
ResearchBench: Evaluating AI Agents on End-To-End Astrophysics Research Paper Replication

Christine Ye¹ Sihan Yuan¹ Suchetha Cooray¹ Steven Dillmann¹ Ian L. V. Roque¹ Dalya Baron¹
Philipp Frank¹ Sergio Martin-Alvarez¹ Nolan Koblishcke² Frank J. Qu¹ Diyi Yang¹ Risa Wechsler¹
Ioana Ciucă¹

Abstract

Frontier AI agents are increasingly being developed as assistants in scientific research. However, their scientific capabilities and instruction-following reliability have not been rigorously evaluated in the context of complex, real-world research tasks. To address this, we introduce ResearchBench, a benchmark for evaluating AI agents on end-to-end reproducibility challenges drawn from peer-reviewed science papers. Focusing initially on astrophysics — where tasks are often fully computational — we select 22 papers from the literature on the basis of reproducibility. We split each paper into a series of tasks which require agents to generate replications of the paper’s core contributions. Each task is co-developed with the original authors and targets a key scientific result, enabling objective evaluation of both **faithfulness** and **correctness**. We benchmark several frontier language models on these tasks and conduct human evaluations of agent traces with domain experts. ResearchBench establishes the first benchmark for paper-scale expert-validated scientific tasks, beginning with astrophysics, and sets a new standard for measuring AI agents’ reliability in scientific research.

1. Introduction

Current AI agents demonstrate expert-level scientific reasoning and autonomous interactions (Kwa et al., 2025) and are being applied across the research lifecycle, from hypothesis generation to coding (Su et al., 2025; Boiko et al., 2023). However, existing experiments on scientific agents demonstrate failures in robustness and domain understanding (Beel et al., 2025).

^{*}Equal contribution ¹Stanford University, Stanford, CA, USA
²University of Toronto, Toronto, ON, Canada. Correspondence to: Christine Ye <christine.ye@stanford.edu>.

To produce high-quality research, agents need both domain knowledge and autonomous capabilities, yet we lack methods to measure their usefulness in such tasks. Existing benchmarks either test static domain knowledge (Hendrycks et al., 2021) or assess agency without requiring domain expertise (Jimenez et al., 2024). Recent work includes TPBench (Chung et al., 2025), which evaluates theoretical physics reasoning from undergraduate to research level, and CURIE (Cui et al., 2025), which assesses LLM understanding and reasoning across six scientific disciplines. No current benchmark simultaneously measures long-horizon agency and expert-level conceptual understanding.

To measure agents’ ability to conduct scientific research, we first seek to measure whether they can execute faithful replications of existing expert-level research, a prerequisite for producing original work. We introduce **ResearchBench**, a benchmark to evaluate AI agents on end-to-end replications of research papers. By enforcing that agents exactly reproduce the original paper, ResearchBench measures agents’ instruction-following **faithfulness** and execution **correctness**. We initially focus on **astrophysics** due to its strong standards for reproducibility and methodology replicable within sandboxed computational environments (Allen et al., 2012); however, the framework can extend to any scientific field.

ResearchBench is designed with the principles of **Sandboxing**, **Realism**, and **Scalability**. To **sandbox** evaluations, replications are self-contained and can be implemented with standard agent execution tools. To construct **realistic** tasks, papers require expert-level reasoning and data analysis on par with the original research. Regarding **scalability**, ResearchBench measures whether agents can reproduce a verifiable objective final result, which significantly reduces expert effort and enables automatic grading.

ResearchBench is the first benchmark to implement an end-to-end evaluation approach for AI agents conducting scientific research, and establishes a rigorous testbed for agents as reliable scientific collaborators. We release the ResearchBench dataset and codebase for use by the community.

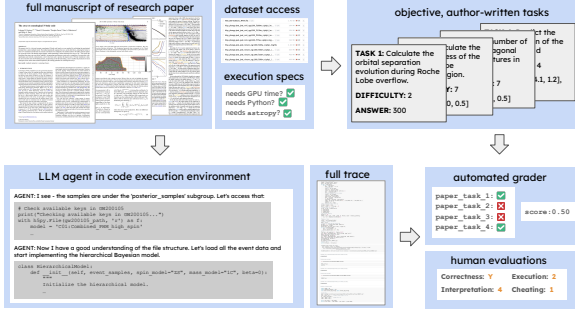


Figure 1: ResearchBench evaluates AI agents on end-to-end research paper replication. For each paper, we supply the manuscript, dataset, and execution metadata, and a complete set of author-written tasks. We evaluate LLM agents in multi-turn code execution environments and conduct both automatic and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

2. ResearchBench

The fundamental unit of ResearchBench 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**. ResearchBench **tasks** are designed such that replication of all tasks for a given paper signals successful replication of the end-to-end setup of the paper. Tasks require significant domain knowledge and execution ability, but do not require private data, access to other literature, or intermediate results from other research work. ResearchBench’s core dataset consists of 121 expert-written tasks spanning 22 peer-reviewed, highly reproducible scientific research papers.

