TxBench - Antibody Discovery

Benchmarking AI Agents on Experimentally Grounded Decisions in Therapeutic Antibody Discovery

Ramesh Ramasamy[1,*], Hannah Le[1,*], Jackson Brougher[1,*], Alex Urrutia[1,*], Arjun Banerjee[1], Sean Poust[1], Kenny Workman[1]

[1]LatchBio, San Francisco, CA, USA
[*]These authors contributed equally.

Correspondence: kenny@latch.bio

ABSTRACT

Biologics discovery requires researchers to integrate heterogeneous experimental evidence and decide whether therapeutic hypotheses, assays, binders, and engineered candidates should advance. We introduce TxBench-Antibody Discovery, a benchmark of 100 experimentally grounded evaluations spanning ten competencies in therapeutic antibody discovery, from target assessment and assay design to cellular pharmacology, antibody engineering, and candidate de-risking. Each evaluation is derived from publicly available experimental studies and uses deterministic grading to assess a consequential research decision. Across 20 model–harness configurations, each scheduled for three attempts per evaluation, Opus 5 with Claude Code achieved the highest evaluable pass rate at 53.0%, with four configurations from xAI and Google performing similarly, within three percentage points. Differences between harnesses showed that execution scaffolds also affected scientific accuracy. Failures were frequently reproducible, although every evaluation was solved by at least one configuration and 89 were solved by models from at least two families. Model rankings varied across biological competencies, decision types, and evidence sources. Greater cost, token usage, and tool use did not consistently improve accuracy. Failure analysis indicated that most incorrect attempts reflected scientific framing rather than analytical execution: agents typically performed internally consistent calculations that answered a scientifically adjacent question. Current agents therefore remain unreliable at turning experimental evidence into defensible biologics discovery decisions, highlighting the need for benchmarks that evaluate scientific reasoning in realistic discovery workflows.

Introduction

Therapeutic biologics discovery is a multiparameter, context-dependent process rather than a sequence of independent assay outcomes. For antibodies, affinity must be considered alongside specificity, stability, solubility, expression, and other properties that ultimately determine whether a molecule can be developed into a therapeutic [1]. Crucially, the meaning of a measurement depends on the experimental system in which it was obtained. Enrichment during phage display, for example, can reflect amplification bias rather than target-dependent selection [2]; measurements against purified or solubilized proteins may not fully capture recognition and signaling properties of the target in its native cell-surface context [3]; and changes in valency or molecular geometry can alter biological activity even when the underlying binding domains remain unchanged [4]. Progress through a discovery program therefore depends not simply on identifying the strongest result in a given assay, but on understanding which experiments are informative for the question at hand, how conflicting measurements should be interpreted, and whether the combined evidence supports advancement, redesign, a change in experimental model, additional experimentation, or program termination.

Existing benchmarks in therapeutic discovery have largely focused on well-defined prediction or design endpoints. Therapeutics Data Commons established standardized datasets and evaluation tasks spanning drug discovery and development [5], while antibody-specific resources such as FLAb2 benchmark models against experimentally measured properties including affinity, stability, expression, aggregation, polyreactivity, pharmacokinetics, and immunogenicity [6]. More recently, the AIntibody challenge introduced a prospective, blinded evaluation in which computationally designed antibodies were tested experimentally for affinity and developability [7]. These efforts provide increasingly realistic measures of predictive and design performance, but they generally evaluate a specified molecular property or the quality of a designed candidate. In our prior work, TxBench-PP, we instead evaluated whether AI agents could recover consequential decisions from small-molecule preclinical pharmacology data spanning target engagement, mechanism of action, safety, pharmacokinetics, and translational efficacy [8]. The present benchmark extends this decision-centered approach to biologics discovery, where interpretation often depends on reconciling evidence across molecular format, antigen context, binding, functional activity, and engineering trade-offs.

We developed this benchmark to evaluate whether AI systems can make these judgments and turn complex experimental evidence into defensible research decisions. The benchmark focuses on therapeutic antibodies, with tasks grounded in public studies and their underlying experimental data. Each evaluation requires the agent to identify the relevant biological context and comparison, preserve experimental relationships that affect interpretation, reconcile discordant measurements, and distinguish intrinsic molecular properties from effects introduced by assay conditions, selection systems, or engineered formats. The goal is therefore not simply to reproduce an experimental result, but to determine what the available evidence supports and what research action should follow from it.

Benchmark Construction

We designed the benchmark around a multidimensional set of discovery decisions rather than as a linear collection of assay-specific questions. Its primary axis places each evaluation within the iterative discovery cycle, spanning target and modality selection, experimental design, binder discovery, molecular characterization, functional testing, and subsequent engineering. Additional annotations describe the type of decision, computational approach, therapeutic modality, evidence source, and difficulty. This organization allows performance to be examined across different aspects of discovery judgment without treating any single assay or analytical method as the scientific endpoint. It also reflects the non-linear nature of biologics discovery, where a failure in a functional or native-context assay may require revisiting an earlier decision about the antigen, experimental model, assay design, binder, or therapeutic modality.

Evaluations are constructed from public experimental evidence and require agents to determine consequential research actions from the underlying data and experimental design. Related molecules, variants, structures, and studies from the same discovery campaign are assigned to the same benchmark stage to limit information leakage across closely related examples. Ground truths are reconstructed from the available data and source methods using analyses that are also available to the agent. Grading is deterministic and assesses the terminal decision together with the supporting quantities or evidence needed to distinguish it from plausible alternatives. Candidate tasks are retained only when the answer is scientifically recoverable, robust to reasonable analytical choices, and distinguishable from plausible but incorrect conclusions. During task development, model rollouts are also examined to separate failures of scientific reasoning from failures caused by formatting, software, or tooling. The resulting unit of evaluation is therefore not an isolated prediction or calculation, but a biologics discovery decision supported by experimental evidence.

Benchmark Anatomy

The benchmark contains 100 evaluations spanning ten competencies across therapeutic biologics discovery (Figure 1). Coverage extends from target opportunity assessment, reagent and assay design, and binder discovery through molecular characterization, cellular pharmacology, engineering, and discovery-stage candidate de-risking. The distribution is deliberately broad rather than uniform. candidate de-risking is the largest category, with substantial representation from assay and screening-funnel design, cellular pharmacology, target-hypothesis assessment, and binding characterization. More specialized areas, including binder generation and sequence or lineage analysis, contain fewer evaluations. These counts are intended to describe the scientific breadth of the benchmark, rather than the relative frequency with which these decisions occur in a typical discovery program.

We also classified each evaluation by its primary evidence or analysis substrate and by the type of decision required from the agent. The benchmark includes expression and single-cell data, dose-response and imaging measurements, assay quality-control and hierarchical analyses, binding kinetics, screening and enrichment data, sequence and repertoire information, structural and epitope evidence, and experimental-design problems. Across these settings, the most common decision involves candidate ranking, selection, or triage, followed by experimental design, comparative assessment, mechanistic inference, and integration of multiple lines of evidence into a discovery gate. Other evaluations focus on quantitative characterization or on diagnosing assay and development failures. Together, these complementary classifications show that the benchmark is not dominated by any single assay type or computational procedure. Instead, it tests whether agents can select and apply the appropriate analysis for the specific scientific decision posed at different stages of biologics discovery.

