FRONTIERCHALLENGE: EVALUATING SCIENTIFIC WORKFLOW COMPLETION

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific agents increasingly analyze data, execute code, and produce research artifacts, yet most benchmarks emphasize final answers, isolated programs, or a single domain. We introduce FRONTIERCHALLENGE, a cross-domain benchmark comprising 300 end-to-end scientific workflows. In this paper, we release and evaluate 97 of these tasks, spanning quantum chemistry, molecular dynamics, materials characterization, analytical chemistry, life science, and electrochemistry/environment. Each task provides fixed inputs and specifies a bundle of required scientific deliverables. We evaluate twelve frontier models in thirteen configurations across three agent scaffolds. Pass Rate measures the fraction of tasks satisfying the rubric-based completion criterion, while Avg. Score captures partial progress. Each of the best-performing configurations completed only 20 of the 97 released tasks, yielding a Pass Rate of 20.6%. Partial progress translated especially poorly into complete delivery in analytical chemistry and electrochemistry/environment: Avg. Scores reached 87.6 and 94.9, but the highest Pass Rates were only 4% and 0%. Among non-passing Claude Code trajectories, 75.5% ended with completion-related language detected by a lexical rule. These findings show that high partial scores and completion-related language can coexist with failure to meet the benchmark's strict completion criterion, highlighting the need to evaluate end-to-end workflow execution and the completeness of scientific deliverables together.

1 INTRODUCTION

Language models are evolving from text generators into agents that can plan, call tools, execute code, and modify persistent files (Liu et al., 2024; Mialon et al., 2024; Xie et al., 2024). Alongside advances in agent scaffolding, continual pre-training has been explored as a way to scale general agent capabilities (Su et al., 2025). Recent systems suggest that agentic support can extend beyond isolated tasks such as literature retrieval or text translation to coordinating multi-stage research workflows with inspectable outputs (Lu et al., 2024). This shift changes what constitutes success on a scientific task. Producing a plausible conclusion is not enough: an agent may need to inspect heterogeneous inputs, select and run an analysis, validate intermediate results, and deliver mutually consistent code, tables, figures, and prose.

Existing benchmarks cover expert knowledge, general tool use, software interaction, code repair, paper replication, and scientific data analysis (Phan et al., 2025; Jimenez et al., 2024; Chen et al., 2025; Siegel et al., 2024; Starace et al., 2025). Their evaluation units, however, are often a final answer, an interaction trace, a single program, or a workflow from one discipline. These settings do not fully characterize whether an agent can complete heterogeneous scientific work whose success depends on several analytical stages and several required deliverables.

FRONTIERCHALLENGE therefore asks a focused question:

> *Given a specified scientific task and fixed data, can an agent independently complete the workflow from input processing to final deliverables and satisfy the complete task contract?*

Reliable execution underpins a trustworthy scientific handoff: analyses and outputs must be inspectable, reproducible, and mutually consistent. FRONTIERCHALLENGE therefore targets a nar-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

rower capability than autonomous science: executing a specified scientific workflow after its objective, inputs, and required outputs are fixed. The agent is not asked to set the research agenda or formulate the problem.

We constructed a pool of 300 end-to-end scientific workflows grouped into six reporting domains: quantum chemistry, molecular dynamics, materials characterization, analytical chemistry, life science, and electrochemistry/environment. In this study, we evaluate and publicly release 97 tasks; the remaining 203 are retained as an internal held-out set. Representative tasks require agents to use domain software, including ORCA, CP2K, LAMMPS, AmberTools, and PLUMED; perform quantitative analysis, quality control, and visualization; and produce executable code and evidence-grounded reports. The unit of evaluation is the complete submitted artifact bundle rather than a single answer.

We evaluated twelve frontier models in thirteen configurations across three agent scaffolds on the 97 tasks. Our primary metric is Pass Rate, defined as the fraction of tasks meeting the frozen rubric-based completion criterion. Avg. Score is reported only as a complementary measure of partial progress; a high-scoring but incomplete submission is not counted as a pass.

We additionally examine process–outcome associations using the HDS6 framework from *Apodex Discovery* (Wang et al., 2026a). This exploratory assessment covers 1,164 frozen trajectories from twelve configurations (Section 4.4).

The central finding is a persistent gap between partial progress and complete scientific delivery. The best-performing configurations (i.e., GPT-5.6 Sol with Codex and Grok 4.6 with Claude Code) completed only 20 of 97 tasks, corresponding to a Pass Rate of 20.6%, despite the highest Avg. Score reaching 87.9. This gap was especially pronounced in analytical chemistry and electrochemistry/environment: Avg. Scores reached 87.6 and 94.9, respectively, while the highest Pass Rates were only 4% and 0%. Thus, although frontier systems can make substantial progress on complex scientific tasks, they remain far from reliable end-to-end execution.

This work makes three contributions:

1. **Cross-domain scientific workflow benchmark.** We construct a pool of 300 end-to-end workflows spanning six reporting domains, evaluate and publicly release 97 tasks, and retain 203 as an internal held-out set. Each task is defined by fixed inputs, a heterogeneous deliverable contract, and a task-specific Grader.
2. **Contract-level evaluation.** We use Pass Rate as the primary metric of complete workflow execution and Avg. Score to characterize partial progress across task-specific scientific rubrics.
3. **Frontier model study and failure analysis.** We evaluate twelve frontier models with multiple agent scaffolds, quantify domain-level completion gaps, compare machine-observable contract-breach signatures across domains, and analyze final-status and error behavior in 970 Claude Code trajectories.

2 Related Work

General capability and long-horizon agent benchmarks. Humanity's Last Exam (HLE) probes the limits of expert knowledge with difficult, checkable questions (Phan et al., 2025). AgentBench, GAIA, OSWorld, and SWE-bench extend evaluation to tool use, computer interaction, and software engineering (Liu et al., 2024; Mialon et al., 2024; Xie et al., 2024; Jimenez et al., 2024). Agents' Last Exam (ALE) evaluates long-horizon professional work, while Frontier-Bench collects difficult, verifiable tasks near the frontier of agent capability (Sun et al., 2026b; Harbor Framework, 2026). These benchmarks establish the importance of evaluating completed work, but they are not centered on cross-domain scientific workflows.

Scientific knowledge and data analysis. LAB-Bench evaluates knowledge and reasoning required for biological research (Laurent et al., 2024). BixBench, BioMysteryBench, and CompBioBench use biological data to construct open-ended or objectively verifiable problems (Mitchener et al., 2025; Anthropic, 2026; Nair et al., 2026). BLADE represents open-ended data analysis through expert-recognized analytical decisions, while ScienceAgentBench asks agents to produce executable analyses derived from the scientific literature (Gu et al., 2024; Chen et al., 2025). These

Table 1: Core design characteristics of related benchmarks. A solid dot indicates that the characteristic is part of the benchmark's primary design; a dash indicates otherwise. Shared features do not imply identical task objectives or acceptance criteria.

Benchmark	Scientific workflow core	Fixed scientific inputs	End-to-end execution	Multiple artifacts	Executable evaluation	Across scientific domains
HLE (Phan et al., 2025)	–	–	–	–	–	•
ALE (Sun et al., 2026b)	–	–	•	•	•	–
Frontier-Bench (Harbor Framework, 2026)	–	–	•	•	•	–
BLADE (Gu et al., 2024)	•	•	–	–	•	•
ScienceAgentBench (Chen et al., 2025)	•	•	•	–	•	•
CORE-Bench (Siegel et al., 2024)	•	•	•	•	•	•
PaperBench (Starace et al., 2025)	•	•	•	•	•	–
BioAgent Bench (Fa et al., 2026)	•	•	•	•	•	–
BiomniBench (Qu et al., 2026)	•	•	•	•	•	–
NatureBench (Wang et al., 2026b)	•	•	•	•	•	•
AstaBench (Bragg et al., 2025)	•	•	•	–	•	•
FRONTIERCHALLENGE	•	•	•	•	•	•

benchmarks substantially increase scientific realism, but many evaluate an answer, a decision set, or a single self-contained program rather than a heterogeneous deliverable set.

