the correct quantitative result. These benchmarks are valuable in that they move beyond static question answering and require models to reason over data of unknown provenance, but their scope is primarily concentrated in computational biology.

LifeSciBench addresses the remaining gap: a combination of expert-level scientific reasoning and data analysis across a broad swathe of applied life science research. Its tasks span multiple subfields of biology and stages of drug discovery. Rather than evaluating only final-answer correctness or performance within a single computational workflow, LifeSciBench uses expert-authored rubrics to assess whether models reach conclusions through scientifically valid reasoning, appropriate consideration of relevant evidence, and operationally useful decision-making.

3 Dataset Creation

3.1. Benchmark Organization & Coverage

The benchmark is primarily organized around seven broad workflow categories:

Table 1: Workflow categories in LifeSciBench.

Workflow	Description
Evidence handling	Extracting, interpreting, comparing, or synthesizing scientific evidence from papers, figures, regulatory documents, experimental outputs, or structured data.
Analysis	Performing quantitative, statistical, computational, or mechanistic analysis of biological, chemical, or experimental data.
Design and optimization	Designing or optimizing molecules, assays, experiments, constructs, protocols, or screening strategies.
Scientific reasoning	Explaining mechanisms, evaluating hypotheses, identifying causal relationships, or reasoning under uncertainty.
Validation and operations	Troubleshooting, quality control, feasibility assessment, operational planning, or translational risk evaluation.
Translation	Connecting preclinical or biological evidence to clinical relevance, therapeutic implications, patient impact, biomarker use, trial design, safety considerations, or advancement toward human studies.
Scientific communication	Summarizing, rewriting, explaining, or communicating scientific findings for a specified audience or decision context.

Life scientists perform a wide range of tasks in day-to-day work, ranging from highly specialized, domain-specific work to activities shared across many subfields of biology. We began by defining a taxonomy of problem types in life sciences by surveying practicing scientists about the workflows they use most often in applied research settings, then grouping their responses into seven central categories. The problems in LifeSciBench were designed around this taxonomy so as to ensure an appropriate balance of broadly relevant research capabilities as well as domain-specific expertise. Moreover, the

benchmark problems include problems which span both computational and experimental contexts, reflecting the need for models to move smoothly between different types of evidence and modes of analysis.

The purpose of this taxonomy is not to reduce scientific work to a fixed set of isolated categories. Many realistic tasks combine multiple capabilities, such as interpretation of evidence, mechanistic reasoning, effective decision-making, and clear scientific communication. Rather, the taxonomy provides a structured way to analyze which kinds of life-science work current models can and cannot perform. The full workflow, biological domain, and data-source taxonomies are provided in Appendix B.

3.2. Expert Writer Cohort

The tasks in LifeSciBench were created by 173 expert scientists spanning a diverse range of life science disciplines chosen to ensure that the benchmark reflects the full breadth of expertise needed to evaluate agentic AI systems across life science research. Experts were required to have completed a Ph.D. in a relevant discipline, such as biochemistry, molecular biology, neuroscience, immunology, pharmacology, medicinal chemistry, computational biology, or a related field, and have at least two years of experience as practicing scientists in the biotechnology or pharmaceutical industries. Moreover, contributors were selected to ensure good coverage across computational, experimental, translational, and clinical domains relevant to life science research and drug discovery, with the goal of producing benchmark tasks that reflect the kinds of problems encountered in applied research settings rather than being purely textbook-style questions.

3.3. Task Formulation

Each LifeSciBench task consists of a prompt, any supporting artifacts needed to answer the question, and a task-specific grading rubric.

Questions. Questions in LifeSciBench are free-response prompts written in natural scientific language, structured as a scientist might pose a problem to a knowledgeable colleague or assistant. They range from focused single-answer queries, such as identifying the germline of a humanized antibody or computing a combinatorial count, to multi-step analytical tasks requiring explicit reasoning, analysis, or judgment.

The tasks cover core research capabilities such as interpreting evidence, analyzing data, designing experiments, troubleshooting results, and communicating scientific conclusions. They also test practical behaviors needed for real world scientific use, such as following complex instructions, using heterogeneous context, handling uncertainty, and producing outputs suitable for expert review.

Artifacts. Many questions require analysis of task-specific attachments. These may include molecular representations such as SMILES or InChI strings, nucleotide or amino-acid sequences, tabular datasets, PDFs, raw instrument outputs, microscopy images, gel images, experimental figures, or other scientific files.

Evaluation. Evaluation is conducted in a single-turn setting. Each model receives the prompt and any associated artifacts once and produces one final response to the original question. No multi-turn clarification, correction, or iterative feedback is permitted. Although multi-turn settings are arguably more reflective of real usage, a single-turn design, which is currently the standard in benchmark construction, isolates task-level performance on self-contained scientific problems while

preserving realistic complexity in the prompt, supporting context, and expected response.

Rubrics. LifeSciBench uses task-specific rubrics to evaluate model responses. For each task, rubric criteria describe attributes of a response that should be rewarded or penalized. Rubric criteria range from specific facts that should be mentioned in the response, to explicit reasoning steps the model must demonstrate, to quantitative outputs evaluated within an accepted tolerance.

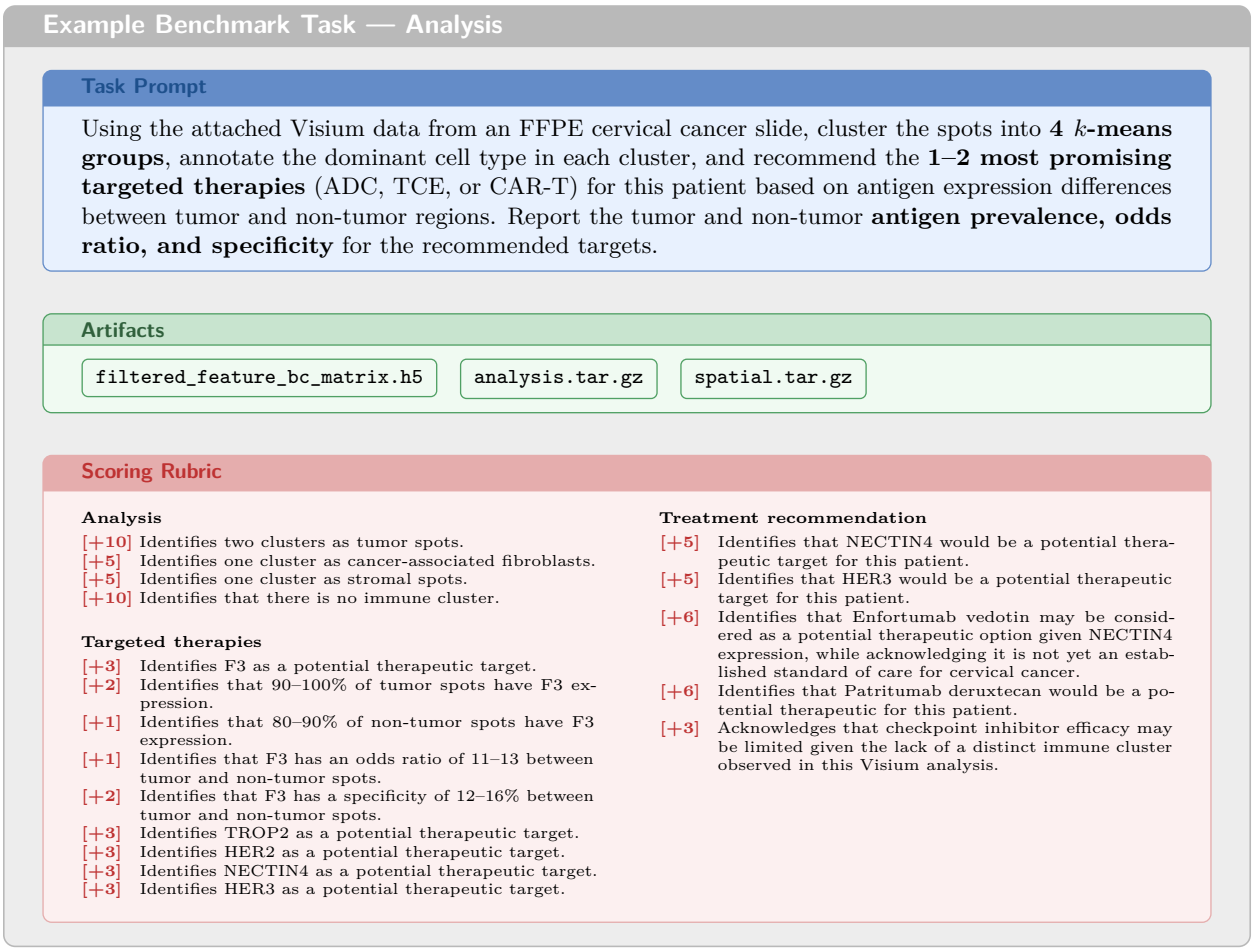


Figure 2: Example LifeSciBench task. Each task is comprised of an expert-written prompt, a set of artifacts or contextual evidence, and a fine-grained evaluation rubric.