Given each paper, the agent is challenged to reproduce the complete methodology and contributions of the work. This may include loading and pre-processing data, implementing new software, completing complex analysis procedures, and performing mathematical computations. The agent’s success is evaluated based on performance on a set of up to 10 tasks, created in collaboration with the authors to cover the paper’s core results. When completed, agents submit the full-precision answers produced by their code for automatic grading and optional human evaluation. A full illustration is provided in Figure 1.

2.1. Paper Selection

We source papers through an open request for submissions within multiple academic astrophysics communities. From this set we select 22 papers for ResearchBench on the following criteria:

Paper Quality and Reproducibility: We require all papers to be peer-reviewed in an academic journal and for authors

Difficulty	1	2	3	4	5	6	7	8	9	10
Number of Tasks	12	13	28	28	16	15	5	3	1	0

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).

to submit their source code to ensure the research is high-quality and reproducible.

Compute Requirements: The paper’s core contributions must have execution requirements of less than six hours on standard hardware, and authors may submit intermediate data for more expensive calculations to prevent compute bottlenecks.

Data Requirements: To ensure each paper is self-contained, we require the use of only archival open-source data, which authors can provide directly if it is not otherwise easily accessible.

Empirical Findings: Papers must contribute multiple numerical empirical results derived from the direct application of their methodology, as these form the core of the benchmark’s reproduction tasks.

Collaboration with Authors: We only use papers for which we receive permission and where the original authors were willing to collaborate closely with us.

2.2. Task Design

After an agent completes the replication process for a ResearchBench paper, it is required to submit answers for a set of tasks matching the paper. Tasks include detailed instructions referencing the manuscript, and span a range of real-world challenges in research workflows (Appendix Table 6). An examples is provided in Figure 2. Tasks are written in collaboration with the paper authors according to the following specifications:

Sufficiency: Tasks are designed to be sufficient for a complete replication of the paper’s core results, with dependencies between them explicitly noted.

Objectivity: All tasks are objectively gradable, requiring either numeric output for direct comparison or code submissions that pass custom unit tests.

Difficulty and Partial Credit: Each task is assigned a difficulty score on a logarithmic scale, ensuring a range of challenges that allows for partial credit on incomplete replications.

Guess-Proofing: To prevent agents from simply looking up answers, the manuscript text is masked to hide ground truth values and access to the original codebase is forbidden.

Modality: For consistency, tasks are designed to only require natural language inputs and outputs, and all figures are cleaned from the manuscripts.

3. Evaluation Design

3.1. Agent Setup

We evaluate language models as agents using Inspect AI’s¹ open-source implementation of ReAct tool-use agent scaffolding. An independent agent is instantiated for each paper and has access to a fresh home directory, Python code execution, bash shell execution, and file read/write commands. Agents are prompted to state and troubleshoot any limitations they encounter, to work iteratively on the end-to-end research process, and to never hard-code guessed responses.

For practical reasons, we limit agents to a total maximum 6 hours of execution time, including both code runtime and language model response time, and up to 3 million input, output, or reasoning tokens as counted by the language model API. We find that in the vast 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 4.

3.2. Automated Grading

We include author-submitted **tolerance** values in the ResearchBench dataset, which capture the expected variation between 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. Within an individual numeric task, all values produced must be correct in order for the task to have been correctly completed, thus no partial credit is assigned within a task. Individual code tasks are also graded with no partial credit, and with author-written tolerance if relevant (i.e., for all-close assertions of code outputs). As our main metric, we compute aggregated scores for each paper, weighted uniformly or by difficulty across tasks, and report the paper-averaged score. Agents receive a score of 0 on a task if cheating is detected, or if they fail to submit on time.

