FRONTIERPHYSICS: EVALUATING AGENTS FOR END-TO-END PHYSICS RESEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly participating in substantial parts of the physics research process. However, current physics benchmarks test fixed questions, isolated subtasks, or paper reproduction with given solutions, and rarely grade research quality alongside correctness. We introduce FRONTIERPHYSICS, a benchmark evaluating the ability of AI agents to conduct end-to-end research across theoretical, computational, and experimental physics and nine domains. The benchmark comprises 56 tasks made by 42 physics researchers, all from task authors' personally conducted research. In each task, agents are asked to write a scientific manuscript as well as structured output files. To succeed in these tasks, agents need to conduct in-depth research on the given topic to formulate proper research plans, select right methodologies, and produce a publishable manuscript of their results. For comprehensive evaluation, FRONTIERPHYSICS combines deterministic verifiers with rubric-based assessment. The former validates structured outputs against reference oracles, while the latter evaluates the agents' produced artifacts and trajectory according to both expert-authored rubrics evaluating the publish-ability of results and universal rubrics checking behavioral alignment. This evaluation protocol assesses the accuracy of the computation, scientific rigor, physics understanding, and scientific manuscript writing. Across eleven frontier agents, the best passes only 22% of rollouts, and seven pass at most 3%, indicating end to end physics research remain challenging to current models. Additionally, we conduct a detailed human study of the agents failure taxonomy by manually analyzing trajectories and identify failure modes. We find lack of rigorous thinking and expert intuition, instead of physics knowledge gap or execution mistakes as being the main weaknesses of frontier LLM agents on our benchmark.

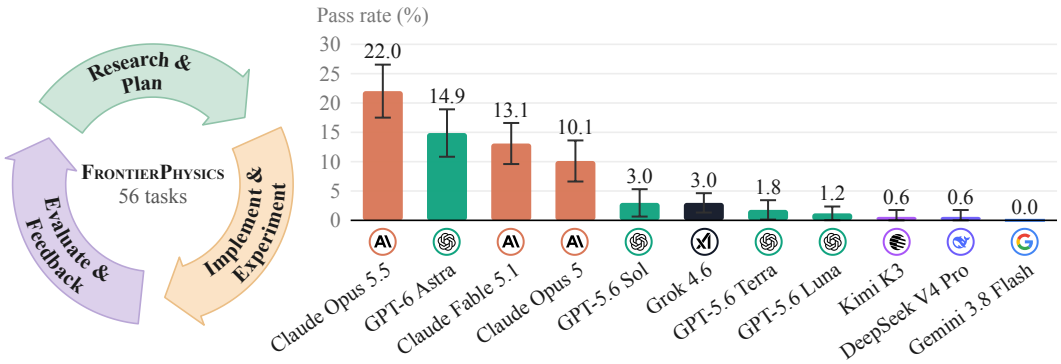

Figure 1: **FRONTIERPHYSICS LEADERBOARD.** Left: End-to-end discovery loop, which spans research and planning, implementation and experimentation, and evaluation and feedback through self-validation and feedback information from environments. Right: overall Avg@3 pass rate under the Table 2 configurations; a pass requires passing all verifier tests and blocker rubrics. Harnesses and inference level: GPT (Codex, xhigh), Claude (Claude Code, max), Grok (Grok Build, xhigh), Gemini (Antigravity, high), Kimi (Claude Code, max), DeepSeek (Claude Code, max). Error bars show approximate 95% confidence intervals for mean pass rate from run-to-run variation.

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.

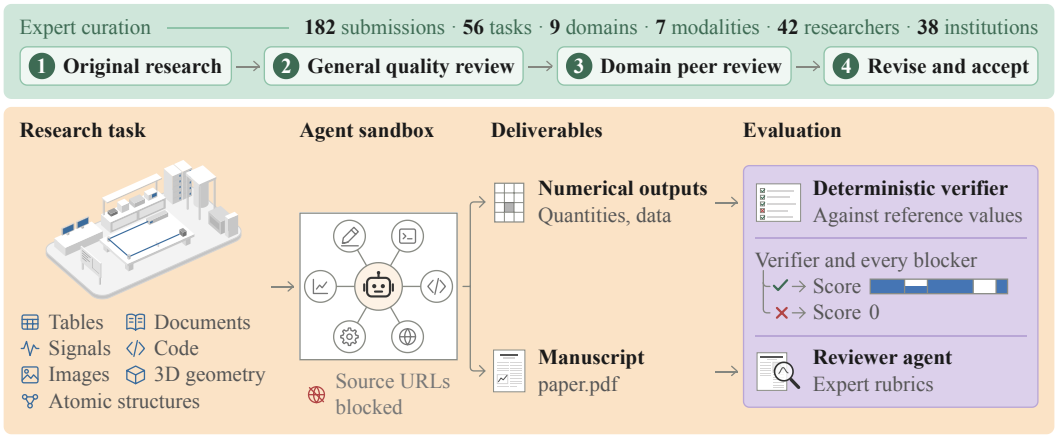

Figure 2: **Overview of FRONTIERPHYSICS.** Top: FRONTIERPHYSICS tasks QA pipeline. Contributors turn their own physics research projects into tasks which go through general quality checks first, and then domain experts peer review to ensure good task quality. Bottom: FRONTIERPHYSICS tasks structure. Given a PI-style research background and task instruction, as well as real lab data and materials in the environment, agents conduct research end-to-end and produce a manuscript and outputs files. Evaluation combines deterministic verification with rubric-guided assessment of physics understanding, literature review, research practices, and the publishability of final manuscripts. A rollout that passes the verifier and every blocker rubric scored with its weighted non-blocker rubrics.

1 INTRODUCTION

Large language models (LLMs) are increasingly evolving from passive scientific assistants into agents capable of participating in substantial parts of the research process. When equipped with external tools and agentic scaffolding, these agents can search and synthesize literature, formulate hypotheses, write and debug scientific code, and produce technical reports. Recent systems have demonstrated tool-augmented reasoning in chemistry, iterative code-based experimentation, multi-agent hypothesis generation, scientific-software optimization, and even interaction with laboratory instruments (M. Bran et al., 2024; Boiko et al., 2023; Lu et al., 2024; Schmidgall et al., 2025; Gottweis et al., 2026; Aygün et al., 2026; Novikov et al., 2025; Swanson et al., 2025; Mandal et al., 2025). Together, these advances suggest a path toward AI agents that can move beyond assisting with isolated subtasks and contribute to increasingly complete scientific workflows.

Evaluating progress toward this goal requires benchmarks that capture the open-ended, multi-stage nature of scientific research. Scientific benchmarks are accordingly shifting from textbook-level questions (Feng et al., 2025; Huang et al., 2024; Xu et al., 2025; Qiu et al., 2025; Wang et al., 2026b) toward research-level problems. Among existing research-level benchmarks, some assess scientific reasoning (Chung et al., 2025; Zhu et al., 2025; Pan et al., 2026), while others evaluate an agent's ability to complete individual tasks within realistic scientific workflows (Wang et al., 2026a; Hu et al., 2026; Mandal et al., 2025; Liu et al., 2026; Cao et al., 2026). Other benchmarks move beyond these isolated capabilities toward end-to-end research by formulating tasks as full-paper reproductions (Qiu et al., 2026; Deng et al., 2026; Ye et al., 2025; Faroughy et al., 2026). However, because agents are given the original publications, these benchmarks substantially specify the research questions and solution structures, limiting their ability to evaluate open-ended scientific inquiry. More recent benchmarks address this limitation by transforming published studies into research problems without directly revealing their solutions (Miao et al., 2026; Wang et al., 2026c; Xu et al., 2026; Jiang et al., 2026). Nevertheless, a fully agentic benchmark made and curated by authentic researchers for their own end-to-end physics research, together with a comprehensive evaluation framework for open-ended research outcomes, remains missing.

In this paper, we introduce **FRONTIERPHYSICS** (Fig. 1), a benchmark for evaluating whether AI agents can conduct **end-to-end** research across theoretical, computational, and experimental physics. For each task, an agent receives a high-level prompt that resembles the instructions provided by a principal investigator (PI) at the beginning of a research project. The agent must then

Table 1: **Comparison of scientific research benchmarks.** We compare scientific research benchmarks across the research authenticity, research components covered in the evaluation, and the evaluation methods. Meanings of labels: **Tasks/Phys.**: total/physics task counts in released task lists. **Own Research**: the tasks come from the contributors' own research. **No Reference Given**: there is no reference (e.g., the original paper) provided to the agent. **Literature Graded**: the quality of agents' literature review and synthesis is evaluated. **Write-up Graded**: agents write up scientific manuscripts and the manuscripts are graded. **Expert Rubric**: at least part of the evaluation protocol uses rubrics created by domain experts. **Deterministic Verifiers**: at least part of the evaluation protocol contains deterministic verifiers. ✓: documented; ○: partial; ×: no;

Benchmark	Tasks	Phys.	Research Authenticity		Research Coverage		Evaluation	
			Own Research	No Reference Given	Literature Graded	Write-up Graded	Expert Rubric	Deterministic Verifiers
Reasoning and QA								
CritPt (Zhu et al., 2025)	71	71	✓	✓	×	×	×	✓
FrontierScience-Research (Wang et al., 2026a)	60	20	✓	✓	×	×	✓	×
Subtasks from scientific workflows								
Terminal-Bench-Science (Terminal-Bench-Science Team, 2026)	70	5	✓	✓	×	×	×	✓
Paper reproduction								
PRBench (Qiu et al., 2026)	30	30	○	×	×	×	✓	×
ReplicationBench (Ye et al., 2025)	111	111	○	○	×	×	✓	✓
QMP-Bench (Deng et al., 2026)	100	100	×	×	×	×	✓	✓
End-to-end research								
NatureBench (Wang et al., 2026c)	90	13	×	✓	×	×	×	✓
PRL-Bench (Miao et al., 2026)	100	100	×	✓	×	×	✓	✓
ResearchClawBench (Xu et al., 2026)	40	4	×	✓	×	✓	✓	×
PhySciBench (Jiang et al., 2026)	200	100	×	✓	○	×	✓	✓
FRONTIERPHYSICS	**56**	**56**	✓	✓	✓	✓	✓	✓

carry out the research and produce both a scientific manuscript and numerical outputs that can be verified deterministically. Much of this process is **open-ended**: the agent must investigate the topic, formulate a research plan, select appropriate methods, validate the results, and establish the significance of its findings. To evaluate these capabilities comprehensively, FRONTIERPHYSICS combines deterministic verification with rubric-based assessment. The rubric-based assessment includes a global rubric and task-specific rubrics. The global rubric evaluates behavior alignment including reward hacking, hallucination and general research practices including literature synthesis, result validation, and the use of evidence to support scientific claims, whereas task-specific rubrics are developed by domain experts to assess the agent's understanding and execution of the underlying physics.

Our contributions are as follows:

- **FrontierPhysics**: A benchmark comprising 56 end-to-end physics research tasks across nine domains, which are designed directly by active physics researchers based on the research projects they have personally conducted. To preserve open-endedness while minimizing false negatives and positives, each task undergoes extensive quality checking, including format checking, domain-expert peer review, and failure mode analysis.
- **Comprehensive evaluation**: We design a comprehensive evaluation scheme that combines deterministic numerical checks with qualitative rubrics. This experts grounded publishability check assesses the accuracy of the computation, scientific rigor, physics understanding, and the quality of scientific manuscript writing.
- **Failure mode analysis**: Our expert-built taxonomy shows that GPT-6 Astra fails mainly for a lack of rigorous thinking and missing expert intuition, instead of physics knowledge gap or incapable execution. (Section 3.3).

2 FRONTIERPHYSICS

2.1 SOURCE

FrontierPhysics tasks are designed by active physics researchers-including Ph.D. students, postdocs, and research scientists-based on the end-to-end research projects they have personally conducted.

To establish the authenticity and difficulty of each task, contributors are asked to provide the project background, a timeline of the original research process, and the amount of time they spent completing it. We require the underlying project to have involved at least two weeks of research effort. Contributors also estimate how long a first-year physics Ph.D. student would require to complete the task given the `task.md`. This estimate is expected to be shorter than the duration of the original project because problem formulation, which can constitute a substantial portion of scientific research, has already been completed and encoded in the task specification. So, we require it to be at least 10 hours.

In total, 56 tasks were accepted from 42 physics researchers from 38 institutions worldwide in 9 different physics domains. A summary can be seen in Fig. 4. For 18 tasks, the underlying research has been published in papers co-authored by the task contributors. These include four tasks based on papers in Nature, with other publication venues including Nature Communications, Physical Review Letters, and Physical Review X. On average contributors report to have spent about 250 days for the original projects.

2.2 Task Design

FrontierPhysics tasks are designed to reconstruct end-to-end research workflows. A task example is provided in appendix Figures 7 and 8. Each task will first provide agents with the research background and motivations, resembling the guidance that a principal investigator (PI) might provide to a graduate student at the beginning of a project. Some important questions to figure out or directions to explore might be given, but no explicit methodology or solution pathways will be specified. The agent must then independently plan and conduct the investigation, including the literature review, experiment planning, implementation and data analysis. When necessary, input data and other supporting files are provided in the task environment to enable the agent to configure and execute the relevant computational or experimental workflow.

Each task requires two types of final deliverables: deterministic artifacts, such as computed quantities, simulation outputs, or processed data, and a paper manuscript submitted as `paper.pdf`. The paper is expected to meet the standards of a peer-reviewed paper by clearly presenting the research context, methodology, results, and conclusions, while faithfully incorporating the evidence and computational artifacts produced during task execution.

2.3 Evaluation

As described in Section 2.2, each task requires two types of final deliverables: deterministic outputs and a scientific manuscript, `paper.pdf`. Our evaluation scheme is designed to assess both the objective correctness of the agent's results and the overall quality of its research. To this end, we combine programmatic verifiers with an LLM-as-judge. The programmatic verifier checks structured outputs against reference values, while the LLM-based judge evaluates the agent's trajectory and artifacts according to an expert-authored `rubric.json`.

