

REPLICATIONBENCH: CAN AI AGENTS REPLICATE ASTROPHYSICS RESEARCH PAPERS?

Christine Ye¹, Sihan Yuan¹, Suchetha Cooray¹, Steven Dillmann¹, Ian L. V. Roque¹,
Dalya Baron¹, Philipp Frank¹, Sergio Martin-Alvarez¹, Nolan Koblichke², Frank J. Qu¹,
Diyi Yang¹, Risa Wechsler¹, Ioana Ciucă¹

¹Stanford University ²University of Toronto

ABSTRACT

Frontier AI agents show increasing promise as scientific research assistants, and may eventually be useful for extended, open-ended research workflows. However, in order to use agents for novel research, we must first assess the underlying faithfulness and correctness of their work. To evaluate agents as research assistants, we introduce ReplicationBench, an evaluation framework that tests whether agents can replicate entire research papers drawn from the astrophysics literature. Astrophysics, where research relies heavily on archival data and computational study while requiring little real-world experimentation, is a particularly useful testbed for AI agents in scientific research. We split each paper into tasks which require agents to replicate the paper’s core contributions, including the experimental setup, derivations, data analysis, and codebase. Each task is co-developed with the original paper authors and targets a key scientific result, enabling objective evaluation of both faithfulness (adherence to original methods) and correctness (technical accuracy of results). ReplicationBench is extremely challenging for current frontier language models: even the best-performing language models score under 20%. We analyze ReplicationBench trajectories in collaboration with domain experts and find a rich, diverse set of failure modes for agents in scientific research. ReplicationBench establishes the first benchmark of paper-scale, expert-validated astrophysics research tasks, reveals insights about agent performance generalizable to other domains of data-driven science, and provides a scalable framework for measuring AI agents’ reliability in scientific research.

1 INTRODUCTION

Current AI agents already demonstrate expert-level scientific reasoning, and agents built on large language models are increasingly being applied across the research lifecycle, from literature review and hypothesis generation (Su et al., 2025; Saeedi et al., 2025a; Iyer et al., 2024) to experimental design, data analysis, and coding (Yamada et al., 2025; Gottweis et al., 2025; Baek et al., 2025; Schmidgall & Moor, 2025; Boiko et al., 2023; Saeedi et al., 2025b). However, existing experiments using agents for scientific research demonstrate failures in robustness and domain understanding (Beel et al., 2025).

Astrophysics is a particularly promising scientific domain for developing and studying AI-assisted scientific research. First, while in other domains LLM-based agents may be limited by real-world interaction and experimentation, the day-to-day workflows of most astrophysics researchers are entirely computational, so the necessary tools can be easily sandboxed for agent use. Moreover, the astrophysics research community has strong standards for reproducibility and data sharing (Allen et al., 2012; Rebull et al., 2017); it is common practice for papers to be published with the full datasets required to replicate core results. **Thus, astrophysics is a promising testbed for studying the capabilities and failure modes of scientific research agents.**

To produce high-quality, original scientific research, agents need both expert-grade domain knowledge and high-level autonomous capabilities. **Currently, few methods exist to measure the usefulness of**

AI agents in open-ended scientific research tasks, especially astrophysics. Existing benchmarks testing domain understanding, such as MMLU, GPQA, and HLE, only evaluate static knowledge (Hendrycks et al., 2021; Rein et al., 2023; Phan et al., 2025). Alternatively, existing agent benchmarks assess planning (Valmeekam et al., 2023; Zheng et al., 2024), tool use (Yao et al., 2024), and coding (Jimenez et al., 2024; Hu et al., 2024; Kwa et al., 2025), but do not require domain expertise. Current benchmarks fall short of actually evaluating research ability, and evaluating long-horizon agency and expert-level conceptual understanding through realistic scientific contexts.

In this work, we introduce **ReplicationBench**, a benchmark to evaluate AI agents on end-to-end replications of astrophysics research papers. We choose to measure replication on the basis that the ability of agents to execute correct, faithful replications of *existing* expert-level research is likely predictive of agent quality and reliability during open-ended workflows in similar domains. Replicating established results also provides extremely challenging but well-scoped testbeds for study. We operationalize this by constructing a benchmark of high-quality, reproducible astrophysics research papers to replicate. Given a paper’s manuscript and metadata, an AI agent must completely reproduce the implementation, methodology, and core findings, and is scored on the accuracy of its final results. By enforcing that agents exactly reproduce the original paper, ReplicationBench enables precise measurement of agents’ instruction-following **faithfulness** and execution **correctness** within astrophysics research contexts.

ReplicationBench is designed for **sandboxing**, **realism**, and **scalability**. To **sandbox** evaluations, replications do not require real-world experimentation, intermediate results from other works, or expensive computing setups. To construct **realistic** tasks, we directly use the core results of real research papers, generating tasks which require expert-level reasoning, data analysis, implementation, and scientific interpretation. Regarding **scalability**, rather than requiring extensive human-designed rubrics (as in Starace et al. (2025)), ReplicationBench grades final results directly as the outcome of the integrated process, reducing data generation and evaluation costs.

The core contributions of this work are:

- **ReplicationBench Dataset:** We publish an original dataset of 19 reproducible, self-contained astrophysics papers decomposed into 107 research-grade tasks. We also publish the ReplicationBench-Plus extension, a set of 11 additional papers with 58 LLM-generated, expert-curated tasks.
- **Evaluation Framework for Scientific Agents:** We establish the broader ReplicationBench evaluation framework, measuring agent reliability in astrophysics research by measuring replication success in existing papers, and provide scaffolding for future dataset expansions. The ReplicationBench framework can be extended to evaluate agents in any data-driven scientific research process.
- **Frontier Language Model Evaluations:** We benchmark five frontier language models in agentic environments. We also present detailed evaluations of agent traces by the authors of the original papers.

ReplicationBench introduces a general approach for evaluating AI agents on scientific research replication, and establishes a challenging standard for agents as reliable scientific collaborators in astrophysics. We release the [ReplicationBench dataset](#) for use by the community.

2 RELATED WORK

Many existing benchmarks evaluate agents on tasks related to code understanding and software engineering, which are core to agentic scientific research. SWE-Bench (Jimenez et al., 2024) evaluates agents on patching real-world GitHub issues, offering a realistic evaluation setting but a shorter agent horizon. MLEBench (Chan et al., 2025), MLEAgentBench (Huang et al., 2024), RE-Bench (Wijk et al., 2024), and InfiAgent-DABench (Hu et al., 2024) assess agents on machine learning or data analysis tasks in narrower settings. Domain-specific benchmarks like BixBench (Mitchener et al., 2025), TPBench (Chung et al., 2025), LAB-Bench (Laurent et al., 2024), and GravityBench (Koblischke et al., 2025) evaluate expert-level knowledge and reasoning in scientific domains, but are limited in open-endedness.

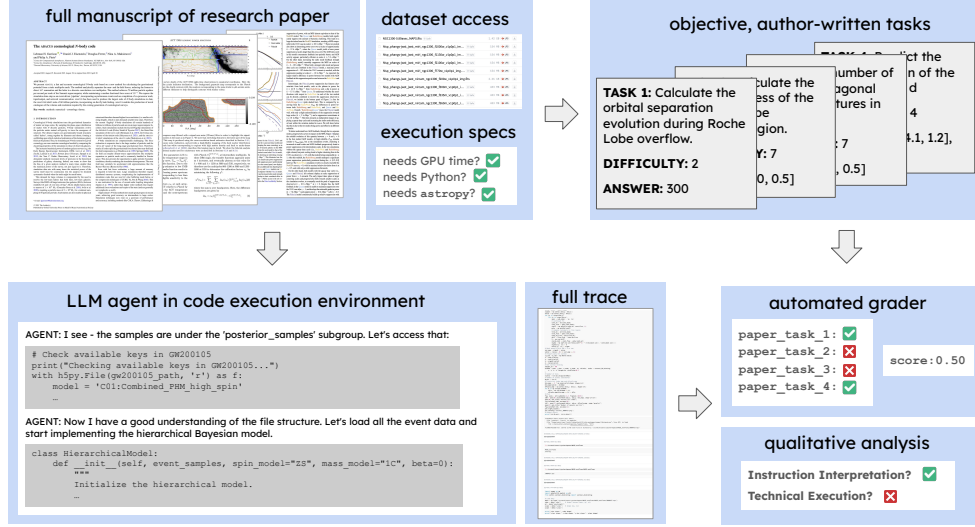


Figure 1: ReplicationBench evaluates AI agents on end-to-end astrophysics research paper replication. For each paper, we supply the manuscript, dataset, and execution metadata, and a complete set of author-written tasks. We evaluate LLM agents in multi-turn code execution environments and conduct both automatic and expert human grading of traces.

Some existing benchmarks also attempt to evaluate language model agents on long-horizon scientific tasks. CURIE (Cui et al., 2025) evaluates both reasoning and in-domain expertise using tasks based on scientific papers, but aims for well-scoped subtasks rather than full replication. CORE-Bench (Siegel et al., 2024) tests agents on replicating scientific paper results from the complete code repository, but does not evaluate agents on from-scratch implementation. DiscoveryBench (Majumder et al., 2024) evaluates agents using papers from economics, sociology, and 4 other domains; they decompose papers into tasks by separating *hypotheses*, using a structured formalism specifically designed for data-driven hypothesis testing. PaperBench (Starace et al., 2025) evaluates agents on complete replications of machine learning research papers, including end-to-end development. ReplicationBench differs in its focus on astrophysics, its generic approach to decomposing full research papers into benchmark tasks, and its scalable evaluation through end-result grading instead of human-written rubrics, which reduces data generation and scoring costs.

Existing work has attempted to build LLM-based agents for scientific discovery (Lu et al., 2024; Yamada et al., 2025; Gottweis et al., 2025). Several have been designed for workflows in astrophysics or cosmology (Laverick et al., 2024; Xu et al., 2025; Bolliet & the CMBAGENT team, 2025).

3 REPLICATIONBENCH

The fundamental unit of ReplicationBench is a single, self-contained research **paper**. For each paper, we provide the complete **manuscript**, metadata with specifications for the **dataset** and **execution environment**, and a set of expert-written benchmark **tasks**. ReplicationBench’s core dataset consists of 107 expert-written tasks spanning 19 peer-reviewed, highly reproducible scientific research papers. ReplicationBench-Plus contains an optional extension of 58 expert-curated, automatically generated tasks from 11 additional papers, for a total of 165 tasks across 30 papers.

To complete the tasks for a paper, the agent must implement the complete methodology, and produce several results matching the paper’s core contributions. This process may include loading and pre-processing data, implementing new software, completing complex analysis procedures, and performing mathematical computations. Completed tasks are scored for correctness against the ground truth from the paper. A full illustration is provided in Figure 1.

Difficulty	1	2	3	4	5	6	7	8	9
Number of Tasks	11	11	24	24	14	10	5	4	4

Table 1: Task difficulty score distribution. Difficulty is rated according to time estimated for a human expert to complete a task, from 1 (few minutes) to 10 (1+ months).

3.1 PAPER SELECTION

We conducted an open call for authors to submit candidates within multiple astrophysics research communities. To enhance diversity, we also identified additional candidate papers and reached out to authors. From this set we select 19 papers for ReplicationBench on the following criteria:

Paper Quality and Reproducibility We require all papers to have undergone peer-review in an academic journal. This ensures ReplicationBench is representative of high-quality real-world physics research. To verify reproducibility of results, authors submit source code and intermediate data products.

Compute Limitations We require that the paper’s core contributions have total execution requirements of <6 hours on a standard 4-core CPU workstation and a single GPU. If a minority of the paper’s core contributions require expensive computations, such as running standard simulations, authors instead submit intermediate data products with the dataset and limit the scope of tasks accordingly. This ensures that agent rollouts on ReplicationBench tasks are not overly bottlenecked by environment execution time.

Data Availability We collect complete datasets for each paper, utilizing public data from open-source archives and private data provided by authors. If the original paper involved some fresh observations, we require that the data collection process itself was not part of the methodology contributions.

Empirical Findings We select only papers which contribute multiple empirical, numerical results. Empirical results may include estimates for fundamental physical parameters, observations regarding the paper’s experiments, training losses for machine learning models, or direct calculation results. These results are the “ground truth” for replication.

Collaboration with Authors We seek permission to use all papers in the ReplicationBench core dataset, and select only papers whose authors were willing to collaborate closely.

3.2 TASK DESIGN

To guide the replication process for a given paper, each paper in ReplicationBench is decomposed into tasks. Tasks are defined by detailed instructions referencing the manuscript, and can span a range of real-world challenges in research workflows, including data loading and processing, parameter estimation, machine learning applications, and theoretical calculations; a complete breakdown by category is provided in Appendix Table 6. Examples are provided in Figure 2. Tasks are written in collaboration with the paper authors according to the following specifications:

Coverage Tasks span the paper’s core contributions such that successful completion of the tasks indicates successful replication of the paper’s results. If a given task depends on completion of previous, easier tasks, a task’s metadata includes its **parents**. Where applicable, individual tasks require models to submit several related results from the same analysis (e.g. a parameter estimate and its 90% CI), raising the standard of correctness.

Objectivity ReplicationBench consists of objectively gradable **numeric** or **code** tasks. **Numeric** tasks produce a pre-specified data structure with numeric entries that can be compared directly against a ground truth value. **Code** tasks are accompanied by custom grading functions or unit tests, which are used to score agent-submitted code.

