

ScreenBench: AI Agents Miss Hidden Errors in Screening-Plate Records

Ward Bradt Leah Damon
Gamow Labs

October 2026

Abstract

We introduce ScreenBench, a benchmark of AI agents’ performance on compound-screening analysis tasks that contain hidden errors. The ScreenBench dataset comprises 182 tasks built from real cell-viability screening plates and their dosing records, into which we plant one to three undisclosed faults. Each task asks for a routine result about compound activity, identity, or measurement location. Answering correctly requires the agent to notice the faults and correct for them. Gemini 4 Argon scored 0.49, Opus 5.5 scored 0.47, and GPT 6.1 Sol scored 0.12. When the task brief mentions the possible fault types, all three agents scored between 0.94 and 0.96.

1 Introduction

In a high-throughput screen, every hit call depends on the plate map: the record of which compound the liquid handler dispensed into which assay well. Automated barcode tracking and transfer logs make errors in that chain rare, but not impossible. A readout filed under the wrong plate barcode or a plate read rotated 180 degrees can still pass quality control because Z' and other control-based statistics [1–3] measure assay performance, not whether each well is matched to the right compound. Hit calling on that data can be arithmetically flawless and still credit activity to the wrong compound, so inactive compounds become hits and real actives disappear. Experienced screening teams guard against this before trusting a plate: they confirm that the controls sit where the plate map puts them and that known actives read out where expected.

We introduce ScreenBench, a collection of 182 compound screening tasks built from real cell-viability screens, to test whether AI agents make the same checks without being told to. Each task gives the agent plate readouts and dosing records and asks a routine question, such as how active each compound is, with one to three undisclosed faults from seven families (Figure 1) planted in the data. Each task also has a *guided* twin version with the same data, question and answer key, plus one paragraph listing the possible fault families. The guided twin method separates an agent’s failure to find the fault from a failure to fix. With the guidance, an oracle ensemble of the three agents (a twin counts as solved if any agent solves it) solved 179 of the 182 guided twins. One author (L.D.), an experienced screening scientist, additionally checked a sample of keys for validity (Section 5).

Agents often miss the faults. We rescale each task’s score so that 0 is the best score of several simple shortcuts, such as copying values from the records, and 1 is a full solve. On this scale Gemini 4 Argon scores 0.49, Opus 5.5 0.47 and GPT 6.1 Sol 0.12. Before rescaling they score 0.66, 0.64 and 0.38, and copying the records scores 0.27. However, Gemini needed up to 3.8 hours and scores 0.30 under the one-hour limit given to the other agents. Faults that disturb a whole plate

are often caught; line swaps, which move one pair of rows or columns, and well-order errors, which leave the controls in place, almost never are. Most wrong answers are the value read straight from the records. Told which fault families to check, all three agents score 0.94–0.96. The answers can be recovered from the data, but the agents rarely check unless asked.

Related work. Benchmarks built on real biological data, such as SpatialBench, scBench and BixBench3, test whether agents can carry out an analysis [4–6]; TxBench-PP and our LabBench (distinct from LAB-Bench [7]) ask agents to separate plate artifacts from real effects [8, 9]. Others plant errors: LABBench2 in protocols [10], GeneBench-Pro in simulated data [11], and BioAgent Bench through corrupted inputs and decoy files [12]. In ScreenBench, the files are well formed and the agent is never told to look. Because the faults are planted, the keys are known by construction, avoiding the broken keys that audits found in BixBench and Humanity’s Last Exam [13, 14]. Statistical methods for systematic plate error [3, 15] assume each well is matched to the right compound.

2 Task definition and construction

Source data. The tasks are built from single-dose compound screens run at Gamow Labs in four non-small-cell lung cancer cell lines. Compounds were dosed into 384-well plates by acoustic transfer (Labcyte Echo) and read with CellTiter-Glo, whose luminescence tracks the number of living cells, so a low signal marks a compound that killed cells or halted their growth. The tasks draw on 467 plates from eight screens (median Z' 0.86); none of the records has been published. Each task holds eight plates from one screen, about 1 MB of text: the readouts, the transfer report with its failed transfers, the source-well map, and each compound’s activity in prior screens of other cell lines. Assay and source plates are relabelled with neutral codes (P01, SP-...).

Example. The records say compound C went into well A01 of plate P01. That well’s signal is 120, and the median compound well on the plate reads 1,000; since viability is a well’s signal as a percentage of that median, C appears to be a strong hit at 12%. But the plate’s readout was stored rotated by 180 degrees: C’s reading is in the opposite corner, well P24, which reads 960, so C is inactive at 96%. Nothing says the plate is rotated, but the data show it: the controls sit on the wrong side, and compounds that killed cells in other screens line up with low signal only once the plate is turned around (Figure 1).

Questions and outputs. Agents answer every entry in a list built from the task’s records, typically about 2,400 compounds, while a hidden subset of 8–12 items is scored (2,150 items overall). The three question types are:

- *Activity* (164 tasks): the compound well’s raw signal as a percentage of the median raw signal of valid compound wells on its plate, or `not_dosed` / `not_measurable`.
- *Locate* (13): the readout label and well that measured the compound, or `not_dosed`.
- *Identify* (5): what produced the signal in each measured well: a catalog ID, `control`, `none`, or `artifact`. Its scored wells are low-signal wells that a screener would follow up.

The brief defines the normalization, exclusions, answer schema and tolerance (Appendix C).

Faults and versions. Each task carries one to three faults from seven families (Table 1; Figure 1). We scale each fault to resemble a real failure in these screens rather than to stand out; Table 1 gives the scale for each family. Edge gradients are decoys: the brief says to report their wells’ raw values.

Before including a task, we check that each planted fault is detectable from the data, that every shortcut answer scores below 0.5 before correction (Appendix B) and that no file reveals the answer; remaining limitations are in Section 5.

The *silent* task names no fault. Its *guided* twin adds a paragraph listing the six families that move or destroy signal and asking the agent to check the readouts. Both versions use identical items and keys.

Table 1. Fault families, how strongly we plant each, and task counts. A task can contain several families.

Family	What the readout shows	Strength	Tasks
Line swap	Two rows or columns are interchanged	Lines differ by a set margin over plate noise	100
Plate mix-up	Two readouts are filed under each other’s labels	Whole plate	73
Rotation	Readout is rotated 180° relative to the dosing layout	Whole plate	63
Failed region	A region retains only part of its signal	5–60% of signal; scatter matched to real failed regions	50
Well order	Compound wells follow row-major instead of column-major order, or the reverse	Whole plate	47
Edge effect	Decoy gradient	Multiple of a measured edge effect	34
Source exchange	Two source plates are logged under each other’s barcodes	Whole source plate	18