Rubric-based grading has been widely adopted in benchmarks involving open-ended exploration or qualitative dimensions of agent performance (Starace et al., 2025; Arora et al., 2025; Fan et al., 2024; Qiu et al., 2026; Cognition, 2026). Whereas deterministic verification establishes whether an output is correct, rubric-based evaluation assesses whether the agent conducted and communicated the research well. Inspired by the rubric design of FrontierCode (Cognition, 2026), we divide our rubric criteria into *blockers* and *non-blockers*. Blockers are binary gates that represent requirements essential to the scientific validity of a task; an agent must satisfy every blocker and the deterministic verifier to pass. Non-blockers capture qualitative aspects of the work. Each non-blocker is scored on a three-level scale: 0 indicates that the requirement is absent or incorrect, 1 indicates partial satisfaction, and 2 indicates complete satisfaction. The precise anchors for these levels are defined separately for each criterion. Every criterion is also assigned an integer *weight* between 1 and 10, reflecting its relative importance. Blocker weights are retained as rubric metadata but do not contribute to the numerical quality score. The final score of a task:

$$S_{\text{final}} = \begin{cases} \dfrac{\sum_{i \in \mathcal{N}} w_i s_i}{2 \sum_{i \in \mathcal{N}} w_i} \times 100, & \text{if } D = 1 \text{ and } b_j = 1 \ \ \forall j \in \mathcal{B}, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

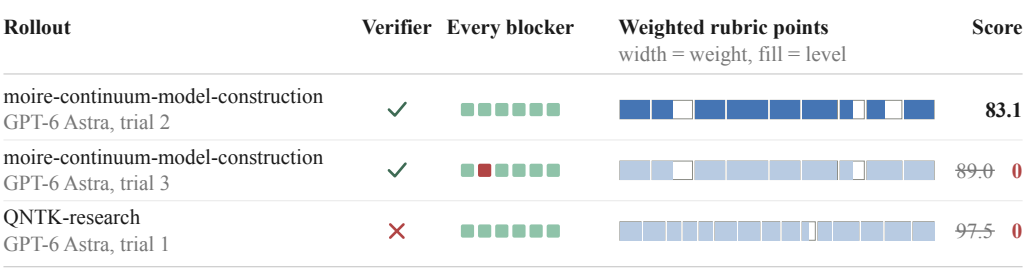

Figure 3: **Scoring three real rollouts** (GPT-6 Astra). A rollout's Score is the filled share of its bar, counted only if the verifier and every blocker pass, and 0 otherwise. Each cell is one weighted rubric criterion, as wide as its weight and filled to the level the reviewer agent awarded (0, 1 or 2). Trial 3 earns more rubric points than trial 2 but fails one blocker, so it scores 0.

Here, $\mathcal{N}$ and $\mathcal{B}$ denotes non-blocker and blockers respectively; $s_i \in \{0, 1, 2\}$, w_i is the criterion weight, D is the deterministic verifier result, and b_j is the blocker result.

$S_{\text{final}} = 0$ indicates that the results are not publishable, while the value of S_{final} when $S_{\text{final}} \neq 0$ measures the quality of the publication. High scores on non-blockers cannot compensate for an incorrect deterministic result or a failed blocker. Conversely, an agent may pass all deterministic checks while producing a poorly motivated, inadequately supported, or badly communicated manuscript. Such a submission passes the deterministic verifier but receives a low S_{final}. Figure 3 applies the rule to three real rollouts.

We note that physics research is highly specialized, so rubric design requires substantial task-specific expertise. We therefore ask each contributor to design a rubric tailored to their task and underlying research project. Depending on the task, the rubric criteria may assess whether the agent identifies and correctly uses key references, discovers the key scientific ideas through its investigation, considers realistic experimental constraints, or provided good motivations for its work in the paper. Each rubric is subsequently reviewed by experts in closely related domains to avoid false negatives/positives. We describe this review process in Section 2.4.

2.4 Task Curation and Quality Review

Rubrics consist largely of qualitative statements authored by task contributors, and then sent into review agents prompt to generate rubrics grading results. Since rubrics grading are subjective, the grading results may contain false negatives or positives. We therefore subject every submitted task to a two-stage quality-assurance process, comprising a general quality review and a domain experts peer review. Only tasks that pass both review stages can be considered to merge.

General quality review. When a contributor submits a task pull request, we first assess its authenticity and provenance: the contributor must provide supporting information (e.g., reference papers or patent links) to demonstrate that they have personally conducted the project. We then check that `task.md` and `rubric.json` are human-readable and contain genuine human input, rejecting tasks that appear purely AI-generated. Finally, we verify that the task conforms to our format specification and that it has been shown to be difficult for state-of-the-art (SOTA) agents instead of being too trivial and contains no valuable failure modes.

Domain peer review. We aim to have every task reviewed by an expert in a closely related field. To scale this, we identify strong contributors, Ph.D.s, postdocs, and researchers with substantial experience, from the general-review stage and invite them to become domain reviewers once their tasks pass the first stage general quality check. Each new reviewer is onboarded through a one-on-one meeting, and we audit the quality of their first review to align standards across reviewers. When a domain reviewer judges a task ready to merge, we perform a final check against our standards.

Each domain reviewer is provided with a domain-review checklist which covers three aspects: First, the reviewer verifies that `task.md` does not over-specify the problem: the prompt should not contain domain knowledge that a competent researcher is expected to know or figure out, so that the

task remains as open-ended as possible. Second, the reviewer confirms that `rubric.json` encodes real physics insight that only someone who has conducted the research would possess, rather than the low-signal criteria typical of AI-generated rubrics. Third, the reviewer examines failure modes after getting agents trajectories: contributors must perform at least three local runs with SOTA agents and an initial failure-mode analysis, and the reviewer inspects these traces for finding potential false positives and false negatives. Common issues include unjustified numerical tolerances that fail agents implementing a correct physical model or ambiguous instructions that lead to agents' misunderstanding etc.

2.5 Assessing Verifier and Rubric Reliability

Despite all the effort to ensure the quality of the verifiers and rubrics, some issues may remain. Designing and reviewing these can be very challenging. One major source of defects is adversarial verifier or rubric engineering designed to suppress task pass rates, such that even correct solutions produced by human experts may fail. To mitigate this, we have a three stage process:

First, we ask every task contributor to do a failure-mode report of the three GPT-6 Astra runs by analyzing the agent trajectories and output files, and explaining where the agent failed and why. Second, a core team of six physics Ph.D. reviewers re-read these contributors' failure-mode reports and double check them against the agent trajectories and output files. After reviewing 20 tasks, reviewers consolidated the recurring failures into the taxonomy in Table 4. More false negatives are found during this process and are labeled in E. Note that, as highlighted in Table 5, an agent may falsely fail one test while legitimately failing others. Therefore, only runs with exclusively false failures (i.e., only FALSE labels) lead to an underestimated pass rate. In total, only nine of the 168 runs incorrectly failed otherwise correct answers due to false negatives.

The non-blocking rubric scores are harder to assess because they involve more subjective judgments. We did not systematically rescore these by hand, so the exact scores should be interpreted with more caution than the failures. As a sanity check of the rubric quality, we asked authors, where possible, to include their own papers, manuscripts, or supporting calculations as reference materials in the oracle. Thirty-three tasks included these materials. For those with available reference materials, we evaluated all the manuscript-level rubric criteria, including blocking and scoring criteria against the authors' reference work. The average score of the LLM based rubric grading for these tasks is 96.2%, so the rubrics and the grading seem to work quite well for the human expert material. The exact data of this analysis can be found in Table 7.

3 Experiments

3.1 Experimental Setup

As in SkillsBench Li et al. (2026b) and ClawsBench Li et al. (2026a), we use BenchFlow CLI as eval harness to run experiments as well as track agent trajectories, verifier and rubrics results. BenchFlow CLI uses the Agent Client Protocol (ACP) to control and communicate with agent harnesses. We use Codex CLI for GPT models, Claude Code CLI for Claude models, Grok build for Grok models, Antigravity CLI for Gemini models. For all open source models we use Claude Code CLI as agent harness, including DeepSeek V4 Pro, and Kimi K3. Each agent runtime is started in parallel cloud sandboxes. For tasks with resources requirements under 4 vCPUs, 16GB RAM, and 10GB storage, we use Daytona as sandbox provider. For tasks with higher resources requirements we use Dockers in Azure VMs with Lambda API access so model can call GPU resources as needed.

After the worker agents finish running, agent trajectory files, verifier results, and output files including paper.pdf are recorded and saved. These files are then copied into another sandbox for a review agent to review and grade the trajectory files and paper.pdf based on rubric.json. Workspace files and trajectory files of reviewer agents are collected the same way to HuggingFace datasets for later analysis.

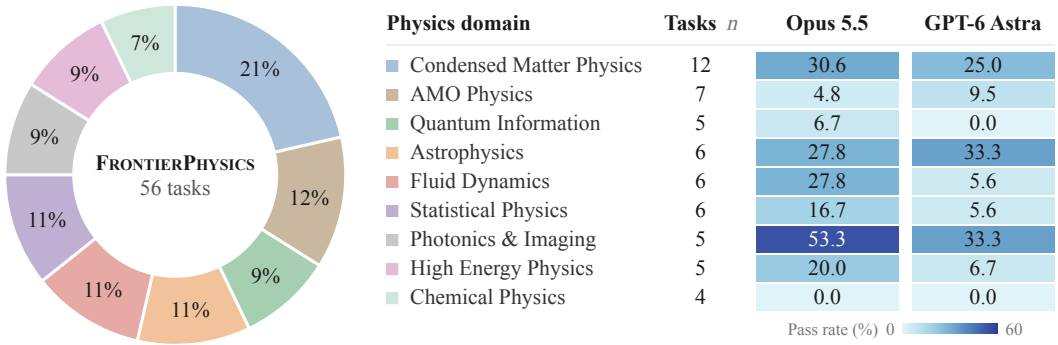

Figure 4: Task composition and overall pass rate (%) by primary physics domain across all 56 tasks. Donut labels give task shares; n gives task counts. Each task belongs to one primary domain and receives equal weight within its domain. Pass rates use overall Avg@3, with unavailable slots contributing zero. Domains contain four to twelve tasks, limiting inference about broader domain expertise.

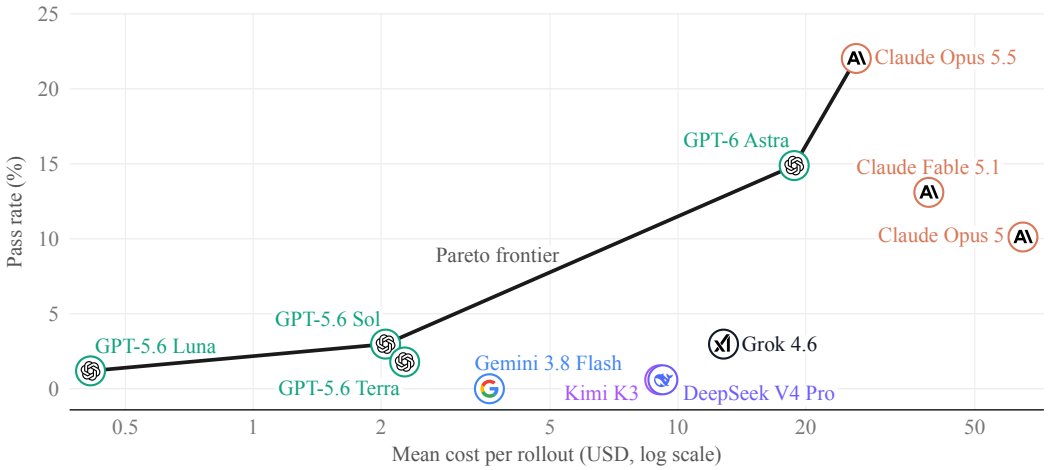

Figure 5: Mean reported inference cost per rollout versus overall pass rate (%) across all 56 tasks, with each model at its highest available inference effort where configurable. The cost axis is logarithmic, the black line marks the empirical Pareto frontier.

3.2 RESULTS

The result (Fig. 1) indicates that the overall pass rates are below 30% for all of the agents tested. Even Claude Opus 5.5, the best performing model, reaches only a pass rate of 22.0%. Our model ranking is consistent with other AI intelligence benchmarks such as the Artificial Analysis Intelligence Index (Artificial Analysis, 2026). Table 2 provides more detailed information on the different runs and scores. A noticeable feature is that Pass@3 is significantly higher than Avg@3. This indicates that agents' performance is very unstable across trials.

Figure 5 plots the pass rate versus the mean cost per rollout. As expected, the higher the cost, the better the performance in general. However, for Claude models, the pass rate decreases as cost increases, in the order Opus 5.5, Fable 5.1, and Opus 5. In addition, ChatGPT models are generally more cost-efficient. The pass-rate per physics domain of Opus 5.5 and Astra can be seen in Fig. 4 but due to the low amounts of tasks per domain it is probably not statistically significant. For more information, Appendix D gives in detailed per-task results.

Table 2: Detailed Frontier-model performance across all 56 tasks. For the experiment we use the highest available inference effort by default, where configurable. The Score column reports the score defined in Equation 1, averaged over the three trials for each task and then across all tasks. Deterministic pass requires all verifier tests to pass and overall pass additionally requires all blocking rubric criteria to pass. Avg@3 gives the average pass rate over the three trials, while Pass@3 gives the percentage of tasks with at least one passing trial. Uncertainties are approximate 95% confidence intervals as in Figure 1.

		Deterministic pass (%)		Overall pass (%)	
Model	**Score**	Avg@3	Pass@3	Avg@3	Pass@3
Claude Opus 5.5 Claude Code	**20.7** ± 4.2	39.9 ± 5.2	58.9	22.0 ± 4.5	37.5
GPT-6 Astra Codex	**14.0** ± 3.7	31.5 ± 4.5	46.4	14.9 ± 4.0	26.8
Claude Fable 5.1 Claude Code	**12.0** ± 3.1	28.6 ± 4.4	41.1	13.1 ± 3.5	19.6
Claude Opus 5 Claude Code	**9.5** ± 3.3	16.1 ± 4.0	26.8	10.1 ± 3.5	17.9
GPT-5.6 Sol Codex	**2.6** ± 2.0	9.5 ± 3.3	17.9	3.0 ± 2.3	7.1
Grok 4.6 Grok Build	**2.6** ± 1.4	4.2 ± 2.3	8.9	3.0 ± 1.6	5.4
GPT-5.6 Terra Codex	**1.4** ± 1.3	9.5 ± 3.1	16.1	1.8 ± 1.6	3.6
GPT-5.6 Luna Codex	**0.9** ± 0.8	9.5 ± 3.5	17.9	1.2 ± 1.2	1.8
Kimi K3 Claude Code	**0.4** ± 0.8	3.0 ± 2.3	7.1	0.6 ± 1.2	1.8
DeepSeek V4 Pro Claude Code	**0.3** ± 0.6	3.0 ± 1.6	5.4	0.6 ± 1.2	1.8
Gemini 3.8 Flash Antigravity	**0.0** ± 0.0	7.7 ± 2.9	14.3	0.0 ± 0.0	0.0

3.3 Failure Mode Analysis

It is important to understand why the agents fail, and how well the agent score approximates what we want to measure. To answer them, we follow the protocol described in section 2.5. Figure 6 summarizes the earlier matched-cohort annotations for 168 GPT-6 Astra runs. Table 5 reports the final review-sheet labels, including locally rerun trials. More detailed information and explanation of the failure modes can be found in table 4.