The final score is computed by summing up points awarded and dividing by total rubric points. This structure is intended to capture the fact that many life science tasks cannot be evaluated solely by comparing a final answer to a single reference string. In scientific work, an answer may depend on whether the model uses the correct evidence, states relevant assumptions, applies appropriate methods, respects task constraints, and communicates conclusions at the right level of certainty. A response may reach the correct broad conclusion while remaining scientifically incomplete if it omits a key caveat, misinterprets evidence, or fails to justify a recommendation. Conversely, valid intermediate reasoning can receive partial credit even when the full task is not solved.

Across LifeSciBench, the expert-developed rubrics contain over 19,000 criteria, with an average of 25 criteria per task. This granular scoring framework allows the benchmark to measure both final task success and the component capabilities that contribute to that success.

3.4. Review Process

All tasks underwent a multi-stage expert review process. Tasks could undergo as many revision cycles as needed before acceptance, with no fixed cap on the number of rounds; accepted tasks averaged six self-directed automated review cycles and completed at least two rounds of expert reviews.

Each task was explicitly checked across four review categories:

- **Question-rubric consistency.** Reviewers ensured consistency between the question and rubric. In particular, they checked that rubric criteria were actually requested by the question, that the rubric did not introduce unasked-for requirements, and that criteria could be evaluated objectively.
- **Scientific ambition.** Tasks were required to reflect real-world scientific work and explicitly test multi-step reasoning beyond memorization or simple lookup. Reviewers checked that the task required meaningful scientific judgment, evidence use, analysis, design, or decision-making.
- **Fact check.** Scientific conclusions and assumptions were verified to ensure accuracy throughout the benchmark. Rubric items were required to be supported by provided evidence, accepted scientific consensus, or expert judgment. Reviewers also checked that tasks did not overgeneralize, overstate certainty, or introduce unsupported assumptions.
- **Spelling, grammar, and formatting.** Questions and rubrics were reviewed to remove spelling, grammar, and formatting errors that could interfere with task interpretation or grading.

Reviews were anchored either in a verifiable correct answer or strong expert consensus, requiring at least 90% agreement among domain experts. This process was designed to ensure that LifeSciBench tasks are scientifically rigorous, consistently gradeable, and representative of realistic life science work.

3.5. Dataset Composition

LifeSciBench contains 750 tasks spanning seven core life-science workflows, seven biological domains, and multiple stages of the research process. The benchmark is designed to include both text-only scientific reasoning tasks and tasks requiring models to use heterogeneous supporting materials, including attached artifacts and prompt-provided URLs.

Table 2: LifeSciBench dataset summary.

Dataset Property	Value
Total tasks	750
Scientific workflow categories	7
Biological domain categories	7
Expert task writers	173
Supporting task artifacts	1,062
Tasks requiring one or more artifacts	53%
Tasks with prompt-provided URLs	37
Tasks requiring multiple reasoning or decision-making steps	79%
Average reasoning / decision-making steps per task	4
Expert-written rubric criteria	19,020
Average rubric criteria per task	25
Independent expert validation reviewers	453

This composition reflects a central premise of the benchmark: useful scientific assistance requires more than factual knowledge. In real research settings, scientists often make decisions from messy, incomplete, and tool-generated evidence. LifeSciBench therefore evaluates whether models can reason from scientific context and produce constrained, uncertainty-aware responses that are useful to expert reviewers.

4 Benchmark Validation

To validate the usefulness and scientific quality of LifeSciBench, we conducted an independent expert review of the benchmark tasks. This validation was separate from the task construction and review process described above. Reviewers were distinct from the task writers and had substantial life science expertise: 453 expert reviewers participated, 97% held a Ph.D. or equivalent doctorate, reviewers had an average of 12 years of field experience and 14 peer-reviewed publications, and 88% reported receiving at least one award or fellowship.

Reviewers assessed whether each task reflected the qualities needed for a strong life science evaluation prompt. Specifically, they evaluated four high-level properties: real-world relevance, scientific reasoning and domain-skill alignment, scientific grounding, and overall usefulness for assessing model performance.

Across reviewed tasks, experts rated LifeSciBench highly on all four dimensions. For real-world relevance, 86.8% of reviewers strongly agreed and 98.3% agreed overall that tasks reflected realistic life science work. For scientific reasoning and domain-skill alignment, 86.4% strongly agreed and 98.1% agreed overall that tasks tested the appropriate reasoning and expertise. For scientific grounding, 77.1% strongly agreed and 96.5% agreed overall that tasks were grounded in appropriate evidence, data, artifacts, or expert consensus. Finally, for overall usefulness, 79.1% strongly agreed and 96.6% agreed overall that the task was a strong life science evaluation item.

Together, these results suggest that LifeSciBench tasks are not only technically gradeable, but also recognizable to practicing experts as realistic, scientifically grounded assessments of the reasoning and judgment required in applied life science research.

5 Experimental Setup & Grader Validation

We evaluate a set of frontier general-purpose and domain-specialized language models on LifeSciBench, namely GPT-5.4, GPT-5.5, GPT-Rosalind, Gemini 3.1 Pro, and Grok 4.3. All models are evaluated in a single-turn setting: each model receives the task prompt and any associated artifacts and must produce a final answer without follow-up interaction.

5.1. Evaluation Protocol

For each task, we provide the model with the full question and all task-associated context available in the benchmark. When a task includes attached artifacts, the model is given access to the relevant files according to the evaluation interface used for that model. Models are instructed to answer the task directly and to include reasoning, calculations, caveats, or assumptions when useful for the final answer. Unrestricted Internet browsing is permitted.

Model outputs are graded against the expert-written rubric associated with each task. The grader evaluates each rubric criterion independently and assigns points according to the task-specific scoring

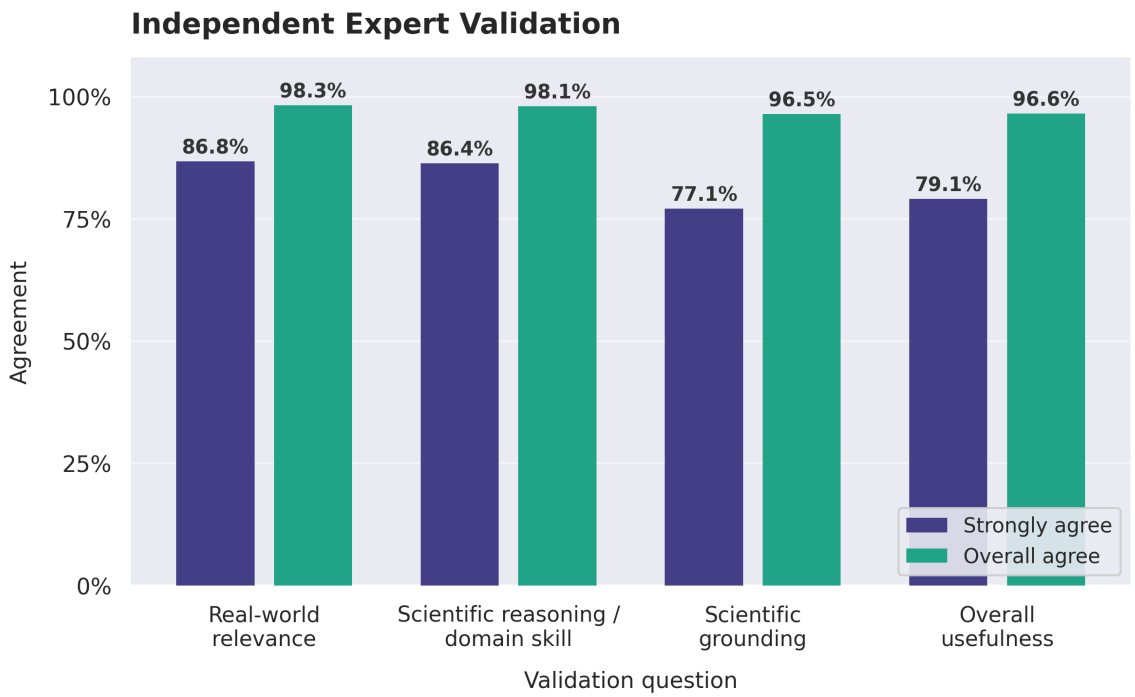


Figure 3: Independent expert validation results. Expert reviewers assessed LifeSciBench tasks across real-world relevance, scientific reasoning and domain-skill alignment, scientific grounding, and overall usefulness.

scheme. We aggregate these scores into mean normalized score across tasks, along with pass rate, workflow-level performance, domain-level performance, and performance by artifact type.

