TxBench-Oligonucleotide Discovery

Benchmarking AI Agents on Experimentally Grounded Decisions in the Discovery of ASO/siRNA Therapeutics

Martin Jacko, Jackson Brougher, Alex Urrutia, Hannah Le,
Arjun Banerjee, Dillon Flood, Kenny Workman*

LatchBio, San Francisco, CA, USA

Abstract

Artificial intelligence (AI) agents promise to accelerate drug discovery by compressing interpretation and decision-making, but practical deployment requires trusted evaluation on realistic programs. We introduce TxBench-Oligonucleotide Discovery, a verifiable benchmark of 113 evaluations that tests whether AI agents, working with no internet access, can recover these decisions from experimental data that would be available to a scientist. Each evaluation is derived from a published or internally curated ASO or siRNA study and graded deterministically against a decision reconstructed from its underlying data, across a 10-section taxonomy spanning target feasibility and drug library design, in vitro pharmacology and safety, and translational readiness. Across the full campaign with 21 model–harness configurations spanning 12 models and 4 execution harnesses, agents passed 38.5% of valid grading runs (2,730/7,098). The strongest configuration that passed 55.5% of endpoint attempts (95% CI 47.2–63.7) was GPT-6 Astra on OpenAI Codex. As drug discovery programs increasingly rely on AI-assisted analysis, these results underscore the practical stakes and the corresponding need for benchmarks like this one to track real capability before broader deployment. Model rankings also vary substantially by evaluation, so no single configuration is uniformly reliable across the ASO/siRNA discovery pipeline, cautioning against deploying any one model as a general-purpose scientific collaborator without task-level validation.

1 Introduction

Progress in drug discovery is driven by an intertwined series of assay results that inform go/no-go calls. At the discovery stage, oligonucleotide therapeutics are designed and characterized across incommensurable types of preclinical work from target and sequence selection, chemical modifications, delivery format design, in vitro potency, off-target profiling, and phenotypic rescue to animal testing of safety/tolerability, pharmacokinetics, pharmacodynamics, and biodistribution. The interpretation often depends on information the measurement itself does not explicitly point to. For example, whether an off-target hit reflects seed-mediated silencing or a hybridization-independent, chemistry-driven effect; whether a hepatotoxicity signal is caused by phosphorothioate backbone loading rather than a specific sequence motif; whether a loss of in vivo knockdown durability reflects target re-expression or a decline in drug exposure. Each of these problems requires a multi-factorial analysis with deep domain expertise. Getting it wrong may advance an unsafe candidate, discard a viable one, or misdirect the next round of experiments, producing a major capital and time inef-

*Corresponding author: kenny@latch.bio

ficiency and ultimately resulting in a missed opportunity to produce a therapeutic product with a positive social impact through extension or improved quality of life.

Recent benchmarks evaluate whether AI agents can derive biological results from experimental data. SpatialBench and scBench assess analysis tasks using spatial and single-cell transcriptomic datasets, respectively, with deterministic grading of the resulting answers (Workman et al., 2025, 2026). EpiBench extends this approach to epigenomic assays (Muralidharan et al., 2026), while VariantBench evaluates genetic variant discovery and interpretation (Bhowmick et al., 2026). SpatialBench-Long and scBench-Long examine more extended analyses in which agents must recover scientific conclusions from raw or near-raw data without prescribed methods (Diks et al., 2026a, 2026b). Together, these benchmarks provide complementary tests of biological data analysis across experimental technologies and task structures.

Within therapeutic discovery, TxBench-PP evaluates decisions involving small-molecule pharmacology, including target engagement, mechanism of action, safety, and translational efficacy (Le et al., 2026). TxBench–Antibody Discovery evaluates decisions involving target assessment, assay design, antibody characterization, engineering, and candidate selection (Ramasamy et al., 2026). Both construct tasks from experimental studies and use deterministic grading to assess whether agents recover the required scientific conclusions. These benchmarks provide the closest methodological precedents for evaluating decisions across oligonucleotide discovery.

Related work in oligonucleotide therapeutics has evaluated LLMs for ASO efficacy prediction (Sundar 2026). Hill et al. introduced ASO Atlas and OligoAI, a deep-learning model that predicts in vitro ASO efficacy using ASO sequence, chemistry, RNA target context, and dosage (Hill 2026).

TxBench-Oligonucleotide Discovery evaluates decisions involving ASO and siRNA sequence, chemistry, efficacy, specificity, safety, and exposure. Its evaluations require agents to interpret experimental evidence relevant to a defined research question and return a structured answer. The benchmark measures performance across selected decisions from the discovery process, with separate analyses by discovery stage and assay or analysis category.

2 Benchmark Construction

TxBench-Oligonucleotide Discovery evaluations are built by independently reconstructing a consequential research conclusions from ASO/siRNA experimental datasets from published articles or patents. As in other TxBench-lineage benchmarks, construction proceeds by identifying a genuine decision point in the source study, staging the subset of data an agent would actually have access to at that point, and grading the terminal decision, not the intermediate arithmetic, deterministically.

The current set comprises 113 evaluations drawn from 20+ discovery programs and ASO Atlas 2.0 (Hill 2026). The individual research studies are spanning chemistry and knockdown catalogs, off-target profiling series (Kuijper 2025; Holgersen 2021), a hepatotoxicity sequence-screen series (Burdick 2014), duplex biophysics and dose-response panels (Hagedorn 2018), phosphorothioate-backbone protein-binding studies (Vickers 2019), subcellular-fractionation and RNA-FISH mechanism studies (Didiot 2018), a CNS-directed SOD1 ASO program (McCampbell 2018), a splice-correction and isoform-quantification study (Klim 2019), an antisense-oligonucleotide liver-disease study (Guo 2014), myotonic-dystrophy off-target studies (Wheeler 2012; Elboujnouni 2023), two KCNT1 studies (a locus-resolution off-target study, Golinski 2026, and a divalent-siRNA seizure-suppression study, Andreone 2025), a tissue-pharmacokinetics study (Backstrom 2024), a 2′-fluoro ASO hepatotoxicity study (Shen 2018), an oligonucleotide cytotoxicity-chemistry study (Janas

2017), an siRNA off-target-specificity study (Lee 2015), and a CACNA1C splice-modulating ASO study for Timothy syndrome (Chen 2024), alongside additional disease-specific programs.

3 Benchmark Anatomy

TxBench-Oligonucleotide Discovery organizes the evaluations along a 10-stage taxonomy spanning three tiers of the discovery-to-development candidate arc: target feasibility and library design (§1–§3), in vitro pharmacology and safety (§4–§6), and translational readiness (§7–§10) (Figure 1A). Coverage is deliberately non-uniform and focuses on stages with key data-driven program decisions. Sequence and physicochemical design (§2), chemistry evaluation (§3), and off-target profiling and cytotoxicity (§6) are all well represented in the earlier and middle tiers. Coverage skews further downstream still: preclinical safety and toxicology (§8) and in vivo PK/PD and biodistribution (§9) together account for 41 of 113 evaluations while candidate benchmarking (§10) has two evaluations.

A second decision-structure classification tags each evaluation by the scientific operation or judgment its passing answer requires, rather than its position in the discovery pipeline (Figure 1B). Safety assessment (32 evaluations) is the largest category, followed by target pharmacology (19), specificity assessment (18), PK/PD characterization (14); program-level integration (12), lead identification (8), and a supporting-analysis category (10) round out the remaining evaluations. While the pipeline tracks when key decision must be made, decision structure tracks what kind of judgment it demands.