The findings in Fig. 6 show that agents mostly do not fail because of physics mistakes (e.g., a wrong calculation, misunderstanding of the physics concepts or lacking of physics knowledge), as one might expect from a physics benchmark. Instead, they often struggle to reason rigorously (42%), apply good physical intuition when making decisions (33%), and synthesize the literature (29%). Typical rigorous thinking failures are: circular self-verification, where the agent's own validation method depends on its wrong assumption so it can never find out its mistakes. Edge cases, where the agent fails to consider all the edge cases of its method or code, so that it fails at some hidden tests. The second frequent failure is about literature. In many cases, agents fail to find the key literature for the research (as identified by the task author), struggle to synthesize information from its literature search and provide a motivation for its project, or struggle to support its argument in the manuscript with literature. Missing physics intuition is another major failure mode. Typical examples of this failure mode include: agents fail to determine the major error source of an experiment, which is not because they do not understand it, but because they do not have the expert intuition to identify what is important; when debugging, agents cannot realize the root of the problem while keep tuning the wrong knob. Besides the top three failure modes, instruction following accounts for 21% of the failures. It could be agents missing some requirements in the instructions or misunderstanding the instructions. In many cases, the task authors believe the meaning should be clear to human domain experts, while the agent cannot get the explicit/implicit message.

In addition, we labeled 6% of the failed trials to be false negatives. This coincides with a recent work Ansari et al. (2026), in which domain experts regrade some leading benchmarks and find that many failures are due to grader errors, incorrect reference solutions, and ambiguous or underspecified questions. This suggests that identifying false negatives is very challenging, which requires both domain expertise and substantial time investment, while our failure mode analysis serves as a good

regrading. Compared to other benchmarks, our auditing results show reasonably low false negative rate.

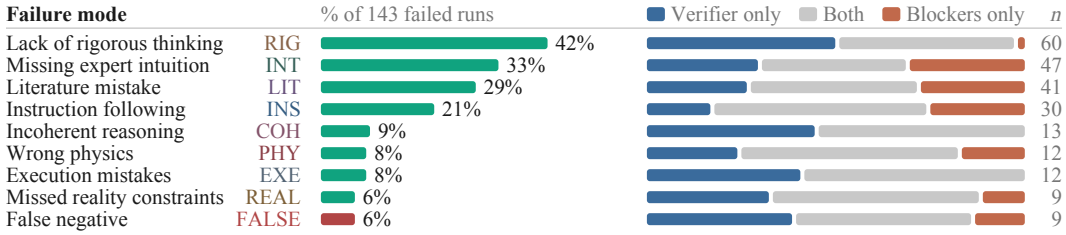

Figure 6: Failure mode analysis of GPT-6 Astra across 56 tasks (168 trials). Counts n and percentages ($n/143$) use the final review labels in Table 5; each trial may have multiple tags. Verifier/rubric proportions are retained from the original recorded assessment. Correcting the nine FALSE outcomes changes the pass rate from 14.9% (25/168) to 20.2% (34/168).

4 Discussion

In our leaderboard results, the GPT-Astra runs used xhigh inference level. Due to the limitation of Codex CLI harness, the agents could not call subagents. While in Claude Code harness calling subagents is natively supported, agents using Codex CLI harness therefore tend to finish the task faster with lower performance, in contrast to Claude Code agents, which cost more to achieve a higher pass rate.

Besides, the current failure mode analysis focuses on the GPT-6 Astra model. However, we might expect that different models, or different models paired with different agent harnesses, could exhibit variations in their failure modes or even give rise to entirely new types of failures. Also because of this, we can only accurately tag all the false negative of these 168 trials and therefore correct the final pass rate of GPT-6 Astra.

5 Conclusion

We have introduced FrontierPhysics, a benchmark dataset with 56 tasks evaluating the ability of AI agents to conduct end-to-end physics research across nine domains. As LLMs take on increasingly substantial roles in physics research and start to discover new scientific findings, there is a growing demand for an authentic and challenging benchmark to evaluate the research capability and quality of an agent. When building FrontierPhysics, we find that checking the quality and correctness of these frontier physics research tasks with rubric grading is extremely challenging, since cross-domain examination is difficult for physics experts. To solve this, we coordinate domain experts to extensively peer review each other's tasks, in a similar way of peer-reviewing physics journal manuscripts, to ensure task quality. In addition, manual human analysis of each agent trial trajectory is still necessary to further improve the rigor of the task design.

FrontierPhysics shows end to end physics research remains challenging to current models, given the best performing agent (Opus5.5 in Claude Code CLI) only achieves up to 37 out of total 168 rollouts. Based on the human experts auditing of 168 SOTA agents rollouts trajectory and output files, we build a comprehensive failure mode taxonomy and label all these 168 agents trials manually. From the taxonomy distribution, we conclude that lack of rigorous thinking (for example, circular self-verification) followed by missing experts initiation, instead of insufficient physics knowledge or execution mistakes, are the main weaknesses of frontier LLM agents.

Looking ahead, truly capable physics research agents should move beyond solving specified tasks toward independently exploring scientific problems, from problem definition to novel findings, with only high-level guidance. Training and evaluating such agents will require more open-ended research tasks and flexible evaluation frameworks that can assess exploratory reasoning, partial progress, and unexpected findings. Building such benchmarks is not just an extension of current evaluation paradigms, but a redefinition of how we measure progress toward complete AI-driven autonomous scientific discovery.

AI USE STATEMENT

We used generative AI methods at every step of the process, from helping to create tasks and find task contributors to reviewing tasks, building the BenchFlow infrastructure, performing statistical analysis, and helping to write this manuscript. However, there were very clear rules. At several stages of the process, we carefully checked that the task.md files were written by humans, that the authors had written the papers and conducted the research themselves, and that the rubrics were written by humans. Several independent people carried out these checks using AI checkers and human judgment. Additionally, the agent results and the authors' failure mode analyses were checked individually by humans. This paper was also mostly written by humans, although LLMs helped with sentence formulation, tables, and figures.

ETHICS STATEMENT

The main authors checked that every task contributor is human, worked on the submitted task themselves, and is a domain expert, and reviewed each task's content for correctness. This includes all the extra reviewing steps explained in section 2. However, despite our best effort, in the end, every contributor and reviewer is themselves responsible for the task they contributed. We minimized false negatives where we could and report the ones that remain.

REPRODUCIBILITY STATEMENT

Section 2 defines the task format, rubric structure, scoring rule and curation process, and Section 3.1 specifies the evaluation protocol: agent harnesses, rollout budget, time limits, sandbox and network policy, the rubric judge, and score aggregation. Appendix D reports per-task results for every model, and Appendix E gives the failure taxonomy, labeling rules and per-trial labels. To limit contamination, we will share an anonymized repository with reviewers and area chairs in a private OpenReview comment once the discussion period opens, and release it publicly upon publication. It will contain the 56 task packages (prompts, Docker environments, input data, oracle solutions, verifiers and rubrics), the evaluation harness, including the rubric-judge prompt, at the versions we used, per-rollout trajectories, verifier outputs, rubric reviews and costs, the failure-mode annotations with their codebook, and the scripts that regenerate every table, figure and reported number. Exact re-runs can still differ because hosted models change, sampling and LLM judging are stochastic, and agents consult the live web; the released trajectories and reviews allow our scores to be recomputed and audited without re-running the agents.

REFERENCES

Ali Ansari, Haoran Sun, Andy Zeyi Liu, Mark Jabbour, et al. How good are frontier models at physics? expert re-grading reveals broken evaluations and near-saturation of leading benchmarks. *arXiv preprint arXiv:2609.13009*, 2026. URL https://arxiv.org/abs/2609.13009.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. HealthBench: Evaluating Large Language Models Towards Improved Human Health. *arXiv preprint arXiv:2505.08775*, 2025. URL https://arxiv.org/abs/2505.08775.

Artificial Analysis. Artificial analysis intelligence index v4.3.2. https://artificialanalysis.ai/evaluations/artificial-analysis-intelligence-index, 2026. Accessed: 2026-09-26.

Eser Aygün, Anastasiya Belyaeva, Gheorghe Comanici, Marc Coram, Hao Cui, Jake Garrison, Renee Johnston, Anton Kast, Cory Y. McLean, Peter Norgaard, Zahra Shamsi, David Smalling, James Thompson, Subhashini Venugopalan, Brian P. Williams, Chujun He, Sarah Martinson, Martyna Plomecka, Lai Wei, Yuchen Zhou, Qian-Ze Zhu, Matthew Abraham, Erica Brand, Anna Bulanova, Jeffrey A. Cardille, Chris Co, Scott Ellsworth, Grace Joseph, Malcolm Kane, Ryan Krueger, Johan Kartiwa, Dan Liebling, Jan-Matthis Lueckmann, Paul Raccuglia, Xuefei Julie Wang, Katherine Chou, James Manyika, Yossi Matias, John C. Platt, Lizzie Dorf-

man, Shibl Mourad, and Michael P. Brenner. An AI system to help scientists write expert-level empirical software. *Nature*, 654:909–916, 2026. doi: 10.1038/s41586-026-10658-6. URL https://doi.org/10.1038/s41586-026-10658-6.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624:570–578, 2023. doi: 10.1038/s41586-023-06792-0. URL https://doi.org/10.1038/s41586-023-06792-0.

Yadi Cao, Sicheng Lai, Jiahe Huang, Yang Zhang, Zach Lawrence, Rohan Bhakta, Izzy F. Thomas, Mingyun Cao, Chung-Hao Tsai, Zihao Zhou, Yidong Zhao, Hao Liu, Alessandro Marinoni, Alexey Arefiev, and Rose Yu. SimulCost: A Cost-Aware Benchmark and Toolkit for Automating Physics Simulations with LLMs. *arXiv preprint arXiv:2603.20253*, 2026. URL https://arxiv.org/abs/2603.20253.

Daniel J. H. Chung, Zhiqi Gao, Yurii Kvasiuk, Tianyi Li, Moritz Münchmeyer, Maja Rudolph, Frederic Sala, and Sai Chaitanya Tadepalli. Theoretical Physics Benchmark (TPBench) – a Dataset and Study of AI Reasoning Capabilities in Theoretical Physics. *arXiv preprint arXiv:2502.15815*, 2025. URL https://arxiv.org/abs/2502.15815.

Cognition. FrontierCode 1.1. Official web page, 2026. URL https://cognition.com/blog/frontier-code-1.1.

Ken Deng, Xiangfei Wang, Guijing Duan, Chen Mo, Junkun Huang, Runqing Zhang, Ling Qian, Zhiguo Huang, Jize Han, and Di Luo. QMP-Bench: A Research-Level Benchmark for Autonomous End-to-End Quantum Many-Body Simulations. Accepted paper under non-archival workshop policy, 2026. URL https://openreview.net/forum?id=9IdnFv2TmQ.

Zhiyuan Fan, Weinong Wang, Xing Wu, and Debing Zhang. SedarEval: Automated Evaluation using Self-Adaptive Rubrics. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16916–16930. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.984. URL https://aclanthology.org/2024.findings-emnlp.984/.

Darius A. Faroughy, Sofia Palacios Schweitzer, Ian Pang, Siddharth Mishra-Sharma, and David Shih. Collider-Bench: Benchmarking AI Agents with Particle Physics Analysis Reproduction. *arXiv preprint arXiv:2605.13950*, 2026. URL https://arxiv.org/abs/2605.13950.

Kaiyue Feng, Yilun Zhao, Yixin Liu, Tianyu Yang, Chen Zhao, John Sous, and Arman Cohan. Physics: Benchmarking Foundation Models on University-Level Physics Problem Solving. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11717–11743, 2025. doi: 10.18653/v1/2025.findings-acl.610. URL https://doi.org/10.18653/v1/2025.findings-acl.610.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, Anil Palepu, Keran Rong, Ryutaro Tanno, Khaled Saab, Fan Zhang, Jacob Blum, Andrew Carroll, Kavita Kulkarni, Nenad Tomašev, Dina Zverinski, Ivor Rendulic, Elahe Vedadi, Florian Hasler, Luka Rimanic, Marina Boia, Ivan Budiselic, Ben Feinstein, Mathias Bellaiche, Tom Sheffer, Jan Freyberg, Jeremy Ratcliff, Ottavia Bertolli, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R. D. Costa, José R. Penadés, Gary Peltz, Yossi Matias, James Manyika, Demis Hassabis, Yunhan Xu, Pushmeet Kohli, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Accelerating scientific discovery with Co-Scientist. *Nature*, 655:487–496, 2026. doi: 10.1038/s41586-026-10644-y. URL https://doi.org/10.1038/s41586-026-10644-y.

Peiyan Hu, Haodong Feng, Hongyuan Liu, Tongtong Yan, Wenhao Deng, Tianrun Gao, Rong Zheng, Haoren Zheng, Chenglei Yu, Chuanrui Wang, Kaiwen Li, Zhi-Ming Ma, Dezhi Zhou, Xingcai Lu, Dixia Fan, and Tailin Wu. RealPDEBench: A Benchmark for Complex Physical Systems with Real-World Data. *arXiv preprint arXiv:2601.01829*, 2026. URL https://arxiv.org/abs/2601.01829.

Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, Yikai Zhang, Yuqing Yang, Ting Wu, Binjie Wang, Shichao Sun, Yang Xiao, Yiyuan Li, Fan Zhou, Steffi Chern, Yiwei Qin, Yan Ma, Jiadi Su, Yixiu Liu, Yuxiang Zheng, Shaoting Zhang, Dahua Lin, Yu Qiao, and Pengfei Liu. OlympicArena: Benchmarking Multi-discipline Cognitive Reasoning for Superintelligent AI. In *Advances in Neural Information Processing Systems 37*, pp. 19209–19253, 2024. doi: 10.52202/079017-0607. URL https://doi.org/10.52202/079017-0607.

Yigeng Jiang, Tengchao Yang, Taoyong Cui, Jiaxing Wan, Yuan Wang, Weida Wang, Zhiyu Liu, Chuyi Peng, Binzhao Luo, Maoli Gao, Huaihai Huang, Yuqianer Zeng, Ziyang Zheng, Dongchen Huang, Chao Chen, Zichao Liu, Weiping Shen, Shuchen Pu, Siyu Zhou, Runmin Ma, Yusong Hu, Fei Chao, Bo Zhang, Xiawu Zheng, Zifu Wang, Lei Bai, Yunqi Cai, and Shufei Zhang. Deep Research in Physical Sciences: A Multi-Agent Framework and Comprehensive Benchmark. *arXiv preprint arXiv:2606.18648*, 2026. URL https://arxiv.org/abs/2606.18648.

Xiangyi Li, Kyoung Whan Choe, Yimin Liu, Xiaokun Chen, Chujun Tao, Bingran You, et al. ClawsBench: Evaluating capability and safety of LLM productivity agents in simulated workspaces. *arXiv preprint arXiv:2604.05172*, 2026a. URL https://arxiv.org/abs/2604.05172. Accepted to COLM 2026.

Xiangyi Li, Yimin Liu, Wenbo Chen, Bingran You, et al. SkillsBench: Benchmarking how well agent skills work across diverse tasks. *arXiv preprint arXiv:2602.12670*, 2026b. URL https://arxiv.org/abs/2602.12670.

Xinge Liu, Terry Jingchen Zhang, Bernhard Schölkopf, Zhijing Jin, and Kristen Menou. Stargazer: A Scalable Model-Fitting Benchmark Environment for AI Agents under Astrophysical Constraints. *arXiv preprint arXiv:2604.15664*, 2026. URL https://arxiv.org/abs/2604.15664.