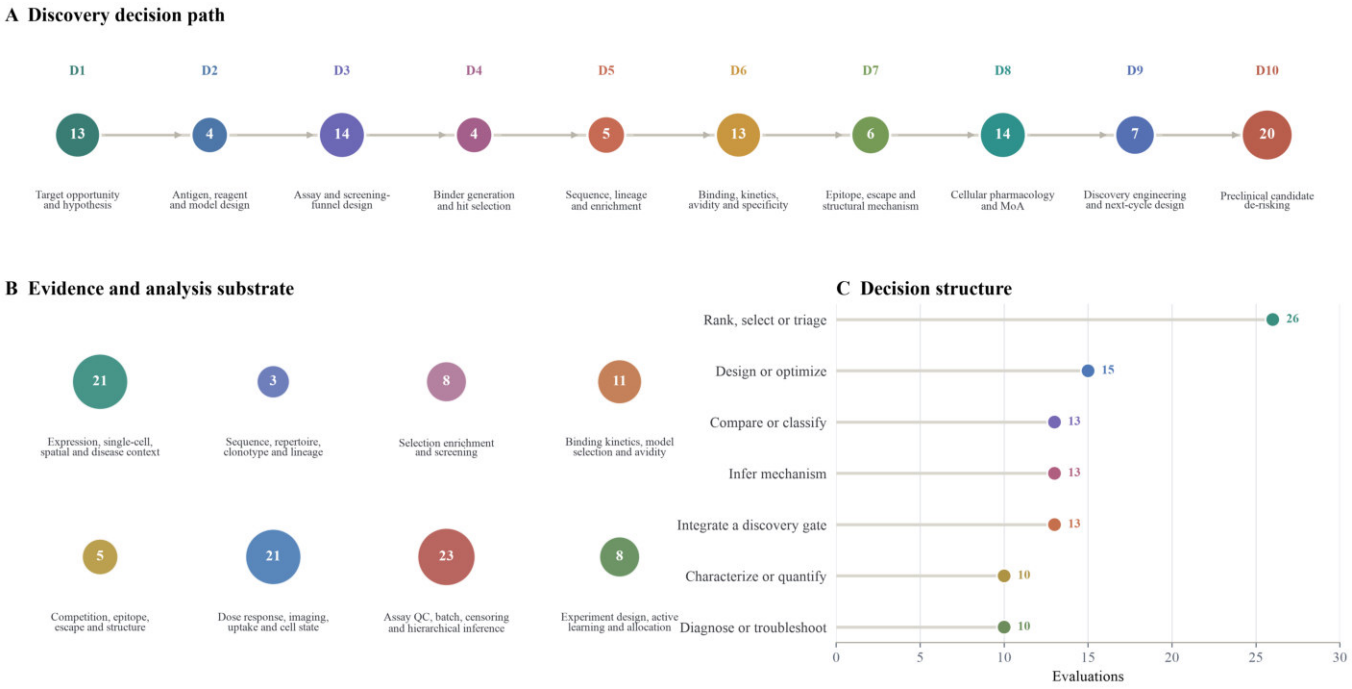

Figure 1: Benchmark anatomy. **A**, Discovery competency labels position evaluations along the therapeutic-biologics discovery path, from target opportunity and experimental design through molecular characterization, cellular pharmacology, engineering, and candidate de-risking. Circle area represents the number of evaluations in each category. **B**, Evidence-and-analysis substrate labels summarize the principal data and analytical families represented in the benchmark; bubble area represents evaluation count. **C**, Decision-structure labels describe the scientific operation or judgment required for a passing answer; horizontal position indicates evaluation count.

Results

Even the strongest model–harness configuration failed on nearly half of the attempts

We found that no model–harness configuration passed more than 53% of evaluable attempts, and even the strongest system failed nearly half of them (Figure 2). Pass rates excluded refusals, platform or execution errors, unsupported answer-only guesses, and unverifiable trajectories. Opus 5 with Claude Code performed best, passing 152 of 287 evaluable attempts (53.0%). Grok 4.6 was similarly competitive with PI (153/296; 51.7%) and Grok Build (153/300; 51.0%), while Opus 5 with PI reached 51.4% and Gemini 3.7 Flash with PI reached 50.0%. Grok 4.6 therefore remained near the top under two harnesses, separated by only 0.7 percentage points. Harness effects were larger for Gemini 3.7 Flash, whose pass rate declined from 50.0% with PI to 45.3% with Mini-SWE-Agent. The OpenAI configurations weren't competitive overall: GPT-5.6 Sol passed 33.3–33.8% of evaluable attempts, GPT-5.5 passed 31.5–33.6%, GPT-5.6 Terra passed 29.7–31.0%, and GPT-5.6 Luna passed 18.2–19.1%.

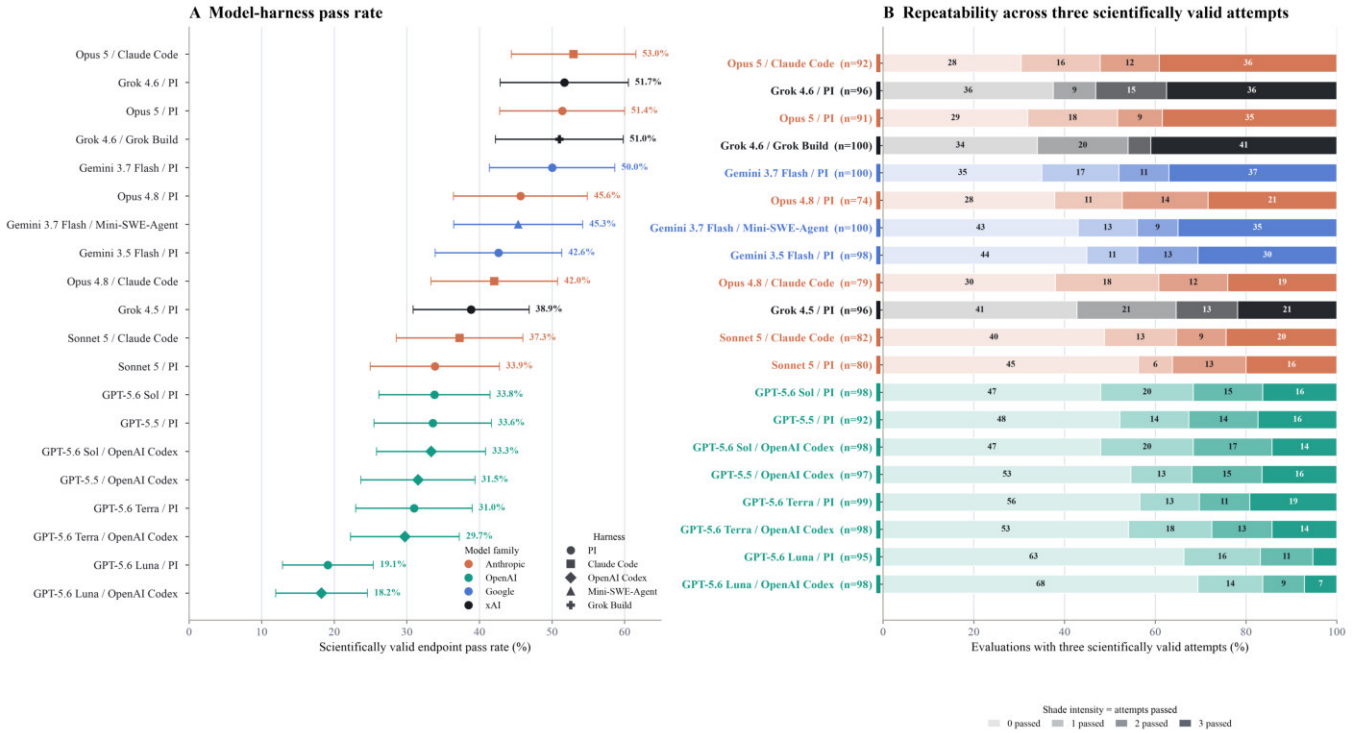

Figure 2: Top systems remain below 55%. Model–harness performance and repeatability across TxBench-Antibody. **A**, Pass rate for each model–harness configuration across 100 evaluations, with three attempts per evaluation. Error bars indicate 95% Student's t intervals calculated across per-evaluation pass rates. Colors indicate model family and marker shapes indicate execution harness. **B**, Evaluation-level consistency across repeated attempts. Analysis was restricted to evaluations with three non-refused attempts for each configuration. Bars are normalized within configuration, with shading from light to dark indicating zero, one, two, or three successful attempts. Numbers within each segment indicate the corresponding evaluation count, and n denotes the number of evaluations with complete non-refused triplicates. Configurations are ordered by their pass rate in panel A.