A third view classifies each evaluation by the principal assay or evidence family its underlying dataset draws from (Figure 1C). PK, biodistribution, and toxicology data is the largest single family (41 of 113), closely trailed by off-target and mechanism-of-action profiling (35); together the two account for two-thirds of the set, reflecting the benchmark's emphasis on safety and exposure judgment over primary potency screening. Duplex biophysics and binding assays (12), cytotoxicity and dose-response panels (10), protein and splice-isoform readouts (9), and expression/knockdown assays (6) make up the balance.

The included indications, targets, and tissues (Figure 1D–F) cluster around a set of programs that are well known reference points in the oligonucleotide field rather than a broad therapeutic-area survey. ALS/FTD-spectrum disease and MAPT-driven tauopathy are the leading indications (Figure 1D), with myotonic dystrophy type 1, Huntington's disease, and KCNT1 epilepsy also represented; the corresponding gene targets, ATXN2, MAPT, DMPK, SOD1, HTT, STMN2, and KCNT1 (Figure 1E), are among the most extensively characterized targets in CNS and neuromuscular ASO/siRNA development, with public chemistry, off-target, and safety datasets mature enough to support deterministic grading. Tissue of origin (Figure 1F) follows the same logic.

4 Evaluation Design and Scoring

We developed evaluation tasks using general approaches to reviewing published research and approaches specific to oligonucleotide development (Figure 2A). The nine general approaches examined scientific decisions, figure reproduction, unexpected findings, evidence across sources, previously validated evaluations, analytical skills, related literature, interview questions, and common analytical errors. The three oligonucleotide-specific approaches focused on development decisions, development pitfalls and optimization, and properties specific to ASOs and siRNAs.

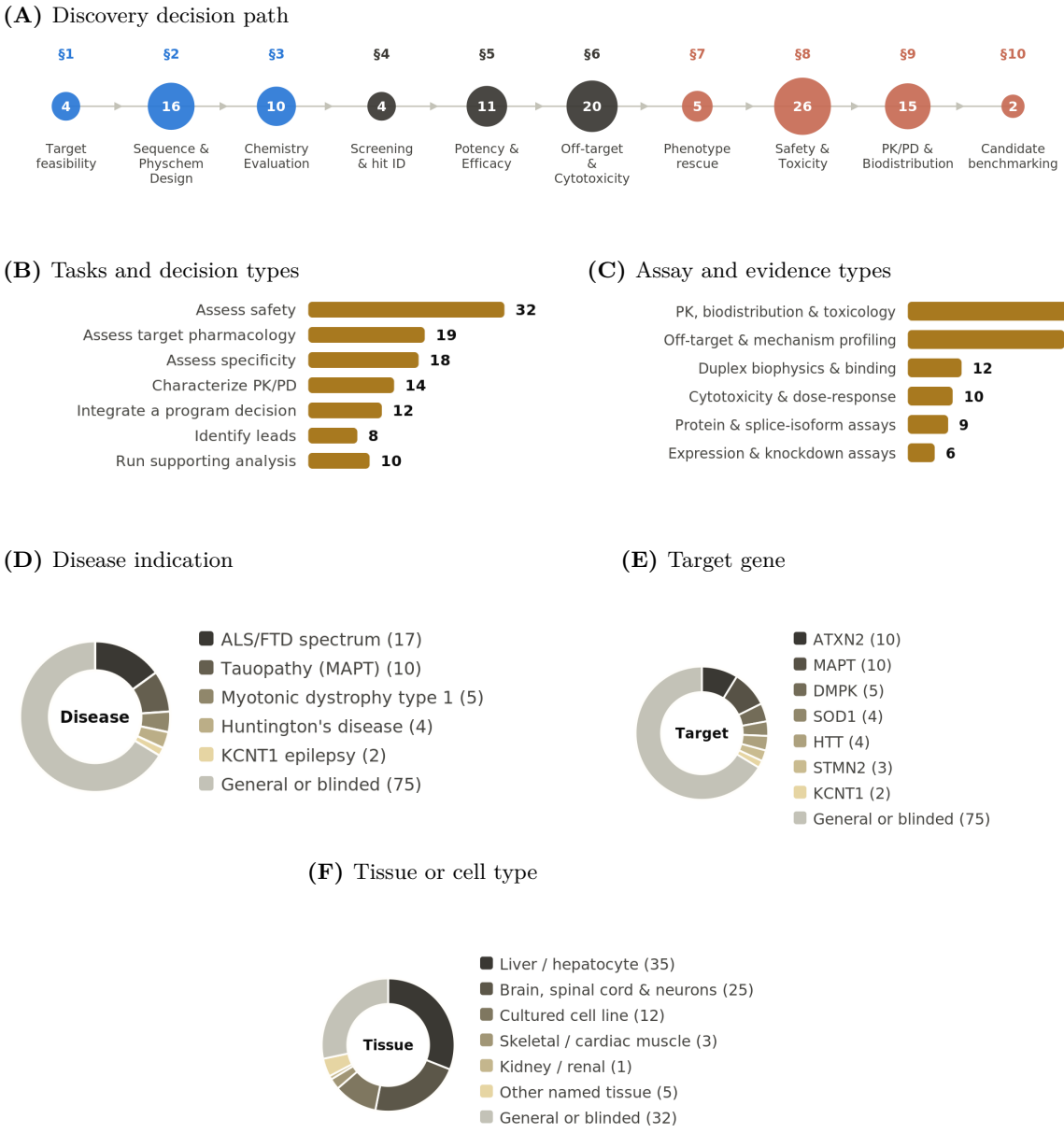


Figure 1: *Benchmark anatomy.* **A,** 10-stage taxonomy of evaluations spanning discovery-to-development candidate (§1–§10); circle area is proportional to evaluation count. **B,** Second classification based on the type of tasks and decisions, independent of pipeline stage. **C,** Third classification based on principal assay or evidence type the underlying dataset draws from. **D,** Statistic on diseases and therapeutic indications analyzed. **E,** Statistic on gene targets included. **F,** Statistic on tissue or cell type each evaluation's dataset was generated in.