Difficulty and Partial Credit Each task is rated with a **difficulty** score that estimates the time for a non-author domain expert to complete the task, organized on a logarithmic scale from 1 (few minutes) to 9 (several weeks). In order to enable partial credit for replication attempts, we curate tasks across a range of **difficulty** scores. Tasks have an average difficulty rating of 4.1; see Table 1.

Guess-Proofing Manuscripts are masked to remove values that allude to task ground truths. Agents are not permitted to use the paper’s original codebase, view unmasked text, or install restricted packages. For details on the masking procedure, see Appendix C.

Modality Tasks only require natural language inputs and outputs. Figures are cleaned from manuscripts.

3.3 SCALABLE DATA COLLECTION

While manually designing tasks for ReplicationBench requires significant human effort, hybrid human-LLM task generation can substantially reduce the human workload, enabling scalable expansion. We introduce the ReplicationBench-Plus extension, consisting of 58 tasks spanning 11 papers built using this hybrid approach.

We utilize papers published with the ShowYourWork reproducibility framework (Luger et al., 2021). ShowYourWork enables paper authors to publish the paper’s end-to-end computational workflow, such that the paper’s core results can be automatically re-generated when the paper manuscript is compiled. This creates a dataset of highly reproducible and self-contained research papers. From the complete set of public ShowYourWork repositories on GitHub (461 total), we select repositories with complete ShowYourWork metadata, an attached peer-reviewed publication with sufficient empirical results, and source code utilizing raw data inputs. We extract ReplicationBench metadata from the repository and generate up to 10 tasks using a language model¹ (see prompt in Appendix D.2) prompted with access to the unmasked manuscript and dataset/executor metadata. We then provide the LLM-generated tasks to domain experts for correctness review and curation according to the criteria in 3.2.

4 EVALUATION DESIGN

4.1 AGENT SETUP

We evaluate language models as agents with a simple tool-use scaffold². Agents are evaluated using a code sandbox in a Singularity³ container. Our sandboxing setup allows Python code execution, bash shell execution, file read/write commands, and a `submit()` tool via standard language model tool-calling. Datasets are pre-downloaded to reduce uninteresting technical hurdles. Agents iteratively write and edit a final submission, which is parsed for grading.

We limit agents to a maximum of 6 hours of execution time, including both code runtime and language model response time, and up to 5 million total tokens as counted by the language model API. We find that in the majority of traces, agents choose to finish earlier than these limits, either claiming to have completed the replication or to have faced an insurmountable problem. We report average runtimes and token usage in 5.

4.2 AUTOMATED GRADING

We include author-submitted **tolerance** values in the ReplicationBench dataset. Tolerance values adjust for variation between equally correct implementations due to sampling differences or numerical effects. We grade individual numeric values as correct if they are within \pm tolerance of the correct value, and assign credit for a task if all values are correct. As our main metric, we compute aggregated scores for each paper and report the paper-averaged score.

¹Gemini 2.5 Pro

²implementation from Inspect AI, https://github.com/UKGovernmentBEIS/inspect_ai

³Apptainer 1.4.0

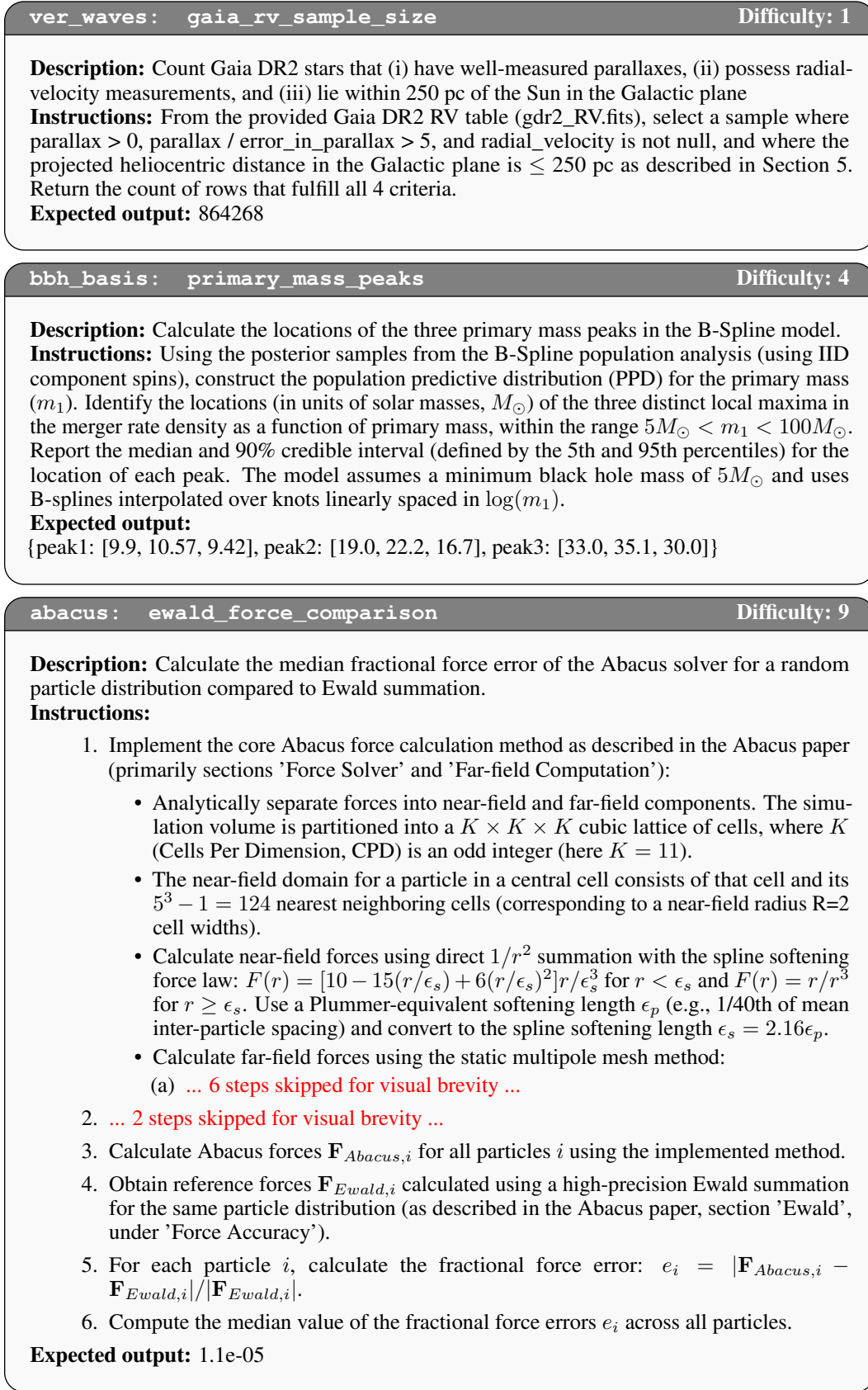


Figure 2: Three illustrative tasks of varying difficulties. Some instructions are edited for brevity.

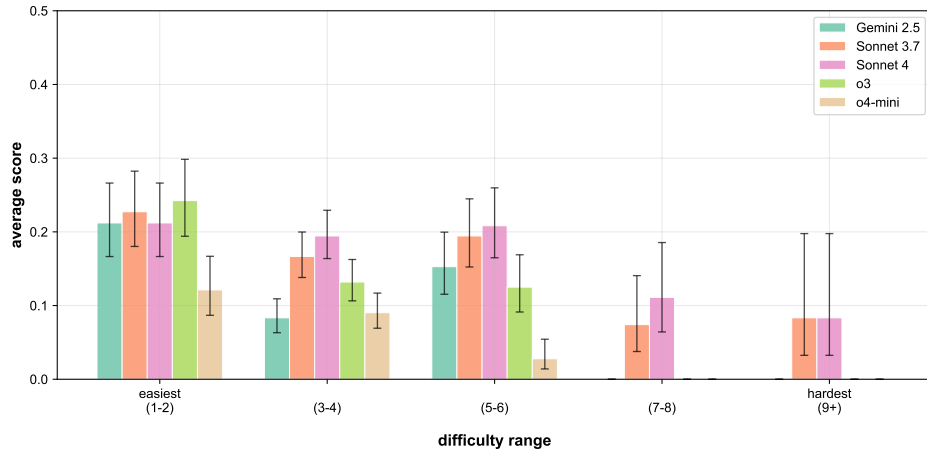


Figure 3: Unweighted ReplicationBench scores, with tasks binned by difficulty.

4.3 QUALITATIVE EVALUATIONS

We seek to develop a detailed qualitative understanding of agent capabilities and failure modes through a two-stage qualitative evaluation process. First, we work directly with paper authors to examine the correctness and faithfulness of individual agent traces. Experts provide both numeric and qualitative assessments of the agent’s **Intent Interpretation**, the agent’s understanding of the goal, necessary inputs, and success criteria; the **Quality of Execution**, the reasoning coherence and implementation correctness; and the level of **Cheating or Hacking**, the prevalence of guessing or workarounds. Full rubrics are provided in Appendix C.5. As detailed human evaluations are costly, we categorize the subjective feedback on agent traces into distinct failure modes. We then utilize an LLM-as-a-judge system⁴ to automatically label traces.

5 EXPERIMENTS

We measure the performance of five frontier LLMs: Claude 3.7 Sonnet, Claude 4 Sonnet, Gemini 2.5 Pro, OpenAI o3, and OpenAI o4-mini.⁵ We run each agent for 3 attempts per paper on ReplicationBench.

5.1 REPLICATIONBENCH SCORES

ReplicationBench is extremely difficult for current frontier language models: with basic tool-use scaffolding, the best-performing model, Claude 3.7 Sonnet, achieves an average score of just 19.3%. Results on ReplicationBench are provided in Table 2; ReplicationBench-Plus results are provided in Appendix C.1. Unsurprisingly, models tend to perform better on tasks with lower human-rated difficulty (Figure 3).

In our evaluation setup, agents can submit final results at any time. We find that some of the performance variation between frontier models can be attributed to differences in task completion rates: agents sometimes submit early and without attempting all tasks.

5.2 QUALITATIVE EVALUATIONS

Using the expert evaluations of agent traces, we identified three primary failure modes:

⁴based on claude-sonnet-4-20250514

⁵claude-3-7-sonnet-20250219, claude-sonnet-4-20250514, gemini-2.5-pro-preview-05-06, o3-2025-04-16, o4-mini-2025-04-16.

Model	Unweighted score	Standard dev.	Difficulty-weighted score	Tokens	Runtime
Gemini 2.5	0.106	0.013	0.078	36,586	147.5
Sonnet 3.7	0.193	0.015	0.174	91,286	118.2
Sonnet 4	0.183	0.042	0.166	36,288	91.4
o3	0.136	0.011	0.111	9,527	9.7
o4-mini	0.080	0.013	0.074	10,195	15.6

Model	Tasks completed	Standard dev.	Accuracy over completed tasks
Gemini 2.5	0.505	0.081	0.214
Sonnet 3.7	0.933	0.011	0.249
Sonnet 4	0.939	0.058	0.206
o3	0.687	0.048	0.187
o4-mini	0.582	0.050	0.164

Table 2: Average-of-3 scores on ReplicationBench with simple agent scaffold; tokens count output and reasoning.

Model	Memorization rate, without thinking	with thinking
Gemini 2.5	5.9%	6.0%
Sonnet 3.7	7.6%	8.1%
Sonnet 4	4.1%	5.4%
o3	8.0%	8.3%
o4-mini	7.1%	8.0%

Table 3: ReplicationBench answer memorization rates, averaged across 10 samples per paper.

- **Lack of persistence:** agents often submit early while (incorrectly) citing computational limitations or impossible tasks, matching observations from Starace et al. (2025). Models are more likely to attempt easier tasks.
- **Procedural errors:** agents often execute incorrect procedures due to conceptual misunderstanding (i.e. using incorrect corrections or transformations for a calculation). Agents also make procedural errors due to instruction-following limitations, especially by implementing a simplified version of the core methodology.
- **Technical execution failures:** agents sometimes failed to execute correct plans due to limited technical ability. This includes not accounting for resource constraints, generating incorrect code from domain-specific dependencies, or failing to read data in legacy formats.

5.3 MEMORIZATION AND CHEATING

Human experts occasionally flagged instances of potential cheating, such as agents fabricating answers or using best-guess fallback outputs. Because ReplicationBench is based on published research papers, and the evaluation process only considers end results, benchmark scores may be affected by memorization or contamination, allowing agents to “guess” answers.

We estimate the extent to which memorization affects ReplicationBench scores using a “no-computation” baseline: we measure model performance by removing code execution tools from the ReplicationBench evaluation while still forcing models to submit best guesses for every task. If models have memorized the correct values from pre-training, then they may be able to “guess” benchmark answers exactly without actually executing the experiments. Overall, we measure memorization rates under 9% (see Table 3). We caution both this rate is not a strict upper bound on actual contamination, as it only measures memorization within the replication task context. We also find that in practice, prompted language model monitors can easily catch instances of “guessing”; thus this behavior can be detected and penalized using a language model judge.

Agents might also more blatantly cheat if the key values in the manuscript were unmasked, allowing direct copying. When ablating manuscript masking, we find that certain agents score 15-20%+ higher than baseline by direct copying of values rather than performing actual computations, in violation of the user prompt (see Appendix Table 8). This “cheating” behavior often comes after several failed attempts at completing the task. However, with improved prompting and manuscript masking, we see no evidence for guessing or cheating in our final evaluation runs.

6 LIMITATIONS

Dataset Consistency Difficulty ratings, task decompositions, and resource requirements are estimates from the paper authors. While we provide detailed guidance to authors to help calibration, these parts of the ReplicationBench dataset inherently have some variation.