Performance differences were also highly repeatable at the level of individual evaluations. Among evaluations with exactly three evaluable attempts, Opus 5 with Claude Code failed all three attempts on 28 of 92 evaluations and passed all three on 36. Grok 4.6 with PI failed all three on 36 of 96 evaluations and passed all three on 36, whereas Grok Build failed all three on 34 of 100 and passed all three on 41. Gemini 3.7 Flash with PI failed all attempts on 35 evaluations and passed all attempts on 37. Lower-performing models had substantially larger sets of consistently failed evaluations: each GPT-5.6 Sol configuration failed all three attempts on 47 of 98 evaluations, while GPT-5.6 Luna did so on 63 of 95 evaluations with PI and 68 of 98 with OpenAI Codex. Solvability was not dependent on isolated successes: 91 evaluations were passed in at least two of three attempts by some configuration, 80 were passed in all three attempts by at least one configuration, and 89 were solved by models from at least two independent families. Aggregate performance differences therefore arose primarily from reproducible strengths and weaknesses on particular scientific problems rather than isolated stochastic errors.

Evaluation-level head-to-head comparisons showed that comparable aggregate pass rates did not mean the models solved identical sets of problems (Figure 3). Upon pooling the two harnesses for each model and restricting the analysis to evaluations containing all six evaluable attempts, Opus 5 outperformed Grok 4.6 on 30 of 84 evaluations. Conversely, Grok 4.6 outperformed Opus 5 on 28 evaluations, and the models tied on the remaining 26. The comparison involving Gemini 3.7 Flash yielded a similarly balanced outcome: Opus 5 performed

better on 32 of 87 evaluations, Gemini 3.7 Flash on 30, and 25 tied. In contrast, Opus 5 maintained a more pronounced advantage over GPT-5.6 Sol, outperforming it on 46 of 86 evaluations; only 22 evaluations favored GPT-5.6 Sol, and 18 ended in ties. Therefore, Grok 4.6 and Gemini 3.7 Flash approached Opus 5's performance by leveraging partially complementary strengths across various scientific problems. Conversely, the lower aggregate score of GPT-5.6 Sol reflected a broader performance deficit at the evaluation level rather than a limited number of catastrophic failures.

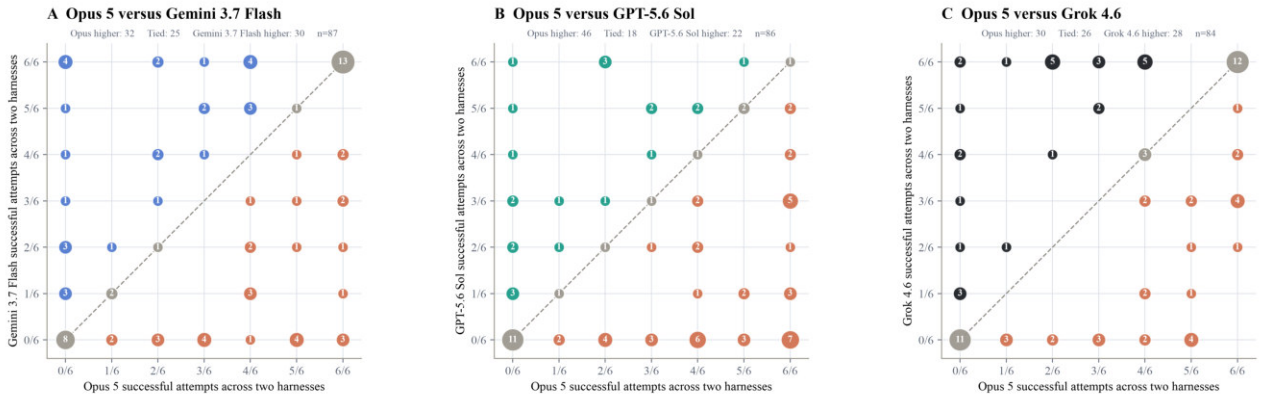

Figure 3: Evaluation-level head-to-head performance across model families. A, Opus 5 versus Gemini 3.7 Flash. **B**, Opus 5 versus GPT-5.6 Sol. **C**, Opus 5 versus Grok 4.6. For each model, we pooled three attempts from each of two execution harnesses, yielding six evaluable attempts per evaluation. We included only evaluations with all six evaluable attempts for both models being compared. Axes show the number of successful attempts from 0/6 to 6/6. Bubble area and the number within each bubble indicate the number of evaluations at that coordinate. Points below the diagonal favor Opus 5, points above it favor the comparator, and points on the diagonal represent ties.

More computation does not consistently improve scientific performance

We observed substantial variation across model–harness configurations in monetary cost (USD), token usage, and tool calls, but greater resource consumption did not consistently improve benchmark performance (Figure 4). We calculated both performance and resource summaries over evaluable runs, excluding refusals, platform or execution errors, unsupported answer-only guesses, and unverifiable trajectories. Opus 5 with Claude Code achieved the highest pass rate (53.0%), averaging approximately $1.70, 1.12 million tokens, and 21 tool calls per evaluable run. Grok 4.6 with PI followed closely at 51.7% while using substantially fewer resources: $0.59, 0.50 million tokens, and 16 tool calls per run. Grok 4.6 with Grok Build reached 51.0% at $0.73, 0.71 million tokens, and 19 tool calls. Gemini 3.7 Flash with PI achieved 50.0% at only $0.46 per run, although it used 1.87 million tokens and 37 tool calls. Pairing Gemini with Mini-SWE-Agent reduced its resource use to $0.33, 1.12 million tokens, and 25 tool calls, but lowered its pass rate to 45.3%. Similar accuracy levels therefore emerged from markedly different execution profiles, and harness choice affected both scientific performance and resource use.

Lower resource use, however, was not sufficient for strong performance. GPT-5.6 Sol used only 0.43–0.55 million tokens and 13–16 tool calls per run, yet passed just 33.3–33.8% of evaluable attempts. Within xAI, Grok 4.5 used fewer resources than Grok 4.6 with PI, averaging 0.39 million tokens and 15 tool calls, but achieved a substantially lower pass rate of 38.9%. GPT-5.6 Luna was the least expensive model, at approximately $0.06–$0.07 per run, but had the lowest pass rate despite using 1.15–1.34 million tokens and 21–24 tool calls. Conversely, the most expensive configurations were not the most accurate: Opus 4.8 cost approximately $1.99–$2.38 per run but passed only 42.0–45.6% of evaluable attempts. Grok 4.6 with PI consequently provided one of the strongest accuracy–resource trade-offs, whereas Opus 5 with Claude Code retained the highest absolute accuracy. Additional cost, token volume, and tool activity did not directly translate to improved agent success.