End-to-end scientific workflows. CORE-Bench and PaperBench evaluate computational reproduction and full research replication (Siegel et al., 2024; Starace et al., 2025). ScienceBoard, SciAgentArena, and SciAgentGym assess scientific software use, interactive research environments, and multi-step tool use (Sun et al., 2026a; Liu et al., 2026; Shen et al., 2026). BioAgent Bench and BiomniBench evaluate end-to-end artifacts or processes in bioinformatics and biomedicine (Fa et al., 2026; Qu et al., 2026). NatureBench and AstaBench broaden the scope to paper-level methods and cross-domain research tasks (Wang et al., 2026b; Bragg et al., 2025). These resources share several design features with FRONTIERCHALLENGE (Table 1), but emphasize different scientific goals. CORE-Bench tests reproduction of published results using supplied code and data (Siegel et al., 2024). NatureBench evaluates whether coding agents can improve on published scientific methods (Wang et al., 2026b). FRONTIERCHALLENGE focuses on fulfilling a specified scientific delivery contract. For example, its calorimetry workflow requires calibration, corrected traces, run-level metrics, a protocol decision, and a mutually consistent report (Appendix B). The contribution is a curated suite for measuring this complete handoff across scientific workflows, together with partial-credit diagnostics.

3 FRONTIERCHALLENGE

3.1 TASK COLLECTION, CURATION, AND PACKAGING

Task collection. We collected tasks from scientific and engineering settings, focusing on realistic workflows that require domain knowledge, specialized software, experimental data, or engineering environments. The tasks originate from analysis, computation, simulation, and research-delivery processes performed in domain practice, rather than from expanded question answering or isolated coding exercises. We prioritized workflows in which an agent must understand a professional objective, use domain tools correctly, execute interdependent steps, and produce verifiable scientific

artifacts. The complete collection contains 300 scientific workflows; in this paper, we release and evaluate 97.

Screening and quality control. We curated tasks according to four design principles:

- **Representativeness.** The workflow, software, and methods must reflect plausible professional practice.
- **Complexity.** The task must require an end-to-end, dependency-aware process culminating in a substantive deliverable, rather than a single command, tool call, or local edit.
- **Diversity.** The collection must vary in scientific knowledge, workflow type, and difficulty, rather than repeat templates with only inputs or parameters changed.
- **Verifiability.** Outputs must be assessable through files, numerical values, quantitative measures, or explicit acceptance criteria, thereby supporting repeatable automated evaluation.



Figure 1: Domain and workflow-family composition of the 97 released tasks; center bars show the retained Hard/Medium split.

We also checked that every included task had fixed inputs, a defined execution environment, a complete deliverable contract, and an executable evaluation procedure. Tasks centered on isolated facts or single-step operations, tasks with purely subjective outputs, and tasks lacking the materials needed for reproducible scoring were excluded before the final 300-workflow collection was formed.

Standardization and packaging. After curation, we organized each task as a self-contained package with five aligned elements: a *task description* defining the scientific objective; fixed *inputs*, including data and necessary context; the *software and tools* available in the execution environment; an *output contract* listing the required deliverables; and an *evaluation procedure* defining successful completion. Deliverables may include scientific reports, structured tables, diagnostic figures, executable analysis code, simulation products, or multiple artifacts that must remain mutually consistent. Completion therefore depends on the entire artifact bundle, not on whether the agent returns a plausible final answer.

The dataset separates agent-visible inputs from evaluator-side materials. Task packages contain instructions, fixed inputs, metadata, and environment definitions; evaluator packages contain the task-specific rubrics, Graders, and reference artifacts. This organization supports inspection of each task's requirements and scoring procedure while keeping reference materials outside the agent's inputs. We will provide complete task packages and a reference solution for each of the 97 evaluated tasks. Appendix C describes the package contents, registry, and runtime requirements.

3.2 Dataset Statistics

All 300 workflows in the curated collection passed the quality-control checks above. For this study, we randomly selected 97 tasks from the subset whose official evaluation does not require GPU resources, publicly released them, and used them for the reported experiments. The remaining 203 tasks form an internal held-out set, which also contains workflows whose official evaluation requires GPUs; none of these held-out tasks is included in the results reported in this paper. The released tasks comprise 74 Hard tasks and 23 Medium tasks. Their fixed inputs comprise public empirical data, public sequence or structural resources, and scientifically constrained synthetic data.

For analysis, we organize the 97 released tasks into six reporting domains: quantum chemistry (20 tasks), molecular dynamics (16), materials characterization (22), analytical chemistry (23), life science (10), and electrochemistry (6). Collectively, these tasks cover 21 workflow families and require heterogeneous deliverables, including scientific reports, structured data, figures, executable code, and simulation products. Figure 1 summarizes the domain counts, all 21 workflow families, and the within-domain Hard/Medium composition.

Domain-level results describe performance on these task groups, whose sizes reflect the composition of the release. They characterize the benchmark's workflows rather than rank the intrinsic difficulty of scientific disciplines.

4 Experiments

We conducted experiments to address four research questions:

1. **RQ1:** How reliably can current frontier models complete specified scientific workflows?
2. **RQ2:** How does scientific workflow performance vary across domains?
3. **RQ3:** Which observable contract breaches and trajectory behaviors most often accompany incomplete scientific handoffs?
4. **RQ4:** What process capabilities do frozen trajectories reveal under a Heavy Duty Solver assessment?

4.1 Agent Scaffolds and Models

Agent scaffolds. We evaluated models with three agent scaffolds: Codex (OpenAI, 2026a), Claude Code (Anthropic, 2026a), and Frontier Agent (ApodexAI, 2026). Codex was used with GPT-5.6 Sol and GPT-5.6 Terra (max), whereas Claude Code served as the common scaffold for ten models. Frontier Agent was evaluated with Apodex 1.1 in the Agent Team configuration.

Models. We evaluated twelve frontier models. These comprised GPT-5.6 Sol and GPT-5.6 Terra (max) (OpenAI, 2026b), Grok 4.6 (xAI, 2026), Kimi K3 (Kimi Team, 2026), Claude Opus 5 (Anthropic, 2026b), and Qwen 3.8 Max (Alibaba Group, 2026). We also included Qwen3.5-397B-A17B (Qwen Team, 2026), DeepSeek V4 Flash-0731 and DeepSeek V4 Pro-0813 (DeepSeek-AI, 2026), Apodex 1.1 (Apodex Team et al., 2026), and GLM-5.2 (Z.ai, 2026), together with Gemini 3.7 Flash (Google DeepMind, 2026) using dynamic thinking. All models received the same 97 task objectives and task-visible inputs. Together, these pairings yield thirteen model–scaffold configurations. Appendix A.1 specifies the configuration coverage of each analysis.

Table 2: Pass Rate for thirteen model–scaffold configurations on 97 tasks, comprising 23 Medium and 74 Hard tasks. Each column uses its own task denominator and the same score threshold of 99.9.