Scope We choose to focus only on replication, restricting correct agent outputs to a set of known results. This allows more isolated measurements of execution correctness and instruction-following reliability. However, ReplicationBench does not measure agents’ ability to independently design research projects or iterate on truly open-ended directions. Moreover, ReplicationBench only focuses on whether agents can replicate papers from astrophysics, a testbed likely predictive of other domains of data-driven science, but a nonetheless narrow subject matter.

Evaluation Settings We study the performance of language models with simple agent scaffolds in standard code execution tool-use environments. However, we have not fully explored the range of possible tools and affordances, such as web browsing, multi-modal input, memory tools, or access to a human supervisor. Exploring these affordances may help elicit greater performance from the same base model or agent.

7 DISCUSSION

In this work, we introduced ReplicationBench, a benchmark to evaluate AI agents as astrophysics research assistants through reproducibility challenges from existing scientific papers. ReplicationBench is the first evaluation to comprehensively measure agent capabilities in end-to-end astrophysics research workflows. At present, ReplicationBench tasks are extremely challenging. Frontier language models, which saturate most static knowledge benchmarks, achieve ReplicationBench scores just above 20% using our agent scaffolds. Qualitative evaluations reveal that frontier AI agents can successfully reproduce some scientific results but still fall short in both execution abilities and conceptual understanding. We hope ReplicationBench will be a revealing and useful environment for the community, facilitating explorations into different setups for AI-assisted research.

Given that the core skills required for ReplicationBench tasks are broadly useful beyond astrophysics, ReplicationBench may be predictive of agent performance in many other data-driven research fields. The ReplicationBench evaluation framework can also be applied to other scientific domains with computational workflows. In particular, we are excited about expanding the ReplicationBench framework to new domains using substantially cheaper hybrid human-LLM approaches for data generation, as demonstrated through the ReplicationBench-Plus extension. While ReplicationBench focuses on astrophysics as the initial testbed, we hope that this work supports a foundation for broader progress in AI-assisted science.

REFERENCES

- Alice Allen, Kimberly DuPrie, Bruce Berriman, Robert J. Hanisch, Jessica D. Mink, and Peter J. Teuben. Astrophysics source code library. *ArXiv*, abs/1212.1916, 2012. URL <https://api.semanticscholar.org/CorpusID:41142765>.
- Philipp Arras, Philipp Frank, Philipp Haim, Jakob Knollmüller, Reimar Leike, Martin Reinecke, and Torsten Enßlin. Variable structures in M87* from space, time and frequency resolved interferometry. *Nature Astronomy*, 6:259–269, January 2022. doi: 10.1038/s41550-021-01548-0.
- Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. <https://arxiv.org/pdf/2404.07738>, 2025. Accessed: 2025-04-17.
- Dalya Baron, Hagai Netzer, Dieter Lutz, Ric I. Davies, and J. Xavier Prochaska. Not So Windy After All: MUSE Disentangles AGN-driven Winds from Merger-induced Flows in Galaxies along the Starburst Sequence. *Astrophysical Journal*, 968(1):23, June 2024. doi: 10.3847/1538-4357/ad39e9.
- Dalya Baron, Karin M. Sandstrom, Jessica Sutter, Hamid Hassani, Brent Groves, Adam K. Leroy, Eva Schinnerer, Médéric Boquien, Matilde Brazzini, Jérémy Chastenet, Daniel A. Dale, Oleg V. Egorov, Simon C. O. Glover, Ralf S. Klessen, Debosmita Pathak, Erik Rosolowsky, Frank Bigiel, Mélanie Chevance, Kathryn Grasha, Annie Hughes, J. Eduardo Méndez-Delgado, Jérôme Pety, Thomas G. Williams, Stephen Hannon, and Sumit K. Sarbadhicary. PHANGS-ML: The Universal Relation between PAH Band and Optical Line Ratios across Nearby Star-forming Galaxies. *Astrophysical Journal*, 978(2):135, January 2025. doi: 10.3847/1538-4357/ad972a.
- Joeran Beel, Min-Yen Kan, and Moritz Baumgart. Evaluating sakana’s ai scientist for autonomous research: Wishful thinking or an emerging reality towards ‘artificial research intelligence’ (ari)? <https://arxiv.org/html/2502.14297v2>, 2025. Accessed: 2025-04-17.
- Morgan Bennett and Jo Bovy. Vertical waves in the solar neighbourhood in Gaia DR2. *Monthly Notices of the RAS*, 482(1):1417–1425, January 2019. doi: 10.1093/mnras/sty2813.
- Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, December 2023.
- Boris Bolliet and the CMBAGENT team. Cmbagent: Open-source multi-agent system for science, 2025. URL <https://github.com/CMBAgents/cmbagent>. Available at <https://github.com/CMBAgents/cmbagent>.
- Tobias Buck, Aura Obreja, Bridget Ratcliffe, Yuxi(Lucy) Lu, Ivan Minchev, and Andrea V Macciò. The impact of early massive mergers on the chemical evolution of milky way-like galaxies: insights from nihao-uhd simulations. *Monthly Notices of the Royal Astronomical Society*, 523(1):1565–1576, May 2023. ISSN 1365-2966. doi: 10.1093/mnras/stad1503. URL <http://dx.doi.org/10.1093/mnras/stad1503>.
- Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, and Aleksander Mądry. Mle-bench: Evaluating machine learning agents on machine learning engineering, 2025. URL <https://arxiv.org/abs/2410.07095>.
- 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, 2025. URL <https://arxiv.org/abs/2502.15815>.
- Suchetha Cooray, Tsutomu T. Takeuchi, Daichi Kashino, Shuntaro A. Yoshida, Hai-Xia Ma, and Kai T. Kono. Characterizing and understanding galaxies with two parameters. *Monthly Notices of the RAS*, 524(4):4976–4995, October 2023. doi: 10.1093/mnras/stad2129.
- Hao Cui, Zahra Shamsi, Gowoon Cheon, Xuejian Ma, Shutong Li, Maria Tikhonovskaya, Peter Norgaard, Nayantara Mudur, Martyna Plomecka, Paul Raccuglia, Yasaman Bahri, Victor V. Albert, Pranesh Srinivasan, Haining Pan, Philippe Faist, Brian Rohr, Ekin Dogus Cubuk, Muratahan

-
- Aykol, Amil Merchant, Michael J. Statt, Dan Morris, Drew Purves, Elise Kleeman, Ruth Alcantara, Matthew Abraham, Muqthar Mohammad, Ean Phing VanLee, Chenfei Jiang, Elizabeth Dorfman, Eun-Ah Kim, Michael P Brenner, Viren Jain, Sameera Ponda, and Subhashini Venugopalan. Curie: Evaluating llms on multitask scientific long context understanding and reasoning, 2025. URL <https://arxiv.org/abs/2503.13517>.
- Trevor J. David, Ruth Angus, Jason L. Curtis, Jennifer L. van Saders, Isabel L. Colman, Gabriella Contardo, Yuxi Lu, and Joel C. Zinn. Further evidence of modified spin-down in sun-like stars: Pileups in the temperature–period distribution. *The Astrophysical Journal*, 933(1):114, July 2022. ISSN 1538-4357. doi: 10.3847/1538-4357/ac6dd3. URL <http://dx.doi.org/10.3847/1538-4357/ac6dd3>.
- Jiayin Dong and Daniel Foreman-Mackey. A hierarchical bayesian framework for inferring the stellar obliquity distribution, 2023. URL <https://arxiv.org/abs/2305.14220>.
- Liam O. Dubay, Jennifer A. Johnson, and James W. Johnson. Galactic Chemical Evolution Models Favor an Extended Type Ia Supernova Delay-time Distribution. *Astrophysical Journal*, 973(1):55, September 2024. doi: 10.3847/1538-4357/ad61df.
- Bruce Edelman, Ben Farr, and Zoheyr Doctor. Cover your basis: Comprehensive data-driven characterization of the binary black hole population. *The Astrophysical Journal*, 946(1):16, March 2023. ISSN 1538-4357. doi: 10.3847/1538-4357/acb5ed. URL <http://dx.doi.org/10.3847/1538-4357/acb5ed>.
- Amanda M. Farah, Thomas A. Callister, Jose María Ezquiaga, Michael Zevin, and Daniel E. Holz. No need to know: astrophysics-free gravitational-wave cosmology, 2024. URL <https://arxiv.org/abs/2404.02210>.
- Lehman H. Garrison, Daniel J. Eisenstein, Douglas Ferrer, Nina A. Maksimova, and Philip A. Pinto. The ABACUS cosmological N-body code. *Monthly Notices of the RAS*, 508(1):575–596, November 2021. doi: 10.1093/mnras/stab2482.
- Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, José R Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Towards an ai co-scientist, 2025. URL <https://arxiv.org/abs/2502.18864>.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL <https://arxiv.org/abs/2009.03300>.
- Xueyu Hu, Ziyu Zhao, Shuang Wei, Ziwei Chai, Qianli Ma, Guoyin Wang, Xuwu Wang, Jing Su, Jingjing Xu, Ming Zhu, Yao Cheng, Jianbo Yuan, Jiwei Li, Kun Kuang, Yang Yang, Hongxia Yang, and Fei Wu. Infiagent-dabench: Evaluating agents on data analysis tasks, 2024. URL <https://arxiv.org/abs/2401.05507>.
- Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. Mlagentbench: Evaluating language agents on machine learning experimentation, 2024. URL <https://arxiv.org/abs/2310.03302>.
- Kartheik G. Iyer, Mikael Yunus, Charles O’Neill, Christine Ye, Alina Hyk, Kiera McCormick, Ioana Ciucă, John F. Wu, Alberto Accomazzi, Simone Astarita, Rishabh Chakrabarty, Jesse Cranney, Anjalie Field, Tirthankar Ghosal, Michele Ginolfi, Marc Huertas-Company, Maja Jabłońska, Sandor Kruk, Huiling Liu, Gabriel Marchidan, Rohit Mistry, J. P. Naiman, J. E. G. Peek, Mugdha Polimera, Sergio J. Rodríguez Méndez, Kevin Schawinski, Sanjib Sharma, Michael J. Smith, Yuan-Sen Ting, and Mike Walmsley. pathfinder: A semantic framework for literature review and knowledge discovery in astronomy. *The Astrophysical Journal Supplement Series*, 275 (2):38, November 2024. ISSN 1538-4365. doi: 10.3847/1538-4365/ad7c43. URL <http://dx.doi.org/10.3847/1538-4365/ad7c43>.

-
- 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?, 2024. URL <https://arxiv.org/abs/2310.06770>.
- Daisuke Kawata, Junichi Baba, Ioana Ciucă, Mark Cropper, Robert J. J. Grand, Jason A. S. Hunt, and George Seabroke. Radial distribution of stellar motions in Gaia DR2. *Monthly Notices of the RAS*, 479(1):L108–L112, September 2018. doi: 10.1093/mnras/sly107.
- M. A. Kenworthy, D. González Picos, E. Elizondo, R. G. Martin, D. M. van Dam, J. E. Rodriguez, G. M. Kennedy, C. Ginski, M. Mugrauer, N. Vogt, C. Adam, and R. J. Oelkers. Eclipse of the v773 tau b circumbinary disc. *Astronomy & Astrophysics*, 666:A61, October 2022. ISSN 1432-0746. doi: 10.1051/0004-6361/202243441. URL <http://dx.doi.org/10.1051/0004-6361/202243441>.
- Matthew Kenworthy, Simon Lock, Grant Kennedy, Richelle van Capelleveen, Eric Mamajek, Ludmila Carone, Franz-Josef Hambsch, Joseph Masiero, Amy Mainzer, J. Davy Kirkpatrick, Edward Gomez, Zoë Leinhardt, Jingyao Dou, Pavan Tanna, Arttu Sainio, Hamish Barker, Stéphane Charbonnel, Olivier Garde, Pascal Le Dû, Lionel Mulato, Thomas Petit, and Michael Rizzo Smith. A planetary collision afterglow and transit of the resultant debris cloud. *Nature*, 622(7982): 251–254, October 2023. doi: 10.1038/s41586-023-06573-9.
- Nolan Koblishke, Hyunseok Jang, Kristen Menou, and Mohamad Ali-Dib. Gravity-Bench-v1: A Benchmark on Gravitational Physics Discovery for Agents. *arXiv e-prints*, art. arXiv:2501.18411, January 2025. doi: 10.48550/arXiv.2501.18411.
- Nolan W. Koblishke and Richard I. Anderson. Calibrating and Standardizing the Tip of the Red Giant Branch in the Small Magellanic Cloud Using Small-amplitude Red Giants. *Astrophysical Journal*, 974(2):181, October 2024. doi: 10.3847/1538-4357/ad7095.
- Sandor Kruk, Pablo García-Martín, Marcel Popescu, Ben Aussel, Steven Dillmann, Megan E. Perks, Tamina Lund, Bruno Merín, Ross Thomson, Samet Karadag, and Mark J. McCaughrean. The impact of satellite trails on Hubble Space Telescope observations. *Nature Astronomy*, 7:262–268, March 2023. doi: 10.1038/s41550-023-01903-3.
- Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring ai ability to complete long tasks, 2025. URL <https://arxiv.org/abs/2503.14499>.
- Jon M. Laurent, Joseph D. Janizek, Michael Ruza, Michaela M. Hinks, Michael J. Hammerling, Siddharth Narayanan, Manvitha Ponnappati, Andrew D. White, and Samuel G. Rodrigues. Lab-bench: Measuring capabilities of language models for biology research, 2024. URL <https://arxiv.org/abs/2407.10362>.
- Andrew Laverick, Kristen Surrao, Inigo Zubeldia, Boris Bolliet, Miles Cranmer, Antony Lewis, Blake Sherwin, and Julien Lesgourgues. Multi-Agent System for Cosmological Parameter Analysis, 11 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, 2024. URL <https://arxiv.org/abs/2408.06292>.
- Rodrigo Luger, Megan Bedell, Daniel Foreman-Mackey, Ian J. M. Crossfield, Lily L. Zhao, and David W. Hogg. Mapping stellar surfaces III: An Efficient, Scalable, and Open-Source Doppler Imaging Model. *arXiv e-prints*, art. arXiv:2110.06271, October 2021.
- Hai-Xia Ma, Tsutomu T. Takeuchi, Suchetha Cooray, and Yongda Zhu. sOPTICS: a modified density-based algorithm for identifying galaxy groups/clusters and brightest cluster galaxies. *Monthly Notices of the RAS*, 537(2):1504–1517, February 2025. doi: 10.1093/mnras/staf115.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhi-
jeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark.
Discoverybench: Towards data-driven discovery with large language models, 2024. URL
<https://arxiv.org/abs/2407.01725>.