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. URL https://arxiv.org/abs/2408.06292.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6:525–535, 2024. doi: 10.1038/s42256-024-00832-8. URL https://doi.org/10.1038/s42256-024-00832-8.

Indrajeet Mandal, Jitendra Soni, Mohd Zaki, Morten M. Smedskjaer, Katrin Wondraczek, Lothar Wondraczek, Nitya Nand Gosvami, and N. M. Anoop Krishnan. Evaluating large language model agents for automation of atomic force microscopy. *Nature Communications*, 16, 2025. doi: 10.1038/s41467-025-64105-7. URL https://doi.org/10.1038/s41467-025-64105-7.

Tingjia Miao, Wenkai Jin, Muhua Zhang, Jinxin Tan, Yuelin Hu, Tu Guo, Jiejun Zhang, Yuhan Wang, Wenbo Li, Yinuo Gao, Shuo Chen, Weiqi Jiang, Yayun Hu, Zixing Lei, Xianghe Pang, Zexi Liu, Yuzhi Zhang, Linfeng Zhang, Kun Chen, Wei Wang, Weinan E, and Siheng Chen. PRL-Bench: A Comprehensive Benchmark Evaluating LLMs' Capabilities in Frontier Physics Research. *arXiv preprint arXiv:2604.15411*, 2026. URL https://arxiv.org/abs/2604.15411.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025. URL https://arxiv.org/abs/2506.13131.

Haining Pan, James Roggeveen, Erez Berg, Juan Alvarez, Debanjan Chowdhury, Surya Ganguli, Federico Ghimenti, Juraj Hasik, Henry Hunt, Hong-Chen Jiang, Mason Kamb, Ying-Jer Kao, Ehsan Khatami, Michael Lawler, Di Luo, Titus Neupert, Xiaoliang Qi, Michael Brenner, and Eun-Ah Kim. CMT-Benchmark: A Benchmark for Condensed Matter Theory Built by Expert

Researchers. In *International Conference on Learning Representations*, volume 2026, pp. 35522–35540, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3bf2d6047e8599e1b7567e5d154d39e9-Abstract-Conference.html.

Shi Qiu, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, Tianyu Luo, Yixuan Yin, Zhang Haoxu, Yi Hu, Chenyang Wang, Chencheng Tang, Haoling Chang, Qi Liu, Ziheng Zhou, Tianyu Zhang, Jingtian Zhang, Zhangyi Liu, Minghao Li, Yuku Zhang, Boxuan Jing, Xianqi Yin, Yutong Ren, Zizhuo Fu, Jiaming Ji, Weike Wang, Xudong Tian, Anqi Lv, Laifu Man, Jianxiang Li, Feiyu Tao, Qihua Sun, Zhou Liang, Yushu Mu, Zhongxuan Li, Jing-Jun Zhang, Shutao Zhang, Xiaotian Li, Xingqi Xia, Jiawei Lin, Zheyu Shen, Jiahang Chen, Qiuhao Xiong, Binran Wang, Fengyuan Wang, Ziyang Ni, Bohan Zhang, Fan Cui, Changkun Shao, Qing-Hong Cao, Ming-xing Luo, Yaodong Yang, Muhan Zhang, and Hua Xing Zhu. PHYBench: Holistic Evaluation of Physical Perception and Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 38*, pp. 1345–1385, 2025. doi: 10.52202/085713-0044. URL https://doi.org/10.52202/085713-0044.

Shi Qiu, Junyi Deng, Yiwei Deng, Haoran Dong, Jieyu Fu, Mao Li, Zeyu Li, Zhaolong Zhang, Huiwen Zheng, Leidong Bao, Anqi Lv, Zihan Mo, Yadi Niu, Yiyang Peng, Yu Tian, Yili Wang, Ziyu Wang, Zi-Yu Wang, Jiashen Wei, Liuheng Wu, Aoran Xue, Leyi Yang, Guanglu Yuan, Xiarui Zhan, Jingjun Zhang, Zifan Zheng, Pengfei Liu, Linrui Zhen, Kaiyang Li, Qichang Li, Ziheng Zhou, Guo-En Nian, Yunwei Xiao, Qing-Hong Cao, Linjie Dai, Xu Feng, Peng Gao, Ying Gu, Chang Liu, Jia Liu, Ming-xing Luo, Yan-Qing Ma, Liang-You Peng, Huichao Song, Shufeng Wang, Chenxu Wang, Tao Wang, Yi-Nan Wang, Chengyin Wu, Pengwei Zhao, and Hua Xing Zhu. PRBench: End-to-end Paper Reproduction in Physics Research. *arXiv preprint arXiv:2603.27646*, 2026. URL https://arxiv.org/abs/2603.27646.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent Laboratory: Using LLM Agents as Research Assistants. *arXiv preprint arXiv:2501.04227*, 2025. URL https://arxiv.org/abs/2501.04227.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. PaperBench: Evaluating AI's Ability to Replicate AI Research. In *International Conference on Machine Learning*, pp. 56843–56873, 2025. URL https://proceedings.mlr.press/v267/starace25a.html.

Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, 646:716–723, 2025. doi: 10.1038/s41586-025-09442-9. URL https://doi.org/10.1038/s41586-025-09442-9.

Terminal-Bench-Science Team. Terminal-Bench-Science: Evaluating AI agents on research workflows across scientific domains. Software release, 2026. URL https://zenodo.org/doi/10.5281/zenodo.22110254.

Miles Wang, Robi Lin, Kat Hu, Joy Jiao, Neil Chowdhury, Ethan Chang, and Tejal Patwardhan. FrontierScience: Evaluating AI's Ability to Perform Expert-Level Scientific Tasks. *arXiv preprint arXiv:2601.21165*, 2026a. URL https://arxiv.org/abs/2601.21165.

Weida Wang, Dongchen Huang, Jiatong LI, Tengchao Yang, Ziyang Zheng, Chuyi Peng, Di Zhang, Dong Han, Benteng Chen, Binzhao Luo, Zhiyu Liu, kunling liu, Zhiyuan Gao, Shiqigeng geng, Wei Ma, Jiaming Su, Xin Li, Shuchen Pu, Yuhan Shui, Qianjia Cheng, Zhihao Dou, Dongfei Cui, Changyong He, Jin Zeng, Zeke Xie, Mao Su, Dongzhan Zhou, Yuqiang Li, Wanli Ouyang, Yunqi Cai, Xi Dai, Shufei Zhang, LEI BAI, Jinguang Cheng, Zhong Fang, and Hongming Weng. CMPhysBench: A Benchmark for Evaluating Large Language Models in Condensed Matter Physics. In *International Conference on Learning Representations*, volume 2026, pp. 3388–3421, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/06e1da559a7227d7cce7364f92976a0e-Abstract-Conference.html.

Yuru Wang, Lejun Cheng, Yuxin Zuo, Sihang Zeng, Bingxiang He, Che Jiang, Junlin Yang, Yuchong Wang, Kaikai Zhao, Weifeng Huang, Kai Tian, Zhenzhao Yuan, Jincheng Zhong, Weizhi Wang, Ning Ding, Bowen Zhou, and Kaiyan Zhang. NatureBench: Can Coding Agents Match the

Published SOTA of Nature-Family Papers? *arXiv preprint arXiv:2606.24530*, 2026c. URL https://arxiv.org/abs/2606.24530.

Wanghan Xu, Shuo Li, Tianlin Ye, Qinglong Cao, Yixin Chen, Hengjian Gao, Yiheng Wang, Qi Li, Kun Li, Sheng Xu, Shengdu Chai, Fangchen Yu, Xiangyu Zhao, Zhangrui Zhao, Weijie Ma, Zijie Guo, Koutian Wu, Haoyu Zhou, Haoxiang Yin, Lixue Cheng, Chaofan Hu, Haoxuan Li, Lu Mi, Xuxuan Xie, Yifan Zhou, Ruizhe Chen, Zhiwang Zhou, Xingjian Guo, Yuhao Zhou, Xuming He, Shengyuan Xu, Xinyu Gu, Jiamin Wu, Mianxin Liu, Chunfeng Song, Fenghua Ling, Dongzhan Zhou, Shixiang Tang, Yuqiang Li, Mao Su, Peng Ye, Siqi Sun, Bin Wang, Xue Yang, Zhenfei Yin, Tianfan Fu, Guangtao Zhai, Wanli Ouyang, Bo Zhang, Lei Bai, and Wenlong Zhang. ResearchClawBench: A Benchmark for End-to-End Autonomous Scientific Research. *arXiv preprint arXiv:2606.07591*, 2026. URL https://arxiv.org/abs/2606.07591.

Xin Xu, Qiyun Xu, Tong Xiao, Tianhao Chen, Yuchen Yan, Jiaxin Zhang, Shizhe Diao, Can Yang, and Yang Wang. UGPhysics: A Comprehensive Benchmark for Undergraduate Physics Reasoning with Large Language Models. *arXiv preprint arXiv:2502.00334*, 2025. URL https://arxiv.org/abs/2502.00334.

Christine Ye, Sihan Yuan, Suchetha Cooray, Steven Dillmann, Ian L. V. Roque, Dalya Baron, Philipp Frank, Sergio Martin-Alvarez, Nolan Koblischke, Frank J Qu, Diyi Yang, Risa Wechsler, and Ioana Ciuca. ReplicationBench: Can AI Agents Replicate Astrophysics Research Papers? *arXiv preprint arXiv:2510.24591*, 2025. URL https://arxiv.org/abs/2510.24591.

Minhui Zhu, Minyang Tian, Xiaocheng Yang, Tianci Zhou, Lifan Yuan, Penghao Zhu, Eli Chertkov, Shengyan Liu, Yufeng Du, Ziming Ji, Indranil Das, Qingzhi Chen, Junyi Cao, Yufeng Du, Jiabin Yu, Peixue Wu, Jinchen He, Yifan Su, Yikun Jiang, Yujie Zhang, Chang Liu, Ze-Min Huang, Weizhen Jia, Yunkai Wang, Farshid Jafarpour, Yong Zhao, Xinan Chen, Jessie Shelton, Aaron W. Young, John Bartolotta, Wenchao Xu, Yue Sun, Anjun Chu, Victor Colussi, Chris Akers, Nathan Brooks, Wenbo Fu, Jinchao Zhao, Marvin Qi, Anqi Mu, Yubo Yang, Allen Zang, Yang Lyu, Peizhi Mai, Christopher Wilson, Xuefei Guo, Juntai Zhou, Daniel Inafuku, Chi Xue, Luyu Gao, Ze Yang, Yaïr Hein, Yonatan Kahn, Kevin Zhou, Di Luo, John Drew Wilson, Jarrod T. Reilly, Dmytro Bandak, Ofir Press, Liang Yang, Xueying Wang, Hao Tong, Nicolas Chia, Eliu Huerta, and Hao Peng. Probing the Critical Point (CritPt) of AI Reasoning: a Frontier Physics Research Benchmark. *arXiv preprint arXiv:2509.26574*, 2025. URL https://arxiv.org/abs/2509.26574.

APPENDIX

A EXAMPLE TASK

Research task
task.md

Read /root/proposal.pdf first. Then carry out the following work.
The implementation is divided into two stages, with each stage handling one key step of this task. All numerics are checked against at least 4 significant figures.
(a) Characterize the apparatus and model the system. First, use the device to measure whatever you need to pin down the interactions of the chain, by driving small, well-chosen atom configurations and analyzing the measured occupations.

Then construct a model of the chain from your measurements (with the drive Omega = 2pi x 6 MHz) and compute the ground-state sigma_mid for a 21-atom and a 27-atom chain at Delta = 10.20 MHz. Because the device supports only a few atoms, these results must come from your own simulation, which needs to be scalable to about 30 atoms with the desired accuracy. Report the value to at least 6 decimal places.

(b) Locate the critical point with the experimental parameters. Using the model you built in (a), compute the ground-state sigma_mid across a range of system sizes and detunings yourself, then perform a finite-size scaling analysis of those curves to determine Delta_c.
(i) Report the finite-size scaling method that you used. For that, report the ratio of the finite-size rescaling factor you apply to sigma_mid for a 7-atom array to that for an 11-atom array.

(ii) Report Delta_c in MHz together with the order-parameter scaling exponent you used, and provide a full analysis of the uncertainty of the obtained Delta_c. Report the value to 6 decimals, including the uncertainty.

When you are done, I want the following:
- /root/result.md: key-value lines following /root/result_template.md.
[The seven keys and their one-line definitions.]
- /root/paper.pdf: a short paper for peer review (abstract, model and methods, results with a finite-size-scaling figure, discussion, references).

Deterministic verifier
test_outputs.py

- ☐ Device log
  $0 < \text{total shots} < 10^6$
- ☐ v2_mhz ±0.002 MHz of the device's V_2
- ☐ sigma_mid_21_delta1020
  $\pm 2.0 \times 10^{-3}$
- ☐ sigma_mid_27_delta1020
  $\pm 2.6 \times 10^{-3}$
- ☐ rescale_ratio_7_11
  0.956346 ± 10^{-6}, so the bulk $(7/11)^{1/8} = 0.945068$ fails
- ☐ delta_c_mhz ±0.010 MHz
- ☐ beta_over_nu 1/8 ± 0.01
- ☐ delta_c_uncertainty_mhz in (0, 0.05] MHz
- ☐ result.md all seven keys
- ☐ paper.pdf 3+ pages; abstract, method, result, reference

Reviewer agent
rubric.json

Rydberg platform understanding (blocker) blockade explained; blockade-violating shots postselected out
Device calibration (weight 6) Ω, V_1, V_2, V_3 measured; V_1 breaks r^{-6}; decay, loss and shot-to-shot noise; traces under 20 µs
Literature grounding (weight 2) sources for the hard-boson model, second-order PXP and chain criticality

Rydberg quantum simulation basics (blocker) blockade (PXP) subspace or postselection, with the reason in the paper
Hamiltonian construction (weight 8) second-order PXP solved by ED or MPS; full-space ED cannot reach 30 atoms

Universality class and CFT basics (blocker) (1+1)D Ising; open-boundary $\sin(\pi/(L+2))^{-1/8}$ σ_{mid}, not the bulk $L^{1/8}$
Plan before analysis (weight 2) σ_{mid} numerics validated before any fit

Extrapolation quality (weight 8) crossings per odd-L pair (L, L+4); beyond-linear extrapolation; fit order and window varied

Artifacts present · No fabrication · No tampering · No reward hacking (blockers)
Format requirements (weight 7) · Paper structure (7) · Paper reproducibility (6) · Synthesize information (5)

Pass all 10 tests and all 7 blockers. **Score** $100 \cdot \Sigma w_i s_i / (2\Sigma w_i)$ over the 9 weighted criteria, $s_i \in \{0, 1, 2\}$; 0 if it fails.