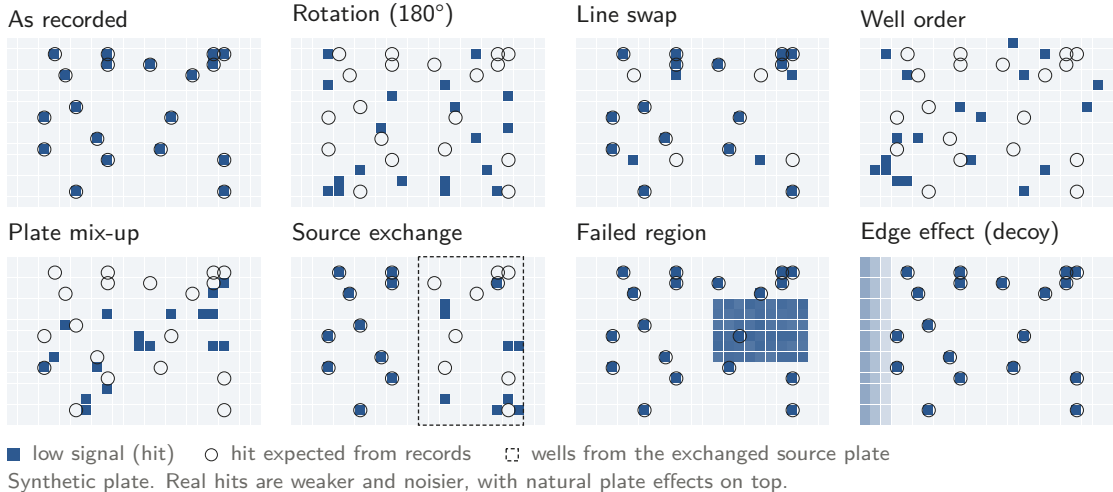


Figure 1. The seven fault families on a synthetic 16×24 plate. Rings mark wells whose recorded compounds are expected to give low signal; fill is the measured signal. In six families the readout’s low wells stop lining up with the rings. In a source exchange, the dashed block of wells was dosed from another source plate, so its low wells follow that plate’s compounds. The edge effect is a decoy that changes no answer. In real tasks we plant the faults into real readouts, at the strengths given in Table 1.

Scoring. An item is correct when a percentage is within 3 points of the key, when a label (a catalog ID or a category such as `not_dosed`) matches the key exactly, or, for Locate, when it names the

right plate and well. The raw weighted score is

$$r_{\text{raw}} = \frac{\sum_i w_i \mathbb{1}[i \text{ correct}]}{\sum_i w_i},$$

where w_i is the weight of scored item i . Clean and not-dosed items together carry at most one third of the weight and faulted items share the rest (Appendix B), so detecting one fault earns partial credit. Because answers that skip the analysis still earn raw credit, the task score (the grader’s reward) is

$$r = \max\left(0, \frac{r_{\text{raw}} - b}{1 - b}\right),$$

where b is the best raw score of any answer in a fixed, pre-specified list of shortcuts, computed per task from the key alone (Appendix B). *Solved* means every hidden graded item is correct, not that the entire requested answer table is correct.

Cohorts. The tasks were built in three successive batches, which we call cohorts (Table 2). For the first, we ran GPT 6.1 Sol four times at high effort on each version of each candidate and kept silent tasks whose mean score was below 0.5 (0.34 on average), with at most 15 tasks per screen. This biases GPT 6.1 Sol’s selected-task scores downward relative to agents not used in selection. Reported scores come from new attempts, not the selection runs. The later cohorts were not selected by model score, but emphasize faults missed in earlier ones.

Table 2. Task cohorts. No task was removed based on its evaluation score.

Cohort	Tasks	Inclusion rule
First	95	Selected from 139 eligible tasks using GPT 6.1 Sol mean score below 0.5.
Second	69	All tasks that passed the inclusion checks above; Activity only, at least two faults, emphasizing line swaps and well order.
Third	18	Chosen from 38 candidates before evaluation; source exchange plus one or two other faults.

3 Evaluation protocol

Each agent ran once on the silent and guided versions of every task between 23 and 25 October 2026: 364 attempts per agent and 1,092 in all. The runs covered 201 candidate tasks; 19 on which a single number given for every compound scores at least 0.5 were then excluded, using the answer keys alone (Appendix B). Every attempt ran in its own container with 2 CPUs, 8 GB of memory and Python 3.11 with common scientific packages. The agent could read the task files and write to a workspace, could reach only its vendor’s API, and could not search the web or install packages. The prompt (Appendix D) asked it to work without asking questions and to write its answers to `answer.json`. Table 3 lists the agents; each pairs a model with its own harness, so we cannot separate the two.

A missing or invalid answer scored zero, which happened in two Gemini attempts, and no attempt reached its time limit. The answer keys never left our machines; we graded every attempt after the runs finished.

4 Results

Opus and Gemini score about the same. Opus and Gemini score within 0.02 of each other, and they lead GPT 6.1 Sol by 0.35 and 0.37. Opus solves 58 silent tasks, Gemini 56 and GPT 6.1 Sol

Table 3. Evaluated agents. Unlisted settings use harness defaults.

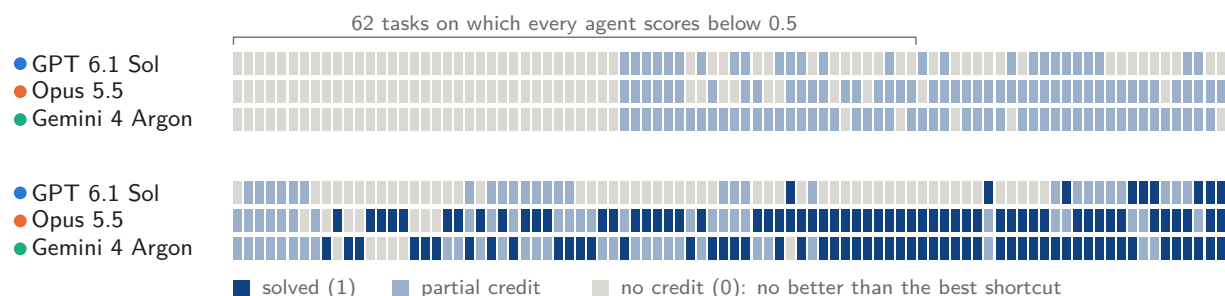
Agent	Model, harness and effort	Time limit
GPT 6.1 Sol	GPT 6.1 Sol; Codex CLI 0.164.0, high effort	1 h
Opus	Opus 5.5; Claude Code 2.1.309–2.1.310, high effort	1 h
Gemini	Gemini 4 Argon-high via Antigravity, version unrecorded (310 attempts), or Argon via Gemini CLI 0.63.0 (54 first-cohort attempts)	6 h

Table 4. Results on the 182 tasks in each version. Score is the mean chance-corrected task score (0 = best shortcut, 1 = solved); Raw is the mean weighted score before correction. Time is median elapsed minutes and cost is estimated mean API dollars per silent attempt. GPT 6.1 Sol participated in selection of 95 of these tasks (Table 2). Gemini ran under a 6 h limit; the capped row scores its attempts longer than 1 h as zero. The best-shortcut row takes each task’s best shortcut answer.

Agent	Silent			Guided			Silent attempt	
	Score	Raw	Solved	Score	Raw	Solved	Min	USD
● Gemini	0.49	0.66	31%	0.96	0.97	84%	53.8	5.43
● Opus	0.47	0.64	32%	0.96	0.98	93%	3.4	0.90
● GPT 6.1 Sol	0.12	0.38	5%	0.94	0.96	90%	2.3	0.12
● Gemini, 1 h cap	0.30	0.39	20%	0.58	0.58	54%	–	–
Record lookup	0.00	0.27	0%	0.00	0.27	0%	–	–
Best shortcut	0.00	0.34	0%	0.00	0.34	0%	–	–

10. Sixty-two tasks leave all three below 0.5 (Figure 2). They solve partly different tasks: Opus solves 25 that Gemini does not, and Gemini solves 23 that Opus does not.