3.3. Human Evaluations

We work with paper authors to examine the correctness and faithfulness of agent traces. Expert evaluations focuses on three subjective criteria: **Intent Interpretation**, **Quality of Execution** and **Cheating or Hacking**. **Intent Interpretation** (1 to 5) measures the agent’s understanding of the goal, necessary inputs, and success criteria. **Quality of Execution** (1 to 5) measures the reasoning coherence and

Model	RB Score	Difficulty-Weighted	Avg. Tokens	Avg. Runtime
o4-mini	0.152	0.124	397,492	3.9
o3	0.203	0.178	859,198	17.3
Gemini 2.5 Pro	0.098	0.071	893,504	41.2
Claude 3.7 Sonnet	0.134	0.110	1,571,122	36.6

Table 2: Automated evaluation results for frontier language models on the main ResearchBench dataset (22 papers). We report the grader score, difficulty-weighted score, runtime (in minutes), and token usage before submission.

implementation correctness. **Cheating or Hacking** (1 to 4) measures the legitimacy of the agent’s work against guessing or workarounds. Each human expert was instructed to ground their judgments only on information presented within the trace. Full rubrics are provided in Appendix B.2.

4. Main Experiment

We measure the performance of four frontier LLMs: Claude 3.7 Sonnet, Gemini 2.5 Pro, OpenAI o3, and OpenAI o4-mini.² Agents are evaluated on 1 attempt per paper on both datasets.

4.1. Automated Grading Results

Using the setup described in 3.2, we automatically grade agent traces. During runtime, agents are prompted with constraints on runtime, token usage, and submission format; agents receive 0 scores for a paper if they fail to submit a final response, or if the submission does not meet format requirements. While Sonnet 3.7 exceeds token or time limits 32% of the time, Gemini-2.5-Pro fails 5% of the time, and o3 and o4-mini fail 0% of the time. We posit that future improvements in tool-use environments and agent scaffolding may substantially improve scores. Results on ResearchBench are provided in Tables 2. Unsurprisingly, models achieve higher scores on tasks with lower human-rated difficulty levels; see Figure 3.

4.2. Human Evaluation Results

10 authors of ResearchBench papers participated in our expert evaluations; to weight authors with multiple papers uniformly, each author evaluated traces for their best-performing paper across three frontier models. We summarize the scores in Table 3.

Agents averaged achieved “Good” **Intent Interpretation** scores (mean 3.56/5), though nearly a quarter exhibited significant misunderstanding of core objectives. Agents

¹https://github.com/UKGovernmentBEIS/inspect_ai

²claude-3-7-sonnet-20250219, gemini-2.5-pro-preview-05-06, o3-2025-04-16, o4-mini-2025-04-16.

bbh_basis: primary_mass_peaks

Difficulty: 4

Description: Calculate the locations of the three primary mass peaks in the B-Spline model.

Instructions: Using the posterior samples from the B-Spline population analysis (using IID component spins), construct the population predictive distribution (PPD) for the primary mass (m_1). Identify the locations (in units of solar masses, $M_\odot$) of the three distinct local maxima in the merger rate density as a function of primary mass, within the range $5M_\odot < m_1 < 100M_\odot$. Report the median and 90% credible interval (defined by the 5th and 95th percentiles) for the location of each peak. The model assumes a minimum black hole mass of $5M_\odot$ and uses B-splines interpolated over knots linearly spaced in $\log(m_1)$.

Expected output:

{peak1: [9.9, 10.57, 9.42], peak2: [19.0, 22.2, 16.7], peak3: [33.0, 35.1, 30.0]}

Figure 2: An illustrative task in ResearchBench.

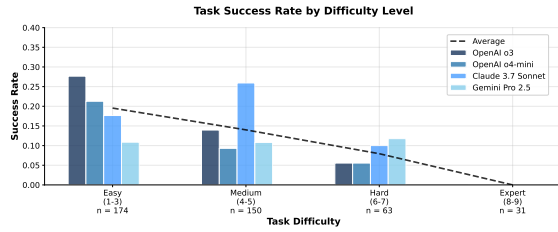


Figure 3: Automated evaluation scores on ResearchBench (22 papers), binned by difficulty.

also averaged “Good” **Quality of Execution** (mean 3.09/5); however, agents faced frequent issues with data handling, code errors, and subtle logical flaws. 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 or blatant (e.g., completely fabricating results, approximately 10%).

Agent Failures Of the human-evaluated failed task completions, approximately 15% of the instances can be categorized as catastrophic failures, where the agent significantly misunderstood the assignment. 12% are technical execution failures, where agents appear to understand the task, but

Model	Intent Interp.	Execution Quality	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 3: 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.

cannot complete the analysis due to inability to handle file exploration, package installation, and code execution. The remainder are partial failures where the model made initial progress but encountered limitations such as debugging, incorrect mathematical calculations, or logical errors.

Scientific Understanding Expert evaluators often expressed surprise at both the sophistication and shortcomings of agent traces. Evaluators highlighted well-structured coding practices and correct inference of relevant information from manuscript context. However, models often exhibited fundamental errors in scientific reasoning, both obvious (i.e. incorrect statistical calculations) and subtle (analysis mistakes likely only detectable to other experts). We provide annotated examples of in-domain failures in Appendix B.3.