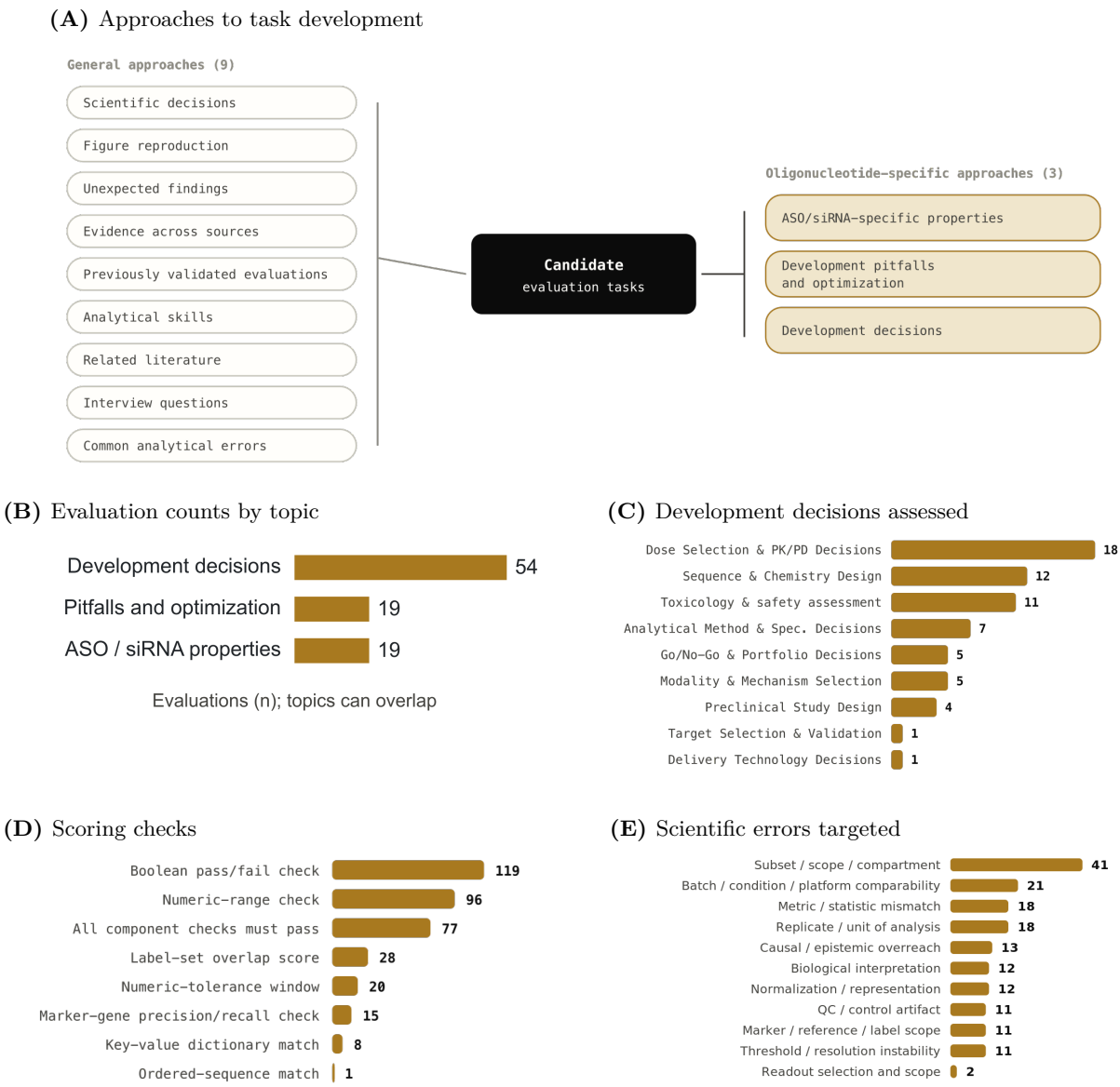


Figure 2: *Evaluation design and scoring.* **A,** Nine general approaches and three oligonucleotide-specific approaches used to develop evaluation tasks. **B,** Evaluation counts for three oligonucleotide-specific topics, assigned retrospectively; categories are not mutually exclusive. **C,** Development decisions assessed; categories with no evaluations are omitted. **D,** Scoring checks across 113 evaluations, including 77 instances in which all component checks must pass. **E,** Scientific errors targeted during construction, with 170 annotations across 55 of 113 evaluations.

Evaluations were retrospectively classified according to their correspondence with these three oligonucleotide-specific categories (Figure 2B). Development decisions were represented in 54 evaluations, development pitfalls and optimization in 19, and ASO/siRNA-specific properties in 19. These categories were not mutually exclusive. Because the approach used to generate each evaluation was not recorded, these assignments describe evaluation content rather than its origin.

The development-decisions category covers 12 types of decisions, from target selection to regulatory strategy. The benchmark includes evaluations in nine of these categories (Figure 2C). Dose selection and PK/PD decisions are the most common, followed by sequence and chemistry design and toxicology and safety assessment.

Evaluations are scored using deterministic checks of individual answer components (Figure 2D). The most common checks assess binary conditions (119 instances) or whether numerical results fall within specified ranges (96 instances). In 77 instances, checks are combined so that all components must pass. Other checks assess label overlap, numerical agreement within a tolerance, marker-gene precision and recall, values associated with specified keys, and sequence order.

During construction, authors annotated the scientific errors that evaluations were designed to detect (Figure 2E). These annotations are available for 55 of 113 evaluations, with 170 annotations in total. The most common concern the selection of biological subsets, compartments, comparators, or baselines. Other frequent categories concern comparability across batches, conditions, or platforms; the choice of metric or statistic; and the treatment of replicates and units of analysis. These annotations describe intended error coverage. They do not measure the frequency of errors in model responses.

5 Results

The highest mean pass rate was 55.5%. We tested 21 combinations of 12 models and four execution harnesses (Pi, Claude Code, OpenAI Codex, and Grok Build) on 113 evaluations, yielding 7,098 valid grading runs. Agents worked in sandboxed environments without internet access. For each configuration, we calculated the pass rate for each evaluation and averaged these rates across evaluations (Figure 3). GPT-6 Astra with OpenAI Codex achieved the highest mean pass rate, at 55.5% (95% CI, 47.2–63.7%). Across all configurations, 2,730 of 7,098 valid runs passed (38.5%).

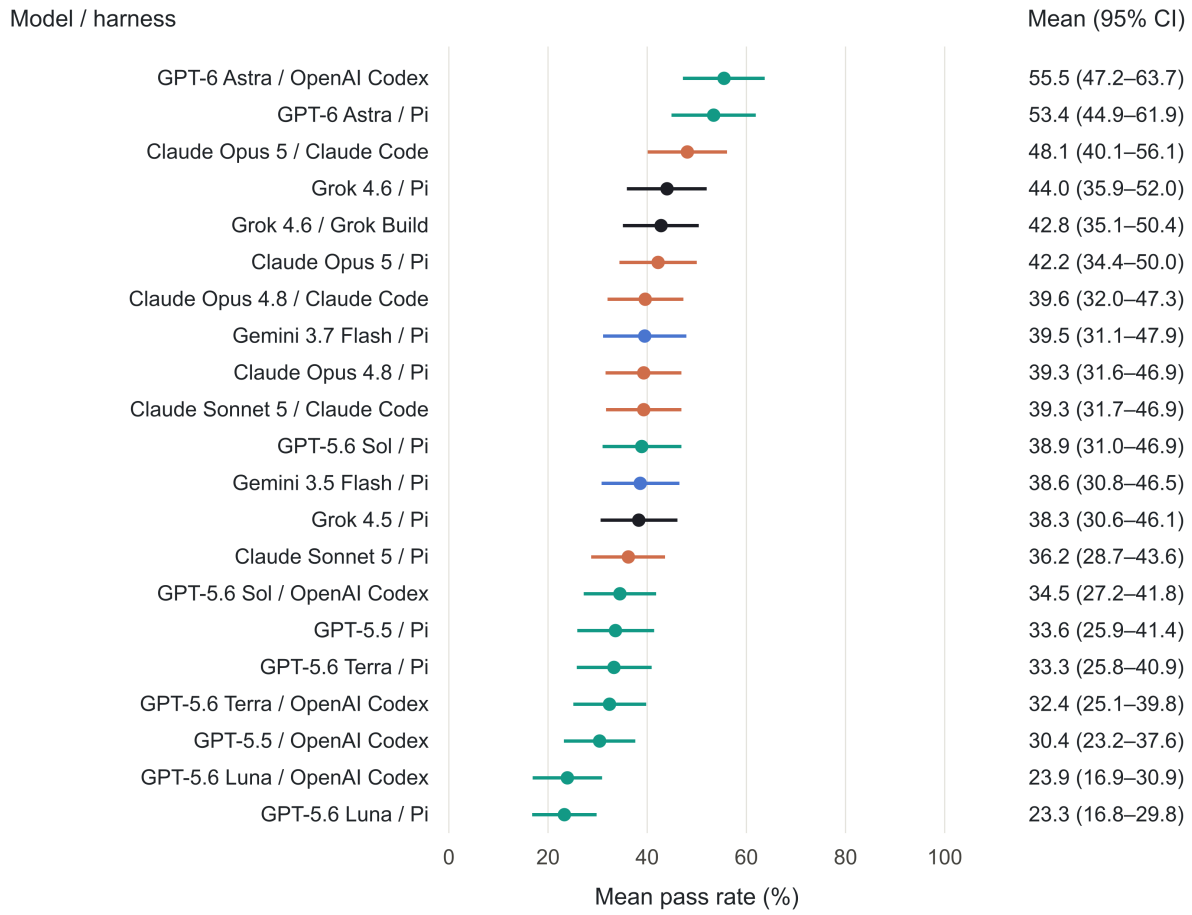


Figure 3: *Mean pass rate by configuration.* Mean pass rates for all 21 model–harness configurations, ordered from highest to lowest. Points show the unweighted mean of evaluation-level pass rates; horizontal lines show 95% Student's t confidence intervals across evaluations. Values at right give the mean and confidence interval. Evaluations without valid grading outcomes were excluded from the corresponding configuration's mean.