Gemini takes far longer. Gemini’s median silent attempt takes 53.8 minutes, compared with 3.4 for Opus and 2.3 for GPT 6.1 Sol (Figure 3). Most Gemini attempts ran on a subscription, so its cost is estimated from logged tokens at API list prices: \$5.43 per silent attempt, six times Opus and 44 times GPT 6.1 Sol. Estimates include cached tokens; rates are in Appendix B. Gemini ran under a six-hour limit, and 80 of its 182 silent attempts (74 guided) ran longer than the one hour given to GPT 6.1 Sol and Opus. Scoring those attempts as zero gives Gemini 0.30 silent and 0.58 guided. This bound is pessimistic: Gemini might work faster under a shorter limit, and its elapsed time may include subscription rate limiting.

**Figure 2.** Per-task results without guidance. Each column is one task and each cell one silent attempt. Columns are sorted by the best score any agent reached, hardest first.

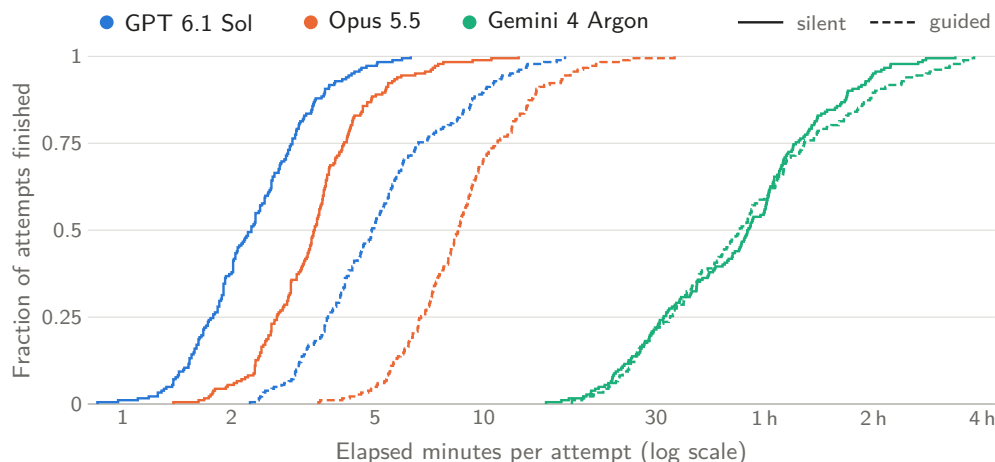


Figure 3. Elapsed wall-clock time per attempt on all 182 tasks, as the fraction of attempts finished by each time. Each agent ran in its own harness; GPT 6.1 Sol and Opus had a 1 h limit and Gemini 6 h.

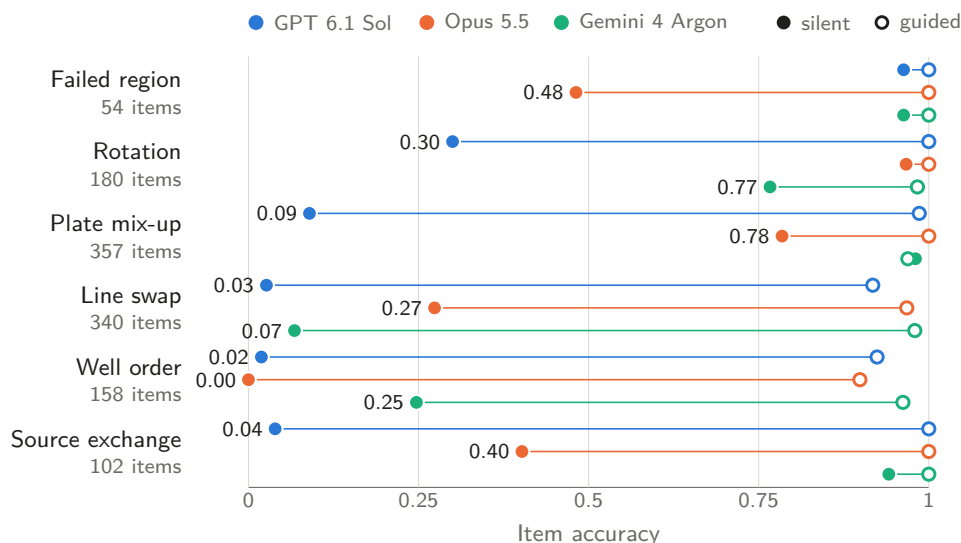


Figure 4. Item accuracy by fault family, pooled over question types. An item is one graded answer (a compound's value or location, or a well's source), grouped by the fault that changes it. Filled markers: silent; rings: guided.

The gap persists on the 87 tasks not selected by score. On those tasks, Opus scores 0.45, Gemini 0.44 and GPT 6.1 Sol 0.13 (raw 0.63, 0.63 and 0.37). These tasks remain adversarial by design (Table 2); their fault mix does not estimate how often these faults occur in real laboratories.

Line swaps and well-order faults are rarely caught. Agents answer 98–100% of items with no fault correctly, but accuracy on faulted items differs sharply by family (Figure 4; exact values in Appendix A). Opus resolves rotations at 0.97 accuracy; Gemini resolves plate mix-ups at 0.98 and source exchanges at 0.94. Neither line-swap nor well-order accuracy exceeds 0.27 for any agent. The two are missed for different reasons. A line swap moves two of 16 rows or 24 columns, so a whole-plate comparison with the records still fits well. A well-order fault scrambles nearly every compound well but leaves the controls in place, so the plate passes a control check and only looks weakly related to prior screens. In ST-104 Opus attributed that pattern to biology (Section 6).

Explicit guidance raises all three scores. Naming the fault families raises mean scores to 0.94–0.96. The paired gains are 0.49 for Opus, 0.47 for Gemini and 0.81 for GPT 6.1 Sol. The passage supplies both an instruction to check and candidate explanations; their effects are not isolated. Guided attempts also take longer for Opus (8.5 median minutes) and GPT 6.1 Sol (4.9), while Gemini’s take about the same time as its silent attempts (51.0 minutes).

Agents mention faults and still return lookup values. A keyword search of the agents’ prose finds possible-fault mentions in 167 Opus, 155 Gemini and 59 GPT 6.1 Sol silent attempts. This is only an upper bound on noticing a fault. Among wrong answers on faulted items, 90% of Opus’s, 79% of GPT 6.1 Sol’s and 58% of Gemini’s match the record-lookup value.

For failed regions, GPT 6.1 Sol and Gemini each mark 34 of 35 affected items `not_measurable`; Opus marks 14. Opus returns numeric values in eight tasks; in sampled transcripts it describes an artifact but invokes the brief’s ban on correction. Twenty of those 21 values fall below a typical hit threshold (about 71% of the plate median). Accepting these numeric answers would raise Opus’s mean only from 0.47 to 0.49 (raw 0.64 to 0.65) and solve two additional tasks.