Model	Agent Scaffold	Overall ↑	Medium ↑	Hard ↑
GPT-5.6 Sol	**Codex**	**20.6%**	39.1%	**14.9%**
Grok 4.6	**Claude Code**	**20.6%**	**43.5%**	13.5%
Kimi K3	Claude Code	17.5%	39.1%	10.8%
Claude Opus 5	Claude Code	17.5%	39.1%	10.8%
GPT-5.6 Terra (max)	Codex	15.5%	39.1%	8.1%
Qwen 3.8 Max	Claude Code	15.5%	34.8%	9.5%
DeepSeek V4 Pro-0813	Claude Code	13.4%	30.4%	8.1%
DeepSeek V4 Flash-0731	Claude Code	12.4%	34.8%	5.4%
Apodex 1.1	Frontier Agent (Agent Team)	12.4%	30.4%	6.8%
Gemini 3.7 Flash	Claude Code	10.3%	21.7%	6.8%
Apodex 1.1	Claude Code	10.3%	30.4%	4.1%
Qwen3.5-397B-A17B	Claude Code	4.1%	13.0%	1.4%
GLM-5.2	Claude Code	3.1%	4.3%	2.7%

4.2 Task-specific Evaluation and Primary Metrics

Each task's Grader evaluates the submitted artifact bundle against its rubric, including required files, numerical results, formats, figures, code execution, and cross-artifact consistency. Task-specific criteria and reference artifacts are provided in the evaluator package (Appendix C). The Grader returns a score $s_{mi} \in [0, 100]$ for configuration m on task i. Rubric-defined semantic criteria are assessed by GPT-5.6 Sol as part of the Grader, with three Judge calls per criterion.

Let $N = 97$ denote the number of evaluated tasks. We set $f_{mi} = 1$ when evaluation is complete and $s_{mi} \geq 99.9$, and $f_{mi} = 0$ otherwise. We report:

$$\text{PassRate}_m = \frac{1}{N} \sum_{i=1}^{N} f_{mi}, \tag{1}$$

$$\text{Avg. Score}_m = \frac{1}{N} \sum_{i=1}^{N} s_{mi}. \tag{2}$$

Pass Rate measures complete delivery under the task-specific rubrics, whereas Avg. Score summarizes partial credit across the suite. Because rubrics differ between tasks, Avg. Score is a descriptive aggregate, not a common scale of scientific accuracy. The evaluation code and reported results use the same pass criterion (Appendix C).

4.3 Overall Performance

Strict completion remained rare across all configurations. Across the evaluated systems, Pass Rate ranged from 3.1% to 20.6%, whereas Avg. Score ranged from 67.5 to 87.9 (Fig. 2 and Table 2). Codex with GPT-5.6 Sol achieved the highest Avg. Score of 87.9 and shared the highest Pass Rate of 20.6% with Claude Code using Grok 4.6, which obtained an Avg. Score of 86.6. Claude Code with Kimi K3 and Claude Opus 5 each achieved a Pass Rate of 17.5%, with Avg. Scores of 85.5 and 84.9, respectively. Codex with GPT-5.6 Terra (max) followed with an Avg. Score of 84.7 and a Pass Rate of 15.5%.

High Avg. Scores did not imply complete delivery. Eight configurations achieved Avg. Scores above 80, but none completed more than 20.6% of tasks under the strict criterion. Thus, high average scores often reflected artifact bundles that still missed at least one requirement. The two Apodex 1.1 configurations also varied substantially. Frontier Agent (Agent Team) obtained a higher Avg. Score and Pass Rate than Claude Code (74.5 and 12.4% versus 71.8 and 10.3%). This comparison is descriptive and does not isolate a pure scaffold effect.

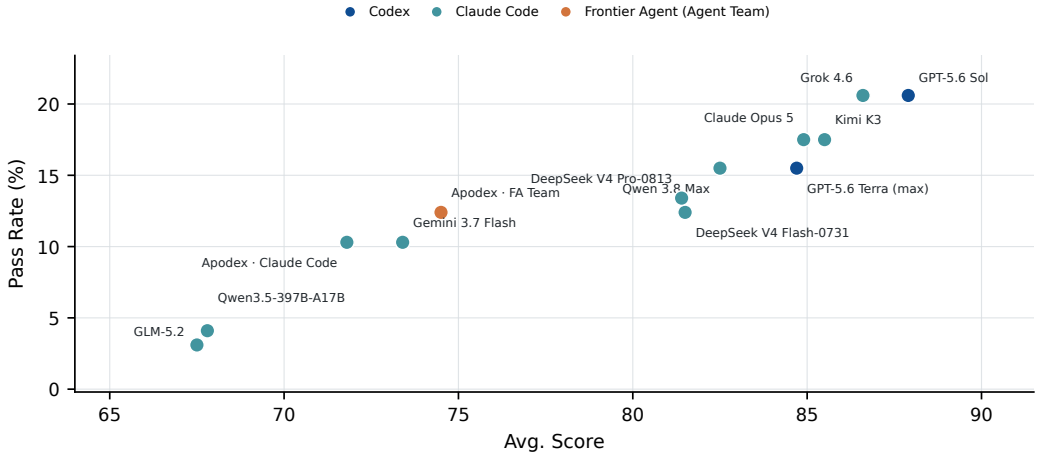


Figure 2: Avg. Score and Pass Rate across thirteen configurations on the same 97-task suite. Each point represents one model–scaffold pairing; color identifies the scaffold. Tasks receive equal weight within each metric.

4.4 Heavy Duty Solver Capability Assessment

To explore associations between artifact-level outcomes and process scores, we applied the HDS6 framework from *Apodex Discovery* (Wang et al., 2026a) to frozen trajectories from 12 model–scaffold configurations on the 97-task suite (1,164 scorecards). HDS6 separately scores Coherence, Evidence, Alternatives, Scope, Tools, and Repair. The assessment excludes Frontier Agent runs, which lack the required structured trajectories, and therefore does not cover every scaffold in the main benchmark.

The process and outcome rankings broadly agreed. Across configurations, HDS6 correlated with Avg. Score (Pearson $r = 0.874$, Spearman $\rho = 0.762$) and Pass Rate ($r = 0.828$, $\rho = 0.799$). The mean within-task Pearson correlation between process score and task score was 0.475 across 94 eligible tasks, with positive correlations on 85% of them. Evidence was the highest or nearly highest dimension for every configuration, whereas Scope received the lowest scores under this rubric (2.185–3.508). These associations are exploratory because the archived rubric was uncalibrated and tool observations were truncated. Appendix A.2 gives the protocol, measurement conditions, aggregation rule, and configuration-level results.

4.5 Domain-level performance

Pass Rates were highest in two domains. Domain-level Avg. Score and Pass Rate were available for every evaluated system (Fig. 3). In quantum chemistry, Claude Code with Grok 4.6 achieved the highest Pass Rate of 60%, followed by Codex with GPT-5.6 Sol and Claude Code with Claude Opus 5 at 55%. In molecular dynamics, GPT-5.6 Sol, GPT-5.6 Terra (max), and Grok 4.6 each achieved a Pass Rate of 38%, while Terra obtained the highest Avg. Score of 93.7.

High domain-level Avg. Scores often coexisted with near-zero Pass Rates. In materials characterization, Avg. Scores reached 88.1, but no configuration exceeded a Pass Rate of 9%. The divergence was stronger in analytical chemistry and electrochemistry/environment. The highest Avg. Score in analytical chemistry was 87.6, yet only DeepSeek V4 Pro-0813 completed any task in that domain, with a Pass Rate of 4%. Electrochemistry/environment reached a maximum Avg. Score of 94.9, but its Pass Rate remained 0% for every configuration. Life-science tasks showed lower Avg. Scores overall; GPT-5.6 Sol achieved the highest Avg. Score of 76.8, whereas Kimi K3 achieved the highest Pass Rate of 20%.

The complete configuration-level domain profiles are reported in Appendix A.3.