Across all 21 configurations, evaluations with three failed attempts were more common than evaluations with both passing and failing attempts (Figure 4). GPT-6 Astra failed all three attempts on 38 of 113 evaluations with OpenAI Codex (33.6%) and 41 of 113 with Pi (36.3%).

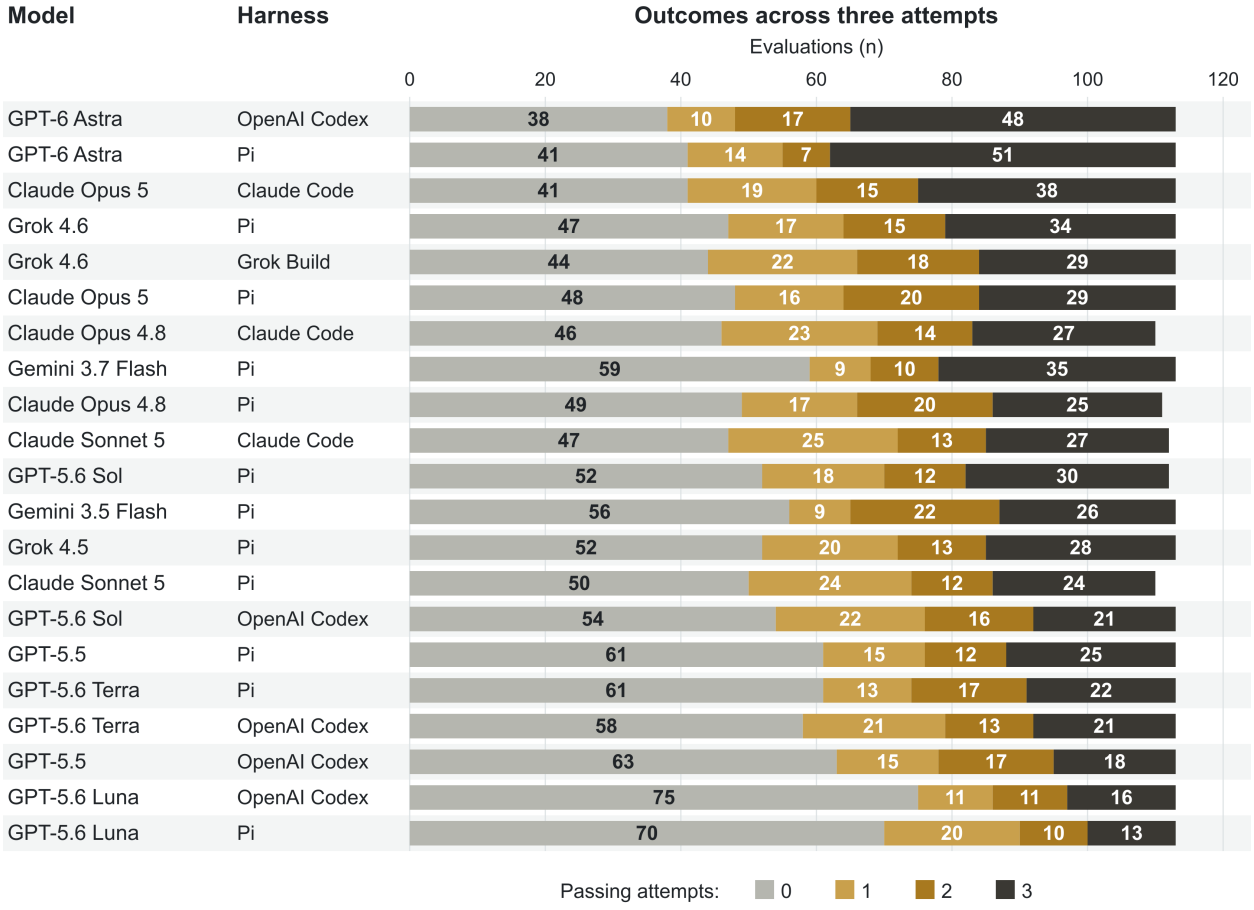


Figure 4: *Repeatability across three attempts.* Number of evaluations with zero, one, two, or three passing attempts for each of 21 model–harness configurations. Only evaluations with three completed grading runs are included (110–113 evaluations per configuration). Configurations are ordered by mean pass rate.

Pass rates varied with execution harness. Nine models were tested with both Pi and a second execution harness: Claude Code for Claude models, OpenAI Codex for GPT models, and Grok Build for Grok 4.6 (Supplementary Figure S1). The two Gemini models and Grok 4.5 were tested only with Pi. All three Claude models had higher mean pass rates with Claude Code, with differences ranging from 0.4 to 5.9 percentage points. Among GPT models, GPT-6 Astra and GPT-5.6-luna performed better with OpenAI Codex, whereas GPT-5.6-sol, GPT-5.5, and GPT-5.6-terra performed better with Pi. Grok 4.6 also performed better with Pi than with Grok Build.

Across the nine comparisons, the mean absolute difference was 2.4 percentage points (median, 2.1). The largest differences favored Claude Code for Claude Opus 5 (5.9 points) and Pi for GPT-5.6-sol (4.4 points). Supplementary Figure S1 shows confidence intervals for each configuration's mean pass rate; uncertainty in the differences between harnesses was not estimated.

Models differed in which evaluations they passed. We compared evaluation-level pass/fail outcomes across 12 models using Pi, classifying an evaluation as passed when at least 50% of valid attempts succeeded. Across the 66 model pairs, agreement ranged from 49.6% to 80.5% (mean, 64.2%; Figure 5). Agreement was highest between GPT-5.5 and GPT-5.6-terra and lowest between GPT-6 Astra and GPT-5.6-luna. The 2D map summarizes these differences in evaluation outcomes; the agreement matrix is provided in Supplementary Figure S2.