5 Validity and limitations

Checking the answer keys. Planting a fault gives a known answer only if the underlying records are correct. Agent attempts on candidate tasks exposed one real labelling shift that a key had missed. We therefore compare each original readout with the records of every usable plate in its screen, exclude readouts that fit another plate better, and leave affected items ungraded. At least one agent solves 179 guided twins and 180 tasks on either version. The remaining two were solved in earlier runs, and those answers are still correct under the current keys.

A PhD-trained screening scientist and author (L.D.) checked three keys: one failed-region task and two well-order tasks. She disputed no item and rated her agreement with the keys 4–5 out of 5. She rated the realism of the planted well-order fault 3 out of 5. A real sort-order error appears as a mismatch between the library map and the transfer log, whereas ours makes the transfer report disagree with the plate. This is a small key review, not a human performance baseline. She noted that a laboratory would confirm a well-order mismatch against the original picklist and re-test if records could not be reconciled.

Shortcuts. The correction subtracts the best listed shortcut. The list covers answers that skip the analysis entirely, but a partial analysis or a shortcut we did not anticipate can still earn credit.

Scope. The data cover single-dose ATP-viability assays at one organization. The mix of faults is ours, not an estimate of how often they occur; our screening reviewer describes these faults as rare in automated, barcode-tracked laboratories. Scores are one-attempt measurements of the configurations in Table 3, not estimates of real-world reliability. Gemini’s attempts mix two harnesses and effort settings, so its scores describe that mix rather than one configuration.

The brief’s no-correction clause is ambiguous for failed regions, and five Activity tasks list multiply dosed compounds without an aggregation rule; those compounds are ungraded. The held-out data limit independent reproduction.

6 Discussion

Agents often did the arithmetic correctly on the wrong plate or well. Three cases show how (task IDs in Appendix E).

Default lookup (ST-086). GPT 6.1 Sol joined the records and normalized the readouts without checking that they matched. It scored 0 (0.33 raw), no better than the record lookup. Opus compared each readout against every plate’s records, found the exchanged pair and solved the task.

Evidence explained away (ST-104). Opus found that prior hits did not line up with low signal. It tested rotation, then concluded that the prior screen *only weakly predicts these results*. It never tried a different well order and scored 0 (0.33 raw).

Wrong remedy (ST-038). Opus found the plate mix-up and noticed controls in row N instead of row H. It concluded that controls had been dispensed over compounds and marked those compounds unmeasurable. The rows had actually been swapped. It scored 0.38 (0.58 raw).

Given the fault checklist, the same agents solve 84–93% of tasks, so the data suffice and the corrections are within their reach. A harder version would add screens whose disagreement with prior screens is real biology, so that an agent must decide whether a mismatch is a record error rather than assume it is.

Data and task availability

Tasks, keys and source records are held out. This paper includes a redacted brief, synthetic illustrations and per-task scores, but no complete proprietary task. Evaluation-access inquiries can be directed to Gamow Labs.

References

- [1] Ji-Hu Zhang, Thomas D. Y. Chung, and Kevin R. Oldenburg. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *Journal of Biomolecular Screening*, 4(2):67–73, 1999.
- [2] Christine Brideau, Bert Gunter, Bill Pikounis, and Andy Liaw. Improved statistical methods for hit selection in high-throughput screening. *Journal of Biomolecular Screening*, 8(6):634–647, 2003.
- [3] Nathalie Malo, James A. Hanley, Sonia Cerquozzi, Jerry Pelletier, and Robert Nadon. Statistical practice in high-throughput screening data analysis. *Nature Biotechnology*, 24(2):167–175, 2006.
- [4] Kenny Workman, Zhen Yang, Harihara Muralidharan, and Hannah Le. SpatialBench: Can agents analyze real-world spatial biology data?, 2025. URL <https://arxiv.org/abs/2512.21907>.
- [5] Kenny Workman, Zhen Yang, Harihara Muralidharan, Aidan Abdulali, and Hannah Le. scBench: Evaluating AI agents on single-cell RNA-seq analysis, 2026. URL <https://arxiv.org/abs/2602.09063>.
- [6] Zane Koch, Asmamaw T. Wassie, Javier Valdes-Aleman, Jason Lee, Michaela M. Hinks, Samuel G. Rodrigues, Andrew D. White, and Jon M. Laurent. BixBench3: Benchmarking AI agents on research-study-scale computational biology tasks, 2026. URL <https://arxiv.org/abs/2608.25286>.
- [7] Jon M. Laurent, Joseph D. Janizek, Michael Ruzo, Michaela M. Hinks, Michael J. Hammerling, Siddharth Narayanan, Manvitha Ponnappati, Andrew D. White, and Samuel G. Rodrigues. LAB-Bench: Measuring capabilities of language models for biology research, 2024. URL <https://arxiv.org/abs/2407.10362>.

- [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, 2026. URL <https://arxiv.org/abs/2606.19245>.
- [9] Ben Reilly, Leah Damon, and Daniel McKinnon. LabBench: Benchmarking AI wet lab experimental design and decision-making for AI agents. Gamow Labs preprint, September 2026. URL <https://gamowlabs.com/assets/labbench-preprint.pdf>. Blog: <https://gamowlabs.com/labbench-benchmarking-ai-wet-lab-decisions>.
- [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, 2026. URL <https://arxiv.org/abs/2604.09554>.
- [11] J. H. Li, S. Shringarpure, E. Wong, and A. J. Ho. GeneBench-Pro: Evaluating multistage statistical reasoning in genomics, quantitative biology, and translational biomedicine. *bioRxiv*, 2026. doi: 10.64898/2026.06.29.735386. URL <https://www.biorxiv.org/content/10.64898/2026.06.29.735386>. v4, 21 July 2026; public package: <https://huggingface.co/datasets/openai/genebench-pro-public-package>.
- [12] Dionizije Fa, Marko Culjak, Bruno Pandza, and Mateo Cupic. BioAgent Bench: An AI agent evaluation suite for bioinformatics, 2026. URL <https://arxiv.org/abs/2601.21800>.
- [13] Laurence Wroe. Auditing BixBench: Broken tasks and faulty judging hide saturation. Generality Labs blog, September 2026. URL <https://generality.org/blog/posts/bixbench-broken-ceiling/>.
- [14] Michael Skarlinski, Jon Laurent, Albert Bou, and Andrew White. About 30% of Humanity’s Last Exam chemistry/biology answers are likely wrong. FutureHouse research blog, July 2025. URL <https://www.futurehouse.org/research/hle-exam>. Posted 23 July 2025, updated 16 September 2025. Data: <https://huggingface.co/datasets/futurehouse/hle-gold-bio-chem>.
- [15] Vladimir Makarenkov, Pablo Zentilli, Dmytro Kevorkov, Andrei Gagarin, Nathalie Malo, and Robert Nadon. An efficient method for the detection and elimination of systematic error in high-throughput screening. *Bioinformatics*, 23(13):1648–1657, 2007.

A Additional results

The 2,150 graded items comprise 904 clean, 1,156 whose readout moved or whose plate median changed, 35 unmeasurable and 55 not-dosed items. The 182 tasks contain 17 single-fault, 80 two-fault and 85 three-fault tasks. The edge-effect decoy counts as a fault; a plate mix-up or source exchange counts as one fault.