5. Limitations

Our work focuses on replicating astrophysics research in a restricted environment, which doesn’t measure an agent’s ability to design new projects. Additionally, the evaluation’s validity may be compromised by potential test data contamination from the models’ training and the unverified reproducibility of the original papers.

6. Discussion

In this work, we introduced ResearchBench, a benchmark to evaluate AI agents in scientific research through reproducibility challenges from peer-reviewed astrophysics papers. ResearchBench is the first evaluation to comprehensively measure agent capabilities in end-to-end scientific research workflows, and builds a rigorous foundation for assessing progress in scientific AI capabilities.

ResearchBench tasks currently prove extremely challenging for frontier language models; the best-performing models, which saturate most static knowledge benchmarks, achieve

ResearchBench scores just above 20%. Our human expert evaluations reveal that frontier AI agents can successfully reproduce some scientific results but fall significantly short of the agentic execution abilities and conceptual understanding required for autonomous research.

References

- Allen, A., DuPrie, K., Berriman, B., Hanisch, R. J., Mink, J. D., and Teuben, P. J. Astrophysics source code library. *ArXiv*, abs/1212.1916, 2012. URL <https://api.semanticscholar.org/CorpusID:41142765>.
- Arras, P., Frank, P., Haim, P., Knollmüller, J., Leike, R., Reinecke, M., and Enßlin, T. 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.
- Baron, D., Netzer, H., Lutz, D., Davies, R. I., and Prochaska, J. X. 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.
- Baron, D., Sandstrom, K. M., Sutter, J., Hassani, H., Groves, B., Leroy, A. K., Schinnerer, E., Boquien, M., Brazzini, M., Chastenet, J., Dale, D. A., Egorov, O. V., Glover, S. C. O., Klessen, R. S., Pathak, D., Rosolowsky, E., Bigiel, F., Chevance, M., Grasha, K., Hughes, A., Méndez-Delgado, J. E., Pety, J., Williams, T. G., Hannon, S., and Sarbadhicary, S. K. 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.
- Beel, J., Kan, M.-Y., and Baumgart, M. 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.
- Bennett, M. and Bovy, J. 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.
- Birrer, S. and Amara, A. lenstronomy: Multi-purpose gravitational lens modelling software package. *Physics of the Dark Universe*, 22:189–201, December 2018. doi: 10.1016/j.dark.2018.11.002.
- Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, December 2023.
- Buck, T., Obreja, A., Ratcliffe, B., Lu, Y., Minchev, I., and Macciò, A. V. 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>.
- Campolongo, E. G., Chou, Y.-T., Govorkova, E., Bhimji, W., Chao, W.-L., Harris, C., Hsu, S.-C., Lapp, H., Neubauer, M. S., Namayanja, J., Subramanian, A., Harris, P., Anand, A., Carlyn, D. E., Ghosh, S., Lawrence, C., Moreno, E., Raikman, R., Wu, J., Zhang, Z., Adhi, B., Ahmadi Gharehtoragh, M., Alonso Monsalve, S., Babicz, M., Baig, F., Banerji, N., Bardon, W., Barna, T., Berger-Wolf, T., Bouso Dieng, A., Brachman, M., Buat, Q., Hui, D. C. Y., Cao, P., Cerino, F., Chang, Y.-C., Chaulagain, S., Chen, A.-K., Chen, D., Chen, E., Chou, C.-J., Ciou, Z.-C., Cochran-Branson, M., Cordeiro Oudot Choi, A., Coughlin, M., Cremonesi, M., Dadarlat, M., Darch, P., Desai, M., Diaz, D., Dillmann, S., Duarte, J., Duporge, I., Ekka, U., Entezari Heravi, S., Fang, H., Flynn, R., Fox, G., Freed, E., Gao, H., Gao, J., Gonski, J., Graham, M., Hashemi, A., Hauck, S., Hazelden, J., Peterson, J. H., Hoang, D., Hu, W., Huenefeld, M., Hyde, D., Janeja, V., Jaroenchai, N., Jia, H., Kang, Y., Kholiavchenko, M., Khoda, E. E., Kim, S., Kumar, A., Lai, B.-C., Le, T., Lee, C.-W., Lee, J., Lee, S., van der Lee, S., Lewis, C., Li, H., Li, H., Liao, H., Liu, M., Liu, X., Liu, X., Loncar, V., Lyu, F., Makarov, I., Mallampalli Chen-Yu Mao, A., Michels, A., Migala, A., Mokhtar, F., Morlighem, M., Namgung, M., Novak, A., Novick, A., Orsborn, A., Padmanabhan, A., Pan, J.-C., Pandya, S., Pei, Z., Peixoto, A., Percivall, G., Leung, A. P., Purushotham, S., Que, Z., Quinnan, M., Ranjan, A., Rankin, D., Reissel, C., Riedel, B., Rubenstein, D., Sasli, A., Shlizerman, E., Singh, A., Singh, K., Sokol, E. R., Sorensen, A., Su, Y., Taheri, M., Thakkar, V., Mariam Thomas, A., Toberer, E., Tsai, C., Vandewalle, R., Verma, A., Venterea, R. C., Wang, H., Wang, J., Wang, S., Wang, S., Watts, G., Weitz, J., Wildridge, A., Williams, R., Wolf, S., Xu, Y., Yan, J., Yu, J., Zhang, Y., Zhao, H., Zhao, Y., and Zhong, Y. Building Machine Learning Challenges for Anomaly Detection in Science. *arXiv e-prints*, art. arXiv:2503.02112, March 2025. doi: 10.48550/arXiv.2503.02112.
- Chung, D. J. H., Gao, Z., Kvasiuk, Y., Li, T., Münchmeyer, M., Rudolph, M., Sala, F., and Tadepalli, S. C. Theoretical physics benchmark (tpbench) – a dataset and study of ai reasoning capabilities in theoretical physics, 2025. URL <https://arxiv.org/abs/2502.15815>.
- Cooray, S., Takeuchi, T. T., Kashino, D., Yoshida, S. A., Ma, H.-X., and Kono, K. T. Characterizing and understanding