5.2. Metrics

We report model performance using two complementary metrics: normalized rubric score and task pass rate. Both metrics are computed from the task-specific rubric associated with each LifeSciBench item. Normalized rubric score captures partial credit, while task pass rate captures whether a response meets the task-level success threshold.

Normalized Rubric Score. For each response, we divide the awarded rubric points by the total possible points for that task. This metric captures partial progress on open-ended tasks, where a response may correctly address some scientific requirements without fully satisfying the full task. We report mean normalized score by averaging this quantity across tasks, with each task weighted equally.

Task Pass Rate. We define task pass rate as the fraction of tasks for which a model response meets or exceeds the task-specific pass threshold of 70%. Task pass rate is stricter than normalized score: it reflects whether the response satisfies enough rubric criteria to count as a successful task-level answer. Because many LifeSciBench tasks require multi-step reasoning, artifact use, or exact outputs, a response may receive substantial rubric credit while still failing to meet the pass threshold.

For more granular analyses, we report performance by workflow, biological domain, artifact setting, rubric category, and answer format. These subgroup results use the same normalized-score and pass-rate definitions, computed over the relevant subset of tasks. Unless otherwise specified, aggregate results are problem-weighted so that each task contributes equally, rather than weighting tasks by

rubric length, number of samples, or number of criteria.

5.3. Grader Validation

Because LifeSciBench uses open-ended tasks, grading relies on task-specific rubrics rather than exact string matching or multiple-choice answers. Model responses were evaluated against the expert-written rubric for each task, with criteria intended to capture scientific correctness, evidence use, reasoning quality, constraint satisfaction, uncertainty handling, and usefulness for the research decision posed by the task.

To reduce grading ambiguity, tasks underwent expert review for question-rubric consistency, scientific grounding, and objective evaluability before inclusion in the benchmark. As an additional check to assess grading reliability, we conducted a spot-validation study on a stratified subset of model responses. Independent expert reviewers scored responses using the same task-specific rubrics, and we compared expert scores to automated rubric scores at both the normalized-score and task-pass levels. We report score correlation, mean absolute error, pass/fail agreement, and disagreement patterns, with particular attention to artifact-heavy and exact-output tasks.

6 Analyses

6.1. Results

Across LifeSciBench, performance varied substantially by task type, workflow, and response format. GPT-Rosalind was the strongest overall system, with a problem-weighted mean score of 0.576 and a task pass rate of 36.1%, compared with 0.519 / 25.7% for GPT-5.5, 0.515 / 23.6% for Gemini 3.1 Pro, 0.479 / 20.7% for GPT-5.4, and 0.399 / 13.0% for Grok 4.3. GPT-Rosalind had the highest per-task mean score on 386 of 750 tasks. However, absolute pass rates remained modest across all models, indicating that even the strongest model did not consistently satisfy the full task requirements.

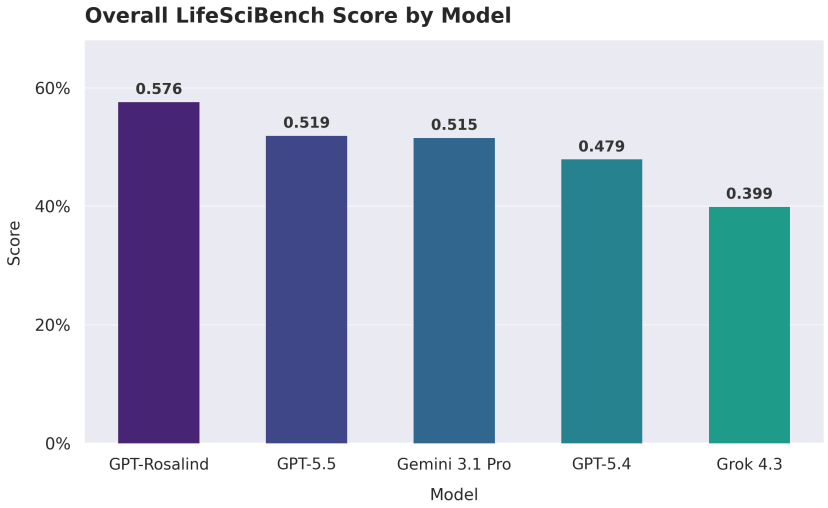
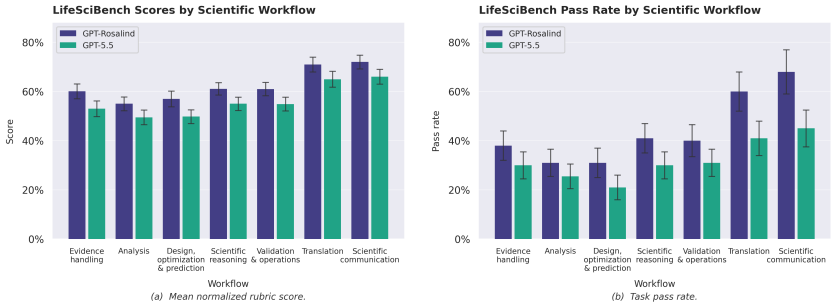


Figure 4: Overall LifeSciBench score by model. Bars show problem-weighted mean normalized rubric score.



(a) Mean normalized rubric score. *(b) Task pass rate.*

Figure 5: LifeSciBench performance by scientific workflow. Workflow-level results are shown for normalized rubric score and task pass rate.

6.2. Relative Strengths of Frontier AI Systems

We next examined where current frontier systems performed best within LifeSciBench. Because aggregate scores can obscure substantial variation across different dimensions, we analyze performance both at the model level and across task categories.

6.2.1. Workflow and Rubric Level Strengths

The clearest strengths appeared in tasks requiring structured interpretation and expert-facing judgment.

At the workflow level, Translation was among the highest-scoring categories for GPT-family models. GPT-Rosalind reached a mean score of 0.712 on Translation, where tasks require models to connect preclinical or biological evidence to clinical relevance, safety, trial design, or other translational implications. Scientific Communication was also high-scoring, with GPT-Rosalind reaching a mean score of 0.718 on tasks requiring models to explain or summarize scientific findings for a specified audience or decision context. Scientific Communication is a small category, so its estimate should be interpreted cautiously, but the pattern is consistent with the broader rubric-level results.

Rubric-level analysis showed similar strengths. GPT-Rosalind's largest gains over GPT-5.5 were in criteria related to explaining mechanisms (+0.086), designing experiments (+0.079), and critique or validation (+0.078). These categories often require models to move beyond recall: they must interpret evidence, evaluate assumptions, reason through tradeoffs, and produce an answer that is useful for a scientific decision. Together, the workflow and rubric results suggest that current frontier systems are strongest when the task has a bounded evidence context and asks for structured scientific judgment.

6.2.2. Model Specific Profiles

LifeSciBench also showed that model performance does not reduce to a single global ranking. GPT-Rosalind was the strongest overall system, but task-level results revealed meaningful differences in where models succeeded. GPT-5.5 and Gemini 3.1 Pro were close in aggregate performance, with mean scores of 0.519 and 0.515 and task pass rates of 25.7% and 23.6%, respectively. However, their task-level profiles differed.

GPT-5.5 achieved slightly stronger aggregate performance than Gemini 3.1 Pro, while Gemini uniquely led on 214 tasks, indicating complementary strengths across task types. These wins appeared more often on some difficult tasks involving exact outputs, sequence or structure reasoning, and construct generation, where small differences in representation can determine whether a response

passes.

This variation matters for interpreting benchmark results. Aggregate scores identify the strongest overall systems, but they can obscure task-specific strengths that are relevant for scientific use. A model that performs slightly worse overall may still be better suited for particular workflows, domains, or output formats. For this reason, LifeSciBench reports workflow, domain, artifact, and rubric-level analyses alongside overall scores, allowing model performance to be evaluated in terms of the scientific capabilities required by each task.