Table 5. Item accuracy by fault family, silent version (guided in parentheses), pooled over question types. Items per family: mix-up 357, rotation 180, line swap 340, well order 158, failed region 54, source exchange 102.

Agent	Failed region	Rotation	Mix-up	Line swap	Well order	Source exch.
Gemini	0.96 (1.00)	0.77 (0.98)	0.98 (0.97)	0.07 (0.98)	0.25 (0.96)	0.94 (1.00)
Opus	0.48 (1.00)	0.97 (1.00)	0.78 (1.00)	0.27 (0.97)	0.00 (0.90)	0.40 (1.00)
GPT 6.1 Sol	0.96 (1.00)	0.30 (1.00)	0.09 (0.99)	0.03 (0.92)	0.02 (0.92)	0.04 (1.00)

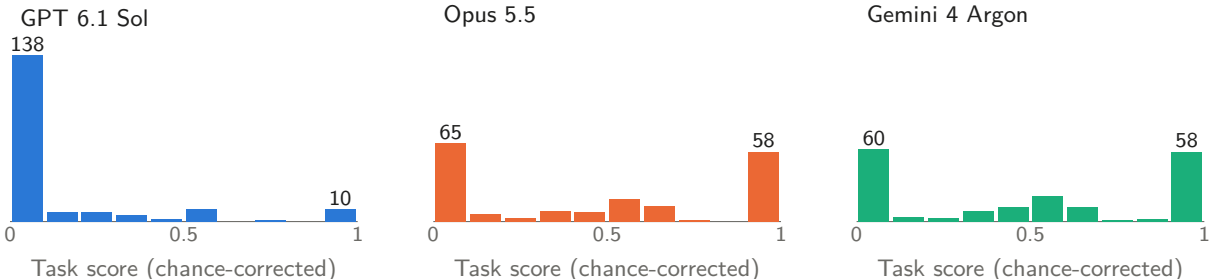


Figure 5. Per-task chance-corrected score on the 182 silent tasks. Multi-fault tasks earn intermediate partial credit when an agent catches one fault and misses another; tasks where no fault is caught score at or near 0.

B Evaluation and audit details

Item selection. Scored items avoid the boundaries of failed regions and the four lines nearest an edge gradient’s side. An edge gradient does not make a well invalid. When an ambiguous exclusion in the brief would shift an item’s value by more than the tolerance, we score the item only if the brief’s literal and intended readings agree.

Item weights. In the first cohort every faulted item carries equal weight; in the other two, each fault carries equal weight, split among its items.

Shortcut list and baseline. Each task’s baseline b is the highest raw score of any answer in a fixed list of shortcuts, graded by the same rule as an agent’s answer (Table 6): an empty answer; the record lookup, which reads every value straight from the records; each compound’s prior-screen value; one label (`not_dosed`, `not_measurable` or `control`) for every item; and one number for every compound. The list uses only the key and the task’s own records, never an agent’s answers. The number is swept over every integer from 0 to 150% and each key value ± 3 . This covers every possible constant, because the best ± 3 window can always slide until one edge meets a key value. The constant sets or ties the baseline on 127 tasks and the record lookup on 77. After correction, every listed shortcut scores 0. A task is kept only if every listed shortcut scores below 0.5 raw in both versions.

Table 6. Raw scores of the listed shortcut answers on the 182 tasks. Numeric answers apply to the 164 Activity tasks. The baseline is each task’s best listed shortcut; after correction every row scores 0.

Answer	Tasks	Mean raw	Max raw
Empty answer (abstention)	182	0.00	0.00
Record lookup, no checks	182	0.27	0.33
Prior-screen value for every compound	164	0.12	0.46
Best single label for every item	182	0.04	0.38
Every compound at 100%	164	0.14	0.41
Best single number for every compound	164	0.33	0.49
Baseline <i>b</i> (best of the above)	182	0.34	0.49

Agents answer every compound in the screen, or every measured well for Identify, and only a hidden subset is scored, so the list itself does not reveal which plates or wells are faulted.

Anonymization and versions. Agents see only the task files; answer keys are stored separately. We checked all 364 distributed task versions for staff names, instrument serials, barcodes, reader timestamps, calendar dates, source file names, file modification times, unexpected files and text from the keys, and found none. Distributed copies replace the calendar dates in 69 second-cohort transfer reports with day offsets, and the cell line named in each brief with a code; agents were evaluated on the earlier versions. Public compounds keep their catalog IDs and structures.

Cost accounting. API costs use logged input, cached and output tokens at the evaluation’s list prices per million tokens: Opus, \$4 input, \$8 cache write, \$0.20 cache read and \$20 output; GPT 6.1 Sol, \$1 input, \$0.10 cache read and \$8 output; Gemini, \$2 input, \$0.10 cache read and \$10 output, including thinking tokens. The 364 attempts cost approximately \$560 for Opus, \$65 for GPT 6.1 Sol and \$2,160 for Gemini. Gemini’s figure is an estimate: 310 of its attempts ran on a subscription, which bills no tokens, and they are priced as if run through the API. Table 4 reports mean silent-attempt API estimates and median elapsed times, not hardware-normalized compute.

C Example brief

The silent activity brief as shipped, with the cell line and library vendor replaced by generic text (the compound list is in `questions.json` and is not reproduced):

Compound read-outs. A cell-viability primary screen (luminescent ATP readout, [cell line] cells, one dose per compound) of a commercial compound library; a drop in signal is a hit. `echo/transfer_report.csv` is the Echo transfer report for the assay plates, whose barcodes have been replaced by codes (P01, P02, ...); source plates carry neutral codes (SP-...) and failed transfers are in its [EXCEPTIONS] section. `library/library_map.csv` maps source plate wells to catalog IDs; control wells come from a control plate and their sample names say what they are (an unnamed control well is DMSO vehicle). `reference/prior_screen.csv` gives each compound’s viability as % of vehicle in sibling screens of the same compounds in other cells. `reads/` holds the plate reads.

`questions.json` lists the compounds of this screen by catalog ID. Answer every one: a subset of them, not shown, is scored. For each one, report its viability in this screen: the raw read of the assay well it was dosed into, as a percentage of the median raw read of all compound wells on the same assay plate. Use the raw reads (no positional or other correction), and leave out of that median every compound well of the plate whose read is not a valid measurement, whether or not it is listed. Answer "not_dosed" if the compound never reached an assay plate,

and "not_measurable" only if the read of its own well is not a valid measurement of it (edge effects and poor controls elsewhere on the plate are not reasons).

Write `answer.json` as `{"answers": {"<catalog ID>": <percentage> | "not_dosed" | "not_measurable"}}`. A percentage counts when it is within 3 points of the true value.

The guided twin adds a paragraph naming the six fault families that move or destroy signal (Table 1) and asking the agent to check each readout against its records. It names no plate, well or answer.

D Agent prompt

You are solving a data-analysis task about a compound screen. Read `/task/brief.md` first; the task's data files are in `/task`, which is read-only. Work in `/work`. You can use Python 3.11 with numpy, pandas, scipy, scikit-learn, statsmodels and matplotlib. There is no internet access and no package installation. Work autonomously and do not ask questions. Run your commands in the foreground and wait for them to finish; do not leave background jobs running. When you are done, write your final answer to `/work/answer.json` exactly as the brief specifies.