Figure 7: **Example task.** `rydberg-ising-critical-point`, the example of Table 4. Left: its `task.md`. The cited proposal (Figure 8) defines σ_{mid} as the order parameter on the central bond; the device simulates a Rydberg apparatus of at most three atoms. The tests that check each step in `task.md` (±: tolerance around withheld reference values) and the summarized rubric criteria that grade it.

Research proposal
proposal.pdf, which task.md says to read first

Research and plan

Quantum phase transition describes a sudden change of quantum system behavior under zero temperature. At the phase transition point, systems display universal behvaiors [sic], and the point of quantum phase transition is called the critical point. Studying universal behavior is therefore an important subject in condensed matter physics.

Recently, Rydberg atom arrays become a competitive candidate in studying such universal behaviors due to its programmability and flexibility. In these setups, an array of neutral atoms are trapped in an equally spaced lattice. Throughout this work, we consider a one-dimensional array of Strontium atoms with open boundary conditions, sitting on the lattice sites [0, 1, 2, …]. A qubit is encoded between its metastable ground state 5s5p $^3P_0$ and a Rydberg state. The system is driven by a global laser with Rabi frequency Omega = 2\pi \times 6 MHz and detuning Delta, and we work in a nearest-neighbor-Rydberg-blockade regime. As the detuning Delta is varied, the chain crosses a quantum phase transition between a disordered phase and an ordered phase at a critical detuning Delta_c.

We want to determine the critical detuning Delta_c of this disorder-to-ordered transition. A simulator of the apparatus is provided as a tool in the workspace (see /root/DEVICE.md for the interface); you can access the device to measure whatever properties of the system you need in order to model it. To locate Delta_c you will need an accurate model of the chain that you can simulate at large sizes; we characterize the transition through the charge-density-wave (CDW) order parameter at the center of the array: on bond j, sigma_j = (-1)^j (n_j - n_{j+1}) (0 indexed) with n being the Rydberg state occupation, and sigma_mid is the order parameter on the central bond, j = floor((N-1)/2).

Research how a critical point of such a chain can be determined from finite-size numerics with open boundary conditions: ❶ what universality class governs this transition and its critical exponents, ❷ how sigma_mid should behave with detuning Delta and system size N near the transition, and ❸ what algorithm and system sizes you need to reach the target precision (as detailed below). ❹ Discuss your reasoning and the relevant literature in the final paper.

	What the literature must supply	Sources the rubric names	Checked by
❶	(1+1)D Ising universality for the disorder-to-Z_2 transition of the blockaded chain; the order parameter's scaling dimension is 1/8, so β/ν = 1/8.	Fendley, Sengupta & Sachdev 2004 (the hard-boson model); Bernien et al. 2017 or Keesling et al. 2019	☐ beta_over_nu Universality class and CFT basics (blocker)
❷	With open boundaries $\sin(\pi/(L+2))^{-1/8}\ \sigma_{mid}$, not the bulk $L^{1/8}\ \sigma_{mid}$, is size-independent at Δ_c; rescaled curves cross for odd-L pairs (L, L+4), and the crossings need a beyond-linear extrapolation.	Fisher & Barber 1972; a boundary-critical-phenomena or crossing-method reference	☐ rescale_ratio_7_11 ☐ delta_c_mhz Extrapolation quality (weight 8)
❸	V_1 is not much larger than Ω, so first-order PXP falls short: a second-order PXP Hamiltonian from a Schrieffer–Wolff expansion, solved by ED in the blockade subspace or by MPS up to about 30 atoms.	Bluvstein et al. 2021	☐ sigma_mid_21_delta1020 ☐ sigma_mid_27_delta1020 Hamiltonian construction (weight 8) Rydberg quantum simulation basics (blocker)
❹	At least three such sources, found and used correctly, one of them for the effective model, and the work placed against them.	—	Literature grounding (weight 2) Synthesize information (weight 5)

Figure 8: **The proposal of the example task.** Top: `proposal.pdf`, the one-page brief that agent need to read first. It sets out the physics and the goal, and describe what are the questions should be answered after literature research. The last paragraph leaves four questions to the agent's own research (numbered). Bottom: for each question, what the literature must supply.

B Task Domain Taxonomy

The 56 tasks are assigned to nine primary domains. Each task contributes to exactly one primary-domain. The "Also spans" column in Table 3 records secondary domains for 15 tasks but these labels do not add extra observations to domain summaries.

Table 3: Primary and secondary domain labels for all 56 FRONTIERPHYSICS tasks. A dash indicates no secondary domain label.

#	Task	Domain	Also spans
1	arpes-sample-alignment	Condensed Matter Physics	Photonics & Imaging
2	baf2-ih001-strain-energy	Condensed Matter Physics	—
3	cpt-frustrated-chain-spectra	Condensed Matter Physics	—
4	eigenstate-optical-response-derivation	Condensed Matter Physics	Photonics & Imaging
5	lattice-band-geometry-discovery	Condensed Matter Physics	—
6	li-graphene-mlip-transferability	Condensed Matter Physics	Chemical Physics
7	llzo-proton-exchange-thermodynamics	Condensed Matter Physics	Chemical Physics
8	lpbf-melt-pool-window	Condensed Matter Physics	Fluid Dynamics
9	mlg-quantum-hall-dot-hartree-fock	Condensed Matter Physics	—
10	moire-continuum-model-construction	Condensed Matter Physics	—
11	rixs-entanglement-witness	Condensed Matter Physics	—
12	smt-reflow-window	Condensed Matter Physics	—
13	ion-trap-3d-vs-2d	AMO Physics	—
14	ion-trap-design	AMO Physics	—
15	multiplexing-ion-chain-qnet	AMO Physics	Quantum Information
16	phase-controlled-optical-tweezer-holography	AMO Physics	—
17	rydberg-cft-modulation-spectroscopy	AMO Physics	—
18	rydberg-cz-phase-noise-infidelity	AMO Physics	Quantum Information
19	rydberg-ising-critical-point	AMO Physics	—
20	QNTK-research	Quantum Information	—
21	cvqkd-calibrated-detector-noise	Quantum Information	—
22	ising-qml-dwave	Quantum Information	—
23	time-bin-qkd-product-povm	Quantum Information	—
24	transmon-cz-pulse-engineering	Quantum Information	—
25	desi-bao-sn-dark-energy	Astrophysics	—
26	ems-scalar-spectral-census	Astrophysics	—
27	nonlinear-boson-star-sideband-threshold	Astrophysics	—
28	ns-eos-multimessenger	Astrophysics	—
29	pulsar-timing-phase-connection	Astrophysics	—
30	solar-wind-turbulence-spectral-break	Astrophysics	—

Continued on next page

Table 3: Task domain labels (continued).

#	Task	Domain	Also spans
31	consistent-two-phase-ls-vof	Fluid Dynamics	—
32	diffuse-ib-rigid-body-fsi	Fluid Dynamics	—
33	iga-branched-incompressible-flow	Fluid Dynamics	—
34	material-agnostic-visual-inverse-dynamics	Fluid Dynamics	—
35	multi-object-visual-inverse-dynamics	Fluid Dynamics	—
36	twisted-slinky-wave-coupling	Fluid Dynamics	—
37	glass-nonequilibrium-landscape-evolution	Statistical Physics	—
38	kinked-ladder-polymer-scattering	Statistical Physics	Chemical Physics
39	learning-credit-assignment	Statistical Physics	—
40	low-dimensional-deep-network-learning	Statistical Physics	—
41	metallic-glass-pel-ruggedness	Statistical Physics	—
42	quantized-ising-3sat-encoding	Statistical Physics	Quantum Information
43	calibrated-cbct-local-reconstruction	Photonics & Imaging	—
44	fbg-rail-axle-inversion	Photonics & Imaging	—
45	fbg-strain-amplifier-design	Photonics & Imaging	—
46	heat-assisted-detection-and-ranging	Photonics & Imaging	—
47	hspic-spectrometer-e2e	Photonics & Imaging	—
48	corsika-air-shower-inverse	High Energy Physics	Astrophysics
49	icecube-pointsource-unbinned	High Energy Physics	Astrophysics
50	o3-monodromy-line-defect-mc	High Energy Physics	Statistical Physics
51	o3-surface-cft-fuzzy-sphere	High Energy Physics	Statistical Physics
52	sipm-spe-gain-breakdown	High Energy Physics	Photonics & Imaging
53	agi-surface-hydrophobicity	Chemical Physics	—
54	hole-bifurcation-ml-landscape	Chemical Physics	—
55	p2d-cell-model-identification	Chemical Physics	—
56	smfret-research-v5	Chemical Physics	Photonics & Imaging

C Input Modalities and Research Activities

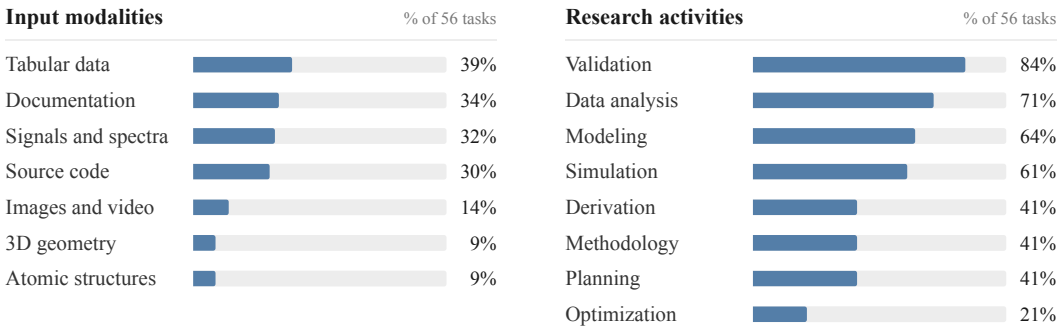

Figure 9: Input modalities and required research activities across the 56 accepted tasks. Each bar gives the percentage of tasks carrying that label. Modality and activity categories are non-exclusive: a single task can contribute to multiple categories, so the percentages within each panel need not sum to 100%.

D Per-Task Results

Figures 10–13 break the results of Table 2 down by task for all 11 model configurations. Each cell summarizes the three designated rollouts of one model on one task. As in Table 2, an unavailable rollout counts as zero. The bottom row of each figure is the model's mean over all 56 tasks. The four following tables show:

- **Verifier pass rate** (Figure 10): the percentage of rollouts in which every applicable deterministic test passes. Its bottom row is the deterministic Avg@3 of Table 2.
- **Rubric blocker pass rate** (Figure 11): the percentage of rollouts in which every rubric blocker passes.
- **Rubric score** (Figure 12): the weighted non-blocker score in Equation 1, before the verifier and blocker gates are applied.
- **Score** (Figure 13): the rubric score (Equation 1) of rollouts that pass the verifier and every blocker, and zero otherwise. Its bottom row is the Score of Table 2.

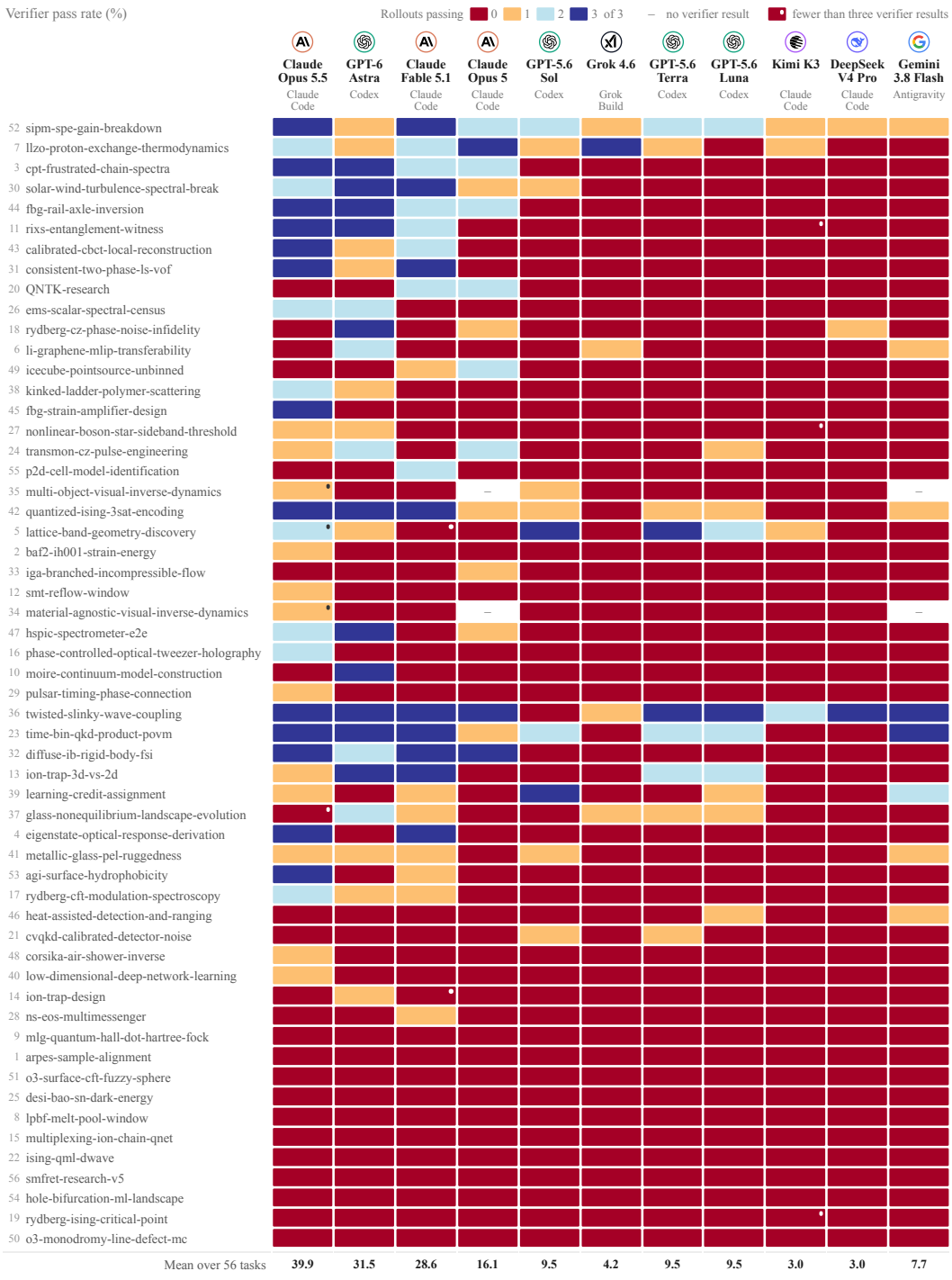

Figure 10: Verifier pass rate per task and model. The bottom row is each model's deterministic Avg@3 over all 56 tasks (Table 2). Unavailable outcomes contribute zero to the aggregate.



Figure 11: Rubric blocker pass rate per task and model. The bottom row is each model's mean over all 56 tasks. Unavailable outcomes contribute zero to the aggregate.