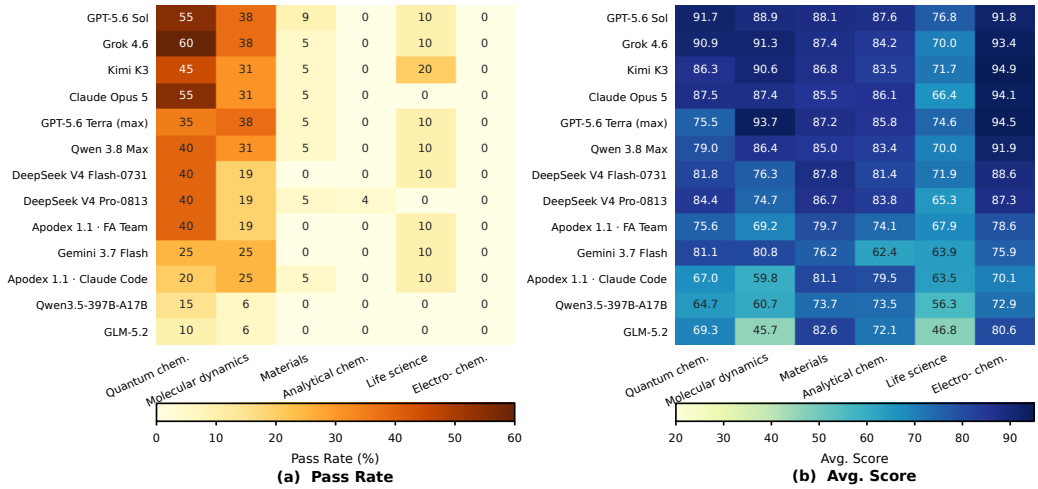


Figure 3: Domain-level Pass Rate and Avg. Score for all thirteen configurations. The six domains contain 20, 16, 22, 23, 10, and 6 tasks, respectively, in the displayed order. Each cell aggregates only tasks within its domain; Pass Rate uses the 99.9 score threshold.

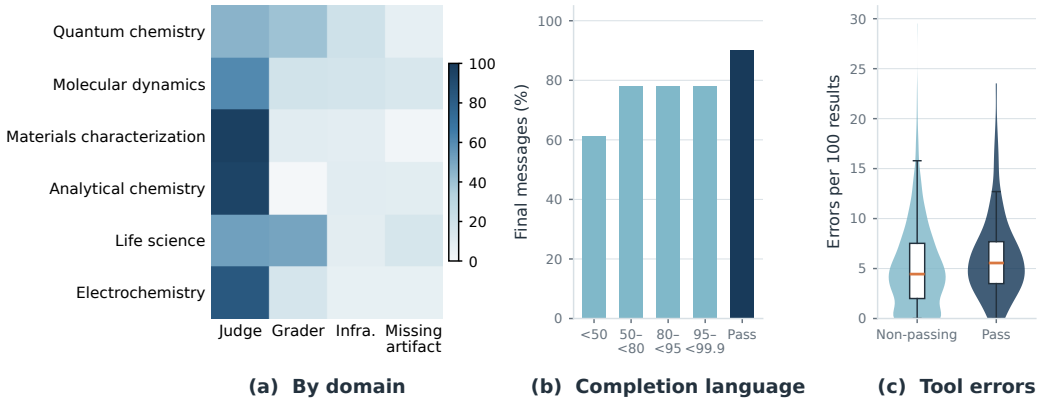


Figure 4: Diagnostic signatures and trajectory behavior. (a) Signature prevalence among non-passing runs within each domain, using twelve configurations and 1,011 non-passing runs; signatures may overlap. (b) Lexical completion-language frequency by final score band among 970 Claude Code runs. (c) Tool errors per 100 tool results for the same runs, split into 849 non-passing and 121 passing submissions. Violins show the distributions; boxes show quartiles and median lines. Analysis coverage and lexical rules appear in Appendix A.1.

Appendix A.4 reports resource use and execution time under the providers' respective accounting conventions.

4.6 Failure Mode Analysis

Figure 4 summarizes incomplete scientific handoffs from three complementary views: domain-level diagnostic signatures, completion-related language in final messages, and raw tool errors. The archived diagnostic analysis covers twelve configurations, excluding GPT-5.6 Terra (max), and 1,011 non-passing runs. Signatures can overlap. The trajectory comparisons cover 970 Claude Code runs with a common event schema (Appendix A.1). Together, these views describe observable shortfalls and behaviors associated with incomplete delivery.

Failure signatures varied substantially by domain. Judge-assessed artifact shortfalls appeared in 96.5% of non-passing materials-characterization submissions, 94.2% of analytical-chemistry submissions, and 83.3% of electrochemistry/environment submissions, compared with 45.0% in quantum chemistry. Deterministic Grader diagnostics were more common in life science (50%) and quantum chemistry (40%) than in the other domains (Fig. 4a). These patterns reflect both the submitted artifacts and differences in task contracts and evaluator composition; the signatures identify observed shortfalls rather than unique failure causes.

Completion-related wording occurred in both passing and non-passing runs. Among the 970 Claude Code trajectories, 641 of the 849 non-passing runs (75.5%) had a final message containing lexical completion language; only 13 (1.5%) matched the separate lexical rule for running, waiting, or in-progress work. Completion language was less frequent among runs scoring below 50 (61.2%) but remained above 78% in each non-passing band above 50; it appeared in 90.1% of passing runs (Fig. 4b). Appendix A.1 specifies the lexical rule used to identify these expressions.

Tool errors also occurred in successful trajectories. At least one tool error occurred in 80.7% of non-passing and 94.2% of passing Claude Code trajectories. Median error rates were 4.4 and 5.6 per 100 tool results, respectively (Fig. 4c). Successful trajectories can encounter and recover from tool errors; the presence of an error alone does not identify the eventual contract breach.

5 Conclusion

FrontierChallenge evaluates scientific workflow completion through the joint assessment of code, data, figures, reports, and their consistency. We release and evaluate 97 workflows across six domains. Across thirteen configurations, Pass Rate ranged from 3.1% to 20.6%, despite Avg. Scores of 67.5 to 87.9. Completion-related language also appeared in 75.5% of non-passing Claude Code trajectories. Substantial partial progress and completion wording therefore coexisted with failure to meet the benchmark's strict completion criterion. These findings concern the released tasks and retained trial records under the specified task rubrics. They motivate evaluating complete scientific delivery alongside partial progress, with explicit contract tracking and cross-artifact verification as practical targets for agent development.

AI use statement

In this work, we used generative AI tools, including OpenAI Codex, primarily to polish and edit manuscript language, identify grammatical errors, assist with the preparation and editing of scientific figures, and organize and refactor code. The authors reviewed all AI-assisted text, verified figure content against the underlying data, and inspected and tested AI-assisted code changes before including them. The authors made the final decisions about the analyses, claims, and presentation, and take responsibility for the complete content of this work.

Reproducibility statement

The 97-task public release provides instructions, input data, environment definitions, metadata, and checksums in a Harbor-compatible format. Separate evaluator packages contain task-specific rubrics, Graders, and reference artifacts for inspecting the scoring procedure. We will provide complete task packages and a reference solution for each evaluated task. Appendix C documents the package structure and integrity checks; Appendix A specifies the analysis coverage and trajectory-processing rules.

REFERENCES

Alibaba Group. Alibaba unveils Qwen3.8-Max: Its largest and most capable flagship model to date. Alibaba Group press release, August 2026. URL https://www.alibabagroup.com/en-US/document-2021044032125272064. Accessed 2026-08-19.