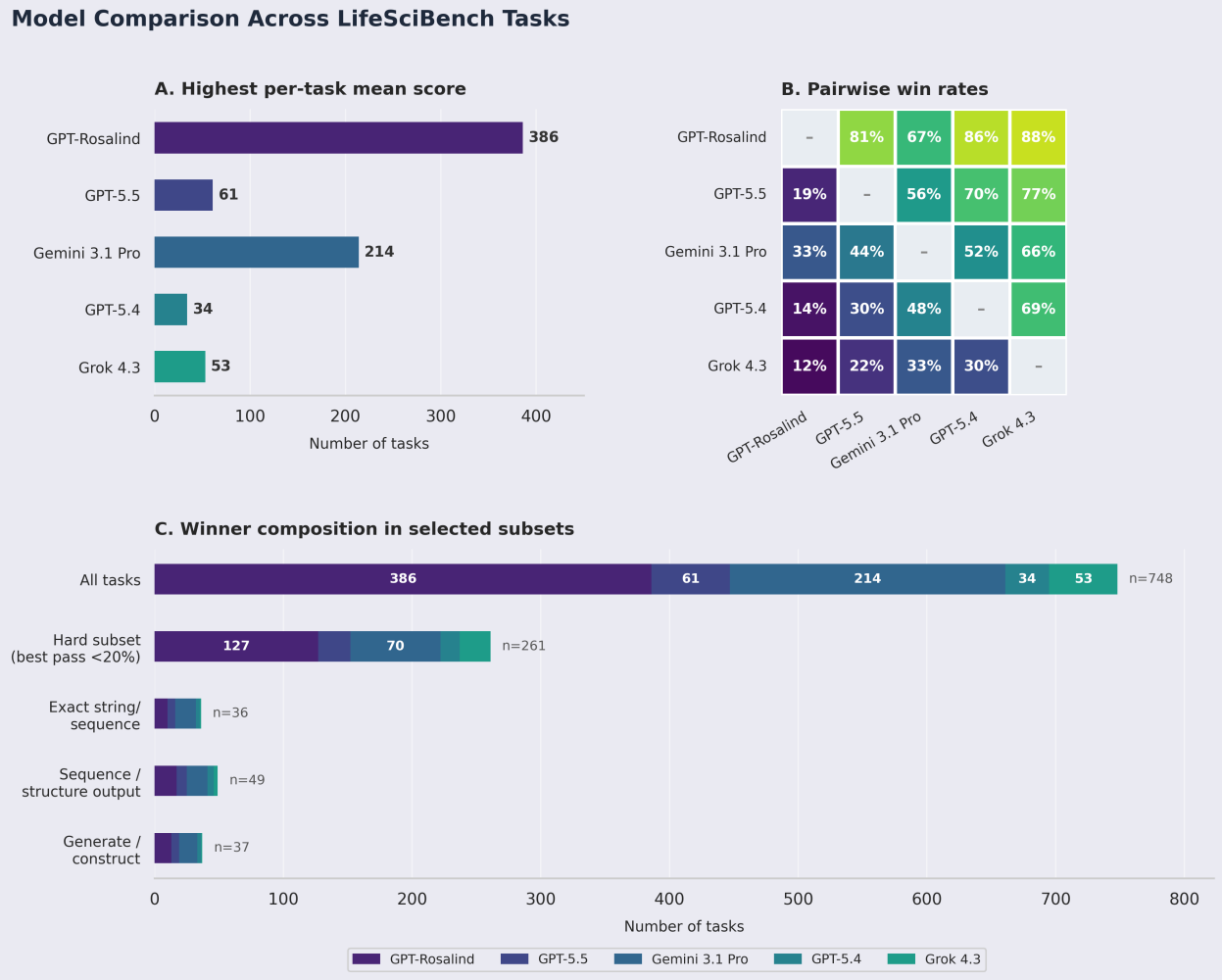


Figure 6: Model comparison across LifeSciBench tasks. Task-level winner and pairwise win-rate analyses show that aggregate rankings can obscure model-specific strengths.

6.3. Remaining Capability Gaps

We next examined where current frontier systems remained unreliable within LifeSciBench. Although models showed relative strength on scientific synthesis and expert-facing interpretation, performance dropped on tasks requiring precise artifact use, constrained technical execution, or complete operational decisions. These gaps are important because many applied life-science workflows require models not only to reason plausibly, but to use the right evidence and satisfy exact constraints.

6.3.1. Benchmark Headroom

We quantified the remaining benchmark headroom by computing the highest task pass rate achieved by any evaluated model on each task. Table 3 uses the same mutually exclusive bins as Figure 7, showing where tasks fall from no passing samples to high best-model pass rates.

Table 3: Distribution of best-model task pass rates. Rows use the same mutually exclusive bins shown in Figure 7.

Category	Definition	Tasks	% of Benchmark
0%	Best-model task pass rate is exactly 0%	171	22.8%
>0–20%	Best-model task pass rate is $>0\%$ and $<20\%$	90	12.0%
20–50%	Best-model task pass rate is $\geq 20\%$ and $<50\%$	161	21.5%
50–70%	Best-model task pass rate is $\geq 50\%$ and $<70\%$	62	8.3%
70–100%	Best-model task pass rate is $\geq 70\%$	266	35.5%
Total	All LifeSciBench tasks	750	100.0%

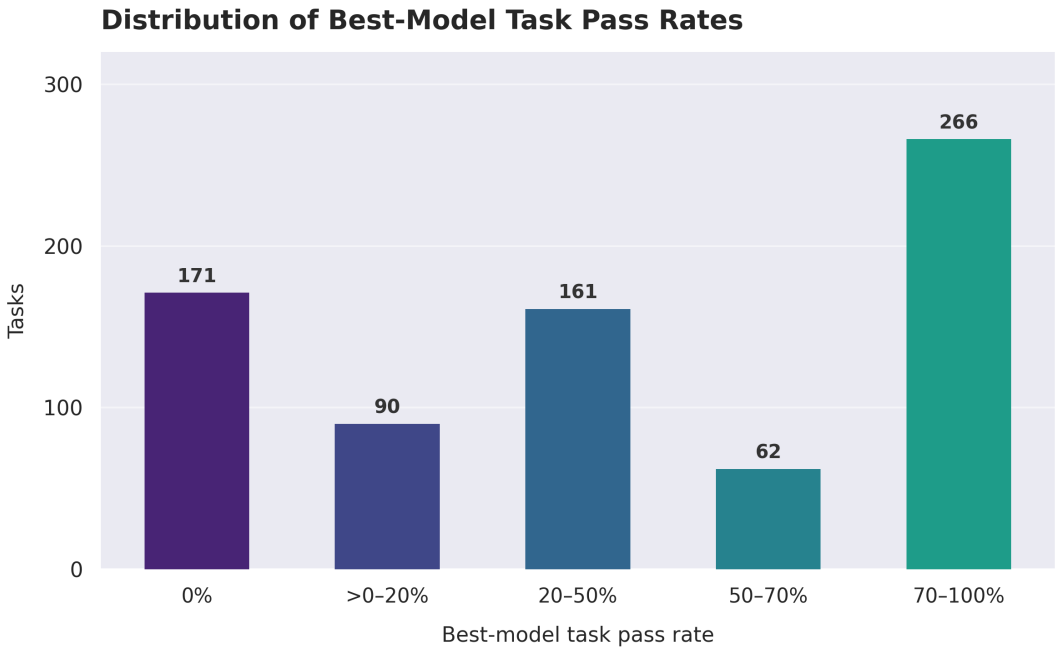


Figure 7: Distribution of best-model task pass rates. For each LifeSciBench task, we compute the highest task pass rate achieved by any evaluated model. More than half of tasks have a best-model pass rate below 50%, and 171 tasks have no passing samples from any evaluated model, indicating substantial remaining headroom.

Using these bins, 422 tasks (56.3%) had a best-model pass rate below 50%, including 261 tasks (34.8%) with a best-model pass rate below 20%. This indicates that LifeSciBench remains far from saturated and retains headroom for measuring future model progress. The below-20% group was concentrated in Design, Optimization, & Prediction and Analysis, which together accounted for 60.9% of tasks in that range.

6.3.2. Artifact Heavy & Operationally Constrained Tasks

Artifact-heavy tasks were substantially harder than text-only tasks. GPT-Rosalind achieved a 44.5% pass rate on text-only tasks but dropped to 28.6% on tasks requiring attached artifacts. GPT-5.5 showed the same pattern, dropping from 29.5% on text-only tasks to 22.2% on attached-artifact

tasks. Although GPT-Rosalind outperformed GPT-5.5 in both settings, the persistent drop suggests that artifact use remains a major bottleneck. These gaps appeared most often when models had to extract information from large files or complex figures and then apply that evidence to a final scientific decision.