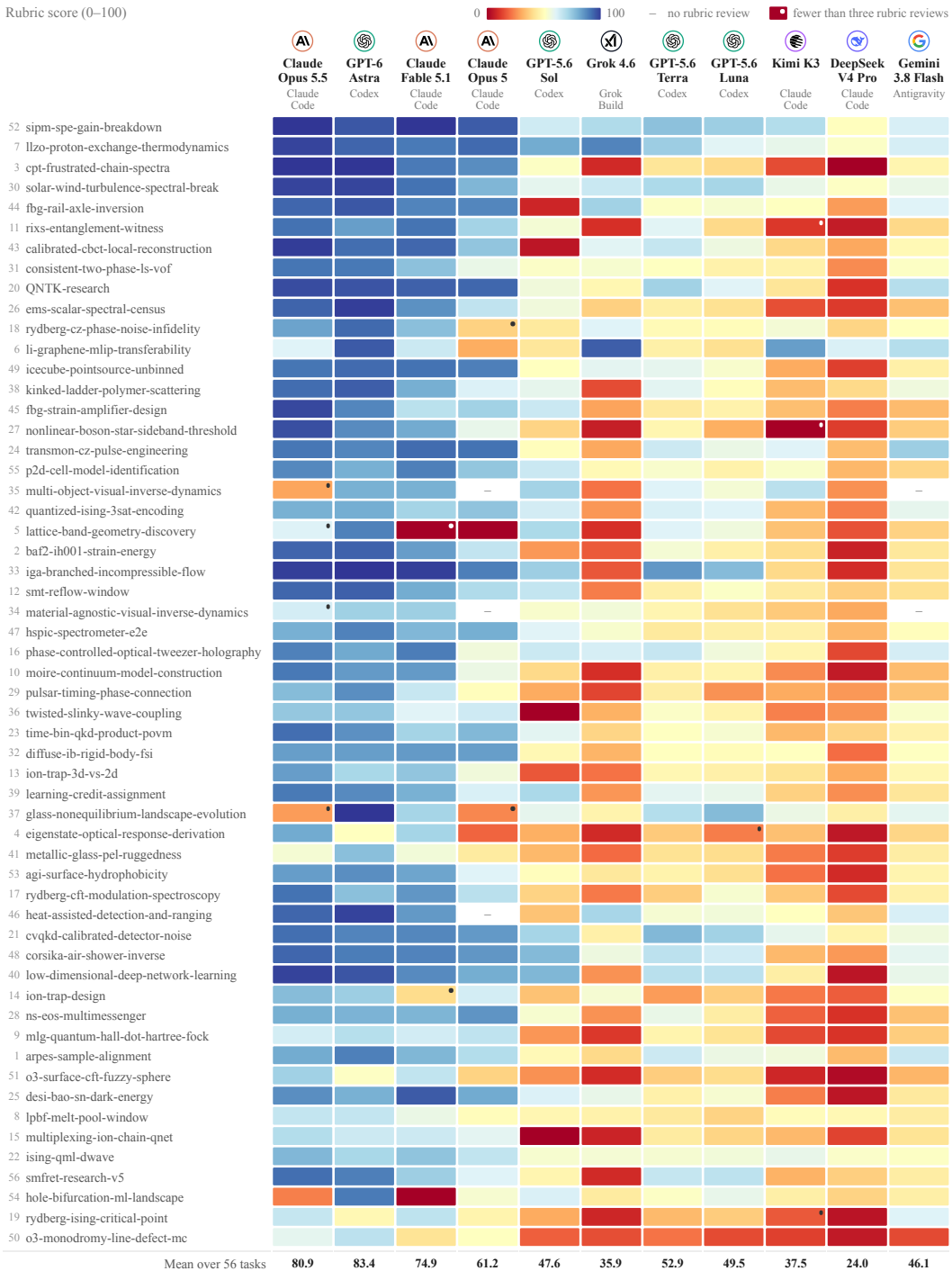

Figure 12: Rubric score per task and model. The weighted non-blocker score (Equation 1) before the verifier and blocker gates, averaged over three rollouts. The bottom row is each model's mean over all 56 tasks.

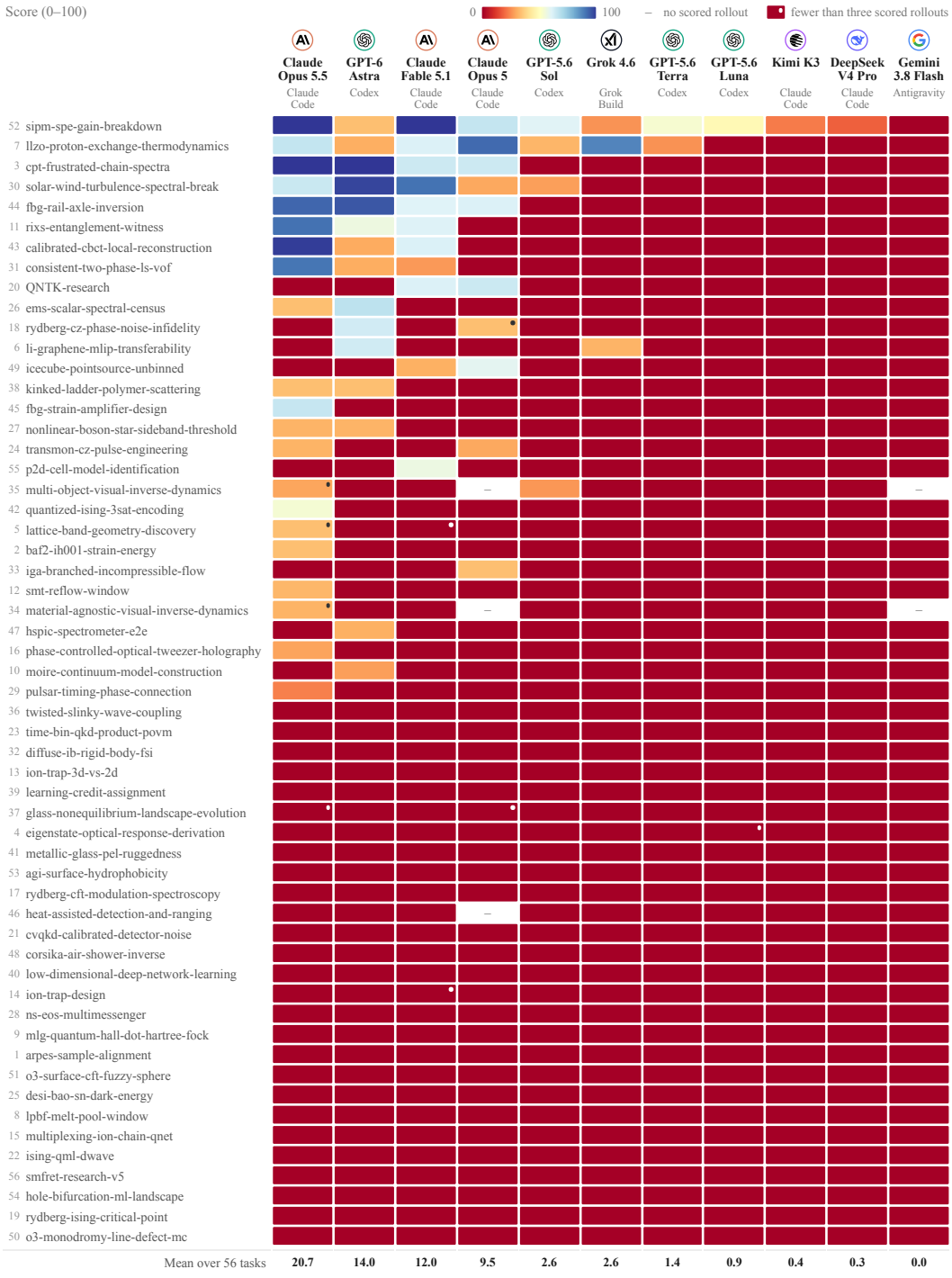

Figure 13: Score (Equation 1) per task and model. The bottom row reproduces Score in Table 2. Rows follow Figure 10. Missing rollouts count as zero.

E FAILURE-MODE ANALYSIS

The following table shows a detailed explanation of the failure mode taxonomy that was extracted on the latest GPT-6 Astra (`xhigh`, `codex-acp`) cohort based on the method described in Section 2.5.

Table 4: Failure mode taxonomy.

Tag	Label	Description	Examples
Get the physics wrong	PHY	• Subtle mistakes in the formulas or theory applied; • Agents do not have such physics knowledge (no signal when checking agents' trajectories).	
Literature mistake	LIT	• Agents cannot find the important papers (specified by the task author in rubrics as important papers that should be found) in the research stage; • In `paper.pdf`, agents cannot use the literature to support their claims correctly.	
Agents do not follow `task.md` properly	INS	• Agents miss certain instructions in `task.md` (do not make the judgement based only on agents' artifacts; check their trajectories to see if agents ever recognize the instructions). • Agents don't understand or misunderstand the instructions. In the example task, agents do not get that "we work in a nearest-neighbor-Rydberg-blockade regime" implies they should make approximations to the Hamiltonian or perform postselection.	`rydberg-ising-critical-point`
Agents do not understand the reality constraints	REAL	In the example task, agents do not get the reality constraints of the simulated device.	`rydberg-ising-critical-point`
Missing physics expert intuition	INT	In the example task rollouts, agents do not consider Rydberg decay when analyzing sources of error. An AMO physicist would naturally consider this mechanism, suggesting a gap in the agents' physical intuition. Also, sometimes in the thinking traces, agents are aware of the instructions but still fail in the end since they lack the expert-level physics intuition needed to apply them correctly.	`rydberg-cft-modulation-spectroscopy` `multiplexing-ion-chain-qnet`
Incoherent reasoning	COH	• Contradictory logic chains or arguments. In the example task rollouts, agents realize that they are wrong, but they later forget this.	`kinked-ladder-polymer-scattering`
Lack of rigorous thinking	RIG	For example, circular self-verification. In the example task rollouts, agents' self-verification method depends on their own assumption, which is wrong, so they can never find their mistakes.	`calibrated-cbct-local-reconstruction` `QNTK-research`
False negative	FALSE	Agents do the task correctly but still fail because of a task-design issue (unclear `task.md`, a verifier or rubric that contradicts the instructions, an overly strict tolerance) or a rubric grading mistake.	
Misgrade	MIS	Certain test or rubric results are not correct for this trial, but they do not influence whether this trial passes or not.	
Agents' execution mistakes	EXE	Agents cannot write proper code or use the tools properly, etc.	

Continued on next page

Table 4: Failure mode taxonomy (continued).

Tag	Label	Description	Examples
Hallucination	**HAL**	Agents fabricate evidence, data, experiments, literature, etc.	
Reward hacking	**HACK**	Agents try to hack to get the full reward by: • Trying to change verifiers or hack rubrics to find answers; • Trying to directly find answers online and copy them to the final workspace to pretend the job has been done, etc.	

E.1 Final Trial Labels and Main Findings

Each trial retains all labels and their order from the final review sheet, including locally rerun trials. PASS may coexist with other annotations. ^{V} preserves a verifier marker from the sheet. An em dash in the finding column means no finding was supplied. For `baf2-ih001-strain-energy`, the three columns correspond to local Trials 4–6. Task numbers follow Table 3.

Table 5: Final review-sheet labels for 56 tasks. Cyan FALSE cells identify the 9 trials whose final pass/fail outcomes should change; MIS denotes verifier or rubric errors that do not change the final outcome.

#	Task	Trial 1	Trial 2	Trial 3	Main finding
Condensed Matter Physics					
1	`arpes-sample-alignment`	RIG LIT MIS	LIT RIG	LIT MIS	Agents recognized tilt–translation coupling but missed pose-recovery tolerances and omitted electron-optics or inverse-optimization references. Reviewers overlooked executed one-coordinate scans in Trials 1 and 3, although both still failed pose recovery.
2	`baf2-ih001-strain-energy`	INT RIG	INT RIG	RIG	Displayed Trials 1–2 insufficiently justified bulk-like ice selection and failed AgI transfer. Displayed Trial 3 passed AgI but did not justify its selected ice population despite crop-dependent results.
3	`cpt-frustrated-chain-spectra`	RIG	PASS	PASS	—
4	`eigenstate-optical-response-derivation`	INS MIS	INS MIS	INS	Agents did not start from the required current-density operator. PDF extraction and Lean environment setup also produced erroneous checks, while other verifier and blocker failures remained.
5	`lattice-band-geometry-discovery`	EXE LIT	FALSE	LIT	Trial 1 omitted a data-manifest row. Trial 2's converged result was rejected by an insufficiently resolved verifier grid. Trial 3 passed all tests but incompletely addressed proof scope and references.
6	`li-graphene-mlip-transferability`	MIS	PASS	PASS	—
7	`llzo-proton-exchange-thermodynamics`	COH	COH	COH	All three papers claim monotonically decreasing octahedral Li and strictly octahedral-first exchange, contradicting their own tables. Trial 2 explicitly prints contradictory values; Trial 3 retains the claim despite passing all verifier tests and publishability review.
8	`lpbf-melt-pool-window`	RIG	RIG	RIG	Agents recommended unreliable process windows, retaining unstable candidates and marking unsafe settings as stable_screening.
9	`mlg-quantum-hall-dot-hartree-fock`	PHY INT MIS	PHY INT MIS	PHY MIS	Agents left external confinement unscreened and missed the low-energy spin-transport channel. Reviewers also deducted points for a Conclusion section not required by the task.
10	`moire-continuum-model-construction`	INS MIS	PASS	INS MIS	Agents used valid physical models and passed every numerical test, but Trials 1 and 3 failed a rubric blocker under strict rules absent from the prompt.
11	`rixs-entanglement-witness`	RIG	PASS	PASS	—

Continued on next page

Table 5: Final review-sheet labels and task findings (continued).

#	Task	Trial 1	Trial 2	Trial 3	Main finding
12	smt-reflow-window	RIG	RIG	RIG	Agents did not validate defect mechanisms using physical stress cases. Bridge and cold-joint scores failed to respond correctly to clearly risky paste or thermal conditions.
AMO Physics					
13	ion-trap-3d-vs-2d	INT MIS	INT	INT	Agents solved the electrostatics correctly but failed to connect static trap depth to experimental ion loss. Trial 1 also received inconsistent grading on pseudopotential validity.
14	ion-trap-design	FALSE	RIG COH	INT RIG COH	Trial 1 met trap-design and operating-configuration requirements; Trial 2 met only the former. Trial 3 applied inappropriate gapless electrostatics to a slitted trap and lacked independent verification despite recognizing its importance.
15	multiplexing-ion-chain-qnet	INT	INT	INT	Agents could not fit Rabi-flop data with a model combining coherent excitation and thermal heating.
16	phase-controlled-optical-tweezer-holography	FALSE	FALSE	FALSE	—
17	rydberg-cft-modulation-spectroscopy	INT REAL PHY	REAL	REAL	All trials omitted device calibration. Trial 1 additionally misunderstood CFT, inadequately assessed Rydberg blockade, and omitted decay as an error source. Its rubric review also contained both false-negative and false-positive judgments.
18	rydberg-cz-phase-noise-infidelity	PASS	PASS	PASS	All three trials passed the verifiers and blockers, with only minor mistakes.
19	rydberg-ising-critical-point	REAL MIS INS LIT	REAL MIS INS LIT	REAL MIS INS LIT	Agents exceeded realistic shot budgets, omitted calibration, misunderstood open-boundary finite-size numerics, and weakly synthesized literature. Grading also relied on rigid section-word checks and an unclear nearest-neighbor blockade assumption.
Quantum Information					
20	QNTK-research	MIS RIG INS	MIS RIG INS	INS MIS RIG	Agents did not use the required circuit model. Their self-verification reused their own assumptions and provided weak checks.
21	cvqkd-calibrated-detector-noise	PHY LIT	PHY LIT	PHY LIT	Agents' model assumptions conflicted with the source reference: they misunderstood task conditions and insufficiently supported the discussion with literature. Detector characterization, CSV schemas, and PDF checks passed in all three trials.
22	ising-qml-dwave	RIG	LIT RIG	LIT RIG	The physics and optimization were correct in all trials; failures concerned the analysis and write-up.