To understand why additional computation did not always improve performance, we manually examined individual execution trajectories. In a cross-species receptor-translation evaluation, Opus 5 with Claude Code passed all three attempts while averaging approximately 0.29 million tokens and nine tool calls. In each successful trajectory, the model used the measured transcriptional response to identify an inconsistency in the treatment record, reconstruct the correct experimental arms, and calculate the required receptor-chain ratios. Gemini 3.7 Flash with PI failed all three attempts despite using more than twice as many tokens, approximately 0.63 million, and three times as many tool calls, averaging 27. The model recognized the same inconsistency but retained the incorrect arm assignment, reversing the relevant expression contrast. Its subsequent calculations propagated this initial error rather than correcting it. We observed similar behavior in other trajectories: additional computation was useful when it resolved a specific uncertainty in the evidence, but it could also extend an analysis organized around an incorrect comparison, estimand, or biological interpretation. Resource consumption therefore measured how much work an agent performed, but not whether that work produced the correct scientific decision.

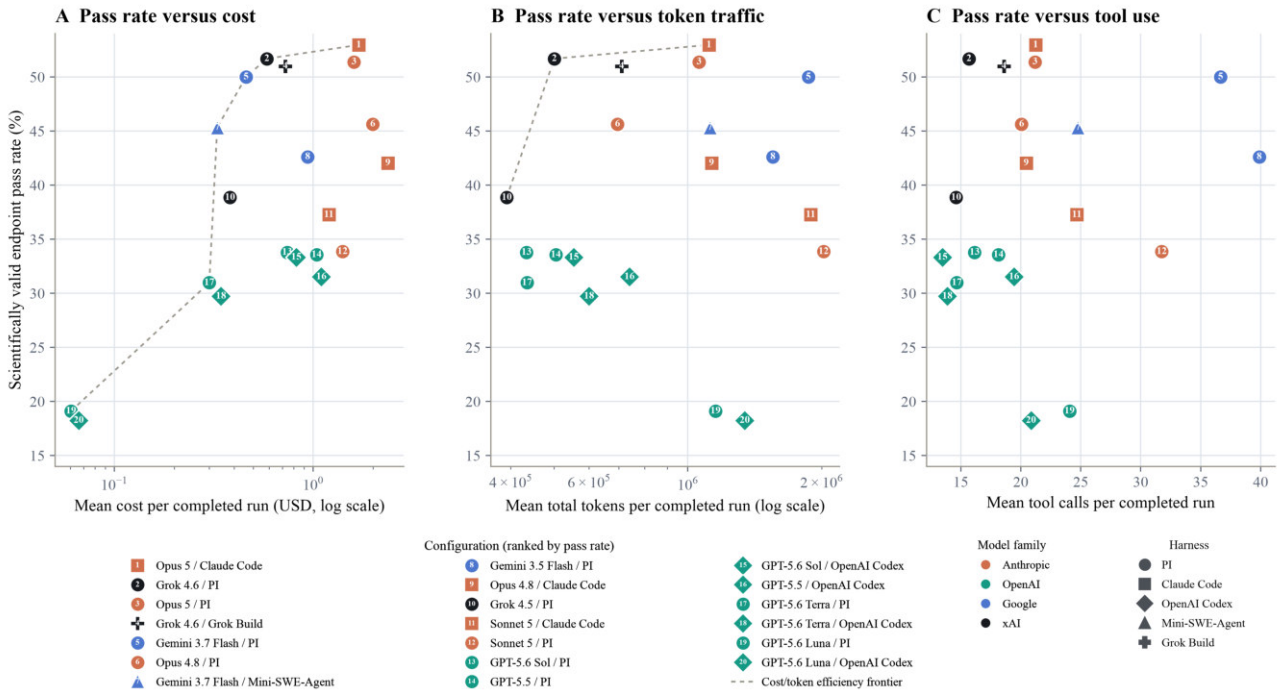

Figure 4: Benchmark performance relative to computational cost, token usage, and tool utilization. **A**, Non-refusal endpoint pass rate plotted against the mean monetary cost per completed run. **B**, Pass rate plotted against the mean total token usage per completed run. The cost and token axes are on logarithmic scales; dashed lines connect configurations along the corresponding accuracy–resource frontier. **C**, Pass rate plotted against the mean number of tool calls per completed run. Each data point represents a unique model–harness configuration. Point numbers indicate the configuration's rank by pass rate; colors denote the model family; and marker shapes signify the execution harness. We calculated pass rates across 100 evaluations, with three attempts per evaluation. We excluded model refusals from the analysis, while scoring execution errors as unsuccessful attempts. We averaged resource metrics over completed, non-refused runs that yielded recoverable usage records.

Scientific judgment, rather than analytical execution, dominates agent failures

Trajectory analyses indicated that performance depended less on whether an agent could execute an analysis than on whether it selected the scientifically appropriate analysis and interpreted its results within the intended decision context. We assigned each evaluation a primary failure class based on its task, grading rationale, and reviewed failure route. We then summarized the 3,517 evaluable incorrect attempts associated with these classes (Figure 5). Only 17.7% of failures arose primarily from selecting or implementing the wrong computational method, metric, model, or aggregation strategy. The remaining 82.3% of failures pertained to drug-discovery interpretation: defining the wrong population or endpoint, extrapolating evidence beyond the context in which it was measured, overlooking an assay-validity constraint, drawing an unsupported causal conclusion, or optimizing the wrong development objective. Agents frequently extracted the relevant records and performed internally consistent calculations, yet they still failed because those calculations answered an adjacent but scientifically distinct question.

Errors related to scope, denominators, or estimands constituted the largest class, accounting for 34.7% of evaluable failures across 36 evaluations. These errors included pooling data across biological compartments that required separate treatment, using an incorrect experimental unit, treating censored or unevaluable measurements as negative results, and substituting an observed association for the decision-relevant contrast. Metric, model, or aggregation errors contributed an additional 17.7%. Common patterns included averaging across sites when reproducibility was the governing criterion, combining correlated measurements as independent evidence, and extrapolating an assay-specific avidity gain directly to a tissue prediction. Cross-context portability issues accounted for 15.6% of failures, arising when agents transferred rankings across species, molecular formats, receptor densities, assay systems, doses, or indications without verifying whether the underlying biological mechanism was preserved. The remaining failures stemmed from flawed evidence integration or an incorrect decision objective (11.5%), causal or biological overreach (10.6%), and the failure to apply an assay-validity, quality-control, or uncertainty veto (9.9%). In these instances, agents often recognized specific limitations but failed to let them influence the final advancement decision.

GPT-5.6 Sol illustrated this dichotomy particularly well. Manual review demonstrated that Sol frequently accessed the correct files, reconstructed the relevant data subsets, and reproduced the same intermediate ratios, contrasts, or rankings as the leading models. In several failed attempts, the model even explicitly identified the caveat that should have governed the final decision. The errors emerged downstream: Sol advanced a strong face-value signal despite an assay-validity concern, transferred an effect across receptor densities or ex-

perimental contexts, altered the relevant denominator when aggregating results, or introduced a causal claim that the measurements did not substantiate. Its failure profile aligned with this pattern. Specifically, Sol failed 85% of attempts involving cross-context portability, 73% involving an assay-validity or uncertainty veto, 68% involving causal overreach, and 64% involving evidence integration, compared to 59% for metric, model, or aggregation tasks. Therefore, Sol's deficit was not simply an inability to process data or perform calculations. While it frequently completed the requisite analytical work, it applied a more permissive decision policy than the stronger models, prioritizing the largest observed effect over the evidentiary boundary necessary for a defensible biologics decision. This behavior also explains why its failures remained reproducible across two execution harnesses and were not ameliorated by additional tokens or tool calls.