Anthropic. Evaluating claude's bioinformatics research capabilities with biomysterybench. Research blog post, Anthropic, 2026. URL https://www.anthropic.com/research/Evaluating-Claude-For-Bioinformatics-With-BioMysteryBench.

Anthropic. Claude code overview. Claude Code documentation, 2026a. URL https://code.claude.com/docs/en/overview. Accessed 2026-08-10.

Anthropic. What's new in Claude Opus 5. Claude Platform documentation, 2026b. URL https://platform.claude.com/docs/en/about-claude/models/whats-new-opus-5. Accessed 2026-08-19.

Apodex Team et al. Apodex 1.1: Scaling agentic intelligence for complex work. *arXiv preprint arXiv:2608.23283*, 2026. doi: 10.48550/arXiv.2608.23283. URL https://arxiv.org/abs/2608.23283.

ApodexAI. FrontierAgent. GitHub repository, 2026. URL https://github.com/ApodexAI/FrontierAgent. Accessed 2026-08-19.

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi, Sergey Feldman, Dany Haddad, Jena D Hwang, Peter Jansen, Varsha Kishore, et al. Astabench: Rigorous benchmarking of ai agents with a scientific research suite, 2025. *URL https://arxiv. org/abs/2510.21652*, 2025.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, volume 2025, pp. 96934–96990, 2025.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Dionizije Fa, Marko Culjak, Bruno Pandza, and Mateo Cupic. Bioagent bench: An ai agent evaluation suite for bioinformatics. *arXiv preprint arXiv:2601.21800*, 2026.

Google DeepMind. Gemini 3.7 Flash model card, August 2026. URL https://deepmind.google/models/model-cards/gemini-3-7-flash/.

Ken Gu, Ruoxi Shang, Ruien Jiang, Keying Kuang, Richard-John Lin, Donghe Lyu, Yue Mao, Youran Pan, Teng Wu, Jiaqian Yu, et al. Blade: Benchmarking language model agents for data-driven science. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 13936–13971, 2024.

Harbor Framework. Frontier-bench: Measuring and evolving with the frontier of agent work. GitHub repository, 2026. URL https://github.com/harbor-framework/frontier-bench. Evolving repository; accessed 2026-08-13.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, volume 2024, pp. 54107–54157, 2024.

Kimi Team. Kimi K3: Open frontier intelligence, 2026. URL https://arxiv.org/abs/2607.24653.

Jon M Laurent, Joseph D Janizek, Michael Ruzo, Michaela M Hinks, Michael J Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D White, and Samuel G Rodriques. Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, 2024.

Tianyu Liu, Allen Xin Wang, Antonia Panescu, Lisa Xinyi Chen, Wenxin Long, Xinyu Wei, Yueqian Jing, Ziyao Zeng, Jihang Chen, Sihan Jiang, et al. Benchmarking ai agents for addressing scientific challenges across scales. *arXiv preprint arXiv:2606.12736*, 2026.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*, volume 2024, pp. 52989–53046, 2024.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *International Conference on Learning Representations*, volume 2024, pp. 9025–9049, 2024.

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 preprint arXiv:2503.00096*, 2025.

Surag Nair, Laura Gunsalus, Brian Orcutt-Jahns, Jordan Rossen, Avantika Lal, Carlo De Donno, Muhammed Hasan Çelik, Kipper Fletez-Brant, Xiaoman Xie, Hector Corrada Bravo, et al. Agentic systems are adept at solving well-scoped, verifiable problems in computational biology. *bioRxiv*, pp. 2026–04, 2026.

OpenAI. Codex cli. Codex documentation, 2026a. URL https://developers.openai.com/codex/cli/. Accessed 2026-08-10.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. OpenAI product release, July 2026b. URL https://openai.com/index/gpt-5-6/. Accessed 2026-08-19.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Yuanhao Qu, Yingzhou Lu, Xinming Tu, Serena Zhang, Tianwei She, Alexander Glenn Shaw, Jou-Ho Shih, Bingqing Zhao, Minjie Shen, Haochen Yang, et al. Biomnibench: Process-level evaluation of llm agents for real-world biomedical research. *bioRxiv*, pp. 2026–05, 2026.

Qwen Team. Qwen3.5-397B-A17B. Hugging Face model repository, 2026. URL https://huggingface.co/Qwen/Qwen3.5-397B-A17B. Accessed 2026-08-20.

Yujiong Shen, Yajie Yang, Zhiheng Xi, Binze Hu, Huayu Sha, Jiazheng Zhang, Qiyuan Peng, Junlin Shang, Jixuan Huang, Yutao Fan, et al. Sciagentgym: Benchmarking multi-step scientific tool-use in llm agents. *arXiv preprint arXiv:2602.12984*, 2026.

Zachary S Siegel, Sayash Kapoor, Nitya Nadgir, Benedikt Stroebl, and Arvind Narayanan. Core-bench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *arXiv preprint arXiv:2409.11363*, 2024.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, et al. Paperbench: evaluating ai's ability to replicate ai research. *arXiv preprint arXiv:2504.01848*, 2025.

Liangcai Su, Zhen Zhang, Guangyu Li, Zhuo Chen, Chenxi Wang, Maojia Song, Xinyu Wang, Kuan Li, Jialong Wu, Xuanzhong Chen, et al. Scaling agents via continual pre-training. *ICLR 2026*, 2025.

Qiushi Sun, Zhoumianze Liu, Chang Ma, Zichen Ding, Fangzhi Xu, Zhangyue Yin, Haiteng Zhao, Zhenyu Wu, Kanzhi Cheng, Zhaoyang Liu, et al. Scienceboard: Evaluating multimodal autonomous agents in realistic scientific workflows. In *International Conference on Learning Representations*, volume 2026, pp. 75694–75731, 2026a.

Yiyou Sun, Xinyang Han, Weichen Zhang, Yuanbo Pang, Tianyu Wang, Yuhan Cao, Yixiao Huang, Chris Duroiu, Haoyun Zhang, Jeffrey Lin, et al. Agents' last exam. *arXiv preprint arXiv:2606.05405*, 2026b.

Brian Wang, Bin Feng, Xiaoman Pan, Chenyang An, Felix Liu, Tangqi Fang, Gongbo Sun, Lingfeng Shen, Ning Wang, Handuo Zhang, Feng Chen, Fuchao Yang, Xiang Wang, Jiacheng Lin, Siting Li, Zixuan Liu, Chi Han, Zhenhailong Wang, Kunlun Zhu, Lawrence Zhao, Yueqi Guo, Kailong Wen, Feng Xing, Yiling Guo, Lidong Bing, David Tan, Bo An, Heng Ji, and Sheng Wang. Apodex Discovery: Reality benchmarks and environments for evaluating and building discoverative artificial intelligence. *arXiv preprint arXiv:2608.11341*, 2026a. doi: 10.48550/arXiv.2608.11341. URL https://arxiv.org/abs/2608.11341.

Yuru Wang, Lejun Cheng, Yuxin Zuo, Sihang Zeng, Bingxiang He, Che Jiang, Junlin Yang, Yuchong Wang, Kaikai Zhao, Weifeng Huang, et al. Naturebench: Can coding agents match the published sota of nature-family papers? *arXiv preprint arXiv:2606.24530*, 2026b.

xAI. Introducing Grok 4.6. xAI product release, August 2026. URL https://x.ai/news/grok-4-6. Accessed 2026-08-19.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024.

Z.ai. GLM-5.2: Built for long-horizon tasks. Z.ai research release, June 2026. URL https://z.ai/blog/glm-5.2. Accessed 2026-08-19.