Continued on next page

Table 5: Final review-sheet labels and task findings (continued).

#	Task	Trial 1	Trial 2	Trial 3	Main finding
23	time-bin-qkd-product-povm	INT LIT	PHY INT INS LIT	INT LIT	Agents computed the coarse-graining matrices correctly but inadequately explained security implications, prior-work limitations, and applicability conditions. Trial 2's initial search violated the literature cutoff and exposed the source paper; the trace does not establish copying.
24	transmon-cz-pulse-engineering	INT	INT LIT	INT INS	Agents implemented the physics correctly but failed line-accuracy tests after recognizing noise in the final third of their results and publishing those values without further smoothing.
Astrophysics					
25	desi-bao-sn-dark-energy	INS INT RIG	INS COH INT RIG	INT RIG	Agents reported a BAO peak location without first checking that a peak existed. Trials 1 and 2 also used the wrong result-template separator.
26	ems-scalar-spectral-census	RIG	INT	PASS	Trial 1 missed hidden edge cases. Trial 2 explicitly considered the edge case; Trial 3 avoided it through a different design.
27	nonlinear-boson-star-sideband-threshold	RIG MIS	EXE^{V} RIG^{V} LIT	PASS	Trial 1's mode classifier missed edge cases and miscounted nodes. Its grading also depended on a precision requirement absent from task.md.
28	ns-eos-multimessenger	FALSE	RIG	PASS	Trial 1 was penalized for an ambiguous prior convention. Trial 2 inconsistently enforced stable-branch selection, and its input guard rejected hidden data. The physics was largely correct; papers omitted equations and caveats present in code.
29	pulsar-timing-phase-connection	RIG MIS	RIG MIS	RIG MIS	Agents applied a standard threshold without adapting or validating it for this specific case.
30	solar-wind-turbulence-spectral-break	PASS	PASS	PASS	All three trials passed.
Fluid Dynamics					
31	consistent-two-phase-ls-vof	FALSE	FALSE	FALSE	Trial 2's verifier failure may stem from an unspecified accuracy requirement that allowed premature stopping.
32	diffuse-ib-rigid-body-fsi	PASS INT MIS	PASS	PASS	Agents implemented surface versus volumetric forcing as superficial point-placement differences rather than using the correct fractional-step stage. Reviewers also penalized coding before a detailed plan, which the task did not require.
33	iga-branched-incompressible-flow	RIG EXE	RIG EXE	RIG EXE	Agents failed hidden cases with different geometries, indicating insufficient consideration of edge cases and incorrect mesh-repair code.
34	material-agnostic-visual-inverse-dynamics	RIG INT	INT	INT	Agents failed to design sufficient supervision for stable constitutive learning.
35	multi-object-visual-inverse-dynamics	INT	INT	INT	Agents failed to distinguish constitutive errors from geometry and contact errors.

Continued on next page

Table 5: Final review-sheet labels and task findings (continued).

#	Task	Trial 1	Trial 2	Trial 3	Main finding
36	twisted-slinky-wave-coupling	INT LIT	INT LIT	INT LIT	Agents treated the spring as a whole instead of analyzing transverse motion at the cross-sectional and coil scales. They therefore failed to explain transverse–torsional coupling and provided only qualitative relations and overall trend fits.
Statistical Physics					
37	glass-nonequilibrium-landscape-evolution	RIG	RIG	RIG	Agents identified the phenomena but failed to construct and validate an abstract numerical model.
38	kinked-ladder-polymer-scattering	INT COH	INT COH	INT COH	Agents retained an incorrect sign rule while tuning other parameters, then claimed validation despite documenting the unresolved problem in the appendix.
39	learning-credit-assignment	INS COH	INS COH	INS COH	All three trials contradicted their stated methodology when defining connection roles and ignored the single-number output requirement.
40	low-dimensional-deep-network-learning	INS LIT	INS INT LIT	INS INT EXE LIT	Agents performed the mathematics correctly but selected the less-efficient, unshared design and did not recognize that their theory required a simplifying assumption.
41	metallic-glass-pel-ruggedness	RIG	RIG	PASS	Agents recognized the described differences but did not identify a quantitative measure with physical implications.
42	quantized-ising-3sat-encoding	INS LIT	INS LIT	INS LIT	All three trials solved the formulas; the failures concerned the papers.
Photonics & Imaging					
43	calibrated-cbct-local-reconstruction	INS LIT	INS LIT	RIG LIT	—
44	fbg-rail-axle-inversion	PASS	PASS	PASS	All three trials passed.
45	fbg-strain-amplifier-design	INS MIS	INS MIS	INS MIS	Agents may have misinterpreted one task requirement, leading them to violate another.
46	heat-assisted-detection-and-ranging	REAL INT RIG EXE	REAL INT RIG EXE	REAL INT RIG EXE	Agents overlooked emissivity structure, real sensing conditions, and range identifiability, then failed to implement stable decomposition of temperature, emissivity, texture, and distance.
47	hspic-spectrometer-e2e	MIS	MIS	PASS	—
High Energy Physics					
48	corsika-air-shower-inverse	INT LIT	INT LIT	INT LIT	—
49	icecube-pointsource-unbinned	EXE	EXE MIS	EXE MIS	Agents installed Numba in their own sandboxes and submitted code importing it, but the grading image lacked Numba. Trial 2's flagged DOI is real.
50	o3-monodromy-line-defect-mc	RIG	RIG	RIG	Agents asserted rather than derived the two analytical results required by the rubric blockers.
51	o3-surface-cft-fuzzy-sphere	RIG LIT	RIG LIT	RIG LIT	Agents misidentified the second and third S_z=1 states and did not establish the ordinary boundary correctly.

Continued on next page

Table 5: Final review-sheet labels and task findings (continued).

#	Task	Trial 1	Trial 2	Trial 3	Main finding
52	sipm-spe-gain-breakdown	INS LIT	PASS	RIG	—
Chemical Physics					
53	agi-surface-hydrophobicity	INT RIG	INT RIG MIS	INT RIG	Agents avoided the intended basal interface by selecting a nonpolar prism. Cross-surface probes and system-size justification were incomplete; Trials 2–3 lacked stable estimates. Trial 3's water model conflicted with the planned study; reviewers overlooked Trial 2's conclusion.
54	hole-bifurcation-ml-landscape	MIS	MIS	PHY RIG	—
55	p2d-cell-model-identification	RIG LIT	RIG	RIG	Agents correctly identified unidentifiable parameters and missing heat, yet their models failed to generalize to unseen conditions.
56	smfret-research-v5	PHY INT LIT	PHY LIT RIG	PHY COH RIG LIT	Agents used incorrect physical models and unsupported literature, and their conclusions did not match the evidence.

F AGENT RESOURCE USE

Figures 14–16 compare overall pass rate (Table 2, overall Avg@3) with agent execution time, agent steps, and output tokens per rollout. Each resource is averaged within each task and then equally across tasks with recorded values, using the same selected rollouts as Table 2. A step denotes one tool invocation. The black lines mark the empirical Pareto frontiers for lower resource use and higher pass rate. Figures 17–19 show how each resource is distributed over tasks and models, in the layout of Figures 10–13. Their colors follow a logarithmic scale, and the few values beyond its ends take the end colors. Appendix F.1 gives the measurement definitions and coverage.

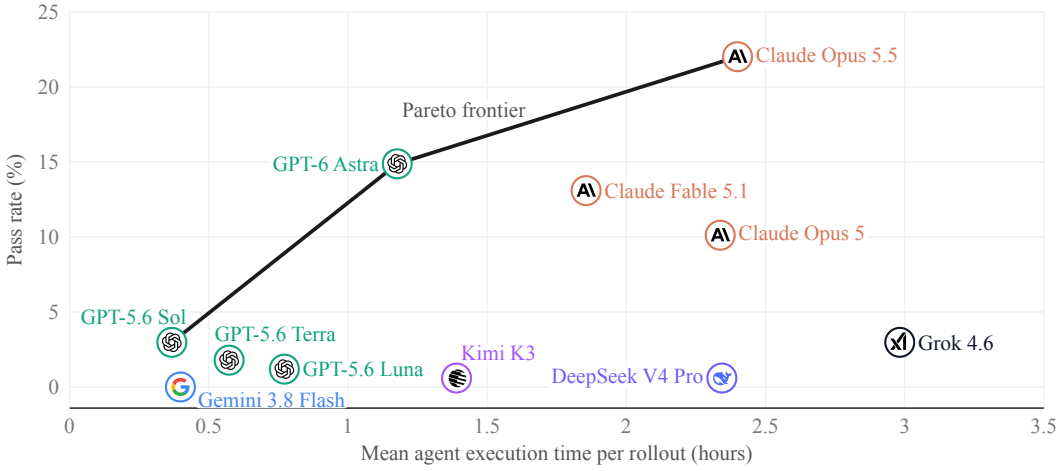

Figure 14: Mean agent execution time per rollout versus overall pass rate (%). Time is measured in hours and excludes environment setup, verification and rubric review.

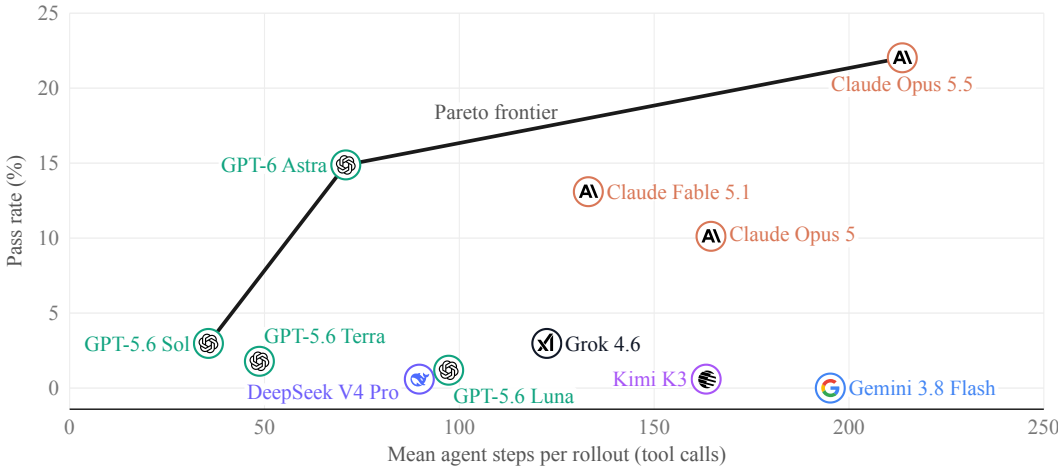

Figure 15: Mean agent steps per rollout versus overall pass rate (%). Steps count recorded tool invocations, including failed and interrupted calls.

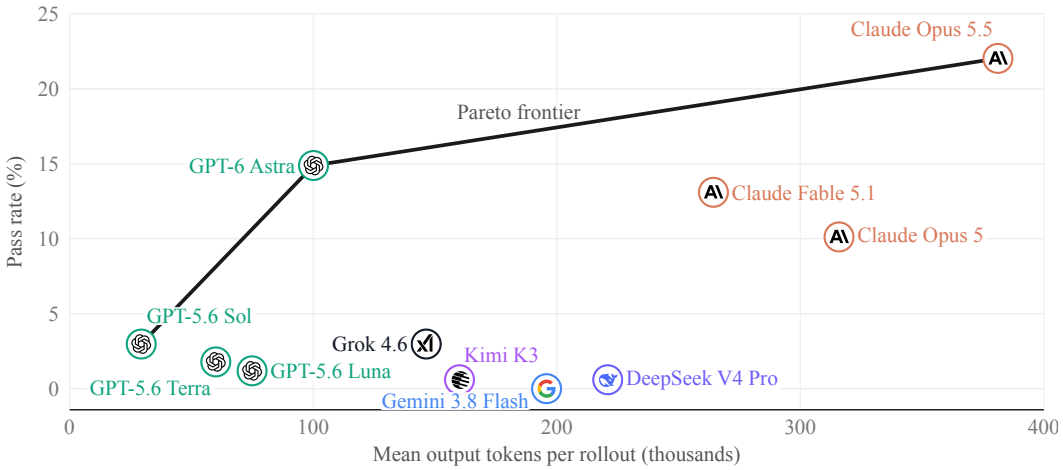

Figure 16: Mean reported output tokens per rollout, in thousands, versus overall pass rate (%). Token counts follow provider or harness usage accounting.

F.1 Agent Resource Accounting

The resource plots use the same selected rollouts and 56-task overall pass rate as the cost comparison. For each resource, available values are averaged within each task and then equally across tasks. Unavailable counters are omitted from resource means, not replaced with zero. The table below reports the number of rollouts and tasks contributing to each resource estimate.

Agent time is the recorded execution-stage wall-clock duration, excluding environment setup, verification and rubric review. When an idle-timeout record lacks that stage timer, its recorded elapsed execution time supplies the duration. Agent steps count recorded tool invocations, including failed and interrupted calls. A partial trace is used only when a termination monitor supplies a tool-call counter. Output tokens use the reported agent output-token count; unavailable usage records are omitted. Tokenization and output-token accounting can differ between providers. In Claude Code runs, agent time and steps include subagent work; output tokens do too, except for Claude Opus 5.5 and Claude Fable 5.1, whose Claude Code usage reports cover only the main agent.

Table 6: Resource coverage: contributing rollouts / tasks for each model.