E Task inventory

Fault codes: LS line swap, MX plate mix-up, EE edge effect, RO rotation, FR failed region, WO well order, SX source exchange. Each agent's silent and guided chance-corrected task scores come from its single evaluation attempt on that version. Base is the task's baseline b ; a corrected score $r > 0$ is a raw score of $b + r(1 - b)$. Gaps in the IDs are the 19 excluded tasks.

Task	Type	Faults	Items	Base	GPT 6.1 Sol		Opus		Gemini	
					Silent	Guided	Silent	Guided	Silent	Guided
ST-001	Activity	FR+LS+RO	12	0.39	0.00	1.00	0.00	1.00	0.45	0.91
ST-002	Activity	LS+MX	12	0.33	0.00	1.00	1.00	1.00	0.44	1.00
ST-003	Activity	LS	12	0.48	0.00	0.00	0.00	1.00	0.00	1.00
ST-004	Activity	EE+FR+MX	10	0.25	0.11	1.00	1.00	1.00	1.00	1.00
ST-005	Activity	EE+WO	12	0.41	0.00	1.00	0.00	1.00	0.00	1.00
ST-007	Locate	EE+WO	12	0.33	0.00	0.78	0.00	1.00	0.00	1.00
ST-008	Activity	WO	12	0.33	0.00	1.00	0.00	1.00	0.00	1.00
ST-009	Locate	WO	10	0.33	0.00	0.00	0.00	1.00	1.00	1.00
ST-010	Activity	LS+RO	12	0.33	0.00	1.00	0.50	1.00	0.50	1.00
ST-011	Activity	EE+FR+MX	12	0.33	0.13	1.00	1.00	1.00	1.00	1.00
ST-012	Activity	LS+WO	12	0.33	0.00	1.00	0.00	1.00	0.00	1.00
ST-013	Activity	EE+FR+LS	12	0.33	0.00	1.00	0.00	1.00	0.00	0.89
ST-014	Activity	LS	12	0.37	0.00	1.00	0.00	1.00	0.00	1.00
ST-015	Activity	LS+RO+WO	12	0.33	0.25	0.62	0.25	1.00	0.25	1.00
ST-016	Activity	LS+RO	12	0.39	0.00	1.00	0.45	1.00	0.45	1.00
ST-017	Activity	FR+LS	12	0.33	0.50	1.00	0.00	1.00	0.50	1.00
ST-018	Activity	LS+MX	12	0.21	0.05	1.00	0.37	1.00	0.37	1.00
ST-019	Identify	WO	8	0.33	0.00	1.00	0.00	0.00	0.00	0.90
ST-020	Activity	LS+WO	12	0.33	0.00	1.00	0.00	1.00	0.00	0.88
ST-021	Activity	LS+WO	12	0.39	0.00	1.00	0.00	1.00	1.00	1.00
ST-022	Activity	LS+WO	12	0.44	0.00	1.00	0.00	1.00	0.00	1.00
ST-024	Activity	EE+LS	12	0.33	0.00	1.00	0.00	1.00	0.00	1.00
ST-025	Activity	FR+MX+WO	12	0.33	0.33	1.00	0.33	1.00	0.67	1.00
ST-026	Activity	LS+MX	12	0.39	0.00	1.00	1.00	1.00	1.00	1.00
ST-027	Activity	LS+MX	12	0.44	0.00	1.00	0.00	1.00	0.40	1.00
ST-028	Activity	EE+RO	12	0.44	0.00	1.00	0.00	1.00	0.00	1.00
ST-029	Activity	LS+MX	12	0.44	0.00	1.00	0.00	1.00	0.00	1.00
ST-030	Activity	LS+RO	12	0.33	0.00	1.00	1.00	1.00	0.00	1.00

Task	Type	Faults	Items	Base	GPT 6.1 Sol		Opus		Gemini	
					Silent	Guided	Silent	Guided	Silent	Guided
ST-031	Activity	LS+MX	12	0.33	0.00	1.00	0.62	1.00	0.62	1.00
ST-032	Activity	FR+LS+MX	12	0.38	0.10	1.00	0.10	1.00	1.00	1.00
ST-033	Locate	LS	12	0.33	0.00	0.00	0.00	1.00	0.00	1.00
ST-034	Activity	FR+LS	12	0.33	0.50	1.00	0.50	1.00	0.50	1.00
ST-035	Activity	LS+MX	12	0.23	0.03	1.00	0.03	1.00	0.35	1.00
ST-036	Activity	FR+LS	10	0.37	0.00	1.00	0.00	1.00	0.00	1.00
ST-037	Activity	EE+FR+MX	10	0.33	0.00	1.00	1.00	1.00	0.50	1.00
ST-038	Activity	FR+LS+MX	12	0.33	0.13	1.00	0.38	1.00	0.75	1.00
ST-039	Activity	LS+RO	12	0.33	0.50	1.00	0.50	1.00	0.50	1.00
ST-040	Activity	LS+MX+RO	12	0.33	0.00	1.00	1.00	1.00	1.00	1.00
ST-041	Activity	MX+WO	12	0.27	0.00	1.00	0.49	1.00	0.49	1.00
ST-042	Activity	RO	12	0.37	0.00	1.00	1.00	1.00	0.00	1.00
ST-043	Activity	LS+WO	12	0.28	0.08	0.54	0.08	1.00	0.08	1.00
ST-044	Activity	FR+MX+WO	12	0.33	0.00	1.00	0.00	1.00	0.62	1.00
ST-045	Activity	LS+WO	12	0.25	0.00	1.00	0.00	1.00	0.00	1.00
ST-046	Activity	LS+RO	12	0.28	0.54	1.00	0.54	1.00	0.54	1.00
ST-047	Activity	FR+WO	12	0.30	0.05	1.00	0.05	0.68	0.05	1.00
ST-048	Activity	RO	12	0.30	0.05	1.00	1.00	1.00	0.05	0.89
ST-049	Activity	LS+RO	12	0.28	0.08	1.00	1.00	1.00	0.08	1.00
ST-050	Activity	MX+RO+WO	12	0.25	0.11	1.00	0.11	1.00	0.67	1.00
ST-051	Activity	EE+MX+RO	12	0.25	0.00	1.00	0.33	1.00	1.00	1.00
ST-052	Activity	LS+RO	12	0.25	0.00	1.00	1.00	1.00	1.00	1.00
ST-053	Activity	FR+LS+MX	12	0.25	0.22	1.00	0.33	1.00	0.67	1.00
ST-054	Activity	EE+MX	12	0.28	0.08	1.00	1.00	1.00	1.00	0.92
ST-055	Activity	LS+MX	12	0.31	0.00	1.00	1.00	1.00	0.45	1.00
ST-056	Activity	RO	12	0.28	0.08	1.00	1.00	1.00	0.08	1.00
ST-058	Activity	LS	12	0.25	0.00	1.00	0.67	1.00	0.00	1.00
ST-059	Activity	EE+MX	12	0.33	0.00	1.00	1.00	1.00	1.00	1.00
ST-060	Activity	FR+RO+WO	12	0.44	0.00	1.00	0.40	1.00	0.00	1.00
ST-061	Locate	FR+WO	12	0.30	1.00	1.00	0.05	1.00	1.00	1.00
ST-062	Activity	EE+MX	12	0.48	0.00	1.00	1.00	0.93	0.93	0.93
ST-063	Activity	LS+MX	12	0.33	0.00	0.58	0.58	1.00	0.17	0.58
ST-064	Activity	EE+RO	12	0.33	1.00	1.00	1.00	1.00	1.00	1.00
ST-065	Activity	EE+MX	12	0.39	0.00	1.00	1.00	1.00	1.00	0.91
ST-066	Activity	RO	12	0.44	0.00	1.00	1.00	1.00	1.00	0.80
ST-067	Activity	LS+RO	12	0.39	0.00	1.00	1.00	1.00	0.45	1.00
ST-068	Activity	EE+MX	12	0.39	0.00	1.00	1.00	1.00	1.00	1.00
ST-071	Activity	EE+MX	12	0.44	0.00	1.00	1.00	1.00	1.00	0.90
ST-072	Activity	LS+MX+WO	12	0.33	0.00	1.00	0.17	1.00	0.17	1.00
ST-073	Activity	FR+WO	12	0.41	0.00	1.00	0.00	1.00	1.00	1.00
ST-074	Activity	LS+MX+RO	12	0.36	0.00	1.00	0.13	1.00	0.57	1.00
ST-075	Activity	LS+MX+RO	12	0.26	0.00	1.00	0.66	1.00	0.66	1.00
ST-076	Activity	EE+MX	12	0.33	0.00	1.00	1.00	1.00	1.00	0.92
ST-078	Activity	FR+LS+MX	12	0.25	0.22	1.00	0.67	1.00	1.00	1.00
ST-079	Activity	LS+MX+RO	12	0.30	0.00	1.00	0.64	1.00	1.00	1.00
ST-081	Activity	FR+RO	12	0.33	0.00	1.00	0.50	1.00	0.50	1.00
ST-082	Activity	LS+MX+RO	12	0.37	0.00	1.00	0.56	1.00	0.56	1.00
ST-083	Activity	LS	12	0.37	0.00	0.00	0.00	0.00	0.00	1.00
ST-084	Activity	LS+MX	12	0.33	0.00	1.00	1.00	1.00	0.50	1.00
ST-085	Activity	LS+MX	12	0.44	0.00	1.00	1.00	1.00	0.40	1.00
ST-086	Activity	EE+MX	12	0.33	0.00	1.00	1.00	1.00	0.92	0.92
ST-087	Activity	MX	12	0.33	0.00	1.00	1.00	1.00	1.00	1.00
ST-089	Activity	FR+LS+RO	12	0.33	0.33	1.00	1.00	0.88	1.00	1.00
ST-091	Activity	FR+LS	11	0.27	0.25	1.00	0.00	1.00	0.25	1.00
ST-092	Activity	FR+LS+WO	12	0.33	0.00	1.00	0.00	1.00	0.00	1.00
ST-093	Activity	MX	12	0.28	0.08	1.00	1.00	1.00	1.00	0.92
ST-094	Locate	LS+WO	12	0.30	0.05	1.00	0.05	1.00	1.00	1.00