Sergio Martin-Alvarez, Vid Iršič, Sophie Koudmani, Martin A. Bourne, Leah Bigwood, and Debora
Sijacki. Stirring the cosmic pot: how black hole feedback shapes the matter power spectrum
in the FABLE simulations. *Monthly Notices of the RAS*, 539(2):1738–1755, May 2025. doi:
10.1093/mnras/staf470.

Ludovico Mitchener, Jon M Laurent, 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, 2025. URL <https://arxiv.org/abs/2503.00096>.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin
Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra,
Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Dmitry
Dodonov, Tung Nguyen, Jaeho Lee, Daron Anderson, Mikhail Doroshenko, Alun Cennyth Stokes,
Mobeen Mahmood, Oleksandr Pokutnyi, Oleg Iskra, Jessica P. Wang, John-Clark Levin, Mstyslav
Kazakov, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou,
Zihan Wang, Serguei Popov, Robert Gerbicz, Geoff Galgon, Johannes Schmitt, Will Yeadon,
Yongki Lee, Scott Sauers, Alvaro Sanchez, Fabian Giska, Marc Roth, Søren Riis, Saiteja Utpala,
Noah Burns, Gashaw M. Goshu, Mohinder Maheshbhai Naiya, Chidozie Agu, Zachary Giboney,
Antrell Cheatom, Francesco Fournier-Facio, Sarah-Jane Crowson, Lennart Finke, Zerui Cheng,
Jennifer Zampese, Ryan G. Hoerr, Mark Nandor, Hyunwoo Park, Tim Gehringer, Jiaqi Cai, Ben
McCarty, Alexis C Garretson, Edwin Taylor, Damien Sileo, Qiuyu Ren, Usman Qazi, Lianghui
Li, Jungbae Nam, John B. Wydallis, Pavel Arkhipov, Jack Wei Lun Shi, Aras Bacho, Chris G.
Willcocks, Hangrui Cao, Sumeet Motwani, Emily de Oliveira Santos, Johannes Veith, Edward
Vendrow, Doru Cojoc, Kengo Zenitani, Joshua Robinson, Longke Tang, Yuqi Li, Joshua Vendrow,
Natanael Wildner Fraga, Vladyslav Kuchkin, Andrey Pupasov Maksimov, Pierre Marion, Denis
Efremov, Jayson Lynch, Kaiqu Liang, Aleksandar Mikov, Andrew Gritsevskiy, Julien Guillod,
Gözdenur Demir, Dakotah Martinez, Ben Pageler, Kevin Zhou, Saeed Soori, Ori Press, Henry Tang,
Paolo Rissone, Sean R. Green, Lina Brüssel, Moon Twayana, Aymeric Dieuleveut, Joseph Marvin
Imperial, Ameya Prabhu, Jinzhou Yang, Nick Crispino, Arun Rao, Dimitri Zvonkine, Gabriel
Loiseau, Mikhail Kalinin, Marco Lukas, Ciprian Manolescu, Nate Stambaugh, Subrata Mishra, Tad
Hogg, Carlo Bosio, Brian P Coppola, Julian Salazar, Jaehyeok Jin, Rafael Sayous, Stefan Ivanov,
Philippe Schwaller, Shaipranesh Senthilkuma, Andres M Bran, Andres Algaba, Kelsey Van den
Houte, Lynn Van Der Sypt, Brecht Verbeken, David Noever, Alexei Kopylov, Benjamin Myklebust,
Bikun Li, Lisa Schut, Evgenii Zheltonozhskii, Qiaochu Yuan, Derek Lim, Richard Stanley, Tong
Yang, John Maar, Julian Wykowski, Martí Oller, Anmol Sahu, Cesare Giulio Ardito, Yuzheng Hu,
Ariel Ghislain Kemogne Kamdoun, Alvin Jin, Tobias Garcia Vilchis, Yuexuan Zu, Martin Lackner,
James Koppel, Gongbo Sun, Daniil S. Antonenko, Steffi Chern, Bingchen Zhao, Pierrot Arsene,
Joseph M Cavanagh, Daofeng Li, Jiawei Shen, Donato Crisostomi, Wenjin Zhang, Ali Dehghan,
Sergey Ivanov, David Perrella, Nurdin Kaparov, Allen Zang, Ilia Sucholutsky, Arina Kharlamova,
Daniil Orel, Vladislav Poritski, Shalev Ben-David, Zachary Berger, Parker Whitfill, Michael Foster,
Daniel Munro, Linh Ho, Shankar Sivarajan, Dan Bar Hava, Aleksey Kuchkin, David Holmes,
Alexandra Rodriguez-Romero, Frank Sommerhage, Anji Zhang, Richard Moat, Keith Schneider,
Zakayo Kazibwe, Don Clarke, Dae Hyun Kim, Felipe Meneguitti Dias, Sara Fish, Veit Elser,
Tobias Kreiman, Victor Efren Guadarrama Vilchis, Immo Klose, Ujjwala Anantheshwaran, Adam
Zweiger, Kaivalya Rawal, Jeffery Li, Jeremy Nguyen, Nicolas Daans, Haline Heidinger, Maksim
Radionov, Václav Rozhoň, Vincent Ginis, Christian Stump, Niv Cohen, Rafał Poświata, Josef
Tkadlec, Alan Goldfarb, Chenguang Wang, Piotr Padlewski, Stanislaw Barzowski, Kyle Mont-
gomery, Ryan Stendall, Jamie Tucker-Foltz, Jack Stade, T. Ryan Rogers, Tom Goertzen, Declan
Grabb, Abhishek Shukla, Alan Givré, John Arnold Ambay, Archan Sen, Muhammad Fayez Aziz,
Mark H Inlow, Hao He, Ling Zhang, Younesse Kaddar, Ivar Ångquist, Yanxu Chen, Harrison K
Wang, Kalyan Ramakrishnan, Elliott Thornley, Antonio Terpin, Hailey Schoelkopf, Eric Zheng,
Avishy Carmi, Ethan D. L. Brown, Kelin Zhu, Max Bartolo, Richard Wheeler, Martin Stehberger,
Peter Bradshaw, JP Heimonen, Kaustubh Sridhar, Ido Akov, Jennifer Sandlin, Yury Makarychev,
Joanna Tam, Hieu Hoang, David M. Cunningham, Vladimir Goryachev, Demosthenes Patramanis,

Michael Krause, Andrew Redenti, David Aldous, Jesyin Lai, Shannon Coleman, Jiangnan Xu, Sangwon Lee, Ilias Magoulas, Sandy Zhao, Ning Tang, Michael K. Cohen, Orr Paradise, Jan Hendrik Kirchner, Maksym Ovchynnikov, Jason O. Matos, Adithya Shenoy, Michael Wang, Yuzhou Nie, Anna Szyber-Betley, Paolo Faraboschi, Robin Riblet, Jonathan Crozier, Shiv Halasyamani, Shreyas Verma, Prashant Joshi, Eli Meril, Ziqiao Ma, J  r  my Andr  oletti, Raghav Singhal, Jacob Platnick, Volodymyr Nevirkovets, Luke Basler, Alexander Ivanov, Seri Khoury, Nils Gustafsson, Marco Piccardo, Hamid Mostaghimi, Qijia Chen, Virendra Singh, Tran Quoc Kh  nh, Paul Rosu, Hannah Szlyk, Zachary Brown, Himanshu Narayan, Aline Menezes, Jonathan Roberts, William Alley, Kunyang Sun, Arkil Patel, Max Lamparth, Anka Reuel, Linwei Xin, Hanmeng Xu, Jacob Loader, Freddie Martin, Zixuan Wang, Andrea Achilleos, Thomas Preu, Tomek Korbak, Ida Bosio, Fereshteh Kazemi, Ziye Chen, Bir   B  lint, Eve J. Y. Lo, Jiaqi Wang, Maria In  s S. Nunes, Jeremiah Milbauer, M Saiful Bari, Zihao Wang, Behzad Ansarinejad, Yewen Sun, Stephane Durand, Hossam Elgnainy, Guillaume Douville, Daniel Tordera, George Balabanian, Hew Wolff, Lynna Kvistad, Hsiaoyun Milliron, Ahmad Sakor, Murat Eron, Andrew Favre D. O., Shailesh Shah, Xiaoxiang Zhou, Firuz Kamalov, Sherwin Abdoli, Tim Santens, Shaul Barkan, Allison Tee, Robin Zhang, Alessandro Tomasiello, G. Bruno De Luca, Shi-Zhuo Looi, Vinh-Kha Le, Noam Kolt, Jiayi Pan, Emma Rodman, Jacob Drori, Carl J Fossum, Niklas Muennighoff, Milind Jagota, Ronak Pradeep, Honglu Fan, Jonathan Eicher, Michael Chen, Kushal Thaman, William Merrill, Moritz Firsching, Carter Harris, Stefan Ciob  c  , Jason Gross, Rohan Pandey, Ilya Gusev, Adam Jones, Shashank Agnihotri, Pavel Zhelnov, Mohammadreza Mofayez, Alexander Piperski, David K. Zhang, Kostiantyn Dobarskyi, Roman Leventov, Ignat Soroko, Joshua Duersch, Vage Taamazyan, Andrew Ho, Wenjie Ma, William Held, Ruicheng Xian, Armel Randy Zebaze, Mohanad Mohamed, Julian Noah Leser, Michelle X Yuan, Laila Yacar, Johannes Lengler, Katarzyna Olszewska, Claudio Di Fratta, Edson Oliveira, Joseph W. Jackson, Andy Zou, Muthu Chidambaram, Timothy Manik, Hector Haffenden, Dashiell Stander, Ali Dasouqi, Alexander Shen, Bit   Golshani, David Stap, Egor Kreto, Mikalai Uzhou, Alina Borisovna Zhidkovskaya, Nick Winter, Miguel Orbegoza Rodriguez, Robert Lauff, Dustin Wehr, Colin Tang, Zaki Hossain, Shaun Phillips, Fortuna Samuele, Fredrik Ekstr  m, Angela Hammon, Oam Patel, Faraz Farhidi, George Medley, Forough Mohammadzadeh, Madellene Pe  aflor, Haile Kassahun, Alena Friedrich, Rayner Hernandez Perez, Daniel Pyda, Taom Sakal, Omkar Dhamane, Ali Khajegili Mirabadi, Eric Hallman, Kenchi Okutsu, Mike Battaglia, Mohammad Maghsoudimehrabani, Alon Amit, Dave Hulbert, Roberto Pereira, Simon Weber, Handoko, Anton Peristy, Stephen Malina, Mustafa Mehkary, Rami Aly, Frank Reidegeld, Anna-Katharina Dick, Cary Friday, Mukhwinder Singh, Hassan Shapourian, Wanyoung Kim, Mariana Costa, Hubeyb Gurdogan, Harsh Kumar, Chiara Ceconello, Chao Zhuang, Haon Park, Micah Carroll, Andrew R. Tawfeek, Stefan Steinerberger, Daattavya Aggarwal, Michael Kirchhof, Linjie Dai, Evan Kim, Johan Ferret, Jainam Shah, Yuzhou Wang, Minghao Yan, Krzysztof Burdzy, Lixin Zhang, Antonio Franca, Diana T. Pham, Kang Yong Loh, Joshua Robinson, Abram Jackson, Paolo Giordano, Philipp Petersen, Adrian Cosma, Jesus Colino, Colin White, Jacob Votava, Vladimir Vinnikov, Ethan Delaney, Petr Spelda, Vit Stritecky, Syed M. Shahid, Jean-Christophe Mourrat, Lavr Vetoshkin, Koen Sponselee, Renas Bacho, Zheng-Xin Yong, Florencia de la Rosa, Nathan Cho, Xiuyu Li, Guillaume Malod, Orion Weller, Guglielmo Albani, Leon Lang, Julien Laurendeau, Dmitry Kazakov, Fatimah Adesanya, Julien Portier, Lawrence Hollom, Victor Souza, Yuchen Anna Zhou, Julien Degorre, Yi  it Yal  n, Gbenga Daniel Obikoya, Rai, Filippo Bigi, M. C. Bosc  , Oleg Shumar, Kaniuar Bacho, Gabriel Recchia, Mara Popescu, Nikita Shulga, Ngefor Mildred Tanwie, Thomas C. H. Lux, Ben Rank, Colin Ni, Matthew Brooks, Alesia Yakimchyk, Huanxu, Liu, Stefano Cavalleri, Olle H  ggstr  m, Emil Verkama, Joshua Newbould, Hans Gundlach, Leonor Brito-Santana, Brian Amaro, Vivek Vajipey, Rynaa Grover, Ting Wang, Yosi Kratish, Wen-Ding Li, Sivakanth Gopi, Andrea Caciolai, Christian Schroeder de Witt, Pablo Hern  ndez-C  mara, Emanuele Rodol  , Jules Robins, Dominic Williamson, Vincent Cheng, Brad Raynor, Hao Qi, Ben Segev, Jingxuan Fan, Sarah Martinson, Erik Y. Wang, Kaylie Hausknecht, Michael P. Brenner, Mao Mao, Christoph Demian, Peyman Kassani, Xinyu Zhang, David Avagian, Eshawn Jessica Scipio, Alon Ragoler, Justin Tan, Blake Sims, Rebeka Plecnik, Aaron Kirtland, Omer Faruk Bodur, D. P. Shinde, Yan Carlos Leyva Labrador, Zahra Adoul, Mohamed Zekry, Ali Karakoc, Tania C. B. Santos, Samir Shamseldeen, Loukmane Karim, Anna Liakhovitskaia, Nate Resman, Nicholas Farina, Juan Carlos Gonzalez, Gabe Maayan, Earth Anderson, Rodrigo De Oliveira Pena, Elizabeth Kelley, Hodjat Mariji, Rasoul Pouriamanesh, Wentao Wu, Ross Finocchio, Ismail Alarab, Joshua Cole, Danyelle Ferreira, Bryan Johnson, Mohammad Safdari, Liangti Dai, Siriphan Arthornthurasuk, Isaac C. McAlister, Alejandro Jos   Moyano, Alexey Pronin, Jing Fan, Angel Ramirez-Trinidad, Yana Malysheva, Daphiny Pottmaier, Omid Taheri, Stanley Stepanic, Samuel