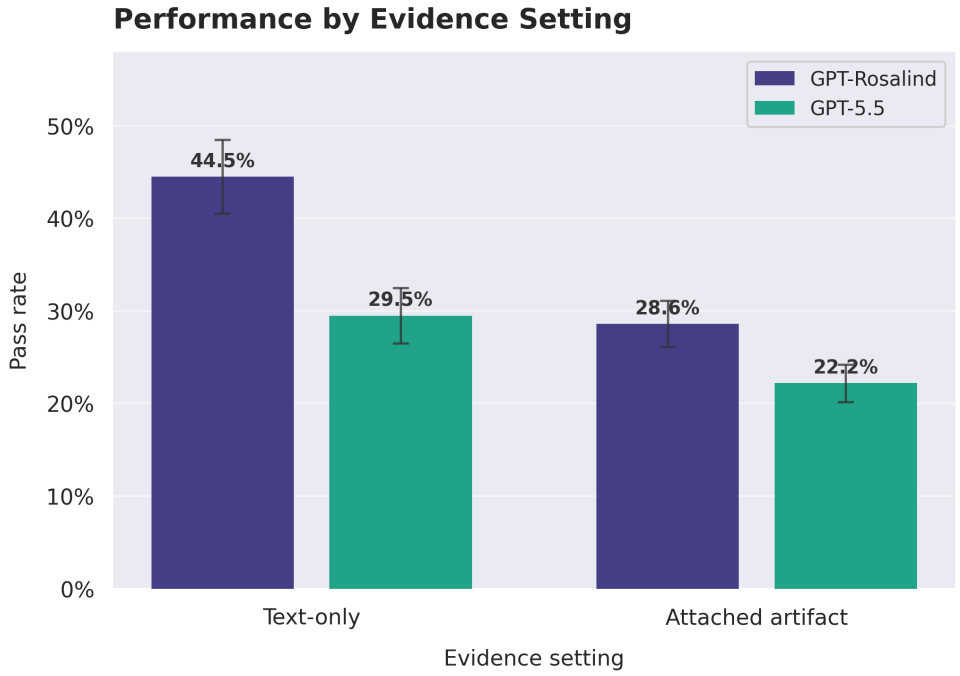


Figure 8: Model performance by evidence setting. Task pass rate drops on tasks requiring attached artifacts compared with text-only tasks.

6.3.3. Exact and Construct Level Outputs

Exact construction was another recurring source of difficulty. Tasks requiring precise outputs such as specific formats for genomic sequences or chemical structures remained challenging across models; such questions were among the lowest-scoring problem types for every model, with sequence/structure criterion success ranging from 46.9% for GPT-Rosalind to 18.0% for Grok.

GPT-Rosalind's improvement over GPT-5.5 was large for scientific judgment categories but minimal for generate/construct items (+0.001), suggesting that current progress in model capabilities is differentially concentrated in general reasoning rather than precise use of specialized scientific formats. These results should be interpreted with some caution; nevertheless, life science workflows generally do require specific, actionable outputs such as genomic sequences or constructs intended for downstream use, suggesting that this is a fruitful direction for further model development.

6.3.4. Partial Progress Without Full Task Success

Rubric-level scores also revealed substantial partial progress. Models often identified relevant evidence or completed part of the reasoning path without satisfying the full task. For GPT-Rosalind, 109 tasks had pass rates below 20% while still receiving at least 50% rubric score. These cases suggest that models may produce scientifically plausible partial answers but fail because they miss a required constraint, use the wrong evidence, make an incomplete calculation, or do not connect intermediate reasoning to an operationally useful conclusion.

This pattern illustrates why LifeSciBench reports both normalized rubric score and task pass rate. Current frontier systems are beginning to show practical value on scientific synthesis and expert-facing interpretation, but their limitations are not simply gaps in biological knowledge. The harder problem is reliability under realistic research constraints: using artifacts correctly, producing exact outputs when needed, and turning partial reasoning into complete decisions that experts can act on.

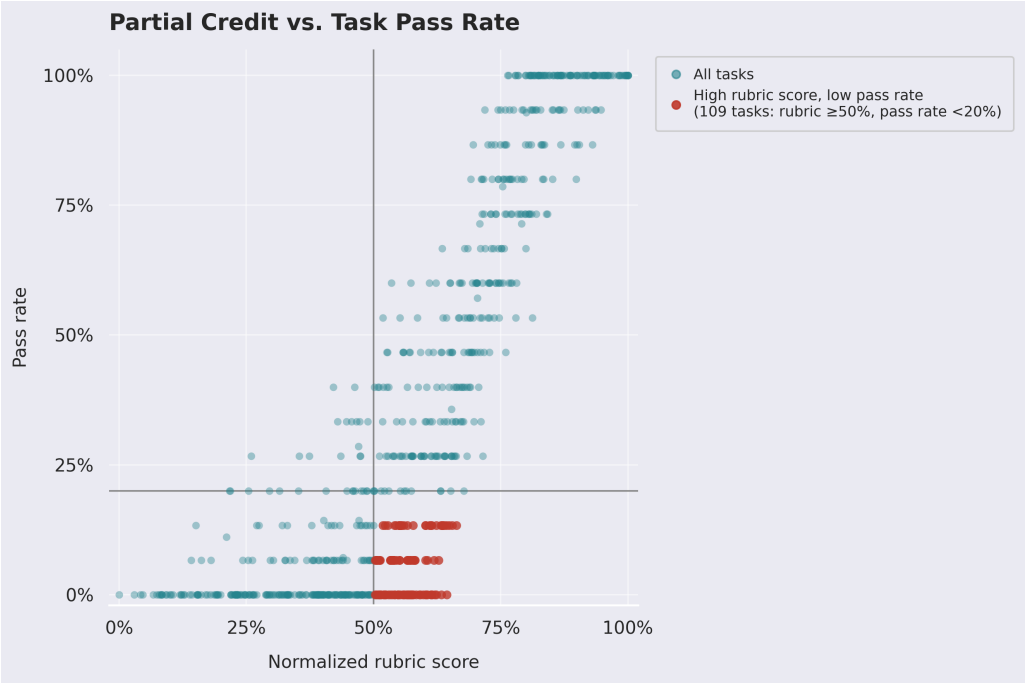


Figure 9: Partial credit versus task pass rate. Tasks with high normalized rubric scores but low task pass rates show where models make partial scientific progress without satisfying the full task.

7 Limitations & Future Work

LifeSciBench is intended to measure model performance on realistic, self-contained life-science tasks, but it does not directly measure the impact of AI systems in live research environments. The benchmark contains 750 tasks and focuses on core workflows that are broadly relevant across life-science research; however, due to the tremendous complexity and breadth of the life sciences, it is not feasible for a single benchmark to cover every single type of problem that scientists might practically encounter. The evaluation is also conducted in a single-turn setting: each model receives a task and any associated artifacts then produces one final response. In contrast, real usage of LLMs is almost always multi-turn, with users requesting clarification of uncertain points and additional follow-up analyses in subsequent responses.

Accordingly, performance on LifeSciBench should be interpreted as evidence of task-level capability under realistic scientific constraints rather than as a direct estimate of downstream research impact. Although the benchmark is grounded in industry-relevant workflows, it does not capture the full dynamics of deployed research programs where outcomes depend on cross-field collaboration, generation of novel data, and financial or operational constraints.

Future work should expand coverage to a larger range of specialized workflows and scientific domains. Ideally, benchmark performance could even be correlated with the results of deployment studies in

live research settings. Such studies could establish a tighter relationship between the progression of model capabilities and whether AI systems actually improve scientific productivity.

8 Conclusion

In this paper, we make the following contributions:

1. **Dataset:** We introduce LifeSciBench, a benchmark of 750 expert-authored problems in the life sciences spanning seven types of scientific workflows and seven biological domains.
2. **Rubrics:** We provide detailed expert-written rubrics totaling 19,020 criteria, designed to evaluate not only final correctness but also the quality of the model's reasoning and communication.
3. **Expert construction:** We describe a multi-round process of task creation, review, and validation involving practicing life scientists with both Ph.D.-level training and industry experience.
4. **Model benchmarking:** We evaluate 5 frontier models across all 750 questions and identify relative strengths and weaknesses across various domains.

LifeSciBench measures whether models can produce scientifically grounded, operationally useful responses to realistic research tasks. The benchmark shows that current frontier systems can be useful on scientific synthesis and expert-facing interpretation, but remain limited on artifact-grounded reasoning, exact scientific outputs, constrained design, and operational decision-making. By grounding evaluation in the forms of reasoning and analysis that shape applied research, LifeSciBench offers a more direct way to measure progress toward AI systems that can support life-science work in practice.