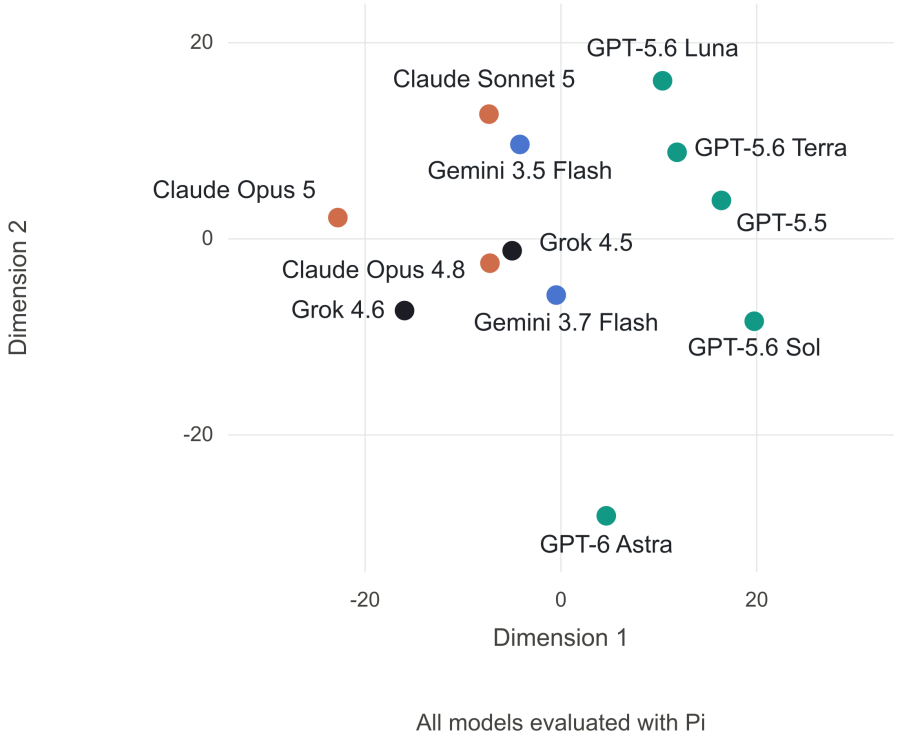


Figure 5: *Similarity in evaluation outcomes.* Models positioned closer together tended to pass and fail the same evaluations. The map shows all 12 models using Pi and a pass threshold of 50% of valid attempts. Positions are a two-dimensional approximation of pairwise disagreement; axis directions have no biological meaning. Exact agreement values are provided in Supplementary Table S2.

Relative performance varied across discovery stages. We compared the configuration with the highest overall mean pass rate from each model family (Figure 6). Among these four configurations, GPT-6 Astra with OpenAI Codex had the highest pass rates for screening and hit identification (75.0%), potency and efficacy (63.6%), and off-target profiling and cytotoxicity (56.7%). Claude Opus 5 with Claude Code had the highest pass rates for target feasibility, sequence design, chemistry evaluation, and phenotype rescue. The largest difference occurred in phenotype rescue, where Claude Opus 5 passed 73.3% of attempts, compared with 0–20% for the other three configurations.

The number of evaluations per stage ranged from two to 26. Screening and hit identification included four evaluations, phenotype rescue included five, and candidate benchmarking included two. Results for these smaller categories should be interpreted cautiously.

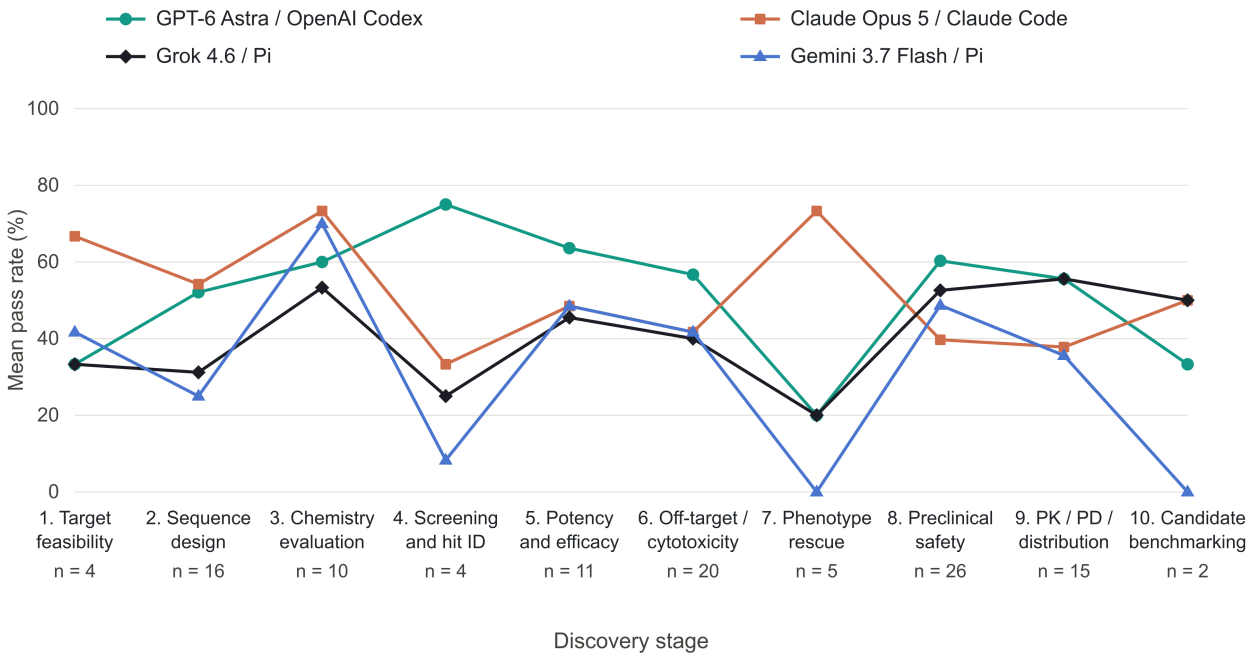


Figure 6: *Mean pass rate by discovery stage.* Stage-specific mean pass rates for the configuration with the highest overall mean pass rate within each model family. Stages follow the classification in Figure 1A; n gives the number of evaluations. Lines connect discrete stages to help readers follow each configuration and do not imply continuous change between stages.

Pass rates varied across assay and analysis categories. We compared 12 models using the Pi harness across 11 categories (Figure 7). Averaged across models, pass rates were highest for oligo sequence design (53.7%), dose response (53.3%), and organ/tissue toxicity (51.9%). The lowest mean pass rates occurred in splicing/isoform analysis (15.2%) and RNA-seq (20.7%).

The highest-performing model differed across categories. Claude Opus 5 achieved the highest pass rate for dose response (72%), Grok 4.6 for biomarker panels (72%), GPT-5.6 Terra for oligo sequence design (66%), and Gemini 3.5 Flash for RNA-seq (47%). Several categories contained few evaluations: splicing/isoform analysis included four, RNA-seq five, and dose response and biomarker panels six each. Comparisons within these categories therefore have limited precision.

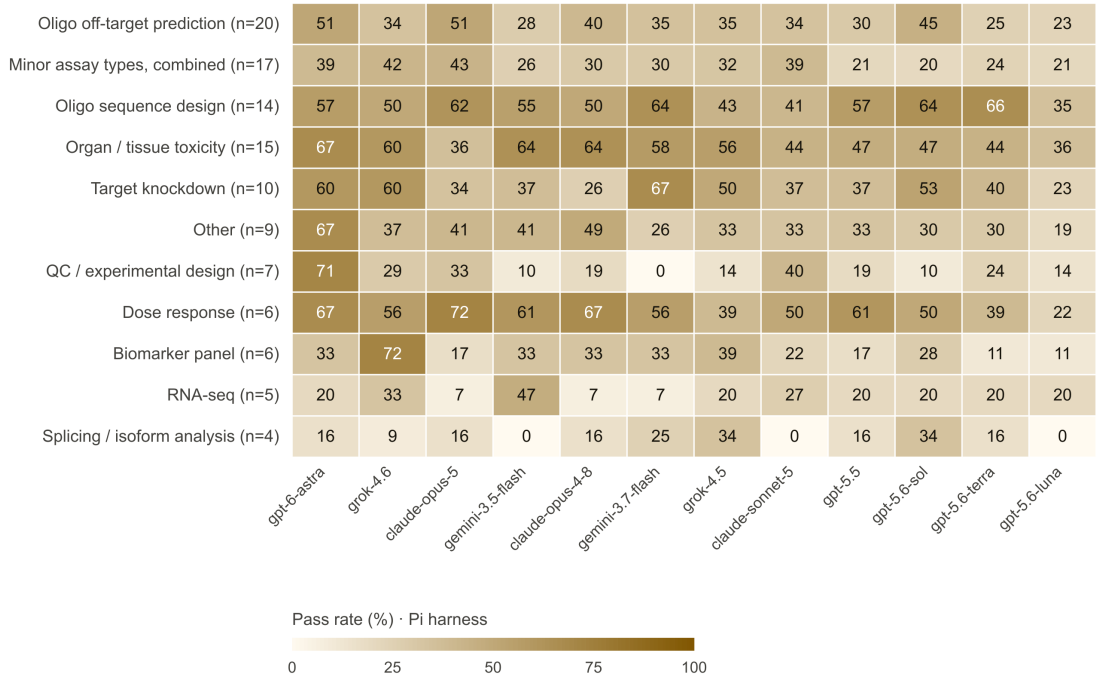


Figure 7: *Pass rate by assay and analysis category.* Pass rates for 12 models evaluated using Pi, grouped by assay and analysis category. Categories were assigned by keyword matching, with less frequent categories combined; n indicates the number of evaluations in each row. Cell values are percentages rounded to the nearest whole number. Color shows pass rate on a 0–100% scale.