A Supplementary Analyses

A.1 Analysis coverage and trajectory definitions

The main results contain thirteen configurations on 97 tasks each (1,261 configuration–task records). HDS6 covers twelve configurations, excluding Frontier Agent, whose structured trajectories are unavailable. The archived failure-signature analysis also contains twelve configurations, but excludes GPT-5.6 Terra (max) instead. Its 1,164 records contain 1,011 non-passing runs. The two twelve-configuration analyses therefore have different membership. Final-message and tool-error analyses use the ten Claude Code configurations only, comprising 970 runs, of which 849 are non-passing and 121 are passing. These coverage definitions apply to the retained analysis records.

Lexical completion rule. The final message is the last nonempty text event in a Claude Code trajectory. Case-insensitive matching detects the English expressions "all done", "all deliverables", "complete", "completed", "finished", "done", and "successfully", together with their recorded Chinese completion expressions. The detector does not resolve negation, quotation, or completion of an individual step. Its frequency therefore measures completion-related wording, not a validated rate of claims that every deliverable is complete. A separate lexical rule flags running, waiting, or queued work; the two flags are not mutually exclusive. All runs remain in the denominator, including those with no nonempty final text event. The exact expressions are provided in `figs/build_failure_analysis.py`.

A.2 Heavy Duty Solver protocol and complete results

Protocol and coverage. Following the HDS6 evaluation framework introduced in *Apodex Discovery* (Wang et al., 2026a), we assess six Heavy Duty Solver capabilities separately from final-task success. We apply this framework through the archived FrontierChallenge HDS6 campaign from September 7, 2026. For each task and configuration, one valid trial was selected for the frozen trajectory archive. This describes the retained analysis unit, not the number of attempts made before selection. The assessment complements the artifact-based completion metrics in Section 4.2. It covers 12 model–scaffold configurations on all 97 tasks (1,164 scorecards). Apodex 1.1 is included only with Claude Code: the Frontier Agent runs lack the structured ATIF trajectories required by this assessment. The partially covered GLM-5.3 configuration is also excluded. These results therefore do not constitute an HDS6 assessment of every scaffold in the main benchmark.

The six dimensions are *Coherence* (long-horizon state and cross-artifact consistency), *Evidence* (fidelity to executed observations), *Alternatives* (hypothesis and method management), *Scope* (boundary and failure reasoning), *Tools* (tool use and execution-state management), and *Repair* (self-correction under verification). Each task uses 13 shared rubric items plus 6–10 task-specific items with 0/1/2 anchors; the archived scorer reports overall and capability scores on a 1–4 base scale. A reviewed opportunity matrix determines which capabilities are applicable. Missing opportunities are excluded rather than assigned zero: Alternatives and Repair are marked not applicable on 47 tasks each. The judge consumes converted trajectories and task packs, without direct access to the verifier's reference outputs. Private reasoning channels are stripped for all configurations. The archived TERRA panel uses GPT-5.6 Terra as judge and reviewer, GPT-5.6 Luna as sweeper, and Claude Opus 5 as arbiter.

Measurement conditions. The archived task-specific rubric packs were not calibrated against expert trajectory ratings. The scoring pipeline capped each tool observation at 8,000 characters, and trajectory conversion omitted inline narration. Consequently, visible evidence can differ with observation length and reporting style. We interpret HDS6 values and their outcome correlations as exploratory summaries under this protocol, rather than calibrated capability estimates. Capability-specific opportunities and anchors also differ, so lower Scope scores do not by themselves establish a larger underlying capability deficit.

Table 3: Exploratory Heavy Duty Solver (HDS6) process assessment. Scores are coverage-adjusted; higher is better. C: Coherence; E: Evidence; A: Alternatives; S: Scope; T: Tools; R: Repair. V/N denotes gate-passing graded runs / available runs.

Model	Scaffold	HDS6	C	E	A	S	T	R	V/N
Claude Opus 5	Claude Code	3.768	3.890	3.951	3.823	3.508	3.704	3.653	97/97
GPT-5.6 Sol	Codex	3.670	3.874	3.919	3.431	3.341	3.641	3.533	96/97
Qwen 3.8 Max	Claude Code	3.624	3.845	3.842	3.472	3.186	3.601	3.607	96/97
GPT-5.6 Terra (max)	Codex	3.612	3.818	3.876	3.413	3.184	3.582	3.560	95/97
Grok 4.6	Claude Code	3.562	3.777	3.853	3.436	3.196	3.523	3.338	92/97
DeepSeek V4 Pro-0813	Claude Code	3.536	3.787	3.857	3.367	3.003	3.536	3.427	97/97
DeepSeek V4 Flash-0731	Claude Code	3.506	3.774	3.767	3.226	3.056	3.488	3.434	92/97
Kimi K3	Claude Code	3.496	3.776	3.815	3.316	2.996	3.523	3.252	95/97
Apodex 1.1	Claude Code	3.293	3.460	3.544	3.262	2.853	3.304	3.300	82/92
GLM-5.2	Claude Code	3.152	3.302	3.532	3.176	2.652	3.232	2.950	89/97
Gemini 3.7 Flash	Claude Code	3.016	3.258	3.451	2.854	2.438	3.098	2.499	90/97
Qwen3.5-397B-A17B	Claude Code	2.643	2.822	2.998	2.341	2.185	2.683	2.410	69/97

Aggregation. Let A_m be the available runs for configuration m, and $V_m \subseteq A_m$ its graded runs passing the HDS6 integrity gate. This gate is distinct from the benchmark's task-completion threshold. For archived overall score h_{mi}, we reproduce the campaign's coverage-adjusted score:

$$H_m = \left(\frac{1}{|V_m|} \sum_{i \in V_m} h_{mi} \right) \sqrt{\frac{|V_m|}{|A_m|}}. \tag{3}$$

Tasks receive equal weight. For each capability, we average its available scores within V_m and apply the same configuration-level multiplier. Gate failures remain in A_m but do not contribute to the score mean. Five Apodex 1.1 runs containing only an API authentication failure have no process grade and are removed from both sets, giving $|A_m| = 92$ for that configuration and 97 for every other configuration. The original outcome metrics retain their original 97-task denominator. Coverage adjustment can reduce a score below the base scale's lower bound; it is not a pass rate.

Complete configuration-level results. Claude Opus 5 achieves the highest HDS6 score (3.768), followed by GPT-5.6 Sol (3.670), Qwen 3.8 Max (3.624), and GPT-5.6 Terra (3.612). Evidence is the highest or nearly highest dimension for every configuration, whereas Scope is consistently the lowest (2.185–3.508). Apodex 1.1 with Claude Code scores 3.293, above Qwen3.5-397B-A17B (2.643), GLM-5.2 (3.152), and Gemini 3.7 Flash (3.016), but below Kimi K3 (3.496) and DeepSeek V4 Flash-0731 (3.506). Its largest gaps relative to Claude Opus 5 are in Scope (2.853 versus 3.508) and Alternatives (3.262 versus 3.823).

Coverage sensitivity. Using the published configuration summaries for the same twelve systems, the unadjusted HDS6 means correlate with Avg. Score at approximately $r = 0.86$ and with Pass Rate at $r = 0.82$. The corresponding coverage-adjusted correlations are approximately 0.87 and 0.83. These calculations use rounded outcome summaries, so they are reported to two decimal places. The positive association persists without the coverage multiplier; this check does not calibrate the rubric or recover excluded trajectories.