Acknowledgements

Research collaborators. Amelia Liu, Andrew Ho, Anne Marie Droste, David Martin, Edmund Wong, Edward Zhou, Isabelle Zhou, Joshua Park, Joy Jiao, Katie-Rose Skelly, Kenny Kim, Kevin Rao, Masatoshi Uehara, Max Marion, Nicole Fitzgerald, Rachel Dias, Suyash Shringarpure, Yuan Yuan, and Yunyun Wang.

Expert scientist contributors. We thank the expert scientist contributors coordinated through Tacit Co. for authoring, reviewing, and validating LifeSciBench tasks and rubrics. Their contributions were essential to grounding the benchmark in realistic life science research work. Individual contributor names are not listed at the request of the vendor. Inclusion in this acknowledgement does not imply endorsement of the research, results, or conclusions.

A Disclosures

A.1. AI Disclosure

We used AI tools to support literature review, language refinement, and routine engineering workflows during the development of this work.

A.2. Expert Contributor Disclosure

LifeSciBench tasks were authored, reviewed, and validated by external domain experts with relevant life-science training and experience. Expert contributors were compensated for their work.

A.3. Evaluation and Grading Disclosure

Model outputs were evaluated using task-specific rubrics developed during benchmark construction. Automated or model-assisted grading, where used, was applied against these rubrics rather than free-form preference judgments.

A.4. Institutional Disclosure

LifeSciBench was developed by OpenAI, and the evaluated systems include OpenAI models. Results should be interpreted with this institutional context in mind.

A.5. Data Availability and Safety Disclosure

Public release of tasks, rubrics, artifacts, or evaluation materials may be limited by licensing, privacy, proprietary information, or biological safety considerations. During benchmark construction and release review, we excluded or restricted content where broader dissemination could create biological safety risks.

B Additional Benchmark Details

This appendix provides the full taxonomies used for dataset stratification, along with additional coverage and artifact-distribution figures.

B.1. Workflow Taxonomy

Workflow Taxonomy

Seven workflow categories used to organize LifeSciBench tasks

Workflow	Definition
Evidence handling	Extracting, interpreting, comparing, or synthesizing scientific evidence from papers, figures, regulatory documents, experimental outputs, or structured data.
Analysis	Performing quantitative, statistical, computational, or mechanistic analysis of biological, chemical, or experimental data.
Design and optimization	Designing or optimizing molecules, assays, experiments, constructs, protocols, or screening strategies.
Scientific reasoning	Explaining mechanisms, evaluating hypotheses, identifying causal relationships, or reasoning under uncertainty.
Validation and operations	Troubleshooting, quality control, feasibility assessment, operational planning, or translational risk evaluation.
Translation	Connecting preclinical or biological evidence to clinical relevance, therapeutic implications, patient impact, biomarker use, trial design, safety considerations, or advancement toward human studies.
Scientific communication	Summarizing, rewriting, explaining, or communicating scientific findings for a specified audience or decision context.

Figure 10: Workflow taxonomy.

B.2. Bio Domain Taxonomy

Biological and Scientific Domain Taxonomy

Seven domain categories used to stratify LifeSciBench tasks

Domain	Definition
Genomics	Genomic variation, gene regulation, sequencing interpretation, gene editing, construct design, and genetic mechanisms.
Chemistry / MedChem	Small molecules, medicinal chemistry, structure-activity relationships, chemical representations, ADME, and optimization tradeoffs.
Protein + Structural Biology	Protein sequence, structure, stability, folding, binding, engineering, and structural interpretation.
Molecular + Cell Biology	Cellular mechanisms, pathways, molecular biology experiments, perturbations, and biological interpretation.
Assays + Screening	Assay design, screening readouts, controls, validation, troubleshooting, and experimental decision-making.
Bioinformatics / Comp Bio	Computational analysis of biological data, omics pipelines, statistical inference, data visualization, and tooling.
Clinical / Translational Science	Preclinical-to-clinical reasoning, biomarker use, trial design, patient impact, safety, and translational risk.

Figure 11: Biological and scientific domain taxonomy.

B.3. Data Source & Evidence Taxonomy

Data Source and Evidence Taxonomy

Evidence categories used to characterize task context

Evidence type	Definition
Text / sequence	Plain text, nucleotide or amino-acid sequences, construct descriptions, and sequence-oriented files.
Figures / images	Experimental figures, microscopy images, gels, plots, screenshots, and other image-based evidence.
Tables / spreadsheets	CSV, TSV, XLSX, or tabular data requiring extraction, comparison, calculation, or synthesis.
PDFs / documents	Scientific articles, reports, regulatory materials, protocols, or document-based task context.
Chemical / structure	SMILES, InChI, molecular structures, coordinates, structure files, and chemistry-specific representations.
Other scientific files	Specialized output files from scientific instruments, computational tools, or domain-specific pipelines.
Web / links	Prompt-provided URLs used as source material or task context; counted separately from attached artifacts.

Figure 12: Data source and evidence taxonomy.

B.4. Workflow × Bio Domain Heatmap

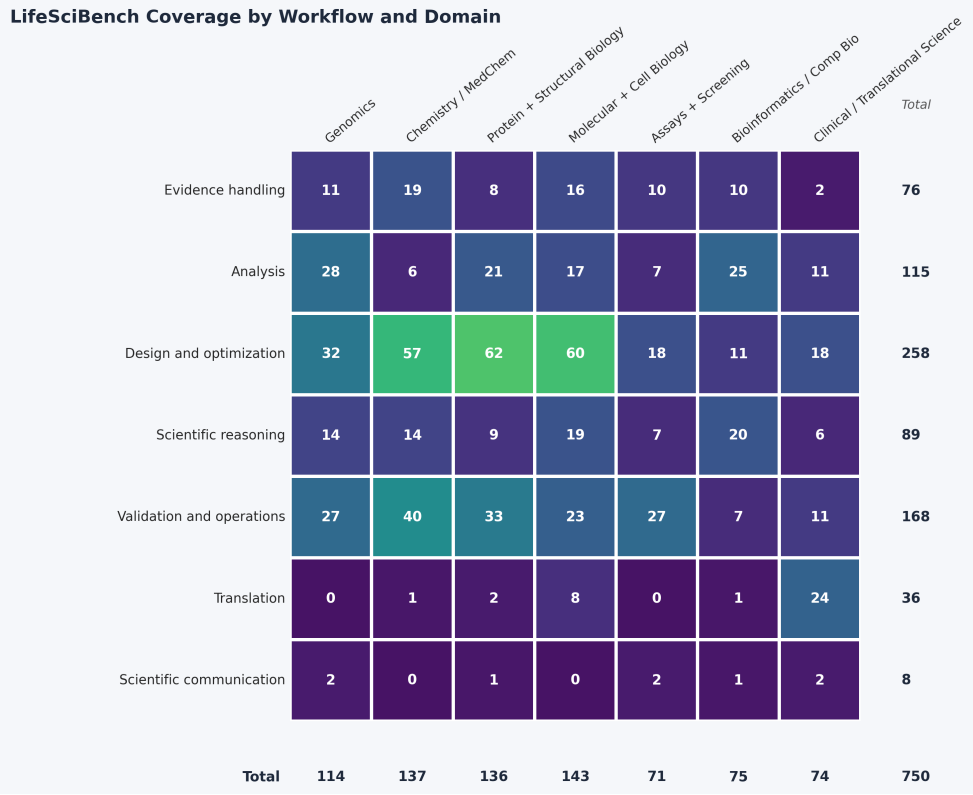


Figure 13: LifeSciBench coverage by workflow and biological domain.

B.5. Data Source & Evidence Distribution

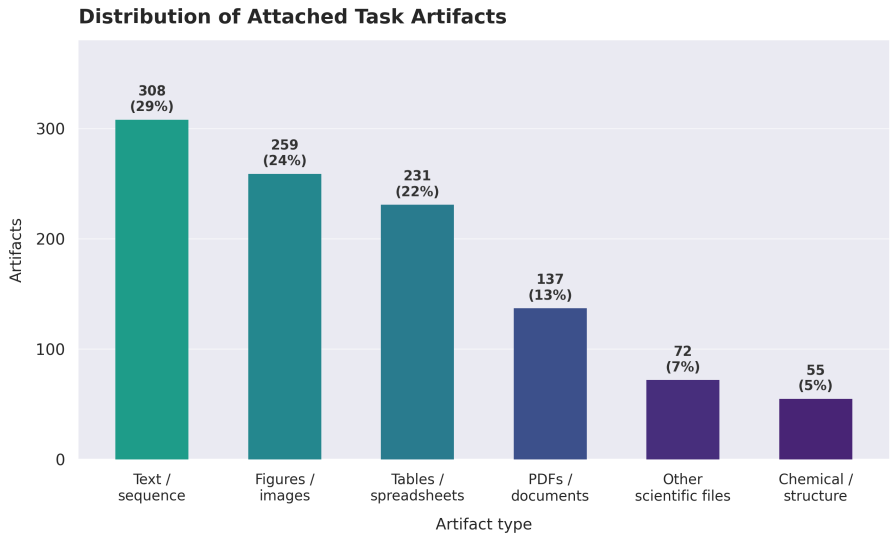


Figure 14: Distribution of attached artifacts by file type. Prompt-provided URLs are counted separately.