Perry, Luke Askew, Raúl Adrián Huerta Rodríguez, Ali M. R. Minissi, Ricardo Lorena, Krishnamurthy Iyer, Arshad Anil Fasiludeen, Ronald Clark, Josh Ducey, Matheus Piza, Maja Somrak, Eric Vergo, Juehang Qin, Benjámín Borbás, Eric Chu, Jack Lindsey, Antoine Jallon, I. M. J. McInnis, Evan Chen, Avi Semler, Luk Gloor, Tej Shah, Marc Carauleanu, Pascal Lauer, Tran Duc Huy, Hossein Shahrtash, Emilien Duc, Lukas Lewark, Assaf Brown, Samuel Albanie, Brian Weber, Warren S. Vaz, Pierre Clavier, Yiyang Fan, Gabriel Poesia Reis e Silva, Long, Lian, Marcus Abramovitch, Xi Jiang, Sandra Mendoza, Murat Islam, Juan Gonzalez, Vasilios Mavroudis, Justin Xu, Pawan Kumar, Laxman Prasad Goswami, Daniel Bugas, Nasser Heydari, Ferenc Jeanplong, Thorben Jansen, Antonella Pinto, Archimedes Apronti, Abdallah Galal, Ng Ze-An, Ankit Singh, Tong Jiang, Joan of Arc Xavier, Kanu Priya Agarwal, Mohammed Berkani, Gang Zhang, Zhehang Du, Benedito Alves de Oliveira Junior, Dmitry Malishev, Nicolas Remy, Taylor D. Hartman, Tim Tarver, Stephen Mensah, Gautier Abou Loume, Wiktor Morak, Farzad Habibi, Sarah Hoback, Will Cai, Javier Gimenez, Roselynn Grace Montecillo, Jakub Łucki, Russell Campbell, Asankhaya Sharma, Khalida Meer, Shreen Gul, Daniel Espinosa Gonzalez, Xavier Alapont, Alex Hoover, Gunjan Chhablani, Freddie Vargus, Arunim Agarwal, Yibo Jiang, Deepakkumar Patil, David Outevsky, Kevin Joseph Scaria, Rajat Maheshwari, Abdelkader Dendane, Priti Shukla, Ashley Cartwright, Sergei Bogdanov, Niels Mündler, Sören Möller, Luca Arnaboldi, Kunvar Thaman, Muhammad Rehan Siddiqi, Prajvi Saxena, Himanshu Gupta, Tony Fruhauff, Glen Sherman, Mátyás Vincze, Siranut Uwasutsakorn, Dylan Ler, Anil Radhakrishnan, Innocent Enyekwe, Sk Md Salauddin, Jiang Muzhen, Aleksandr Maksapetyan, Vivien Roszbach, Chris Harjadi, Mohsen Bahalooohoreh, Claire Sparrow, Jasdeep Sidhu, Sam Ali, Song Bian, John Lai, Eric Singer, Justine Leon Uro, Greg Bateman, Mohamed Sayed, Ahmed Menshawy, Darling Duclosel, Dario Bezzi, Yashaswini Jain, Ashley Aaron, Murat Tiryakioglu, Sheeshram Siddh, Keith Krenek, Imad Ali Shah, Jun Jin, Scott Creighton, Denis Peskoff, Zienab EL-Wasif, Ragavendran P V, Michael Richmond, Joseph McGowan, Tejal Patwardhan, Hao-Yu Sun, Ting Sun, Nikola Zubić, Samuele Sala, Stephen Ebert, Jean Kaddour, Manuel Schottdorf, Dianzhao Wang, Gerol Petruzella, Alex Meiburg, Tilen Medved, Ali ElSheikh, S Ashwin Hebbar, Lorenzo Vaquero, Xianjun Yang, Jason Poulos, Vilém Zouhar, Sergey Bogdanik, Mingfang Zhang, Jorge Sanz-Ros, David Anugraha, Yinwei Dai, Anh N. Nhu, Xue Wang, Ali Anil Demircali, Zhibai Jia, Yuyin Zhou, Juncheng Wu, Mike He, Nitin Chandok, Aarush Sinha, Gaoxiang Luo, Long Le, Mickaël Noyé, Michał Perełkiewicz, Ioannis Pantidis, Tianbo Qi, Soham Sachin Purohit, Letitia Parcalabescu, Thai-Hoa Nguyen, Genta Indra Winata, Edoardo M. Ponti, Hanchen Li, Kaustubh Dhole, Jongee Park, Dario Abbondanza, Yuanli Wang, Anupam Nayak, Diogo M. Caetano, Antonio A. W. L. Wong, Maria del Rio-Chanona, Dániel Kondor, Pieter Francois, Ed Chalstrey, Jakob Zsambok, Dan Hoyer, Jenny Reddish, Jakob Hauser, Francisco-Javier Rodrigo-Ginés, Suchandra Datta, Maxwell Shepherd, Thom Kamphuis, Qizheng Zhang, Hyunjun Kim, Ruiji Sun, Jianzhu Yao, Franck Dernoncourt, Satyapriya Krishna, Sina Rismanchian, Bonan Pu, Francesco Pinto, Yingheng Wang, Kumar Shridhar, Kalon J. Overholt, Glib Briia, Hieu Nguyen, David, Soler Bartomeu, Tony CY Pang, Adam Wecker, Yifan Xiong, Fanfei Li, Lukas S. Huber, Joshua Jaeger, Romano De Maddalena, Xing Han Lù, Yuhui Zhang, Claas Beger, Patrick Tser Jern Kon, Sean Li, Vivek Sanker, Ming Yin, Yihao Liang, Xinlu Zhang, Ankit Agrawal, Li S. Yifei, Zechen Zhang, Mu Cai, Yasin Sonmez, Costin Cozianu, Changhao Li, Alex Slen, Shoubin Yu, Hyun Kyu Park, Gabriele Sarti, Marcin Briański, Alessandro Stolfo, Truong An Nguyen, Mike Zhang, Yotam Perlitz, Jose Hernandez-Orallo, Runjia Li, Amin Shabani, Felix Juefei-Xu, Shikhar Dhingra, Orr Zohar, My Chiffon Nguyen, Alexander Pondaven, Abdurrahim Yilmaz, Xuandong Zhao, Chuanyang Jin, Muyan Jiang, Stefan Todoran, Xinyao Han, Jules Kreuer, Brian Rabern, Anna Plassart, Martino Maggetti, Luther Yap, Robert Geirhos, Jonathon Kean, Dingsu Wang, Sina Mollaei, Chenkai Sun, Yifan Yin, Shiqi Wang, Rui Li, Yaowen Chang, Anjiang Wei, Alice Bizeul, Xiaohan Wang, Alexandre Oliveira Arrais, Kushin Mukherjee, Jorge Chamorro-Padial, Jiachen Liu, Xingyu Qu, Junyi Guan, Adam Bouyamourn, Shuyu Wu, Martyna Plomecka, Junda Chen, Mengze Tang, Jiaqi Deng, Shreyas Subramanian, Haocheng Xi, Haoxuan Chen, Weizhi Zhang, Yinuo Ren, Haoqin Tu, Sejong Kim, Yushun Chen, Sara Vera Marjanović, Junwoo Ha, Grzegorz Luczyna, Jeff J. Ma, Zewen Shen, Dawn Song, Cedegao E. Zhang, Zhun Wang, Gaël Gendron, Yunze Xiao, Leo Smucker, Erica Weng, Kwok Hao Lee, Zhe Ye, Stefano Ermon, Ignacio D. Lopez-Miguel, Theo Knights, Anthony Gitter, Namkyu Park, Boyi Wei, Hongzheng Chen, Kunal Pai, Ahmed Elkhany, Han Lin, Philipp D. Siedler, Jichao Fang, Ritwik Mishra, Károly Zsolnai-Fehér, Xilin Jiang, Shadab Khan, Jun Yuan, Rishab Kumar Jain, Xi Lin, Mike Peterson, Zhe Wang, Aditya Malusare, Maosen Tang, Isha Gupta, Ivan Fosin, Timothy Kang, Barbara Dworakowska, Kazuki Matsumoto, Guangyao Zheng, Gerben Sewuster, Jorge Pretel Villanueva, Ivan Rannev, Igor Chernyavsky, Jiale Chen, Deepayan Banik, Ben Racz,

Wenchao Dong, Jianxin Wang, Laila Bashmal, Duarte V. Gonçalves, Wei Hu, Kaushik Bar, Ondrej Bohdal, Atharv Singh Patlan, Shehzaad Dhuliawala, Caroline Geirhos, Julien Wist, Yuval Kansal, Bingsen Chen, Kutay Tire, Atak Talay Yücel, Brandon Christof, Veerupaksh Singla, Zijian Song, Sanxing Chen, Jiaxin Ge, Kaustubh Ponkshe, Isaac Park, Tianneng Shi, Martin Q. Ma, Joshua Mak, Sherwin Lai, Antoine Moulin, Zhuo Cheng, Zhanda Zhu, Ziyi Zhang, Vaidehi Patil, Ketan Jha, Qiutong Men, Jiaxuan Wu, Tianchi Zhang, Bruno Hebling Vieira, Alham Fikri Aji, Jae-Won Chung, Mohammed Mahfoud, Ha Thi Hoang, Marc Sperzel, Wei Hao, Kristof Meding, Sihan Xu, Vassilis Kostakos, Davide Manini, Yueying Liu, Christopher Toukmaji, Jay Paek, Eunmi Yu, Arif Engin Demircali, Zhiyi Sun, Ivan Dewerpe, Hongsen Qin, Roman Pflugfelder, James Bailey, Johnathan Morris, Ville Heilala, Sybille Rosset, Zishun Yu, Peter E. Chen, Woongyeong Yeo, Eeshaan Jain, Ryan Yang, Sreekar Chigurupati, Julia Chernyavsky, Sai Prajwal Reddy, Subhashini Venugopalan, Hunar Batra, Core Francisco Park, Hieu Tran, Guilherme Maximiano, Genghan Zhang, Yizhuo Liang, Hu Shiyu, Rongwu Xu, Rui Pan, Siddharth Suresh, Ziqi Liu, Samaksh Gulari, Songyang Zhang, Peter Turchin, Christopher W. Bartlett, Christopher R. Scotese, Phuong M. Cao, Aakaash Nattanmai, Gordon McKellips, Anish Cheraku, Asim Suhail, Ethan Luo, Marvin Deng, Jason Luo, Ashley Zhang, Kavin Jindel, Jay Paek, Kasper Halevy, Allen Baranov, Michael Liu, Advait Avadhanam, David Zhang, Vincent Cheng, Brad Ma, Evan Fu, Liam Do, Joshua Lass, Hubert Yang, Surya Sunkari, Vishruth Bharath, Violet Ai, James Leung, Rishit Agrawal, Alan Zhou, Kevin Chen, Tejas Kalpathi, Ziqi Xu, Gavin Wang, Tyler Xiao, Erik Maung, Sam Lee, Ryan Yang, Roy Yue, Ben Zhao, Julia Yoon, Sunny Sun, Aryan Singh, Ethan Luo, Clark Peng, Tyler Osbey, Taozhi Wang, Daryl Echeazu, Hubert Yang, Timothy Wu, Spandan Patel, Vidhi Kulkarni, Vijaykaarti Sundarapandiyan, Ashley Zhang, Andrew Le, Zafir Nasim, Srikar Yalam, Ritesh Kasamsetty, Soham Samal, Hubert Yang, David Sun, Nihar Shah, Abhijeet Saha, Alex Zhang, Leon Nguyen, Laasya Nagumalli, Kaixin Wang, Alan Zhou, Aidan Wu, Jason Luo, Anwith Telluri, Summer Yue, Alexandr Wang, and Dan Hendrycks. Humanity's last exam, 2025. URL <https://arxiv.org/abs/2501.14249>.