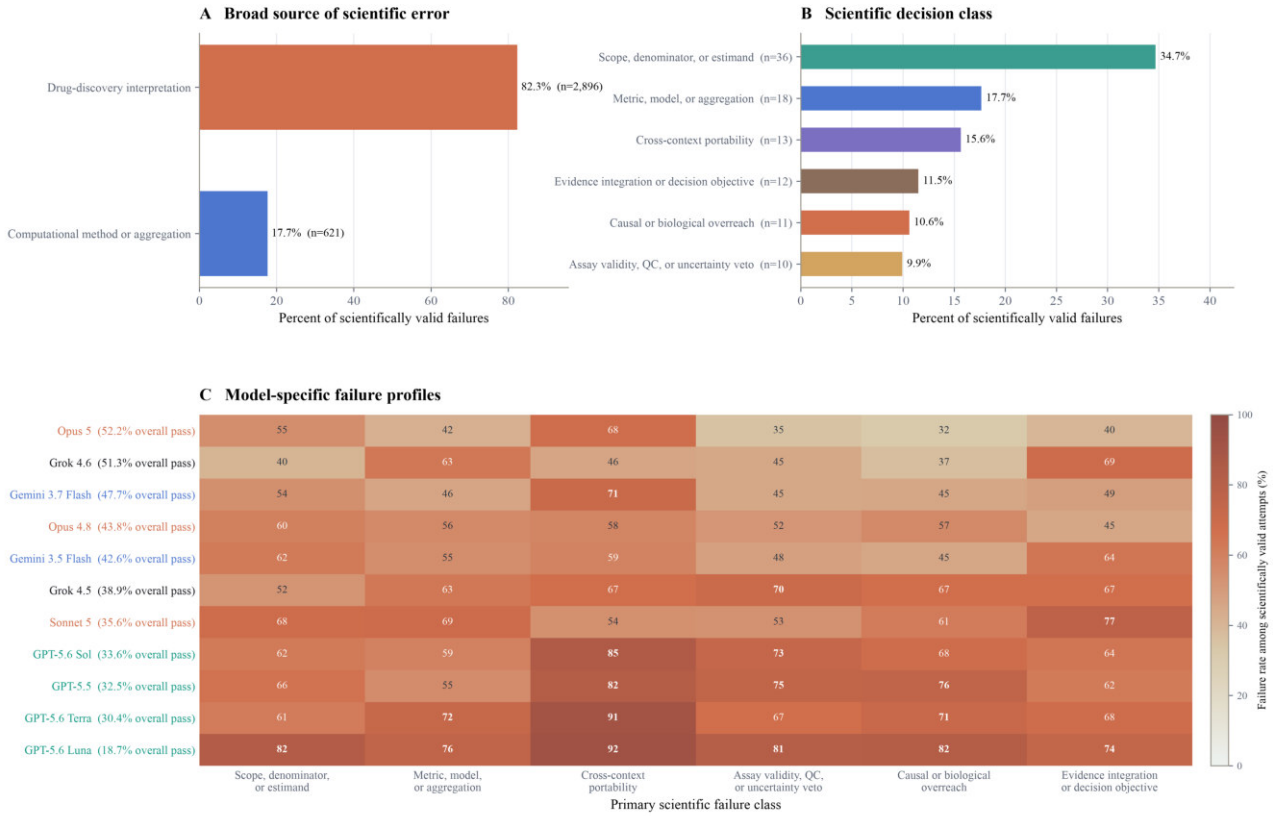

Figure 5: Errors in scientific judgment dominate agent failures. A, Broad sources of error among 3,517 evaluable incorrect attempts. We distinguished failures attributed primarily to computational methods, metrics, models, or aggregations from those requiring drug-discovery interpretation, including scope definition, evidence transport, assay-quality adjudication, causal reasoning, and decision integration. **B**, Distribution of incorrect attempts across six primary scientific failure classes. We assigned each evaluation a single primary class based on its task, deterministic grading rationale, author notes, and reviewed failure route. Percentages are weighted by the number of evaluable incorrect attempts, whereas n indicates the number of evaluations assigned to each class. **C**, Model-specific failure rates for each scientific class, pooled across execution harnesses. Rows represent model families and their overall evaluable pass rates, while columns represent the primary failure classes.

Model rankings changed across biologics discovery competencies

We found substantial variation across the ten biologics discovery competencies, and the overall model ranking did not hold consistently across categories (Figure 6). After pooling evaluable attempts across two harnesses per model, assay and screening-funnel design remained difficult for every family: Gemini 3.7 Flash achieved the highest pass rate at 40.5%, compared with 26.2% for both Opus 5 and Grok 4.6 and 21.4% for GPT-5.6 Sol. These evaluations required agents to determine whether an assay supported the requested conclusion, rather than merely extract its largest effect. Examples included recognizing when a dose-response experiment did not establish a reliable potency estimate, separating target-dependent enrichment from clone growth or amplification during display, and refusing to collapse assays with different biological meanings into one candidate ranking. Binding, kinetics, avidity, and specificity produced a wider separation: Grok 4.6 reached 59.0%, followed by Opus at 44.1%, Sol at 32.1%, and Gemini at 30.8%. These tasks often required distinguishing intrinsic molecular affinity from bivalent binding on high-density cell surfaces and deciding whether the resulting avidity advantage could be transported to the intended tissue. Discovery-stage candidate de-risking was also challenging but favored Grok, which reached 55.8%, compared with 46.7% for Gemini, 45.4% for Opus, and 40.0% for Sol.

The models showed different areas of relative strength.

Opus performed best on target opportunity and therapeutic-hypothesis decisions (71.3%) and cellular pharmacology and mechanism of action (68.8%). It also led binder generation and primary hit selection at 82.5%, compared with 75.0% for both Grok and Gemini and 12.5% for Sol. These evaluations included deciding whether a tissue-restricted receptor perturbation supported ligand neutralization or receptor blockade and whether functional activity was better explained by antibody geometry, valency, or receptor organization than by affinity alone. Grok showed a different profile, leading antigen, reagent, and model design (87.5%), binding and kinetics (59.0%), discovery engineering and next-cycle design (69.0%), and candidate de-risking (55.8%). Gemini performed best on epitope, escape, and structural-mechanism evaluations at 55.6%, with Sol close behind at 50.0%, while Opus and Grok reached approximately 32% and 31%. Success in this category depended on combining competition experiments, mutational scans, escape profiles, and structural evidence while separating directly observed interactions from modeled assemblies or mechanisms.

Sol's strongest category was sequence, lineage, and enrichment analysis, where it reached 56.7%, compared with 46.7% for Gemini and 43.3% for both Opus and Grok. It was also competitive in discovery engineering and next-cycle design at 64.3%, behind Grok at 69.0% and Gemini at 66.7% but ahead of Opus at 52.4%. These tasks involved reconstructing relationships among antibody sequences or screening populations and allocating subsequent experiments across candidate families, molecular formats, or receptor-binding arms. Taken together, the category-level reversals show that biologics discovery performance cannot be represented by a single model hierarchy. Opus was strongest on several biological-hypothesis and functional-mechanism decisions, Grok on multiple design, biophysical, and de-risking categories, Gemini on structural-mechanism interpretation, and Sol on sequence-centered analysis. Harness choice changed absolute performance, but it did not eliminate these broader differences in how model families reasoned across biological evidence types.