Pass rates by task structure and anticipated failure mode. Short-form evaluations required a decision from a single experimental snapshot; long-form evaluations required integrating information across timepoints or experiments. Across all 21 configurations, pooled pass rates were 38.7% for short-form evaluations (99 evaluations; 2,408/6,221 valid attempts) and 36.7% for long-form evaluations (14 evaluations; 322/877 valid attempts; Figure 8A).

Pass rate did not consistently increase or decrease with the number of checks used to score an evaluation (Figure 8B). Category means ranged from 32.6% to 57.6%, with substantial variation among evaluations within several categories.

Mean pass rates also varied among groups defined by anticipated failure-mode tags (Figure 8C). Evaluations tagged for readout selection and scope had the lowest mean pass rate (14.3%; two evaluations), while those tagged for causal or epistemic overreach had the highest (37.0%; 13 evaluations). These tags describe errors the evaluations were designed to test. They do not establish which errors occurred in the model responses. Evaluations could carry multiple tags, so the groups overlap.

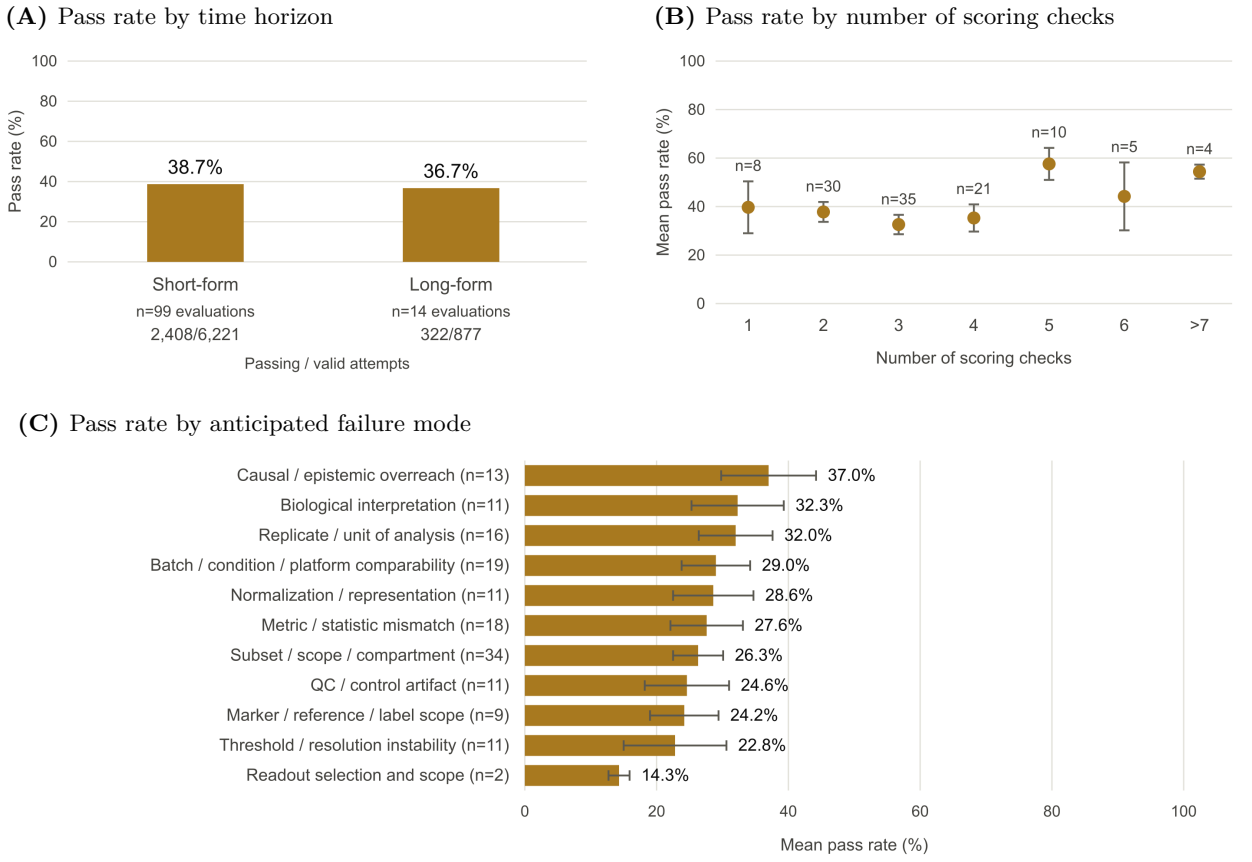
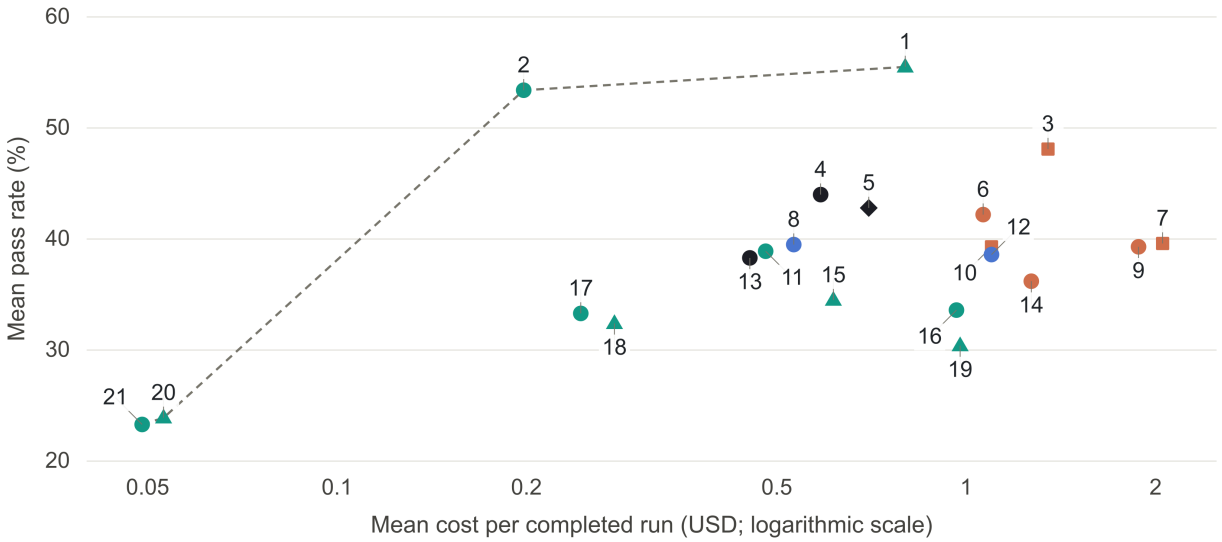


Figure 8: *Pass rates by task structure and anticipated failure mode.* **A,** Pooled pass rates across all 21 configurations for short-form and long-form evaluations. Labels give evaluation counts and passing/valid attempts. **B,** Mean pass rate by number of scoring checks, combining runs across all 21 configurations. **C,** Mean pass rate for evaluations with each anticipated failure-mode tag. In B and C, error bars show ±1 standard error of the mean across evaluations; n gives the number of evaluations. Failure-mode groups overlap because an evaluation can carry multiple tags.