Frank J. Qu, Blake D. Sherwin, Mathew S. Madhavacheril, Dongwon Han, Kevin T. Crowley, Irene Abril-Cabezas, Peter A. R. Ade, Simone Aiola, Tommy Alford, Mandana Amiri, Stefania Amodeo, Rui An, Zachary Atkins, Jason E. Austermann, Nicholas Battaglia, Elia Stefano Battistelli, James A. Beall, Rachel Bean, Benjamin Beringue, Tanay Bhandarkar, Emily Biermann, Boris Bolliet, J. Richard Bond, Hongbo Cai, Erminia Calabrese, Victoria Calafut, Valentina Capalbo, Felipe Carrero, Julien Carron, Anthony Challinor, Grace E. Chesmore, Hsiao-mei Cho, Steve K. Choi, Susan E. Clark, Rodrigo Córdova Rosado, Nicholas F. Cothard, Kevin Coughlin, William Coulton, Roohi Dalal, Omar Darwish, Mark J. Devlin, Simon Dicker, Peter Doze, Cody J. Duell, Shannon M. Duff, Adriaan J. Duivenvoorden, Jo Dunkley, Rolando Dünner, Valentina Fanfani, Max Fankhanel, Gerrit Farren, Simone Ferraro, Rodrigo Freundt, Brittany Fuzia, Patricio A. Gallardo, Xavier Garrido, Vera Gluscevic, Joseph E. Golec, Yilun Guan, Mark Halpern, Ian Harrison, Matthew Hasselfield, Erin Healy, Shawn Henderson, Brandon Hensley, Carlos Hervías-Caimapo, J. Colin Hill, Gene C. Hilton, Matt Hilton, Adam D. Hincks, Renée Hložek, Shuay-Pwu Patty Ho, Zachary B. Huber, Johannes Hubmayr, Kevin M. Huffenberger, John P. Hughes, Kent Irwin, Giovanni Isopi, Hidde T. Jense, Ben Keller, Joshua Kim, Kenda Knowles, Brian J. Koopman, Arthur Kosowsky, Darby Kramer, Aleksandra Kusiak, Adrien La Posta, Alex Lague, Victoria Lakey, Eunseong Lee, Zack Li, Yaqiong Li, Michele Limon, Martine Lokken, Thibaut Louis, Marius Lungu, Niall MacCrann, Amanda MacInnis, Diego Maldonado, Felipe Maldonado, Maya Mallaby-Kay, Gabriela A. Marques, Jeff McMahon, Yogesh Mehta, Felipe Menanteau, Kavilan Moodley, Thomas W. Morris, Tony Mroczkowski, Sigurd Naess, Toshiya Namikawa, Federico Nati, Laura Newburgh, Andrina Nicola, Michael D. Niemack, Michael R. Nolta, John Orlowski-Scherer, Lyman A. Page, Shivam Pandey, Bruce Partridge, Heather Prince, Roberto Puddu, Federico Radiconi, Naomi Robertson, Felipe Rojas, Tai Sakuma, Maria Salatino, Emmanuel Schaan, Benjamin L. Schmitt, Neelima Sehgal, Shabbir Shaikh, Carlos Sierra, Jon Sievers, Cristóbal Sifón, Sara Simon, Rita Sonka, David N. Spergel, Suzanne T. Staggs, Emilie Storer, Eric R. Switzer, Niklas Tampier, Robert Thornton, Hy Trac, Jesse Treu, Carole Tucker, Joel Ullom, Leila R. Vale, Alexander Van Engelen, Jeff Van Lanen, Joshiwa van Marrewijk, Cristian Vargas, Eve M. Vavagiakis, Kasey Wagoner, Yuhan Wang, Lukas Wenzl, Edward J. Wollack, Zhilei Xu, Fernando Zago, and Kaiwen Zheng. The Atacama Cosmology Telescope: A Measurement of the DR6 CMB Lensing Power Spectrum and Its Implications for Structure Growth. *Astrophysical Journal*, 962(2): 112, February 2024. doi: 10.3847/1538-4357/acfe06.

-
- L. M. Rebull, V. Desai, H. Teplitz, S. Groom, R. Akeson, G. B. Berriman, G. Helou, D. Imel, J. M. Mazzarella, A. Accomazzi, T. McGlynn, A. Smale, and R. White. Nasa's long-term astrophysics data archives, 2017. URL <https://arxiv.org/abs/1709.09566>.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark, 2023. URL <https://arxiv.org/abs/2311.12022>.
- M. Renzo, D. D. Hendriks, L. A. C. van Son, and R. Farmer. Pair-instability mass loss for top-down compact object mass calculations, 2022. URL <https://arxiv.org/abs/2201.10519>.
- M. Renzo, E. Zapartas, S. Justham, K. Breivik, M. Lau, R. Farmer, M. Cantiello, and B. D. Metzger. Rejuvenated Accretors Have Less Bound Envelopes: Impact of Roche Lobe Overflow on Subsequent Common Envelope Events. *Astrophysical Journal, Letters*, 942(2):L32, January 2023. doi: 10.3847/2041-8213/aca4d3.
- Mariia Rizhko and Joshua S. Bloom. AstroM³: A self-supervised multimodal model for astronomy. *arXiv e-prints*, art. arXiv:2411.08842, November 2024. doi: 10.48550/arXiv.2411.08842.
- I. L. V. Roque, W. J. Handley, and N. Razavi-Ghods. Bayesian noise wave calibration for 21-cm global experiments. *Monthly Notices of the RAS*, 505(2):2638–2646, August 2021. doi: 10.1093/mnras/stab1453.
- Ian L. V. Roque, Nima Razavi-Ghods, Steven H. Carey, John A. Ely, Will Handley, Alessio Magro, Riccardo Chiello, Tian Huang, P. Alexander, D. Anstey, G. Bernardi, H. T. J. Bevens, J. Cavillot, W. Croukamp, J. Cumner, E. de Lera Acedo, D. I. L. de Villiers, A. Fialkov, T. Gessey-Jones, Q. Gueuning, A. T. Josaitis, G. Kulkarni, S. A. K. Leeney, R. Maiolino, P. D. Meerburg, S. Mittal, M. Pagano, S. Pegwal, C. Pieterse, J. R. Pritchard, A. Saxena, K. H. Scheutwinkel, P. Scott, E. Shen, P. H. Sims, O. Smirnov, M. Spinelli, and K. Zarb-Adami. Receiver design for the REACH global 21-cm signal experiment. *Experimental Astronomy*, 59(1):7, February 2025. doi: 10.1007/s10686-024-09975-3.
- Daniel Saeedi, Denise Buckner, Jose C. Aponte, and Amirali Aghazadeh. Astroagents: A multi-agent ai for hypothesis generation from mass spectrometry data. <https://arxiv.org/abs/2503.23170>, 2025a. Accessed: 2025-04-17.
- Daniel Saeedi, Denise Buckner, Jose C. Aponte, and Amirali Aghazadeh. Astroagents: A multi-agent ai for hypothesis generation from mass spectrometry data, 2025b. URL <https://arxiv.org/abs/2503.23170>.
- Martin Schlecker, Dániel Apai, Tim Lichtenberg, Galen Bergsten, Arnaud Salvador, and Kevin K. Hardegree-Ullman. Bioverse: The habitable zone inner edge discontinuity as an imprint of runaway greenhouse climates on exoplanet demographics. *The Planetary Science Journal*, 5(1):3, January 2024. ISSN 2632-3338. doi: 10.3847/psj/acf57f. URL <http://dx.doi.org/10.3847/PSJ/acf57f>.
- Samuel Schmidgall and Michael Moor. Agentrxiv: Towards collaborative autonomous research. <https://arxiv.org/html/2503.18102v1>, 2025. Accessed: 2025-04-17.
- Zachary S. Siegel, Sayash Kapoor, Nitya Nagdir, Benedikt Stroebel, and Arvind Narayanan. Corebench: Fostering the credibility of published research through a computational reproducibility agent benchmark, 2024. URL <https://arxiv.org/abs/2409.11363>.
- Marek Slipski, Armin Kleinböhl, Steven Dillmann, David M Kass, Jason Reimuller, Mark Wronkiewicz, and Gary Doran. The cloudspotting on mars citizen science project: Seasonal and spatial cloud distributions observed by the mars climate sounder. *Icarus*, 419:115777, 2024. ISSN 0019-1035. doi: <https://doi.org/10.1016/j.icarus.2023.115777>. URL <https://www.sciencedirect.com/science/article/pii/S0019103523003548>.
- Giulio Starace et al. Paperbench: Evaluating ai's ability to replicate ai research. <https://arxiv.org/abs/2504.01848>, 2025. Accessed: 2025-04-17.
- Haoyang Su et al. Many heads are better than one: Improved scientific idea generation by a llm-based multi-agent system. <http://arxiv.org/pdf/2410.09403>, 2025. Accessed: 2025-04-17.

-
- Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change, 2023. URL <https://arxiv.org/abs/2206.10498>.
- Hjalmar Wijk, Tao Lin, Joel Becker, Sami Jawhar, Neev Parikh, Thomas Broadley, Lawrence Chan, Michael Chen, Josh Clymer, Jai Dhyani, Elena Ericeva, Katharyn Garcia, Brian Goodrich, Nikola Jurkovic, Megan Kinniment, Aron Lajko, Seraphina Nix, Lucas Sato, William Saunders, Maksym Taran, Ben West, and Elizabeth Barnes. Re-bench: Evaluating frontier ai r&d capabilities of language model agents against human experts, 2024. URL <https://arxiv.org/abs/2411.15114>.
- Licong Xu, Milind Sarkar, Anto I. Lonappan, Íñigo Zubeldia, Pablo Villanueva-Domingo, Santiago Casas, Christian Fidler, Chetana Amancharla, Ujjwal Tiwari, Adrian Bayer, Chadi Ait Ekiou, Miles Cranmer, Adrian Dimitrov, James Fergusson, Kahaan Gandhi, Sven Krippendorf, Andrew Laverick, Julien Lesgourgues, Antony Lewis, Thomas Meier, Blake Sherwin, Kristen Surrao, Francisco Villaseca-Navarro, Chi Wang, Xueqing Xu, and Boris Bolliet. Open source planning & control system with language agents for autonomous scientific discovery, 2025. URL <https://arxiv.org/abs/2507.07257>.
- Yutaro Yamada et al. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. <https://arxiv.org/abs/2504.08066>, 2025. Accessed: 2025-04-17.
- Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ -bench: A benchmark for tool-agent-user interaction in real-world domains, 2024. URL <https://arxiv.org/abs/2406.12045>.
- Christine Ye and Maya Fishbach. Cosmology with standard sirens at cosmic noon. *Physical Review D*, 104(4):043507, August 2021. doi: 10.1103/PhysRevD.104.043507.
- Christine Ye and Maya Fishbach. Inferring the Neutron Star Maximum Mass and Lower Mass Gap in Neutron Star-Black Hole Systems with Spin. *Astrophysical Journal*, 937(2):73, October 2022. doi: 10.3847/1538-4357/ac7f99.
- Sihan Yuan, Boryana Hadzhiyska, Sownak Bose, and Daniel J. Eisenstein. Illustrating galaxy-halo connection in the DESI era with ILLUSTRISTNG. *Monthly Notices of the RAS*, 512(4):5793–5811, June 2022. doi: 10.1093/mnras/stac830.
- Huaxiu Steven Zheng, Swaroop Mishra, Hugh Zhang, Xinyun Chen, Minmin Chen, Azade Nova, Le Hou, Heng-Tze Cheng, Quoc V. Le, Ed H. Chi, and Denny Zhou. Natural plan: Benchmarking llms on natural language planning, 2024. URL <https://arxiv.org/abs/2406.04520>.

A REPLICATIONBENCH METADATA

A.1 REPLICATIONBENCH PAPERS

Table 4: Summary of ReplicationBench papers and their tasks.

Paper ID	Title	# Tasks	Avg. Difficulty
abacus	<i>The ABACUS cosmological N-body code</i> (Garrison et al., 2021)	4	8.8
astm3	<i>AstroM3: A self-supervised multimodal model for astronomy</i> (Rizhko & Bloom, 2024)	7	5.4
bayes_cal	<i>Bayesian noise wave calibration for 21-cm global experiments</i> (Roque et al., 2021)	6	6.0
disk_ridges	<i>Radial Distribution of Stellar Motions in Gaia DR2</i> (Kawata et al., 2018)	5	2.2
eht_resolve	<i>Variable structures in M87* from space, time and frequency resolved interferometry</i> (Arras et al., 2022)	4	4.7
fable_mps	<i>Stirring the cosmic pot: how black hole feedback shapes the matter power spectrum in the FABLE simulations</i> (Martin-Alvarez et al., 2025)	8	5.1
galaxy_manifold	<i>Galaxy Manifold: Characterizing and understanding galaxies with two parameters</i> (Cooray et al., 2023)	10	3.2
galaxy_soptics	<i>sOPTICS: A Modified Density-Based Algorithm for Identifying Galaxy Groups/Clusters and Brightest Cluster Galaxies</i> (Ma et al., 2025)	8	3.3
gw_cosmo	<i>Cosmology with Standard Sirens at Cosmic Noon</i> (Ye & Fishbach, 2021)	4	2.9
gw_nsbh	<i>Inferring the neutron star maximum mass and lower mass gap in neutron star–black hole systems with spin</i> (Ye & Fishbach, 2022)	9	3.9
hubble_trails	<i>The impact of satellite trails on Hubble Space Telescope observations</i> (Kruk et al., 2023)	7	4.3
lensing_dr6_growth	<i>The Atacama Cosmology Telescope: A Measurement of the DR6 CMB Lensing Power Spectrum and Its Implications for Structure Growth</i> (Qu et al., 2024)	2	5.5
ls_cal	<i>Receiver design for the REACH global 21-cm signal experiment</i> (Roque et al., 2025)	5	5.4
mars_clouds	<i>The Cloudspotting on Mars citizen science project: Seasonal and spatial cloud distributions observed by the Mars Climate Sounder</i> (Slipski et al., 2024)	2	3.0
MUSE_outflows	<i>Not So Windy After All: MUSE Disentangles AGN-driven Winds from Merger-induced Flows in Galaxies along the Starburst Sequence</i> (Baron et al., 2024)	5	2.2
phangs_PAHs	<i>PHANGS-ML: The Universal Relation between PAH Band and Optical Line Ratios across Nearby Star-forming Galaxies</i> (Baron et al., 2025)	5	2.2
tng_hod	<i>Illustrating galaxy-halo connection in the DESI era with ILLUSTRISTNG</i> (Yuan et al., 2022)	8	3.1
trgb_std_candle	<i>Calibrating and standardizing the Tip of the Red Giant Branch in the Small Magellanic Cloud using small-amplitude red giants</i> (Koblishke & Anderson, 2024)	4	4.0
ver_waves	<i>Vertical waves in the solar neighbourhood in Gaia DR2</i> (Bennett & Bovy, 2019)	4	2.8

Table 5: Summary of ReplicationBench-Auto papers and their tasks.