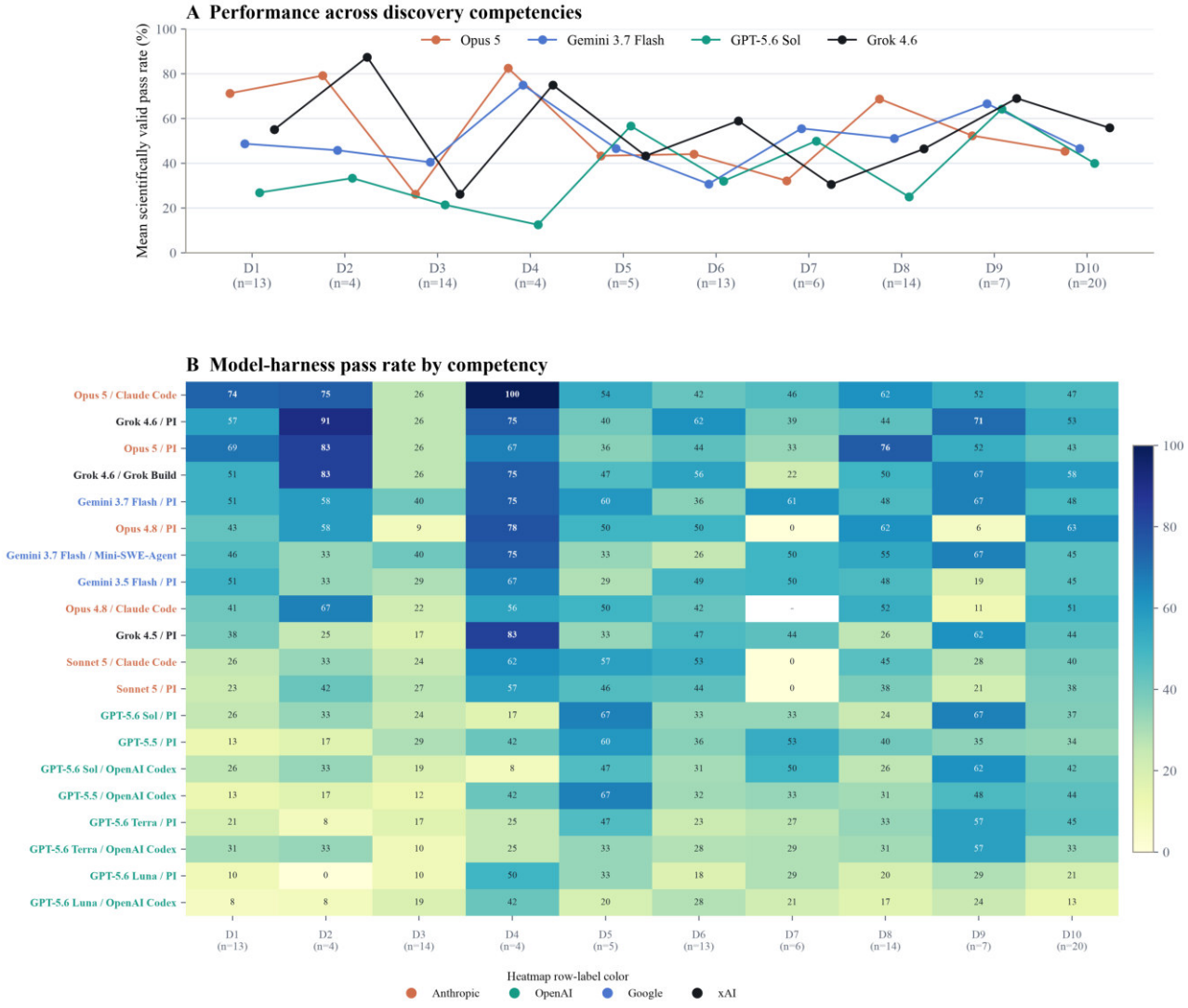

Figure 6: No single model leads across discovery competencies. A, Mean non-refusal pass rate across evaluations assigned to each roadmap category (D1–D10) for Opus 5, Grok 4.6, Gemini 3.7 Flash, and GPT-5.6 Sol. Values for each model combine its two evaluated harnesses and are averaged at the evaluation level. **B**, Non-refusal endpoint pass rate for every model–harness configuration within each category; rows are ordered by overall benchmark performance, and row-label colors indicate model family. Numbers in cells are percentages. Category sample sizes are shown beneath each axis. Refusals were excluded from pass-rate denominators. Each configuration was attempted three times per evaluation.

Performance varied with both decision type and evidence source

We next decomposed performance using two annotations spanning the ten discovery competencies: the type of decision required and the primary evidence used to support it (Figure 7). Quantitative characterization was among the most difficult decision types, with pass rates of 22–33% across the four model families. The challenge was often not arithmetic but identifying the quantity that the experiment could legitimately support. Agents had to distinguish systemic exposure from peak concentration, tissue concentration from a measurement made on bulk lysate, and directly observed assay values from pharmacokinetic quantities that would require extrapolation beyond the available data. Comparison and classification tasks also separated the models: Grok 4.6 reached 48%, compared with 32% for Gemini 3.7 Flash, 31% for Opus 5, and 25% for GPT-5.6 Sol. These decisions often depended on first establishing whether measurements were comparable across assay formats, species, sample matrices, or treatment contexts. Performance was more similar for diagnosis and troubleshooting, where Opus, Gemini, and Sol reached 43–46%, although Grok was lower at 35%.

Tasks requiring evidence integration produced a different ordering. Opus passed 67% of mechanism-inference evaluations and 70% of integrated discovery-gate evaluations. These tasks required agents to connect observations across biological levels, such as determining whether receptor engagement produced downstream function or whether exposure, species biology, assay quality, and pharmacodynamic response jointly supported advancing a candidate. Grok reached 63% on ranking and triage tasks, slightly above Opus and Gemini at 60%, but performed worse on mechanism inference at 49%. Sol reached only 32% on ranking and triage, 33% on mechanism inference, and 36% on integrated discovery gates. Design and optimization decisions were more balanced, ranging from 42% for Sol to 53% for Grok. This decomposition distinguishes evaluations that were difficult because several valid measurements had to be combined into one program decision from those in which the main challenge was establishing whether the comparison itself was scientifically admissible.

Grouping evaluations by their primary evidence source revealed further specialization. Opus reached 60% on expression, single-cell, spatial, and disease-context data and 62% on dose-response, imaging, uptake, and cell-state measurements. Gemini performed best on selection and enrichment screens at 75% and reached 63% on competition, epitope, escape, and structural evidence. Grok showed a marked advantage on binding kinetics, model selection, and avidity, reaching 70% compared with 52% for Opus, 27% for Gemini, and 24% for Sol. Sol was relatively stronger on sequence and lineage analysis (44%), assay quality control and hierarchical inference (44%), and experiment design and allocation (50%), but reached only 20% on dose-response and cellular-state evidence. These results show that benchmark difficulty was determined not only by where a decision occurred in the discovery process but also by the structure of the evidence and the scientific inference that the evidence was expected to support. Similar aggregate pass rates could therefore arise from substantially different capabilities in screening analysis, biophysical interpretation, cellular pharmacology, and evidence integration.