A.3 Extended domain profiles

GPT-5.6 Sol had the broadest leading profile, with the highest Avg. Score in quantum chemistry, materials characterization, analytical chemistry, and life science. Grok 4.6 achieved the highest quantum-chemistry Pass Rate (60%) and shared the highest molecular-dynamics Pass Rate (38%). The strongest Apodex 1.1 configuration, Frontier Agent (Agent Team), was more competitive in quantum chemistry (40% Pass Rate) than in molecular dynamics (19%) and did not complete a task in materials characterization, analytical chemistry, or electrochemistry/environment. Terra's aggregate deficit to Sol was concentrated in quantum chemistry: it scored 75.5 versus 91.7 and completed

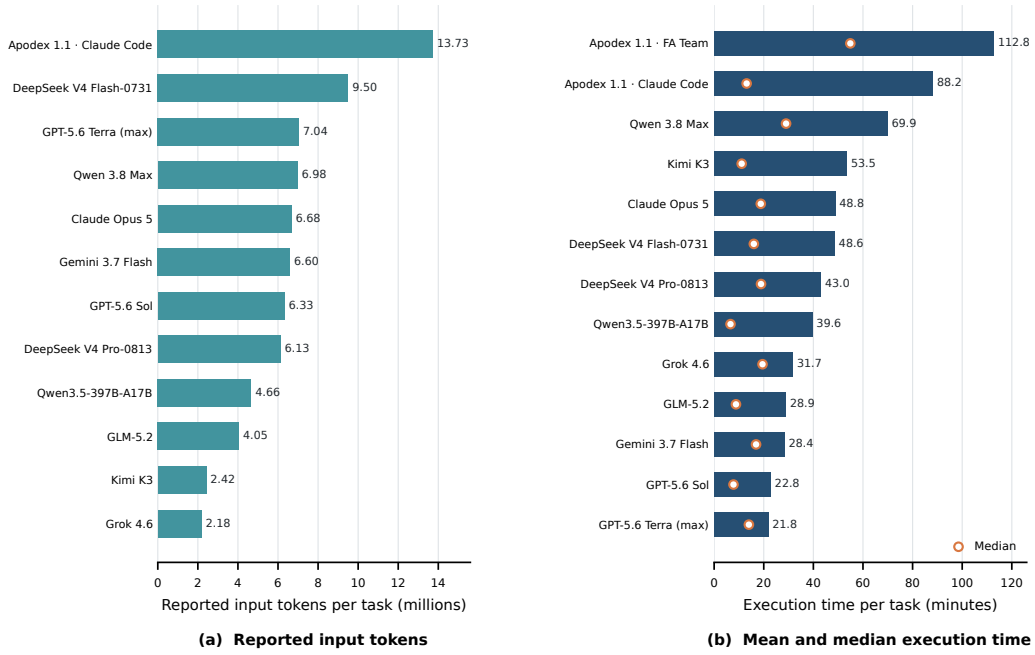


Figure 5: Reported resource use and observed execution time. (a) Token accounting for twelve configurations with available records. (b) Execution-time summaries across evaluated configurations, including means and medians. Provider-specific tokenization, caching, and reporting prevent a hardware-normalized efficiency interpretation.

35% versus 55% of tasks, whereas Terra exceeded Sol in molecular-dynamics Avg. Score (93.7 versus 88.9) and electrochemistry/environment Avg. Score (94.5 versus 91.8). Gemini 3.7 Flash also varied across domains: its Avg. Score ranged from 62.4 in analytical chemistry to 81.1 in quantum chemistry, and it completed 25% of tasks in both quantum chemistry and molecular dynamics but none in materials characterization, analytical chemistry, or electrochemistry/environment. These profiles show that the aggregate ranking did not capture domain-specific strengths and weaknesses.

A.4 Resource use and execution time

Reported token use varied by more than sixfold. Among the twelve configurations with token records, reported input use ranged from 2.183 million tokens per task for Grok 4.6 with Claude Code to 13.730 million for Apodex 1.1 with Claude Code. GPT-5.6 Sol used 6.327 million reported input tokens and 23.1 thousand output tokens per task. GPT-5.6 Terra (max) used 7.039 million input tokens and 38.6 thousand output tokens per task. Their reported cache shares were 98.8% and 98.5%, respectively. Provider-specific tokenizers, caching, and reporting conventions preclude a hardware-normalized efficiency interpretation. Gemini 3.7 Flash used 6.599 million reported input tokens and 25.8 thousand output tokens per task, with a 75.3% cache share (Fig. 5a).

Execution times varied substantially and showed long tails. Across the evaluated systems, mean execution time ranged from 21.8 to 112.8 minutes per task. Frontier Agent (Agent Team) had the longest mean time at 112.8 minutes, with a median of 54.9 minutes and a 90th percentile of 272.2 minutes. Terra (max) had a higher median than Sol (14.0 versus 7.8 minutes) but a lower mean (21.8 versus 22.8), lower 90th percentile (35.0 versus 46.1), fewer trajectories over 60 minutes (5 versus 7), and fewer total machine hours (35.3 versus 36.9), indicating a shorter observed tail. Gemini 3.7 Flash averaged 28.4 minutes per task, with a median of 16.9 minutes, a 90th percentile of 71.9 minutes, and 11 trajectories exceeding 60 minutes. These aggregate summaries describe observed runtimes rather than isolated model speed.

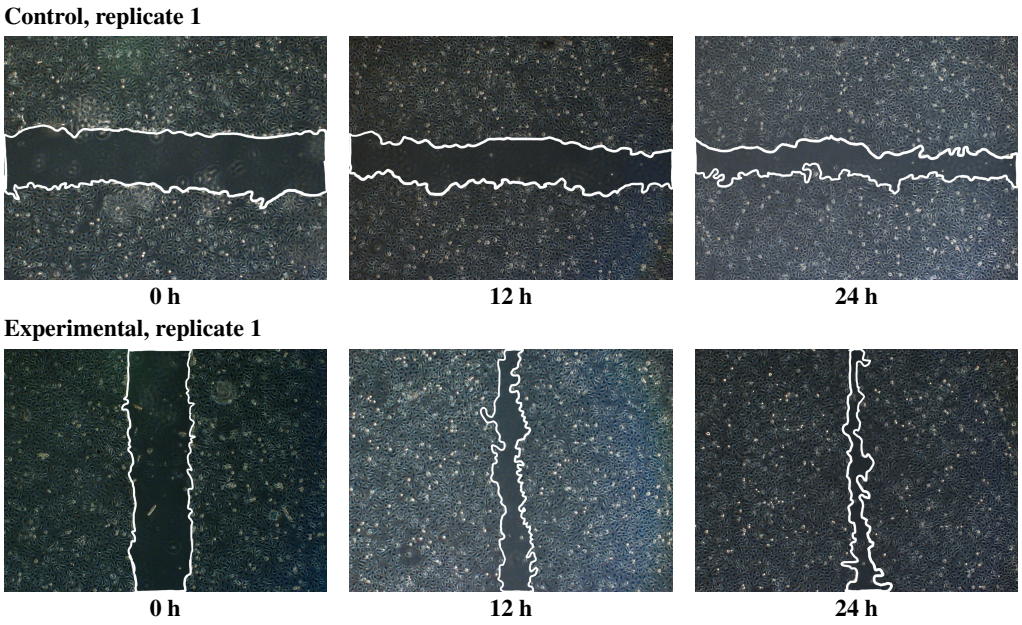


Figure 6: Agent-visible input from the wound-healing case. Columns are the three supplied time points; rows show one biological replicate from each group. White contours were drawn by the experimenter in the original images to delineate the unmigrated region. The remaining twelve images supply two additional replicates for every group–time combination.

B Illustrative Task Case

The following three cases illustrate the benchmark unit at the level visible to an evaluated system: a scientific objective, a frozen input bundle, a required workflow, and an artifact contract. They intentionally disclose neither model submissions nor scores, evaluator code, hidden checks, expected outputs, or reference artifacts. Accordingly, the examples characterize what must be done and handed off, not the answer to any task.