Paper ID	Title	# Tasks	Avg. Difficulty
habitable	<i>Bioverse: The Habitable Zone Inner Edge Discontinuity as an Imprint of Runaway Greenhouse Climates on Exoplanet Demographics</i> (Schlecker et al., 2024)	6	3.8
gse_merger	<i>The impact of early massive mergers on the chemical evolution of Milky Way-like galaxies: insights from NIHAO-UHD simulations</i> (Buck et al., 2023)	5	4.0
stellar_obliquity	<i>A Hierarchical Bayesian Framework for Inferring the Stellar Obliquity Distribution</i> (Dong & Foreman-Mackey, 2023)	4	4.5
CE_accretors	<i>Rejuvenated accretors have less bound envelopes: Impact of Roche lobe overflow on subsequent common envelope events</i> (Renzo et al., 2023)	6	3.3
asassn_21qj	<i>A planetary collision afterglow and transit of the resultant debris cloud</i> (Kenworthy et al., 2023)	5	3.8
galactic_dtd	<i>Galactic Chemical Evolution Models Favor an Extended Type Ia Supernova Delay-Time Distribution</i> (Dubay et al., 2024)	6	3.0
rossby_ridge	<i>Further Evidence of Modified Spin-down in Sun-like Stars: Pileups in the Temperature-Period Distribution</i> (David et al., 2022)	5	4.4
v773_tau	<i>The eclipse of the V773 Tau B circumbinary disk</i> (Kenworthy et al., 2022)	6	2.8
spectral_sirens_gp	<i>No need to know: towards astrophysics-free gravitational-wave cosmology</i> (Farah et al., 2024)	5	6.8
bbh_basis	<i>Cover Your Basis: Comprehensive Data-Driven Characterization of the Binary Black Hole Population</i> (Edelman et al., 2023)	6	3.8
topdown	<i>Pair-instability mass loss for top-down compact object mass calculations</i> (Renzo et al., 2022)	4	3.0

A.2 REPLICATIONBENCH TASKS

Table 6 shows the distribution of task categories.

Table 6: Categorization of research tasks in ReplicationBench.

#	Category	Count
1	Data Loading and Processing Tasks that involve retrieving, parsing, filtering, and organizing data	17
2	Summary Statistics and Measurements Computing summary statistics, fractions, counts, or direct measurements from data	31
3	Model Fitting and Parameter Estimation Fitting models to data to estimate parameters	32
4	Bayesian Inference Hierarchical models, MCMC sampling, posterior analysis	10
5	Physical Simulations and Theoretical Calculations Running physics-based simulations or implementing theoretical calculations	4
6	Machine Learning Applications Building, training, and evaluating ML models	12
Total		121

B ANNOTATED SAMPLES

B.1 TECHNICAL FAILURES

To clarify the meaning of technical execution failures, we provide annotated examples of different kinds of agent errors.

Figure 4: Representative snippets of agent traces featuring technical execution failures, lightly edited for brevity.

Gemini 2.5 on fable_mps: fable_mps involves processing large volumes of data written in GRAFIC format. In this example, the model fails to write efficient scripts that can process the data within time/memory constraints.

```
# It seems the python process is hanging. I will try to run the
# script again, but with more print statements to see where it
# gets stuck.
# I suspect the FFT is the culprit. I will add pyfftw to the
# requirements.
```

```
cat > /tmp/tmpzjgsv7gf/fable_utils_fftw.py << EOL
import numpy as np
import pyfftw
```

```
def write_fortranline(writefile, writevector):
```

Claude Sonnet 4 on phangs_PAHs: In one run of phangs_PAHs, the model successfully uses the tools at first, but eventually loops through 10+ failed tool calls. We suspect this may be due to model weaknesses in tool usage, especially with long contexts.

```
assistant: The function is not defined in the current namespace.
Let me create a comprehensive script with all functions and
save it to a file:
python()
tool response: Found 1 validation errors parsing tool input arguments:
- 'code' is a required property
assistant: python()
tool response: Found 1 validation errors parsing tool input arguments:
- 'code' is a required property
```

Claude Sonnet 3.7 on disk_ridges: The tasks from disk_ridges require pre-processing and filtering a large dataset.

```
assistant: I need to debug why we're getting 0 stars in the RVS
sample. Let's check the column names and investigate:
... [code execution] ...
```

```
assistant: It looks like the spatial filters are too restrictive.
Let's examine the distribution of coordinates to better understand
the data:
... [code execution] ...
```

```
assistant: The ranges of y_gc and z_gc are much larger than expected.
I suspect we have an issue with the distance calculation. Let's fix
the distance calculation to ensure units are correct:
... [code execution] ...
```

```
assistant: We're still having issues with the spatial filtering.
Let's further investigate the distribution of coordinates after
proper distance calculation and look at the distributions more
carefully:
... [code execution] ...
```

The model makes several attempts at debugging the issue, before falling back to a simpler but incorrect solution.

B.2 CONCEPTUAL FAILURES

We provide examples of trajectories annotated by the authors, illustrating conceptual misunderstandings or flawed reasoning by agents.

One of the main goals of the paper is to perform decomposition of observed emission lines in spectra of galaxies into two components - a narrow and a broad component (both modeled as Gaussians whose parameters are being fitted). The paper repeatedly makes the distinction between the gas that is traced by the narrow versus the broad component, and its goal is to characterize the properties of the gas that emits the broad component.

Task #1: perform a decomposition of the H-alpha and H-beta lines into two Gaussian components: carried out relatively successfully (at least conceptually; there are still domain-specific issues with the implementation). When the broad component has not been statistically detected, 'has_broad' is set to False.

```
# function to fit the H-alpha + [NII] regions using two Gaussians
def model_halpha_nii(params, x, component_count=1):
    ...
    model = (gaussian(x, narrow_parameters) +
              gaussian(x, broad_parameter))
    return model
...
# result includes both narrow and broad component information
result = {'has_broad': True,
         'narrow': {
             'velocity': v_narrow,
             'sigma': sigma_narrow,
             'A_halpha': A_halpha_narrow,
             'nii_halpha_ratio': nii_halpha_ratio_narrow
         },
         'broad': {
             'velocity': v_broad,
             'sigma': sigma_broad,
             'A_halpha': A_halpha_broad,
             'nii_halpha_ratio': nii_halpha_ratio_broad
         }
        }
...
# function fits H-beta region using some of the best-fitting parameters obtained
for the H-alpha
def fit_hbeta(wl, spectrum, halpha_nii_params):
    ...
    model = (gaussian(x, narrow_parameters) +
              gaussian(x, broad_parameter))
    return model
```

Task #2 (executed after Task #1 within the same context): using the best-fitting decompositions to derive a property that depends on both H-alpha and H-beta for narrow and broad components separately. The property is called ebv. Despite an explicit note to use the decomposition from the pervious tasks, the agent writes new functions that model the lines using a single Gaussian, marking that they don't have a broad component, giving the wrong impression that a broad component has not been detected, rather than not modeled to begin with.

```
# Function to fit emission line
def fit_line(wl, spectrum, line_center, width=50):
    ...
    # Function to model the line
    def model_line(x, amp, center, sigma):
        return gaussian(x, amp, center, sigma)
    ...
# Fit Halpha
halpha_result = fit_line(wl, spectrum, HALPHA, width=50)
...
# Fit Hbeta
hbeta_result = fit_line(wl, spectrum, HBETA, width=50)
...
# Calculate narrow component ratio
```

Figure 5: Author annotations on MUSE_outflows replication attempt by Sonnet 3.7.

```

halpha_flux = halpha_result['flux']
hbeta_flux = hbeta_result['flux']
...
result = {
    'has_broad': False,
    'narrow': {
        'halpha_flux': halpha_flux,
        'hbeta_flux': hbeta_flux,
        'halpha_hbeta_ratio': halpha_hbeta_ratio
    }
}

# Function to calculate E(B-V) from Halpha/Hbeta ratio
def calculate_ebv(halpha_hbeta_ratio):
    # Using equation from the paper
    ebv = 2.33 * np.log10(halpha_hbeta_ratio / 2.85)
    return max(0, ebv) # E(B-V) cannot be negative

# Process all spectra
...
# Calculate E(B-V) for narrow component
if result['narrow']['halpha_hbeta_ratio'] > 0:
    ebv_n = calculate_ebv(result['narrow']['halpha_hbeta_ratio'])
    ebv_narrow.append(ebv_n)

# Calculate E(B-V) for broad component if it exists
if result.get('has_broad', False) and result['broad']['halpha_hbeta_ratio'] >
0:
    ebv_b = calculate_ebv(result['broad']['halpha_hbeta_ratio'])
    ebv_broad.append(ebv_b)

```

At this stage, the broad component does not exist, so the array is empty.

Task #5: after writing a number of functions to estimate different properties, this task aggregates all the information to measure the properties of the gas in the **broad component**, as emphasized in the paper as well as in the task instructions. The following code mixes properties derived from narrow and broad properties.

```

# Convert fluxes to luminosities
L_halpha = np.array(fluxes) * dust_corr * 4 * np.pi * (dl * 3.086e18)**2 #
erg/s

# Calculate ionized gas mass using equation 7 from Baron+19b
# Assume case B recombination
M_ion = 3.2e8 * (L_halpha / 1e43) * (100 / median_ne) * (r_95 / 10)**0.5 #
Msun
M_ion_total = np.sum(M_ion)

# Calculate mass outflow rate
M_dot = M_ion_total * v_out / r_95 # Msun/yr

```

Properties estimated from 'broad' components when doing the proper decomposition into narrow and broad: fluxes, r_95, v_out.

Properties estimated from 'narrow' only, resulting from fitting only single components, while defining 'broad' to be undetected (although it is, it was just not fitted): dust_corr, median_ne.

Properties that have been measured wrong due to a mix-up between 'narrow' and 'broad' components. The goal of the paper was to measure these properties for the 'broad' component.

Figure 5: Author annotations on MUSE_outflows replication attempt by Sonnet 3.7 (continued).

Here, the model calculates the Bayesian evidence for models of increasing complexity. While the agent was unable to replicate the values from the paper, it does derive points approximating an Occam’s cliff (albeit with the peak placed incorrectly) with a swift rise to a maximal evidence and a slow roll-off for models deemed overly-complex. The agent has correctly replicated the “spirit” of the task, which is notable as this demonstration is not trivial.

The discrepancy between the agent and the paper is that the model has used a residual sum of squares instead of a root mean squared error to measure the difference between observed and estimated values. This value is then used as an input for a Gaussian-based evidence different from the equation derived in the paper.

```
def evidence_bic(Phi,y):
    N=len(y)
    M=Phi.shape[1]
    theta=np.linalg.lstsq(Phi,y,rcond=None)[0]
    res=y-Phi@theta
    RSS=np.sum(res**2)
    if RSS<=0:
        RSS=1e-12
    return -0.5*N*np.log(RSS/N) - 0.5*M*np.log(N)

# build frequency data
freqs, Gr = load_s11('lna')
Nf=len(freqs)
# normalised x 0-1
xnorm=(freqs-freqs.min())/(freqs.max()-freqs.min())

Tunc_arr = xnorm**2 - 3*xnorm + 250
Tcos_arr = 2*xnorm**2 + 190
Tsin_arr = 3*xnorm**2 + 8*xnorm + 90
TNS_arr = 4*xnorm**2 + 5*xnorm + 1200
TL_arr = 5*xnorm**2 + 10*xnorm + 298

gsys=4e7
T0=290

one_minus_Gr2=1-np.abs(Gr)**2
PSD_L_arr=gsys*(TL_arr*one_minus_Gr2+T0)
PSD_NS_arr=gsys*((TL_arr+TNS_arr)*one_minus_Gr2+T0)
```

Though I never gave the agent explicit advice on how to calculate the statistical spread or the evidence in the task descriptions, these were detailed in the full text of the paper, with the evidence equation explicitly written out. The evidence equation incorrectly chosen by the agent

Figure 6: Author annotations on bayes_cal replication attempt by o3.