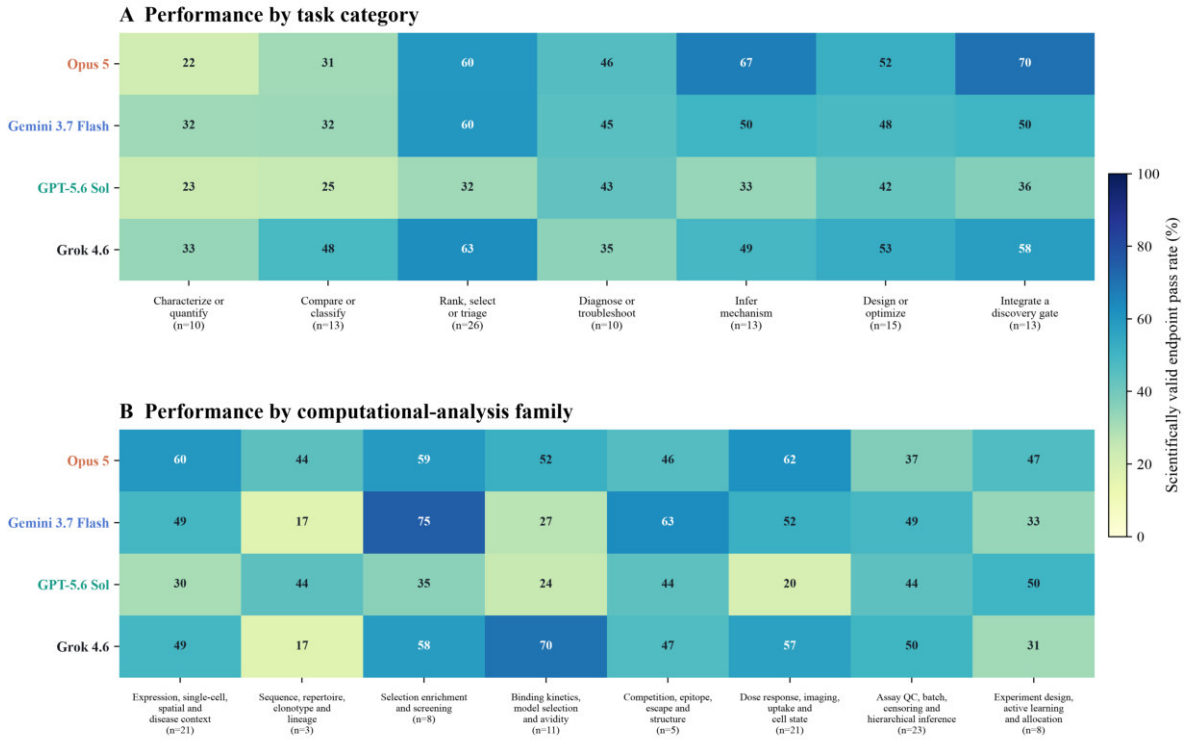

Figure 7: Performance varied with both the scientific decision and the evidence substrate. **A**, Non-refusal pass rates for Opus 5, Grok 4.6, Gemini 3.7 Flash, and GPT-5.6 Sol across seven categories of scientific decision. **B**, Pass rates for the same models across eight categories representing the primary evidence and analysis substrate. Each evaluation was assigned one primary category on each axis. Results for each model are pooled across the two harnesses tested. Cell labels show pass rates, and n indicates the number of evaluations in each category. Each model–harness configuration was run three times per evaluation.

Discussion

Our results show that frontier agents are not yet reliable decision-makers in biologics discovery. The strongest configuration passed 50.3% of attempts, and failures were often reproducible across the same evaluations. Because every evaluation was solved by at least one configuration, and 85 were solved by models from at least two families, the observed ceiling is unlikely to reflect a uniformly impossible subset. Instead, it points to stable, capability-specific gaps. This extends TxBench-PP from small-molecule preclinical pharmacology to biologics [8] and is consistent with ScienceAgentBench, BixBench, and LABBench2, which report limitations in realistic scientific work [9–11]. Strong performance on knowledge, coding, or narrowly specified prediction benchmarks should not be taken as evidence that an agent can reliably recover research decisions.

Model specialization was equally important. Rankings reversed across competencies, decision types, and evidence sources. Similar limitations have emerged in antibody modeling. FLAb2 found that no evaluated model predicted all developability properties or performed consistently across related datasets [6], while AbBiBench emphasized evaluating affinity in the context of the antibody–antigen complex [12]. CHIMERA-Bench tests generalization to unseen epitopes, antigen folds, and temporally held-out targets [13], and the blinded AIntibody challenge found that success in one antibody-design task often did not transfer to another [7]. Our results extend this pattern from molecular prediction and design to experimental reasoning. A model that handles sequence or structural evidence well may still misinterpret avidity, cellular context, species translation, or a translational gate. Aggregate scores should therefore be accompanied by stratified capability profiles.

Trajectory analysis showed that more computation alone is unlikely to close these gaps. Additional tokens and tool calls helped when they resolved the decisive ambiguity, but often propagated an incorrect comparator or estimand. BioDesign-Bench similarly traced much of the protein-design agent gap to shallow candidate evaluation and showed that compute-matched, multi-metric evaluation improved performance [14]. ScienceAgentBench found that greater inference-time computation improved accuracy, but at substantially higher cost and without achieving reliable task completion [9]. These findings motivate harnesses that explicitly check biological scope, comparator validity, contradictory evidence, and decision sufficiency rather than simply allowing longer trajectories. The benchmark remains retrospective, uses public studies, and lacks human baselines and prospective wet-lab outcomes. Future work should include blinded evaluations of unpublished campaigns, expert comparisons, and experimental validation, following AIntibody [7]. Decision-centered benchmarks can complement predictive and generative benchmarks [5–7, 12, 13]

by testing whether a model can determine what the evidence supports and what should happen next.

Methods

Benchmark composition and data

The Antibody Discovery Benchmark comprised 100 evaluations derived from publicly available experimental studies. Each evaluation included an agent-facing task, deterministic graders, benchmark annotations, references to staged input data, and private notes documenting the scientific decision and the ground-truth rationale. The benchmark primarily covers therapeutic antibodies and includes engineered antibody formats and other protein-based modalities.

Input materials included source-derived tables, sequences, images, and assay outputs. Identifiers that could reveal the answer were blinded, while decision-relevant experimental structure, including dose, time, donor, replicate, batch, and molecular format, was retained. Evaluations were annotated with discovery competency, decision type, computational-analysis family, therapeutic modality, evidence type, difficulty, and primary failure mode. Related evaluations from the same research campaign were grouped to reduce information leakage.

Task format and grading

Agents received a scientific decision question, a structured response schema, and access to the staged data. Tasks generally did not prescribe the analytical workflow, as selecting an appropriate comparison, representation, or method was part of the skill being evaluated.

All evaluations used deterministic grading. Categorical outputs were assessed using exact matches or predicates; set-valued outputs were evaluated with precision and recall against canonical identifiers; and numeric outputs were assessed using task-specific tolerances or admissible ranges. A run passed only if all required checks passed. Missing required fields, invalid structured responses, and unparseable answers were counted as incorrect.

We reconstructed ground truths from the same evidence available to the agent. We checked candidate tasks for scientific recoverability, information leakage, robustness to reasonable analytical choices, and separation between valid approaches and plausible but incorrect alternatives.

Agent runs and execution

Each evaluation was run three times across 20 model-harness configurations, yielding 6,000 runs. Each configuration paired a language model with the execution scaffold that provided its tools and interaction protocol. The evaluated harnesses included Claude Code, PI, Mini-SWE-Agent, OpenAI Codex, and Grok Build.

Models were run at their maximum available reasoning-effort

setting in isolated environments without internet access. Agents could inspect and analyze only the files staged for the evaluation. Complete trajectories were retained, including model messages, tool calls, execution outputs, and final responses. Token usage, execution time, and estimated cost were recorded when available.