Higher-cost configurations did not consistently achieve higher pass rates. Across the 21 configurations, mean cost ranged from $0.049 to $2.052 per completed run (Figure 9). GPT-6 Astra with Pi achieved a mean pass rate of 53.4% at $0.198 per run. With OpenAI Codex, its pass rate was 55.5% at $0.800 per run: a 2.1-percentage-point increase at approximately four times the cost.

Four configurations offered distinct trade-offs between cost and pass rate: GPT-5.6 Luna and GPT-6 Astra, each with Pi and OpenAI Codex (Figure 9). For each of these configurations, no alternative achieved an equal or higher pass rate at a lower cost. GPT-5.6 Luna had the lowest costs, at $0.049–0.053 per run, with pass rates of 23.3–23.9%.

GPT-6 Astra also used the fewest tokens per run: approximately 135,000 with Pi and 238,000 with OpenAI Codex (Supplementary Figure S3; Supplementary Table S1). Every other configuration used more tokens and had a lower mean pass rate. These comparisons describe the tested configurations; they do not isolate the effect of increasing a model's computation budget.



1 GPT-6 Astra / Codex
2 GPT-6 Astra / Pi
3 Claude Opus 5 / Code
4 Grok 4.6 / Pi
5 Grok 4.6 / Build
6 Claude Opus 5 / Pi
7 Claude Opus 4.8 / Code
8 Gemini 3.7 Flash / Pi
9 Claude Opus 4.8 / Pi
10 Claude Sonnet 5 / Code
11 GPT-5.6 Sol / Pi
12 Gemini 3.5 Flash / Pi
13 Grok 4.5 / Pi
14 Claude Sonnet 5 / Pi
15 GPT-5.6 Sol / Codex
16 GPT-5.5 / Pi
17 GPT-5.6 Terra / Pi
18 GPT-5.6 Terra / Codex
19 GPT-5.5 / Codex
20 GPT-5.6 Luna / Codex
21 GPT-5.6 Luna / Pi

Codex: OpenAI Codex; Code: Claude Code; Build: Grok Build.

Figure 9: *Cost and pass rate.* Mean pass rate versus mean cost per completed run for all 21 configurations. Numbers identify configurations in the key. The dashed line connects configurations for which no alternative has an equal or higher mean pass rate at lower cost, or a higher mean pass rate at the same cost. The cost axis is logarithmic; the pass-rate axis spans 20–60%. Confidence intervals and token usage are provided in Supplementary Figure S3 and Supplementary Table S1.

6 Discussion

TxBench-Oligonucleotide Discovery evaluates whether AI agents can recover scientific decisions from experimental data across the ASO and siRNA discovery process. The results show substantial limitations in the tested configurations. GPT-6 Astra with OpenAI Codex achieved the highest mean pass rate, at 55.5%, but failed all three attempts on 38 of 113 evaluations. Thus, even the highest-scoring configuration left approximately one-third of evaluations unanswered correctly after three attempts. These findings establish a baseline for measuring improvements in experimentally grounded scientific reasoning.

Performance varied across discovery stages and analysis categories. The configuration with the highest overall pass rate did not achieve the highest pass rate in every category, and models differed in which evaluations they passed. These differences suggest that model selection should account for the scientific tasks an agent will perform. However, several categories contained few evaluations, limiting the precision of comparisons. Larger evaluation sets and repeated comparisons would be needed to determine whether the observed differences reflect reproducible strengths in particular areas of oligonucleotide discovery.

Execution harness and resource use also affected the practical comparison between configurations. For GPT-6 Astra, OpenAI Codex produced a mean pass rate 2.1 percentage points higher than Pi at approximately four times the mean cost per run. Several other configurations cost more or used more tokens while achieving lower pass rates. These results show why cost and harness should be reported alongside model performance. They do not establish how performance would change if the computation budget of a fixed configuration were increased.

Several limitations constrain interpretation. The benchmark covers a selected set of scientific decisions, with uneven representation across discovery stages and analysis categories. Deterministic grading measures whether final answers satisfy specified requirements; it does not by itself identify the reasoning process that produced an answer. Similarly, the anticipated failure-mode tags describe errors targeted during evaluation design, rather than errors demonstrated in model responses. Establishing why agents fail will require systematic review of their analyses and outputs. The study also does not establish how performance compares with that of scientists or whether agent assistance improves decisions in a research workflow.

The immediate implication is that agents intended for a particular scientific application should be evaluated on tasks relevant to that application. Future work should expand underrepresented categories, characterize observed errors, and compare unaided scientists with scientists using agent assistance. These studies would help determine where agents provide useful support and where their outputs require additional verification.

7 Methods

Benchmark composition and grading. TxBench-Oligonucleotide Discovery comprises 113 evaluations derived from 21 ASO or siRNA discovery programs. Each evaluation includes a task, associated input data, a deterministic grader, and metadata describing its discovery stage and task category. Anticipated failure-mode annotations are available for 55 evaluations. Graders assess specified components of a structured final answer, including binary conditions, numerical ranges and tolerances, label overlap, and sequence order. Where checks are combined under an all-must-pass rule, every component must pass for the answer to be scored as correct.

Model execution. We tested 21 configurations comprising 12 models and four execution harnesses: Pi, Claude Code, OpenAI Codex, and Grok Build. Each configuration was scheduled for three attempts per evaluation. Agents ran in isolated environments without internet access, using the supplied task files and installed scientific analysis tools. Models were run at their maximum available reasoning-effort setting.

Pass rates and exclusions. For each evaluation and configuration, we calculated pass rate as the number of passing attempts divided by the number of valid grading outcomes. Runs without a valid grading outcome were excluded. Evaluations with no valid outcomes were omitted from that configuration's mean. Configuration-level pass rates were calculated as the unweighted mean of evaluation-level pass rates. We calculated 95% Student's t confidence intervals across evaluation-level rates. The pooled campaign pass rate was calculated separately by dividing total passing attempts by total valid attempts.

Repeatability and model comparisons. Repeatability analyses included evaluations with three valid attempts and classified them by the number of passing attempts. Pairwise model comparisons used Pi for all 12 models. An evaluation was classified as passed when at least 50% of valid attempts passed; agreement was the percentage of evaluations with matching pass/fail classifications. For the 2D similarity map, we used classical multidimensional scaling on distances defined as 100 minus the pairwise percentage agreement. The first two positive-eigenvalue dimensions accounted for approximately 47% of the sum of positive eigenvalues. The projection approximates the full agreement matrix and was plotted with equal scaling on both axes.

Category analyses. Discovery-stage comparisons used the configuration with the highest overall mean pass rate within each model family. Assay and analysis-category comparisons used all 12 models with Pi. Time-horizon rates pooled passing and valid attempts across all configurations. Scoring-complexity and anticipated failure-mode analyses summarized evaluation-level pass rates within each category, with error bars showing one standard error of the mean. Failure-mode categories overlapped because evaluations could have multiple annotations.

Resource use. We compared mean cost and token usage per run with configuration-level mean pass rate. The cost–pass-rate frontier included configurations for which no alternative had an equal or higher mean pass rate at lower cost, or a higher mean pass rate at the same cost. Confidence intervals, costs, and token counts for every configuration are reported in Supplementary Table S1.