Model	Agent time	Agent steps	Output tokens
Claude Opus 5.5	162 / 56	162 / 56	162 / 56
GPT-6 Astra	168 / 56	168 / 56	168 / 56
Claude Fable 5.1	166 / 56	166 / 56	166 / 56
Claude Opus 5	162 / 54	162 / 54	162 / 54
GPT-5.6 Sol	168 / 56	168 / 56	168 / 56
Grok 4.6	168 / 56	168 / 56	168 / 56
GPT-5.6 Terra	168 / 56	168 / 56	168 / 56
GPT-5.6 Luna	168 / 56	168 / 56	168 / 56
Kimi K3	168 / 56	168 / 56	168 / 56
DeepSeek V4 Pro	166 / 56	161 / 56	168 / 56
Gemini 3.8 Flash	162 / 54	162 / 54	162 / 54

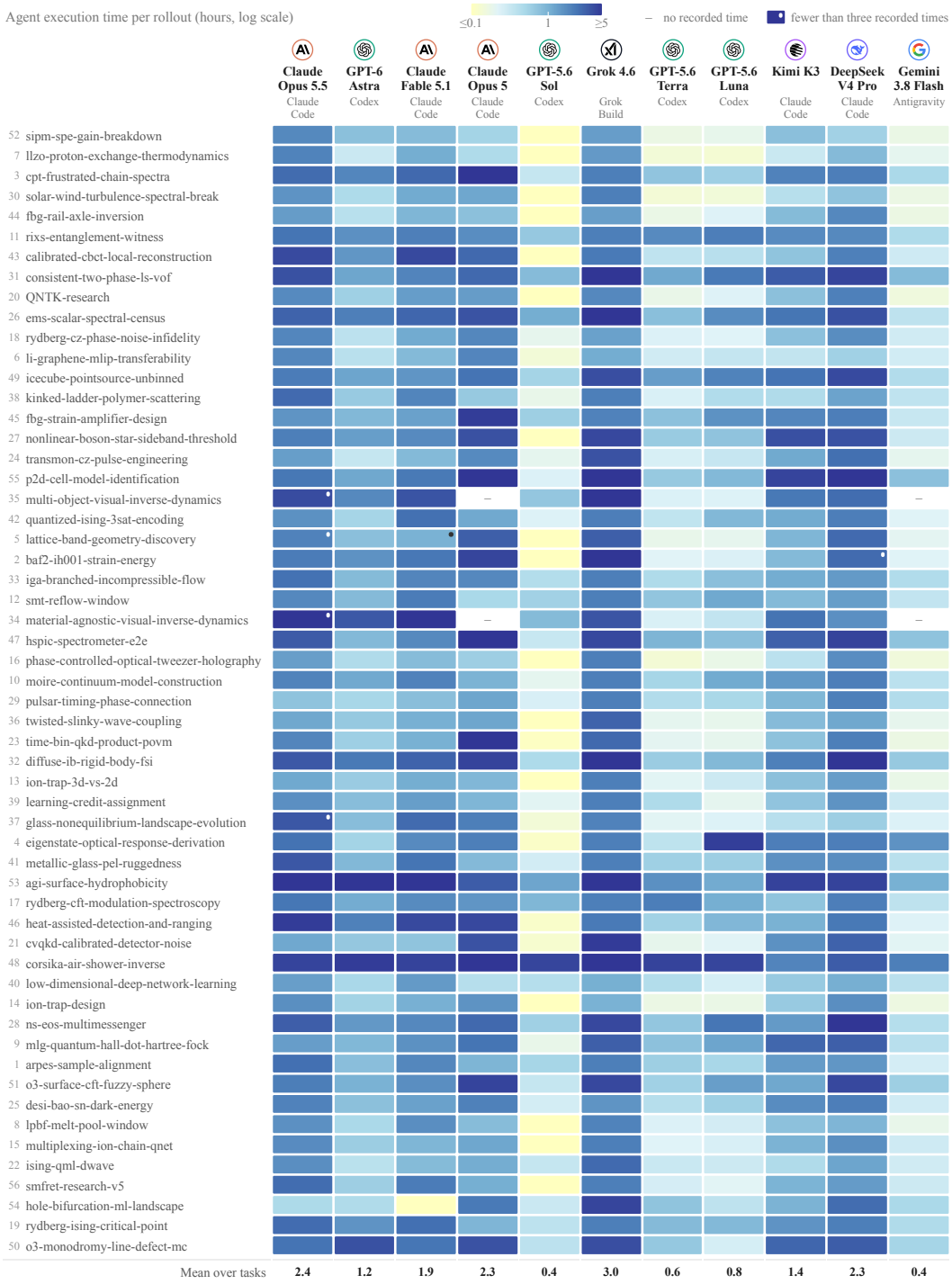

Figure 17: Agent execution time per rollout, in hours, per task and model: the mean over the recorded rollouts in each cell. The bottom row gives each model's mean over tasks with records, as plotted in Figure 14. A dash marks a task without a recorded time and a dot one with fewer than three.

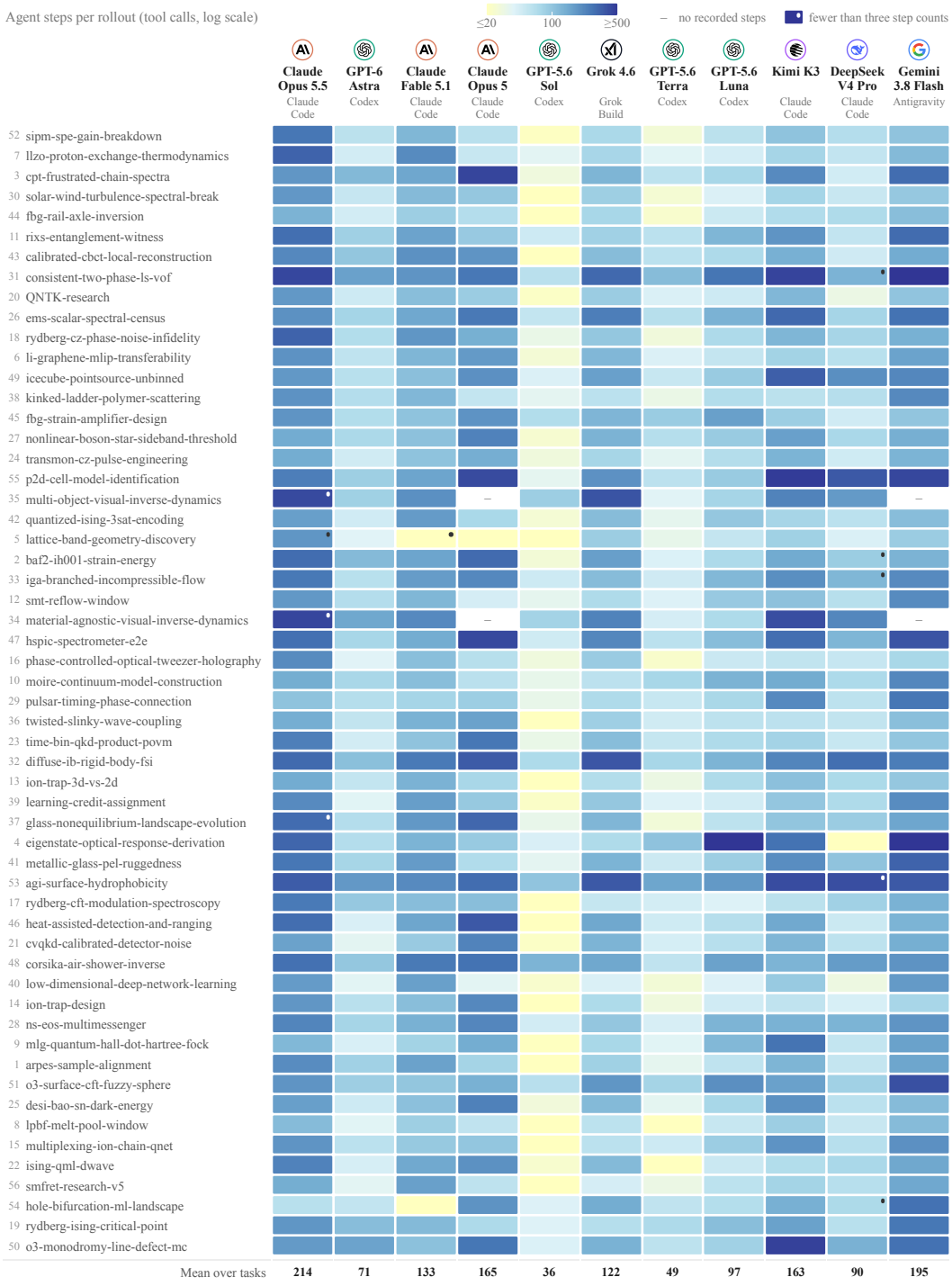

Figure 18: Agent steps (tool calls) per rollout, per task and model: the mean over the recorded rollouts in each cell. The bottom row gives each model's mean over tasks with records, as plotted in Figure 15. Rows follow Figure 10. A dash marks a task without a recorded step count and a dot one with fewer than three.

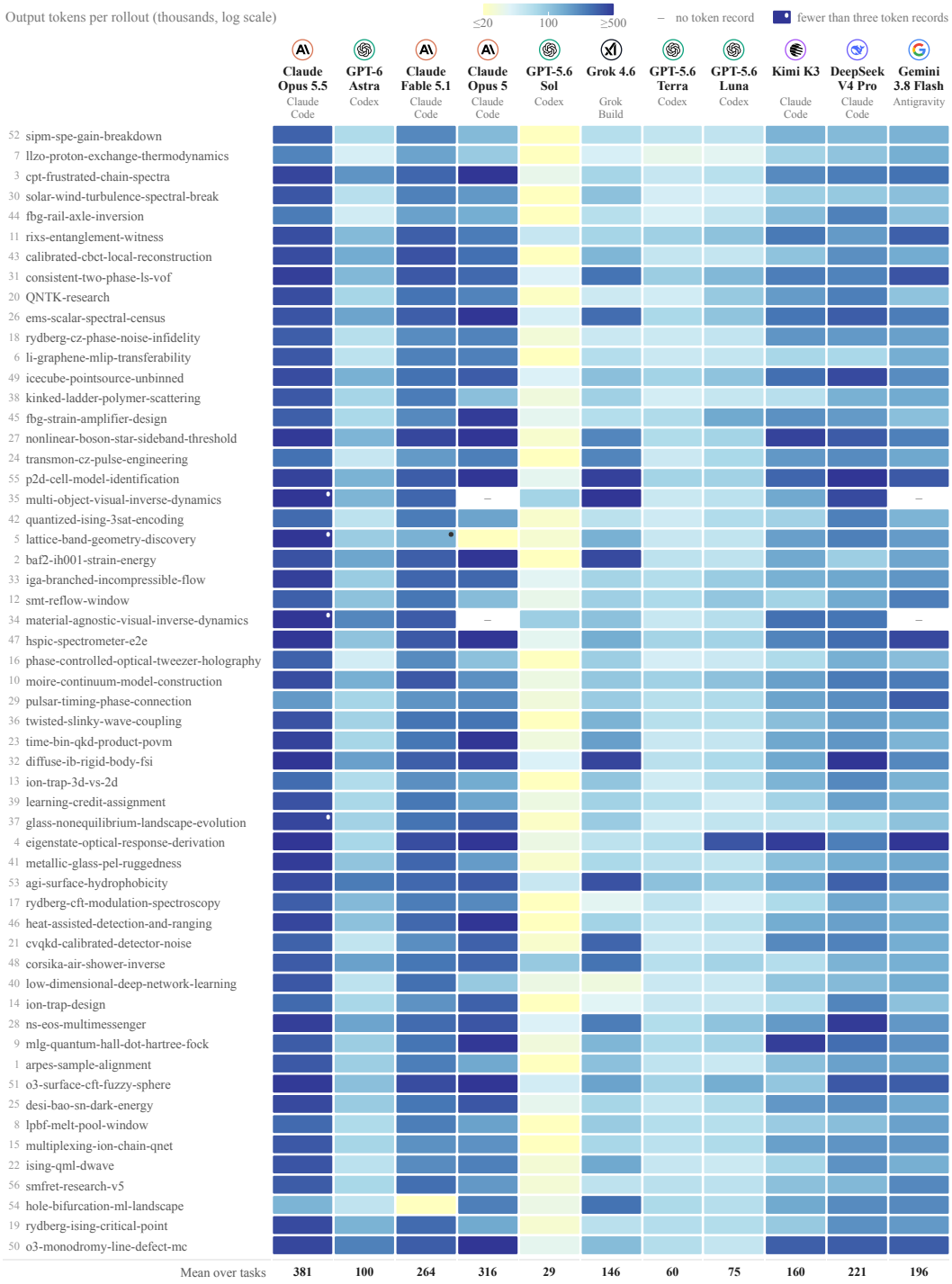

Figure 19: Output tokens per rollout, in thousands, per task and model: the mean over the recorded rollouts in each cell. The bottom row gives each model's mean over tasks with records, as plotted in Figure 16. Rows follow Figure 10. A dash marks a task without an output-token record and a dot one with fewer than three.

G REFERENCE-MATERIAL EVALUATION

Table 7: Evaluating rubric scores for all 33 tasks where the authors provided a manuscript or paper as a reference. Only the rubrics, that actually evaluated/ needed the present manuscript were evaluated. If no blocking rubric applied to the manuscript, it is shown as a dash. If no scoring rubric applied to the manuscript it is shown as NA.

Task / reference	Weighted score	Blockers
QNTK-research	100%	2/2
consistent-two-phase-ls-vof	100%	1/1
cpt-frustrated-chain-spectra	N/A	**0/1 - fail**
cvqkd-calibrated-detector-noise	100%	**2/3 - fail**
diffuse-ib-rigid-body-fsi	100%	-
eigenstate-optical-response-derivation	100%	5/5
fbg-rail-axle-inversion	100%	-
fbg-strain-amplifier-design	100%	2/2
glass-nonequilibrium-landscape-evolution - bundled report	**91.94%**	2/2
glass-nonequilibrium-landscape-evolution - original paper	**50%**	-
heat-assisted-detection-and-ranging	100%	-
hole-bifurcation-ml-landscape	**81.58%**	-
hspic-spectrometer-e2e	100%	3/3
icecube-pointsource-unbinned	100%	-
ion-trap-3d-vs-2d - bundled report	100%	4/4
ion-trap-3d-vs-2d - original paper	100%	1/1
ion-trap-design - bundled report	100%	-
ion-trap-design - original paper	100%	-
lattice-band-geometry-discovery	100%	1/1
learning-credit-assignment	100%	4/4
li-graphene-mlip-transferability	**93.52%**	2/2
llzo-proton-exchange-thermodynamics	100%	-
low-dimensional-deep-network-learning	100%	7/7
material-agnostic-visual-inverse-dynamics	100%	-
metallic-glass-pel-ruggedness	**72.73%**	1/1
moire-continuum-model-construction	**80%**	-
multi-object-visual-inverse-dynamics	100%	-
multiplexing-ion-chain-qnet	100%	-
o3-surface-cft-fuzzy-sphere	100%	-
phase-controlled-optical-tweezer-holography	N/A	**0/1 - fail**
rixs-entanglement-witness	100%	-
rydberg-cft-modulation-spectroscopy	100%	4/4
rydberg-cz-phase-noise-infidelity	100%	6/6
rydberg-ising-critical-point	100%	3/3
sipm-spe-gain-breakdown	100%	-
time-bin-qkd-product-povm	100%	5/5