Outcome classification and aggregation

Each run was classified as correct, incorrect, or refused. A run was correct only if every required grader passed. Gradable answers that failed any required check, along with timeouts and execution errors, were classified as incorrect. Infrastructure-level refusals were detected via provider or harness signals, while model-initiated refusals were identified from the recorded trajectories. Tool denials that did not prevent a gradable final answer were not counted as refusals.

Pass rates were calculated from correct and incorrect runs, with refusals reported separately. Configuration-level performance was averaged across evaluations so that each scientific decision contributed equally. Uncertainty was summarized using 95% Student's t intervals across per-evaluation pass rates. Resource-use summaries included only completed, non-refused runs.

Data availability

The public TxBench-Antibody repository is available at https://github.com/latchbio/. It contains a subset of evaluations and trajectories for manual inspection, along with task definitions, graders, and available data artifacts for those evaluations. Source assay data are distributed or linked according to the original data licenses and access terms.

References

[1] Tushar Jain, Tingwan Sun, Stéphanie Durand, Amy Hall, Nga Rewa Houston, Juergen H Nett, Beth Sharkey, Beata Bobrowicz, Isabelle Caffry, Yao Yu, Yuan Cao, Heather Lynaugh, Michael Brown, Hemanta Baruah, Laura T Gray, Eric M Krauland, Yingda Xu, Maximiliano Vásquez, and K Dane Wittrup. Biophysical properties of the clinical-stage antibody landscape. *Proc. Natl. Acad. Sci. U. S. A.*, 114(5):944–949, 31 January 2017.

[2] Wadim L Matochko, S Cory Li, Sindy K Y Tang, and Ratmir Derda. Prospective identification of parasitic sequences in phage display screens. *Nucleic Acids Res.*, 42(3):1784–1798, February 2014.

[3] Eilyn R Lacy, Rupesh Nanjunda, Scott L Klakamp, Deborah Kwok, Jennifer F Nemeth, Gordon D Powers, H Hugo Caicedo, and Steven A Jacobs. A novel label-free method to determine equilibrium dissociation constants of antibodies binding to cell surface proteins. *Sci. Rep.*, 15(1):3352, 27 January 2025.

[4] Steffen Dickopf, Guy J Georges, and Ulrich Brinkmann. Format and geometries matter: Structure-based design defines the functionality of bispecific antibodies. *Comput. Struct. Biotechnol. J.*, 18:1221–1227, 14 May 2020.

[5] Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv* [*cs.LG*], 18 February 2021.

[6] Michael Chungyoun and Jeffrey Gray. Fitness landscape for antibodies 2: Benchmarking reveals that protein AI models cannot yet consistently predict developability properties. *bioRxivorg*, page 2025.12.27.696706, 27 December 2025.

[7] M Frank Erasmus, Daniel Bedinger, Elizabeth Hopkins, Ginger Ferguson, Justine Strickler, Christilyn P Graff, Samantha R Summers, Stacy L Capehart, Joshua D Slocum, Crystal Richardson, Sumit Kumar, Zhifei Sun, Yujie Shang, Jixian Zhang, Ming Gu, Lixia Yi, Alon Wellner, Shuangjia Zheng, Wei Lu, Pietro Sormanni, Matthew Greenig, Haiping Zhang, Brendan T Mann, Mahdi Baghbanzadeh, Ali Rahnavard, Gregory L Moore, Huaiyu Sun, Ying Ding, Alex Nisthal, Jitendra Kanodia, Matthew J Bernett, Aurélien Pélissier, Yanjun Shao, Maria Rodriguez Martinez, Karthik Ramesh, Horacio Nastri, Andreas Evers, Anhar Abdelatif, Andrew J Bordner, Mykola Bordyuh, Lim Heo, Brian A Kidd, H Serhat Tetikol, Shuai Wei, Jung-Eun Shin, Ryan Peckner, Leigh Manley, Ajitesh Lunge, Yashas Devasurmutt, Bora Guloglu, Liviu Copoiu, Miles McGibbon, Monica L Fernandez-Quintero, Nitesh Mishra, Sean M Callaghan, Olivia M Swanson, Daniel L V Bader, James A Ferguson, Sai S R Raghavan, Benjamin Nemoz, Colleen A Maillie, Charles Bowman, Bryan Briney, Andrew B Ward, Paolo Marcatili, Rahmad Akbar, Bing He, Fandi Wu, Jianhua Yao, Bin Hu, Michal Kucer, Kaetlyn Rose Gibson, Rahul Somasundaram, Li-Wei Hung, Tomasz Kaszuba, Daved H Fremont, Hyeongsun Jeong, Vinodh Babu Kurella, Shipra Malhotra, Satyendra Kumar, Yanyun Liu, Lingling Xu, Joshua Misa, Alexander Nicholas St John, Jeff Vogt, Fátima A Dávila-Hernández, Da Xu, Michael Chungyoun, Zyaja D Huggan, Jeffrey J Gray, Jonathan Parkinson, Young Su Ko, Wei Wang, Franziska Geiger, Jonathon D Ziegler, Nikhil Haas, Chance Challacombe, Ahmad Qamar, Akshita Singh, Yi-Ching Tang, Zhiqiang An, Xiaoqian Jiang, Yejin Kim, Xinyan Zhao, Erik Swanson, Jürgen Klattig, Karsten Winkler, Tschimegma Bataa, Volker Sandig, Lilian Denzler, Chunan Liu, Randall J Brezski, Laura Spector, Katheryn Perea-Schmittle, Sara D'Angelo, Fortunato Ferrara, and Andrew R M Bradbury. A blinded, prospective benchmark of in silico antibody discovery anchored to experimental affinity and developability. *Nat. Biotechnol.*, pages 1–13, 19 August 2026.

[8] Hannah Le, Ramesh Ramasamy, Alex Urrutia, Mahsa Yazdani, Tim Proctor, and Kenny Workman. TxBench-PP: Analyzing AI agent performance on small-molecule preclinical pharmacology. *arXiv* [*cs.AI*], 17 June 2026.

[9] Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. *arXiv* [*cs.CL*], 7 October 2024.

[10] Jon M Laurent, Albert Bou, Michael Pieler, Conor Igoe, Alex Andonian, Siddharth Narayanan, James Braza, Alexandros Sanchez Vassopoulos, Jacob L Steenwyk, Blake Lash, Andrew D White, and Samuel G Rodriques. LABBench2: An improved benchmark for AI systems performing biology research. *arXiv* [*cs.AI*], 4 February 2026.

[11] Ludovico Mitchener, Jon M Laurent, Alex Andonian, Benjamin Tenmann, Siddharth Narayanan, Geemi P Wellawatte, Andrew White, Lorenzo Sani, and Samuel G Rodriques. BixBench: A comprehensive benchmark for LLM-based agents in computational biology. *arXiv* [*q-bio.QM*], 28 February 2025.

[12] Xinyan Zhao, Yi-Ching Tang, Akshita Singh, Victor J Cantu, Kwanho An, Junseok Lee, Adam E Stogsdill, Ibraheem M Hamdi, Ashwin Kumar Ramesh, Zhiqiang An, Xiaoqian Jiang, and Yejin Kim. AbBiBench: A benchmark for antibody binding affinity maturation and design. *arXiv* [*q-bio.QM*], 23 May 2025.

[13] Mansoor Ahmed, Nadeem Taj, Imdad Ullah Khan, Hemanth Venkateswara, and Murray Patterson. CHIMERA-bench: A benchmark dataset for epitope-specific antibody design. *arXiv* [*cs.LG*], 13 March 2026.

[14] Jeonghyeon Kim and Philip Romero. Benchmarking and behavioral characterization of LLM agents for protein design. *bioRxivorg*, page 2026.05.06.723381, 28 June 2026.