Data availability. Four example evaluations and selected agent trajectories are publicly released. The full evaluation set, graders, reference answers, internal solution notes, and input data remain under restricted access.

References

1. Ramasamy, R., Le, H., Brougher, J., Urrutia, A., Banerjee, A., Poust, S. & Workman, K. TxBench – Antibody Discovery: Benchmarking AI Agents on Experimentally Grounded Decisions in Therapeutic Antibody Discovery. LatchBio, 2026.

2. Le, H., Ramasamy, R., Urrutia, A., Yazdani, M., Proctor, T. & Workman, K. TxBench-PP: Analyzing AI Agent Performance on Small-Molecule Preclinical Pharmacology. *arXiv:submit/7725456 [cs.AI]*, 17 June 2026.

3. Bhowmick, A., Rahmatpour, N., Sharma, S., Xu, Q., Gupta, A., Lagwankar, A., Banerjee, A. & Zou, C. VariantBench: An Agentic Benchmark for Genetic Variant Discovery and Interpretation. LatchBio, 2026.

4. Hill, B., Jaques, M.R., Nair, R.R., Whiffin, N., Wood, M.J.A., Sanders, S.J., Oliver, P.L., Hill, A.C. & Rinaldi, C. Accurately modeling RNase H-mediated antisense oligonucleotide efficacy. *Molecular Therapy Nucleic Acids* 37(3):103004, 2026.

5. Sundar, A., Wei, Z. & Griesmer, S. Benchmarking Large Language Models for Predicting Therapeutic Antisense Oligonucleotide Efficacy. *bioRxiv* 2026.02.17.706455, 2026.

6. Andreone, B.J. et al. Durable suppression of seizures in a preclinical model of KCNT1 genetic epilepsy with divalent small interfering RNA. *Epilepsia* 66(5):1677–1690, 2025.

7. Bäckström, E., Bonetti, A., Johnsson, P., Öhlin, S., Dahlén, A., Andersson, P., Andersson, S. & Gennemark, P. Tissue pharmacokinetics of antisense oligonucleotides. *Molecular Therapy Nucleic Acids* 35(1):102133, 2024.

8. Burdick, A.D. et al. Sequence motifs associated with hepatotoxicity of locked nucleic acid-modified antisense oligonucleotides. *Nucleic Acids Research* 42(8):4882–4891, 2014.

9. Chen, X., Birey, F., Li, M.-Y., Revah, O., Levy, R., Thete, M.V., Reis, N., Kaganovsky, K., Onesto, M., Sakai, N., Hudacova, Z., Hao, J., Meng, X., Nishino, S., Huguenard, J. & Pașca, S.P. Antisense oligonucleotide therapeutic approach for Timothy syndrome. *Nature* 628(8009):818–825, 2024.

10. Didiot, M.C. et al. Nuclear localization of huntingtin mRNA is specific to cells of neuronal origin. *Cell Reports* 24(10):2553–2560, 2018.

11. El Boujnouni, N., van der Bent, M.L., Willemse, M., 't Hoen, P.A.C., Brock, R. & Wansink, D.G. Block or degrade? Balancing on- and off-target effects of antisense strategies against transcripts with expanded triplet repeats in DM1. *Molecular Therapy Nucleic Acids* 32:622–636, 2023.

12. Golinski, S.R., Soriano, K., Briegel, A.C. et al. RNA targeting therapy for a prenatally enriched potassium channel associated with severe childhood epilepsy and premature death. *Nature Communications* 17:5864, 2026.

13. Guo, S., Booten, S.L., Aghajan, M. et al. Antisense oligonucleotide treatment ameliorates alpha-1 antitrypsin-related liver disease in mice. *Journal of Clinical Investigation* 124(1):251–261, 2014.

14. Hagedorn, P.H., Pontoppidan, M., Bisgaard, T.S. et al. Identifying and avoiding off-target effects of RNase H-dependent antisense oligonucleotides in mice. *Nucleic Acids Research* 46(11):5366–5380, 2018.

15. Holgersen, E.M., Gandhi, S., Zhou, Y. et al. Transcriptome-wide off-target effects of steric-blocking oligonucleotides. *Nucleic Acid Therapeutics* 31(6):392–403, 2021.

16. Janas, M.M., Jiang, Y., Schlegel, M.K., Waldron, S., Kuchimanchi, S. & Barros, S.A. Impact of oligonucleotide structure, chemistry, and delivery method on in vitro cytotoxicity. *Nucleic Acid Therapeutics* 27(1):11–22, 2017.

17. Klim, J.R., Williams, L.A., Limone, F. et al. ALS-implicated protein TDP-43 sustains levels of STMN2, a mediator of motor neuron growth and repair. *Nature Neuroscience* 22(2):167–179, 2019.

18. Kuijper, E.C., van der Graaf, L., Pepers, B.A. et al. Determining off-target effects of splice-switching antisense oligonucleotides using short read RNAseq in neuronally differentiated human induced pluripotent stem cells. *Human Molecular Genetics* 34(22):1912–1925, 2025.

19. Lee, H.-S., Seok, H., Lee, D.H., Ham, J., Lee, W., Youm, E.M., Yoo, J.S., Lee, Y.-S., Jang, E.-S. & Chi, S.W. Abasic pivot substitution harnesses target specificity of RNA interference. *Nature Communications* 6:10154, 2015.

20. McCampbell, A., Cole, T., Wegener, A.J. et al. Antisense oligonucleotides extend survival and reverse decrement in muscle response in ALS models. *Journal of Clinical Investigation* 128(8):3558–3567, 2018.

21. Shen, W., De Hoyos, C.L., Sun, H., Vickers, T.A., Liang, X.H. & Crooke, S.T. Acute hepatotoxicity of 2′ fluoro-modified 5–10–5 gapmer phosphorothioate oligonucleotides in mice correlates with intracellular protein binding and the loss of DBHS proteins. *Nucleic Acids Research* 46(5):2204–2217, 2018.

22. Vickers, T.A., Rahdar, M., Prakash, T.P. & Crooke, S.T. Kinetic and subcellular analysis of PS-ASO/protein interactions with P54nrb and RNase H1. *Nucleic Acids Research* 47(20):10865–10880, 2019.

23. Wheeler, T.M., Leger, A.J., Pandey, S.K. et al. Targeting nuclear RNA for in vivo correction of myotonic dystrophy. *Nature* 488(7409):111–115, 2012.

24. Popper, K. The Logic of Scientific Discovery. Basic Books, 1959.

25. Duhem, P. The Aim and Structure of Physical Theory. Princeton University Press, 1954.

26. Quine, W.V.O. Two Dogmas of Empiricism. *The Philosophical Review* 60(1):20–43, 1951.

27. Workman, K., Yang, Z., Muralidharan, H. & Le, H. SpatialBench: Can Agents Analyze Real-World Spatial Biology Data? *arXiv:2512.21907*, 2025.

28. Workman, K., Yang, Z., Muralidharan, H., Abdulali, A. & Le, H. scBench: Evaluating AI Agents on Single-Cell RNA-seq Analysis. *arXiv:2602.09063*, 2026.

29. Muralidharan, H., Baskar, R., Lee, S.H., Proctor, T. & Workman, K. EpiBench: Verifiable Evaluation of AI Agents on Epigenomics Analysis. *arXiv:2606.13602*, 2026.

30. Diks, I., Muralidharan, H., Proctor, T. & Workman, K. Verifiable Benchmarking of Long-Horizon Spatial Biology. *arXiv:2605.28065*, 2026a.

31. Diks, I., Yang, Z., Banerjee, A., Proctor, T. & Workman, K. scBench-Long: Verifiable Benchmarking of Long-Horizon Single-Cell Biology. LatchBio, 2026b.