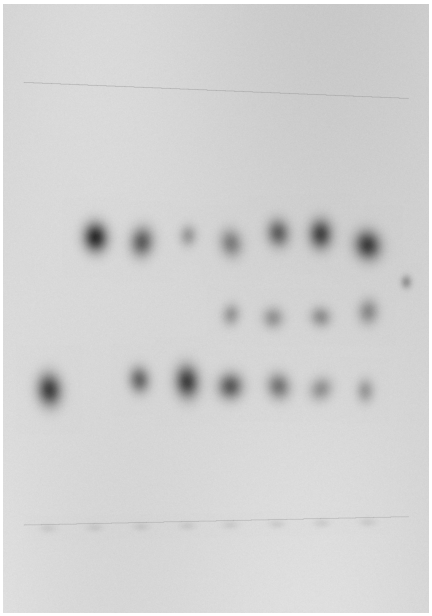
B.1 Cell-migration wound-healing assay

Task and visible evidence. This case is an image-analysis and statistical workflow built from 18 supplied bright-field microscopy images. The frozen input crosses two groups (control and experimental), three time points (0, 12, and 24 hours), and three biological replicates. In every image, the experimenter has already marked the closed boundary of the unmigrated region in white. Figure 6 shows one complete time course from each group exactly as provided to the system; the outlines are part of the source images, not generated results.

Expected handoff. The system must segment the marked unmigrated region in all 18 images, inspect segmentation quality, pair time points by group and replicate, and compute the 12-hour and 24-hour migration rates relative to the matching 0-hour image. It then performs separate between-group tests at both follow-up times. The checkable handoff consists of a reproducible analysis script, an image-level table, a group-statistics table, two labeled summary plots, and a methodological report. No measured area, migration rate, test statistic, or conclusion is shown here.

B.2 TLC monitoring of a Suzuki coupling

Task and visible evidence. This case is an analytical-chemistry image workflow centered on a supplied UV254 thin-layer chromatography plate. Its eight lanes comprise starting- material and product standards, a co-spot, and reaction samples collected at 0, 15, 30, 60, and 120 minutes. The frozen input also supplies the plate map, compound-reference Rf windows and response factors, sampling metadata, image geometry, integration settings, and the endpoint rule. The spotting line



Agent-visible plate map

Lane	Sample	Time
L01	Starting-material standard	–
L02	Product standard	–
L03	Co-spot	–
L04	Reaction sample	0 min
L05	Reaction sample	15 min
L06	Reaction sample	30 min
L07	Reaction sample	60 min
L08	Reaction sample	120 min

Dark UV-quenching spots are the measured image signal. Lane centers, approximate line endpoints, compound search windows, response factors, and the endpoint criterion are supplied separately as structured input.

Figure 7: Unmodified agent-visible TLC input for the Suzuki- coupling case, shown with the supplied lane map. The plate contains standards, a co-spot, and a five-time-point reaction series. No detected spot boundary, Rf assignment, corrected intensity, composition, or endpoint decision is displayed.

and solvent front are intentionally tilted, so Rf must be computed from the local geometry at each lane.

Expected handoff. The system must detect and integrate spots after local background correction, assign compounds using the standards and co-spot, calculate lane-specific Rf values, apply the declared response correction, reconstruct reaction progress, and recommend an endpoint under the supplied rule. Its handoff includes a reproducible script; plate-QC, spot-level, lane-composition, reaction-progress, and endpoint tables; an annotated plate; lane profiles; and a report. Figure 7 is the unmodified agent-visible plate image and contains no derived annotations, integrated intensities, conversion values, or endpoint.

B.3 Reaction-calorimetry safety assessment

Task and visible evidence. This case is a hard process-chemistry and thermal-safety workflow. Its frozen input contains electrical calibration pulses, blank-dosing experiments, reaction runs, run and protocol metadata, and a configuration containing the screening rules. The left side of Figure 8 plots three representative raw input traces exactly as supplied; no baseline correction, integration, derived metric, or protocol decision is displayed.

Expected handoff. The analysis must calibrate and correct heat flow, derive run-level metrics, summarize replicates by protocol, and apply the declared screening criteria. A reproducible script must produce mutually traceable calibration, correction, time-series, run-metric, protocol-summary, and decision tables, together with diagnostic figures and a report. The right side of Figure 8 shows this audit chain rather than its outcome.

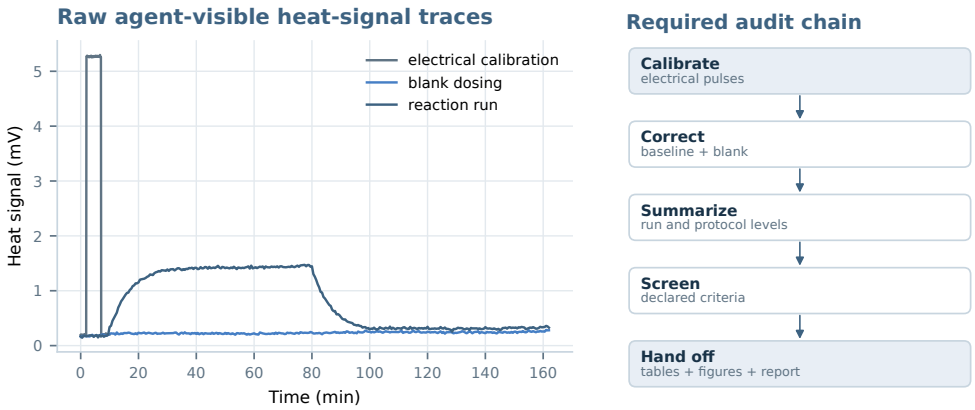


Figure 8: Answer-free view of the reaction-calorimetry case. Raw agent-visible signal traces provide visual context, while the schematic records the required path from calibration evidence to a complete handoff.

C Dataset Package and Reference Solutions

Task package. Each of the 97 task directories supplies instructions, metadata, an environment definition, and required inputs. The machine-readable manifest indexes tasks by identifier, domain, workflow family, difficulty, and runtime track, with content hashes for version identification. The registry links each task package to its evaluator archive, which contains the task-specific Grader and reference artifacts.

Integrity verification. All 97 task packages passed checksum, identifier, required-file, and manifest–instruction consistency checks. We also verified checksums and authenticated decryption for all 97 evaluator archives. Each contains a Grader and a nonempty reference-output directory; 77 contain a Judge runner. These checks verify package integrity and exclude the runtime image archive; they are separate from scientific validation and reference-solution execution.

Runtime requirements. The release has 81 open-runtime tasks and 16 requiring a locally provisioned ORCA runtime; ORCA binaries are not redistributed. Task definitions specify resource, network, timeout, and verifier defaults. Reproduction records should identify any overrides and the software versions used.

Reference solutions and evaluator separation. We will provide complete task packages and a reference solution for each of the 97 evaluated tasks, documenting the workflow from inputs to scientific deliverables. Reference solutions and evaluator artifacts are organized separately for auditing and reproduction and remain outside the evaluated agent's visible inputs.

Archived Judge controls. The evaluator archives retain paired positive and negative Judge controls for 63 tasks, using `qwen/qwen3.7-flash`. The archived positive-control scores are all 100/100; negative-control scores range from 0 to 3.5/100. These are packaging-time controls from a different Judge, rather than independent expert assessments of the main experiment. In 14 of the 63 tasks, at least one control record's prompt hash differs from the distributed prompt. The archived scores therefore describe the recorded model and prompt versions, not calibration of the current Grader–Judge pipeline.

Pass criterion. A task passes when evaluation is complete and its score is at least 99.9 out of 100 (equivalently, a normalized score of at least 0.999). The evaluation code and reported Pass Rate use this criterion.