Task	Type	Faults	Items	Base	GPT 6.1 Sol		Opus		Gemini	
					Silent	Guided	Silent	Guided	Silent	Guided
ST-095	Activity	FR+LS+MX	12	0.30	0.29	0.64	0.64	0.64	0.64	0.64
ST-096	Identify	LS+MX+WO	12	0.33	0.00	1.00	0.00	1.00	0.00	1.00
ST-097	Activity	FR+WO	12	0.39	0.45	1.00	0.45	1.00	0.38	1.00
ST-098	Locate	EE+FR+LS	12	0.33	0.00	1.00	0.00	0.00	0.00	1.00
ST-099	Locate	FR+LS+WO	12	0.33	0.00	1.00	0.00	1.00	0.00	1.00
ST-100	Activity	LS+MX+RO	12	0.33	1.00	1.00	0.62	1.00	0.62	1.00
ST-101	Activity	MX+RO	12	0.33	1.00	1.00	1.00	1.00	1.00	1.00
ST-102	Activity	FR+MX	12	0.38	0.20	1.00	1.00	1.00	0.80	1.00
ST-103	Activity	LS+WO	12	0.25	0.00	1.00	0.00	1.00	0.00	1.00
ST-104	Activity	WO	11	0.44	0.00	0.00	0.00	1.00	0.00	0.92
ST-105	Locate	LS+WO	12	0.25	0.00	0.67	0.00	0.67	0.00	0.67
ST-107	Activity	MX+RO	12	0.33	0.00	1.00	1.00	1.00	1.00	1.00
ST-109	Activity	LS+WO	12	0.42	0.00	1.00	0.14	0.57	0.57	0.57
ST-110	Identify	EE+WO	10	0.33	0.00	0.93	0.00	0.93	1.00	0.86
ST-111	Activity	RO+WO	12	0.17	0.10	1.00	0.70	1.00	1.00	0.60
ST-112	Activity	FR+RO	12	0.48	0.00	1.00	1.00	1.00	1.00	1.00
ST-113	Activity	LS+RO	12	0.33	0.00	1.00	0.50	1.00	0.50	1.00
ST-114	Activity	LS+MX	12	0.42	0.00	1.00	1.00	1.00	0.57	1.00
ST-115	Activity	FR+LS+MX	12	0.25	0.22	1.00	0.67	1.00	0.67	1.00
ST-116	Activity	FR+LS+RO	12	0.31	0.35	1.00	0.68	1.00	0.68	1.00
ST-117	Activity	EE+MX	12	0.33	0.00	1.00	1.00	1.00	1.00	1.00
ST-119	Activity	LS	12	0.33	0.00	1.00	0.00	0.94	0.00	1.00
ST-120	Activity	FR+LS	12	0.33	0.00	1.00	0.00	1.00	0.00	1.00
ST-121	Activity	LS+MX	12	0.46	0.00	1.00	1.00	1.00	0.08	1.00
ST-122	Activity	WO	12	0.28	0.08	1.00	0.08	1.00	0.08	1.00
ST-123	Locate	LS	12	0.33	0.00	1.00	0.00	1.00	0.00	0.89
ST-124	Activity	EE+LS+RO	12	0.33	0.50	1.00	0.50	1.00	0.50	1.00
ST-125	Activity	EE+RO	12	0.48	1.00	1.00	1.00	1.00	0.00	1.00
ST-126	Activity	FR+LS	10	0.37	0.00	1.00	0.00	1.00	0.00	1.00
ST-127	Activity	WO	12	0.33	0.00	1.00	0.00	1.00	0.00	1.00
ST-128	Activity	FR+LS	10	0.29	0.07	1.00	1.00	1.00	0.07	1.00
ST-129	Activity	FR+MX	12	0.28	0.08	1.00	1.00	1.00	1.00	1.00
ST-130	Activity	MX	12	0.28	0.08	1.00	0.08	1.00	1.00	1.00
ST-131	Activity	LS+RO	12	0.33	0.00	1.00	0.50	1.00	0.50	1.00
ST-132	Activity	LS+MX	12	0.25	0.78	1.00	0.00	1.00	0.44	1.00
ST-133	Activity	FR+MX	10	0.31	0.04	1.00	0.04	1.00	1.00	1.00
ST-135	Activity	EE+MX+WO	12	0.42	0.00	1.00	0.57	1.00	1.00	1.00
ST-136	Activity	RO+WO	12	0.33	0.00	1.00	0.50	1.00	0.50	0.42
ST-137	Activity	FR+LS+RO	12	0.43	1.00	1.00	0.23	1.00	1.00	1.00
ST-140	Activity	FR+MX+WO	12	0.25	0.00	1.00	0.67	1.00	0.89	1.00
ST-141	Activity	EE+MX+RO	12	0.25	1.00	1.00	0.33	1.00	0.67	0.78
ST-142	Activity	EE+FR+RO	10	0.37	0.00	1.00	1.00	1.00	1.00	1.00
ST-143	Locate	FR+WO	12	0.33	0.00	0.00	0.00	0.00	0.00	1.00
ST-144	Activity	EE+MX+RO	12	0.42	0.00	1.00	1.00	1.00	1.00	1.00
ST-145	Activity	MX+RO	12	0.33	0.00	1.00	0.25	1.00	1.00	1.00
ST-146	Activity	EE+RO	12	0.41	0.00	1.00	1.00	1.00	0.00	1.00
ST-147	Activity	EE+MX	12	0.33	0.00	1.00	1.00	1.00	1.00	0.00
ST-148	Activity	LS+RO	12	0.30	0.05	1.00	0.05	1.00	0.05	1.00
ST-149	Identify	EE+WO	11	0.33	0.00	1.00	0.00	1.00	1.00	1.00
ST-150	Activity	LS	12	0.33	0.00	1.00	1.00	1.00	0.00	1.00
ST-151	Activity	LS	12	0.33	0.00	1.00	0.00	1.00	0.00	1.00
ST-152	Locate	WO	12	0.33	0.00	1.00	0.00	0.00	0.00	1.00
ST-153	Activity	EE+FR+RO	10	0.49	1.00	1.00	1.00	1.00	1.00	1.00
ST-154	Activity	FR+LS	12	0.46	0.17	1.00	0.00	1.00	0.17	1.00
ST-155	Activity	FR+LS	12	0.28	0.08	1.00	0.08	1.00	0.08	1.00
ST-156	Activity	FR+LS	12	0.33	0.50	0.50	0.50	1.00	0.50	1.00
ST-157	Activity	EE+LS+RO	12	0.33	0.50	1.00	0.50	1.00	0.50	1.00