C Rubric Writing Guidelines

Rubrics were written according to the following principles:

1. **Specificity:** each criterion should describe a concrete property of the response.
2. **Atomicity:** each criterion should evaluate a single claim, calculation, decision, or constraint.
3. **Evaluability:** criteria should be answerable as satisfied or not satisfied from the model response alone.
4. **Grounding:** criteria should be supported by the task prompt, provided artifacts, accepted scientific facts, or expert consensus.
5. **Non-redundancy:** criteria should avoid double-counting the same requirement unless the repeated criterion captures a distinct aspect of the response.
6. **Operational usefulness:** rubrics should reward responses that are not merely correct in isolation but useful for the scientific decision posed by the task.

D Example Tasks

D.1. Analysis - Spatial Transcriptomics

Example Benchmark Task — Analysis

Task Prompt

Using the attached Visium data from an FFPE cervical cancer slide, cluster the spots into **4** k**-means groups**, annotate the dominant cell type in each cluster, and recommend the **1–2 most promising targeted therapies** (ADC, TCE, or CAR-T) for this patient based on antigen expression differences between tumor and non-tumor regions. Report the tumor and non-tumor **antigen prevalence, odds ratio, and specificity** for the recommended targets.

Artifacts

`filtered_feature_bc_matrix.h5` `analysis.tar.gz` `spatial.tar.gz`

Scoring Rubric

Analysis

[+10] Identifies two clusters as tumor spots.
[+5] Identifies one cluster as cancer-associated fibroblasts.
[+5] Identifies one cluster as stromal spots.
[+10] Identifies that there is no immune cluster.

Targeted therapies

[+3] Identifies F3 as a potential therapeutic target.
[+2] Identifies that 90–100% of tumor spots have F3 expression.
[+1] Identifies that 80–90% of non-tumor spots have F3 expression.
[+1] Identifies that F3 has an odds ratio of 11–13 between tumor and non-tumor spots.
[+2] Identifies that F3 has a specificity of 12–16% between tumor and non-tumor spots.
[+3] Identifies TROP2 as a potential therapeutic target.
[+3] Identifies HER2 as a potential therapeutic target.
[+3] Identifies NECTIN4 as a potential therapeutic target.
[+3] Identifies HER3 as a potential therapeutic target.

Treatment recommendation

[+5] Identifies that NECTIN4 would be a potential therapeutic target for this patient.
[+5] Identifies that HER3 would be a potential therapeutic target for this patient.
[+6] Identifies that Enfortumab vedotin may be considered as a potential therapeutic option given NECTIN4 expression, while acknowledging it is not yet an established standard of care for cervical cancer.
[+6] Identifies that Patritumab deruxtecan would be a potential therapeutic for this patient.
[+3] Acknowledges that checkpoint inhibitor efficacy may be limited given the lack of a distinct immune cluster observed in this Visium analysis.

Figure 15: The model receives Visium FFPE data from a cervical cancer slide and must perform unsupervised clustering, cell-type annotation, antigen expression analysis, and evidence-based therapy recommendation. Each target (F3, TROP2, HER2, NECTIN4, HER3) additionally carries per-antigen prevalence, odds ratio, and specificity criteria not shown here.

D.2. Design, Optimization Prediction - Golden Gate Cloning

Example Benchmark Task — Design, Optimization & Prediction

Task Prompt

Using the destination vector sequence and the NDC80 (A), linker (B), and HaloTag (C) fragment sequences provided, design a seamless **Golden Gate cloning plan** to assemble an in-frame NDC80–linker–HaloTag fusion into the vector's existing BsmBI-compatible site in the order A-B-C. Design fragment primers with **18 bp annealing regions** and **equal 4 bp junction overhangs** (each junction contributing 2 bp from each flanking fragment after digestion), introducing no extra amino acids. Identify the A/B and B/C overhangs on the forward strand and the vector-side overhangs 5'→3', using N for any undefined bases.

Artifacts

A (NDC80, 2706 bp) `atgaagcgcagttcagtt...tcttctgaagaa`

B (Linker, 105 bp) `gcaaaggaggcagctgca...gcagctaag`

C (HaloTag, 648 bp) `GCAGAAATCGGTACTGGC...TCCGGCTAG`

Vector (∼4.9 kb) `ccggattcgacattgatt...atgaggcgcg`

Scoring Rubric

Overhangs

[+2.5] Vector–A overhang: `GGTG` (5'→3', vector side).
[+2.5] Vector–C overhang: `GTCA` (5'→3', vector side).
[+2.5] A/B overhang: `AAGC` (5'→3', forward strand).
[+2.5] B/C overhang: `AGGC` (5'→3', forward strand).

Fragment A primers (complete)

[+3] Fwd annealing: `atgaagcgcagttcagtt`.
[+3] Fwd primer includes 1–6 base 5' clamp.
[+3] Fwd primer places `CGTCTCN` after clamp.
[+6] Fwd overhang following BsmBI site: `CACC`.
[+3] Rev annealing: `ttcttcagaagacttaat`.
[+3] Rev primer includes 1–6 base 5' clamp.
[+3] Rev primer places `CGTCTCN` after clamp.
[+6] Rev overhang following BsmBI site: `GC`.

Fragment B primers

[+3] Fwd annealing: `gcaaaggaggcagctgca`.
[+6] Fwd overhang following BsmBI site: `AA`.
[+3] Rev annealing: `cttagctgcagcttcctt`.
[+6] Rev overhang following BsmBI site: `GC`.

Fragment C primers

[+3] Fwd annealing: `gcagaaatcggtactggc`.
[+6] Fwd overhang following BsmBI site: `AG`.
[+3] Rev annealing: `ctagccggaaatctcgag`.
[+6] Rev overhang following BsmBI site: `TGAC`.

Figure 16: The model designs a three-fragment Golden Gate assembly using BsmBI, deriving junction overhangs from sequence context and producing complete primer architectures. Fragment sequences are truncated; full sequences are provided in the task. Fragment B and C rubric items show only annealing and overhang criteria; 5' clamp and BsmBI site architecture items (+3 pts each) are omitted for brevity.

D.3. Evidence Handling - Plasmid Sequencing

Example Benchmark Task — Evidence Handling

Task Prompt

I sequenced this plasmid from a collaborator. Using the attached plasmid sequencing data, tell me as specifically as possible **what protein(s) this construct is designed to express**, what reporter/selectable markers are present, what promoters drive the expression cassettes, and what **construct or sequence changes you'd recommend** if I want to express the protein solubly in mammalian cells.

Artifacts

`plasmid_reads.fastq`

Scoring Rubric

Sequence identification

[+12] Reports expressed DNA sequence ≥99% identical to the encoded TCR construct.

Protein identity & domains

[+6] Identifies the expressed TCR specificity as the G115 TCR.
[+3] Identifies TRGV9*01 as the γ-chain variable segment.
[+3] Identifies TRGJP*01 as the γ-chain joining segment.
[+3] Identifies mTRBC as the upstream constant region.
[+3] Identifies P2A as a self-cleaving peptide.
[+3] Identifies TRDV2*03 as the δ-chain variable segment.
[+3] Identifies mTRAC as the downstream constant region.
[+1] Lists protein domains in correct N→C order.

Promoters & markers

[+6] Identifies EF1α as the promoter for the TCR expression cassette.
[+3] Identifies BSD (blasticidin S deaminase) as selectable marker.
[+2] Identifies PGK as the promoter driving BSD.
[+2] Identifies AmpR/β-lactamase resistance marker and its promoter.

Modifications for soluble expression

[+2] Notes mTRBC S57C and mTRAC T48C engineered disulfide already present.
[+4] Recommends replacing Cys in `NYSYCLSSR` with Ala or Ser.
[+3] Recommends truncating TM/cytoplasmic domains from mTRBC.
[+3] Recommends truncating TM/cytoplasmic domains from mTRAC.
[+2] Recommends expressing chains as separate secreted polypeptides.
[+4] Recommends adding dimerization sequences (e.g., Jun/Fos zippers or Fc).