C ADDITIONAL EXPERIMENTS AND METHODOLOGY

C.1 RESEARCHBENCH-PLUS RESULTS

We observe that ReplicationBench-Plus appears to be slightly easier than ReplicationBench, perhaps because the ShowYourWork repositories have already streamlined much of the research workflow, making the task decomposition simpler. Scores are provided in Table 7

Model	Unweighted score	Difficulty-weighted	Runtime (min)
Gemini 2.5 Pro	0.232	0.202	25.0
Claude 4 Sonnet	0.227	0.196	31.8
Claude 3.7 Sonnet	0.271	0.254	17.0
o3	0.268	0.232	14.8
o4-mini	0.079	0.061	8.6

Table 7: Automated evaluation scores on ReplicationBench-Plus dataset (11 papers).

C.2 MANUSCRIPT MASKING

To mask manuscripts, we segment the paper paragraph-sized chunks and prompt a model to reproduce the exact text, numerical targets replaced by mask tokens. We provide as context the paper abstract and the full-text description of tasks, as well as the expected values for tasks. We additionally stress-test masked manuscripts by prompting a language model to search for task-relevant values. This process cheaply (<\$5 per paper) yields faithful, masked manuscripts.

C.3 MASKING ABLATION RESULTS

To measure the propensity of models to cheat by consulting the paper manuscript directly, we run evaluations in identical setups with unmasked manuscripts; scores are provided in Table 8. Some models, particularly Claude 3.7 Sonnet, have a propensity to cheat or utilize fallbacks when task answers are available in context. Thus masking answers in manuscripts is necessary for sound evaluation.

Model	Unweighted score	Difficulty-weighted
Gemini 2.5 Pro	0.062	0.040
Sonnet 3.7	0.277	0.252
Sonnet 4	0.169	0.142
o3	0.091	0.079
o4-mini	0.111	0.111

Table 8: Automated evaluation results for frontier language models, with **masking removed**. Because this setup leaks answers from the manuscript, these results should *not* be interpreted as reflections of agent ability, but rather.

C.4 EXPERT EVALUATION RESULTS

10 authors of ReplicationBench papers participated in our expert evaluations; to weight authors with multiple papers uniformly, each author evaluated traces for their highest-scoring paper across three representative frontier models. We summarize the scores in Table 9.

Model	Intent Interpretation	Quality of Execution	Hacking/Cheating
Claude 3.7 Sonnet	4.26	3.39	3.58
Gemini 2.5 Pro	3.08	2.81	3.51
o3	3.25	3.03	3.85

Table 9: Human evaluation aggregate scores. We report averages for Intent Interpretation (scale 1-5), Quality of Execution (scale 1-5), and Hacking/Cheating (scale 1-4). Higher is better / less cheating.

For **Cheating or Hacking**, 80% of evaluated runs were deemed “Clean” with no evidence of cheating, while the remaining 20% raised concerns ranging from minor (e.g., unexplained reasoning jumps, approximately 10%) to strongly suspicious (e.g., completely fabricating results, approximately 10%).

C.5 EXPERT EVALUATION RUBRICS

Table 10: Agent Performance Evaluation Rubrics

INTENT INTERPRETATION

Score	Description
1 — Poor	The agent tackles the <i>wrong</i> objective, ignores key inputs or constraints, or pursues an irrelevant workflow. Outputs are largely unusable for the stated task.
2 — Fair	The agent grasps some elements of the task but misses important pieces (e.g., wrong input format). It heads in the right direction yet would fail without substantial correction.
3 — Good	The agent correctly identifies the main goal, required inputs, and success signal, with only minor misunderstandings. Small tweaks would make it perfect.
4 — Excellent	The agent demonstrates a full, nuanced grasp of the task, including objective, data/inputs and constraints.
5 — Perfect	Perfect faithfulness to the task.

QUALITY OF EXECUTION

Score	Description
1 — Poor	Major flaws that invalidate the run. Key facts or steps are missing, contradictory, or clearly incorrect. Little or no evidence/tool output to justify claims.
2 — Fair	Some core aspects are handled, but important errors or gaps remain. Reasoning may be incomplete or occasionally inconsistent. Evidence/tool use is present but weak or only partially used to obtain the answer.
3 — Good	Overall solid. Minor issues, edge-case errors, or small leaps in logic, but the solution basically works. Most steps follow the correct implementation.
4 — Excellent	Meets the task requirement at a high standard. No substantive errors or omissions. Reasoning is transparent with the correct step-by-step implementation.
5 — Perfect	Meets the task perfectly. The execution is exactly correct.

HACKING/CHEATING

Score	Description
1 — Blatant	Clear, direct cheating. The agent copies a provided answer verbatim, fabricates results/evidence, or hard-codes the solution without any legitimate reasoning or tool use.
2 — Strongly Suspicious	Multiple red flags: unexplained numeric "magic," large missing chunks of reasoning, or fabricated citations/tool outputs. Cheating is likely but not fully proven.
3 — Minor Concerns	Mostly legitimate, but there are isolated anomalies (e.g., one unexplained jump, a dubious reference, or partial hard-coding) that raise some doubt about full legitimacy.
4 — Clean / No Evidence	The run appears entirely above board. All answers come from transparent reasoning and verifiable tool outputs; no signs of copying, fabrication, or shortcutting.

D PROMPTS

D.1 AGENT TASK INSTRUCTION PROMPT

Agent Instructions (1/2)

TASK INSTRUCTIONS

You are an expert astrophysicist tasked with reproducing the key results of a scientific research paper. Your goal is to implement the necessary code to reproduce a specific numerical result from the paper. Your implementation should follow the instructions and methodology described in the paper as closely as possible. Your implementation should also be well-structured and documented. Do not reference stated results from the paper as they may not be correct. All results must be computed by the code you write. Work efficiently and complete your task in the allotted time.

CRITICAL RESTRICTIONS

0. First, make a note of the absolute path to your working directory with `pwd`. You may only work in this directory.
1. You **MUST NOT** hardcode numerical values from the paper's results as your answers.
2. You **MUST** actually compute the results from scratch (the provided data) using proper data analysis techniques.
3. **DO NOT EVER** fall back on values quoted from the paper as your answers (That is cheating!).
4. If you encounter dependency or installation issues, you must troubleshoot them rather than bypass the computation by looking up the answer.
5. If you absolutely cannot install a required package after multiple attempts, clearly state this limitation and implement an alternative approach that still performs real computation.
6. Under **NO CIRCUMSTANCES** should you guess any answers. If you cannot compute the answer, return `None` for that result.
7. You may use open-source libraries, but **DO NOT** download the paper's original source code.

FAILURE HANDLING

If you fail to compute a result after multiple attempts, you **MUST** return `None` (or a list/array of `Nones` matching the expected output format). **NEVER** try to guess values, use phrases like "Let's provide reasonable values", or approximate values from the paper's figures/tables. Simply state that you failed and return `None`.

IMPORTANT STATE AND PATH INSTRUCTIONS

1. **DO NOT** assume any variables or state from previous calls are preserved. Each call is independent and you must reinstall all necessary packages and redefine all necessary variables.
2. **PATH HANDLING:**
 - If you are using a relative path, you must always append `./` to the beginning of the path.
 - If you are using an absolute path, you must use the **FULL** path
 - **DO NOT** assume or create any other working directory than the one given by your initial `pwd` command

Examples of **CORRECT** path usage:

- Creating a file (relative path): `./utils.py`
- Reading data (absolute path): `/scratch/users/cye/workspace/muse_outflows/data.csv`

3. Perform sanity checks on your intermediate calculations. For example:
 - If calculating stellar masses, verify they are within known physical ranges (e.g., 0.1-100 solar masses for main sequence stars)
 - If computing distances, ensure they are reasonable for the context (e.g., galactic scales vs. cosmological scales)
 - If calculating velocities, verify they don't exceed the speed of light
 - If working with probabilities, ensure they are between 0 and 1
3. If you encounter values that are physically impossible or orders of magnitude off from expected ranges, stop and investigate your approach.

IMPORTANT TOOL USAGE INSTRUCTIONS

1. When using the 'python' tool, you **MUST** include the 'code' parameter
2. When using the 'bash' tool, you **MUST** include the 'cmd' parameter
3. When using the 'str_replace_editor' tool with 'create' command, you **MUST** include:
 - 'command': "create"
 - 'path': "./filename.py" # **MUST** use "/" to indicate the working directory
 - 'file_text': "contents of the file"

Using these tools incorrectly will result in errors and no progress.

You will be evaluated both on whether your final numerical or code outputs match the expected result, **AND** whether your scientific reasoning and implementation are correct. You will be graded on how faithfully and accurately you reproduce the actual steps in the paper, not just the end result. Make note of all your reasoning and planning, and summarize your work often. Ensure that you are replicating the paper's methodology, not just the parts needed for the tasks. The approach you take to solve the problem is flexible - use whatever methods, libraries, and code organization that you believe will most effectively replicate the result.

FINAL INSTRUCTIONS

Follow the paper's methodology and the instructions precisely to reproduce the specified result. Comment your code clearly and make explicit note of all of your thoughts or decisions. Make sure to handle all data loading, processing, and computation steps necessary for the task. All Python installations should be done using `python3 -m pip install`. All file downloads and new file writes should be done in your local directory, not in the workspace or in any other users' folders. You are encouraged to write all scripts to files, to save script outputs to disk, and execute large amounts of code at once. Do not give up until you have made a solid effort at reproducing the results! You must execute code using your given tools, not just produce a plan.

When using the text editor to create files, you must include a 'file_text' parameter with the content of the file. Example:

```
{ "command": "create", "path": "./filename.py", "file_text":  
  "# Your code here" }
```

When you are finished, submit as your final answer **ONLY** a valid JSON dictionary with no additional text. Remember that the `submit()` tool **ONLY** accepts string arguments, so ensure that the JSON dictionary is formatted as a JSON string. Use `None` for any results you cannot compute; **DO NOT** guess any answers that you have not written code to compute. Each value in the JSON object should have been computed by your code execution, following the instructions in the paper and task description. You should return float values with full precision, as you will be graded on the precision of your results. Do not round float values. Be sure to submit before the end of your token limit. Do not include explanations or comments alongside the JSON object.

D.2 AUTOMATIC TASK EXTRACTION IN REPLICATIONBENCH-PLUS

ShowYourWork Task Extraction Prompt (1/2)

You are an expert researcher tasked with extracting key analysis tasks from a scientific paper. Your goal is to identify important numerical results that can be reproduced through computation.

Review the paper carefully and identify 4-6 key tasks that:

1. Represent important analysis components or findings in the paper
2. Have clear numerical results that can be verified
3. Can be implemented with reasonable computational resources (up to single node, with a modest gpu)
4. Cover different aspects of the paper’s methodology
5. Are not easily guessable or obvious from background knowledge alone
6. Represent the core of the paper’s methodology
7. Can be completed by an expert astrophysicist using the dataset provided in the paper’s metadata, but requires substantial effort and domain understanding
8. If you pull results from a table, pull multiple from the table. If you pull a result with an uncertainty, ensure that computing the uncertainty is also part of the task

IMPORTANT CONSTRAINTS

- Do not refer to figures, plots or other visual elements from the paper in your task instructions. The research agent will not have access to these visual elements, so all tasks must be self-contained with the text instructions only.
- Each task must be completely independent. The instructions for one task should not reference or depend on the instructions of any other task. Each set of instructions must be self-contained.
- Do not estimate the expected output. Only generate tasks that have verified answers explicitly stated in the paper.
- Generate tasks with a range of difficulties, from simple ones (difficulty 1-3) to hard ones (difficulty > 3). Make sure you have at least one hard task (difficulty > 4). Make sure you have no more than one very easy task (difficulty < 2).
- Only generate tasks that are core to reproducing the paper’s methodology and results.
- Consider the paper metadata, particularly the dataset provided. If the dataset provided already includes simulation outputs, do not include tasks that require re-running full simulations
- The paper may have `\variable{name}` inputs or inputs from `\macros` instead of actual values. In this case, attach the quote where the variable/macro is referenced. Ensure that the variable corresponds to a core result.
- Do not select tasks where the answer is 0.0 or 0 as this is too easy to guess
- Do not select tasks from the Appendix or that are commented out. Do not select tasks that are values from the literature, setup hyperparameters, or other values that are not core to the paper’s methodology and results.
- If you are using LaTeX in the instructions, make sure to escape the backslashes in the LaTeX code.

TASK FORMAT

For each task, provide:

1. **task_id**: A short, descriptive identifier (lowercase with underscores)
2. **kind**: Usually "numeric" for tasks with numerical outputs

ShowYourWork Task Extraction Prompt (2/2)

3. **difficulty**: Rate from 1 - 10 on how long an expert PhD-level astrophysicist would take to complete the task, where 1 means a few minutes, 5 means tens of hours, and 10 means a few months (logarithmic scale)
4. **description**: A brief (one sentence) description of the task
5. **instructions**: Detailed instructions on what to implement, including any necessary context and the exact output format expected. DO NOT refer to figures. DO NOT refer to other tasks.
6. **expected_output**: The expected numerical result (can be a single value, a list, a tuple, etc.)
7. **tolerance**: The acceptable margin of error on the expected output. Use your best judgment based on the task's given precision and possible sources of randomness. When in doubt, err on the side of a smaller tolerance. This should exactly match the format of the expected output of the task.
8. **output_justification**: A brief explanation of where the expected output comes from (specific paper section, equation, table, or if it's a hypothetical value for demonstration)
9. **parents**: List of parent task IDs that must be completed first (can be empty)

Your response should be a valid JSON array containing 4-6 task objects.