Task	Type	Faults	Items	Base	GPT 6.1 Sol		Opus		Gemini	
					Silent	Guided	Silent	Guided	Silent	Guided
ST-159	Activity	RO	12	0.37	0.00	1.00	1.00	1.00	0.00	1.00
ST-160	Activity	FR+RO	12	0.29	0.53	1.00	1.00	1.00	1.00	0.87
ST-161	Activity	LS+WO	12	0.33	0.00	1.00	0.50	1.00	0.00	1.00
ST-162	Activity	EE+FR+MX	10	0.33	0.00	1.00	1.00	1.00	1.00	1.00
ST-163	Activity	FR+LS	12	0.33	0.33	1.00	0.00	1.00	0.33	1.00
ST-164	Activity	LS+WO	12	0.28	0.08	0.46	0.54	1.00	0.08	1.00
ST-165	Activity	LS+RO+WO	12	0.25	0.00	1.00	0.67	1.00	0.33	1.00
ST-166	Activity	LS+MX	12	0.43	0.00	1.00	1.00	1.00	0.02	1.00
ST-167	Locate	EE+FR+MX	12	0.33	0.00	0.00	1.00	1.00	1.00	1.00
ST-168	Activity	LS+MX+WO	12	0.27	0.00	1.00	0.00	1.00	0.31	0.91
ST-169	Activity	FR+LS+RO	12	0.33	0.25	1.00	0.62	1.00	0.62	1.00
ST-170	Identify	MX+RO	12	0.33	1.00	1.00	1.00	1.00	0.31	0.31
ST-171	Locate	LS	11	0.33	0.00	1.00	0.00	1.00	0.00	1.00
ST-172	Activity	LS	12	0.39	0.00	1.00	0.00	1.00	0.00	1.00
ST-173	Activity	LS+MX	12	0.17	0.10	1.00	0.40	1.00	0.40	1.00
ST-174	Activity	MX	12	0.48	0.00	1.00	0.00	1.00	1.00	1.00
ST-175	Activity	MX+RO	12	0.40	0.38	1.00	0.38	1.00	1.00	1.00
ST-176	Activity	LS+RO	12	0.33	0.50	1.00	0.50	1.00	0.00	1.00
ST-177	Activity	FR+RO	12	0.33	0.50	1.00	0.50	1.00	1.00	1.00
ST-178	Activity	FR+LS+RO	10	0.33	0.00	1.00	0.00	1.00	0.00	1.00
ST-179	Activity	MX	12	0.33	0.00	1.00	1.00	1.00	1.00	1.00
ST-181	Activity	LS+WO	12	0.39	0.00	1.00	0.00	1.00	0.00	1.00
ST-182	Activity	MX+SX	10	0.38	0.00	0.20	0.00	1.00	0.20	0.20
ST-183	Activity	LS+SX	12	0.27	0.00	1.00	1.00	1.00	0.49	1.00
ST-184	Activity	LS+SX	12	0.43	0.00	1.00	1.00	1.00	0.02	1.00
ST-186	Activity	LS+MX+SX	12	0.31	0.00	1.00	0.20	1.00	0.60	1.00
ST-187	Activity	LS+RO+SX	12	0.39	0.09	1.00	0.09	1.00	0.55	1.00
ST-188	Activity	RO+SX+WO	12	0.43	0.03	1.00	0.03	1.00	1.00	1.00
ST-189	Activity	MX+RO+SX	12	0.29	1.00	1.00	0.61	1.00	1.00	1.00
ST-190	Activity	RO+SX	12	0.31	0.00	1.00	1.00	1.00	1.00	1.00
ST-191	Activity	RO+SX	12	0.27	0.00	1.00	0.49	1.00	0.00	1.00
ST-192	Activity	SX+WO	12	0.40	0.00	1.00	0.38	1.00	0.38	1.00
ST-193	Activity	RO+SX	12	0.36	0.00	1.00	1.00	1.00	1.00	1.00
ST-194	Activity	RO+SX	12	0.25	0.00	1.00	0.50	1.00	0.50	1.00
ST-195	Activity	MX+SX	12	0.28	0.00	1.00	1.00	1.00	1.00	1.00
ST-197	Activity	MX+SX	12	0.25	0.44	1.00	1.00	1.00	1.00	1.00
ST-198	Activity	LS+SX	12	0.33	0.00	1.00	0.44	1.00	0.44	1.00
ST-199	Activity	RO+SX	12	0.33	0.00	1.00	0.44	1.00	1.00	1.00
ST-200	Activity	LS+SX	12	0.42	0.00	1.00	0.00	1.00	0.14	1.00
ST-201	Activity	LS+SX	12	0.46	0.00	0.85	0.31	1.00	0.31	0.85