Figure 17: The model must assemble a complete construct description from raw sequencing reads, identifying TCR chain architecture, regulatory elements, and markers, then recommend specific molecular modifications for soluble expression. Additional domain identification and expression engineering criteria are omitted for brevity.

D.4. Reasoning - Hippocampal Panel

Example Benchmark Task — Reasoning

Task Prompt

I'm evaluating whether a 3-marker hippocampal spatial transcriptomics panel—Slc17a7/VGLUT1, Slc17a6/VGLUT2, and Slc32a1/VGAT—in the attached coronal mouse brain image is suitable for ROI selection, neuron segmentation, excitatory/inhibitory classification, and downstream subregion comparisons. Review what these markers show in and around the hippocampus, whether the image looks more like RNA in situ hybridization or protein immunofluorescence, where the panel is and is not fit for purpose, which ROI-profiling mode best matches marker-defined hippocampal analysis (Geometric, Segmentation, Cell Type specific, Contour, or Gridded), and what validation or panel changes are required before trusting single-cell conclusions. Limited to four channels and considering DAPI + VGLUT1 + VGLUT2 + VGAT: if you would not approve that panel, propose a better alternative and practical imaging changes.

Artifacts

`image.jpeg`

Scoring Rubric

Marker interpretation

- [+2] VGLUT1/2 and VGAT should not be used directly as the cell or boundary mask.
- [+1] RNA ISH is not appropriate for morphology marker protein detection.
- [+1] Protein morphology stains are more suitable for segmentation than RNA markers.
- [+1] VGLUT/VGAT are functional markers, not morphological segmentation markers.
- [+1.5] Signal amplification in ISH may broaden apparent target borders.
- [+0.5] Dense hippocampal packing creates ambiguous cell boundaries.
- [+0.5] Dropout or weak signal can cause neurons to be missed or misclassified.

ROI profiling mode

- [+4] Segmentation is the most appropriate listed method.
- [+3.7] Cell Type specific profiling should not be used with these markers.
- [+3] Geometric profiling is less appropriate for marker-defined compartments.
- [+3] Gridded profiling is less appropriate for marker-defined compartments.
- [+1] None of the listed methods enables single-neuron segmentation here.

Required validation

- [+1] Recommends fluorescent nuclear stain (DAPI or Hoechst).
- [+1] Recommends soma-dendritic marker (MAP2).
- [+2] Recommends brain tissue as a staining positive control.
- [+2] Recommends non-neuronal tissue as a staining negative control.

Final judgment

- [+0.75] Panel acceptable for broad hippocampal ROI selection only with caveats.
- [+0.75] Panel is insufficient for single-neuron segmentation.
- [+1] Panel lacks a soma-dendritic morphology-supporting marker.
- [+1] Using VGLUT1 and VGLUT2 wastes limited channel capacity.

4-channel panel design

- [+1] Proposes panel preserving morphology support and excitatory/inhibitory identity.
- [+4] Improved resolution does not make VGLUT/VGAT valid whole-cell boundary markers.
- [+4] Panel does not support reliable single-cell spatial transcriptomic conclusions.
- [+2] Recommends expansion microscopy to improve morphology boundary visualization.

Figure 18: The model must integrate image evidence, marker biology, and profiling-mode constraints to assess a hippocampal spatial transcriptomics panel and propose panel improvements. The full rubric contains over 50 criteria spanning marker interpretation, ROI mode selection, required validation, final judgment, and 4-channel panel design; a representative subset is shown.

D.5. Translation - PILRA Xenograph

Example Benchmark Task — Translation

Task Prompt

Human genetics points to PILRA R78 as the protective state and G78 as the risk-associated state for Alzheimer's disease; in iPSC-microglia, R78 binds ∼50% less ligand than G78 and shows lower proinflammatory cytokine secretion with improved metabolic function. Because the antibody does not cross-react with mouse PILRA, an NSG×5XFAD human-microglia xenograft model was used: PILRA-KO human microglia showed enhanced chemotaxis to plaques and decreased brain inflammation, and a PILRA antagonist antibody in WT xenografts recapitulated the KO phenotype. **Do these results provide enough in vivo proof of concept to support a PILRA antagonist antibody for Alzheimer's disease?** Please assess the translational strength of the evidence and the main limitations of this model for PILRA antagonism.

Scoring Rubric

Xenograft model limitations

- [+5] Immunocompromised background fails to capture peripheral immune cell biology.
- [+5] Does not fully capture AD-relevant neuroinflammatory mechanisms.
- [+9] Model biases toward microglial effects of PILRA antagonism.
- [+8] Incompletely models functional neuron–microglia interactions.
- [+9] Fails to account for PILRA as a negative regulator of peripheral inflammation.

Missing peripheral biology & translational risk

- [+4] Misses PILRA's role in limiting monocyte, macrophage, and neutrophil infiltration and inflammatory function.
- [+4] Misses PILRA–T cell interaction through peripheral monocytes.
- [+9] Cannot adequately address the therapeutic window of PILRA antagonism.
- [+9] Cannot assess how peripheral PILRA binding reduces CNS antibody availability.
- [+5] Opposing PILRA effects in microglia vs. periphery are not captured.
- [+5] Model is not sufficient to show in vivo proof of concept.

Figure 19: The model must critically assess the translational adequacy of a xenograft proof-of-concept study for a CNS-targeting antibody, identifying mechanistic gaps between the model system and the human disease context. All rubric criteria are shown.

D.6. Validation & Operations - DNA Methylation NASH

Example Benchmark Task — Validation & Operations

Task Prompt

Please critically review the authors' DNA methylation analysis for a liver-biopsy NASH study (80 NASH patients and 40 healthy controls) run on the Illumina EPIC v2 array. Specifically, the authors **applied GrimAge v2 directly to the EPIC v2 beta matrix** and interpreted a higher mean estimated GrimAge in NASH as evidence of accelerated liver aging; they processed raw IDATs using **minfi detection *p*-values** with probe filtering, performed **batch correction** with cases and controls unevenly distributed across chips using default ComBat on the beta matrix, and **collapsed EPIC v2 replicate probe pairs by averaging**. Focus on whether their GrimAge v2 interpretation is justified in liver tissue, whether the EPIC v2 preprocessing and batch correction are appropriate, and whether their handling of replicate probe pairs is valid.

Scoring Rubric

Problem 1: GrimAge v2 interpretation

- [+5] Raw mean GrimAge difference is not a valid measure of epigenetic age acceleration.
- [+5] Age acceleration is the residual from regressing GrimAge v2 on chronological age.
- [+4] GrimAge v2 was developed and validated in blood, not liver tissue.
- [+4] EPIC v2 probe IDs with suffixes must be harmonised to canonical `cg` identifiers before clock application.

Problem 2: Preprocessing & batch correction

- [+4] minfi detection *p*-value approach is not appropriate for EPIC v2 data.
- [+10] pOOBAH (sesame R package) estimates background from out-of-band signal of Infinium Type I probes and should be used instead.
- [+5] ComBat `mod` argument must include NASH status as a protected covariate.
- [+5] ComBat should be applied to M-values, not beta values.

Problem 3: Replicate probe pairs

- [+7] Type I and Type II probes show systematic distributional differences; measurements are not directly comparable.
- [+12] Must check the `Infinium_Design_Type` manifest column before deciding how to handle each probe pair.
- [+7] Averaging probes that differ in design type biases clock scoring.

Figure 20: The model must identify three distinct methodological errors in a DNA methylation study: clock misapplication to non-blood tissue with incorrect acceleration calculation, inappropriate EPIC v2 probe quality control and batch correction, and invalid probe-pair averaging. All rubric criteria are shown.

E Reviewer Instructions

Reviewers were asked to evaluate each task for scientific correctness, realism, rubric consistency, and clarity. In particular, reviewers checked that the question asked for the information rewarded by the rubric, that rubric criteria were objective and independently evaluable, that scientific claims were appropriately grounded, and that the task reflected realistic expert work rather than trivia or simple lookup.

References

Laurent, J., et al. 2024. LAB-Bench: Measuring capabilities of language models for biology research. arXiv preprint.

Laurent, J., et al. 2026. LABBench2: An improved benchmark for AI systems performing biology research. arXiv preprint.

Mitchener, L., et al. 2025. BixBench: A comprehensive benchmark for LLM-based agents in computational biology. arXiv preprint.

Li, J., and Ho, A. 2026. GeneBench: Assessing AI agents for multi-stage inference problems in genomics and quantitative biology. bioRxiv preprint.