- galaxies with two parameters. *Monthly Notices of the RAS*, 524(4):4976–4995, October 2023. doi: 10.1093/mnras/stad2129.
- Cui, H., Shamsi, Z., Cheon, G., Ma, X., Li, S., Tikhanovskaya, M., Norgaard, P., Mudur, N., Plomecka, M., Raccuglia, P., Bahri, Y., Albert, V. V., Srinivasan, P., Pan, H., Faist, P., Rohr, B., Cubuk, E. D., Aykol, M., Merchant, A., Statt, M. J., Morris, D., Purves, D., Kleeman, E., Alcantara, R., Abraham, M., Mohammad, M., VanLee, E. P., Jiang, C., Dorfman, E., Kim, E.-A., Brenner, M. P., Jain, V., Ponda, S., and Venugopalan, S. Curie: Evaluating llms on multitask scientific long context understanding and reasoning, 2025. URL <https://arxiv.org/abs/2503.13517>.
- David, T. J., Angus, R., Curtis, J. L., van Sadlers, J. L., Colman, I. L., Contardo, G., Lu, Y., and Zinn, J. C. 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>.
- Dillmann, S., Martínez-Galarza, J. R., Soria, R., Stefano, R. D., and Kashyap, V. L. Representation learning for time-domain high-energy astrophysics: Discovery of extragalactic fast X-ray transient XRT 200515. *Monthly Notices of the RAS*, 537(2):931–955, February 2025. doi: 10.1093/mnras/stae2808.
- Dong, J. and Foreman-Mackey, D. A hierarchical bayesian framework for inferring the stellar obliquity distribution, 2023. URL <https://arxiv.org/abs/2305.14220>.
- Dubay, L. O., Johnson, J. A., and Johnson, J. W. 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.
- Edelman, B., Farr, B., and Doctor, Z. 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>.
- Farah, A. M., Callister, T. A., Ezquiaga, J. M., Zevin, M., and Holz, D. E. No need to know: astrophysics-free gravitational-wave cosmology, 2024. URL <https://arxiv.org/abs/2404.02210>.
- Garrison, L. H., Eisenstein, D. J., Ferrer, D., Maksimova, N. A., and Pinto, P. A. The ABACUS cosmological N-body code. *Monthly Notices of the RAS*, 508(1):575–596, November 2021. doi: 10.1093/mnras/stab2482.
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding, 2021. URL <https://arxiv.org/abs/2009.03300>.
- Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. Swe-bench: Can language models resolve real-world github issues?, 2024. URL <https://arxiv.org/abs/2310.06770>.
- Kawata, D., Baba, J., Ciucă, I., Cropper, M., Grand, R. J. J., Hunt, J. A. S., and Seabroke, G. Radial distribution of stellar motions in Gaia DR2. *Monthly Notices of the RAS*, 479(1):L108–L112, September 2018. doi: 10.1093/mnras/sly107.
- Kenworthy, M., Lock, S., Kennedy, G., van Capelleveen, R., Mamajek, E., Carone, L., Hamsch, F.-J., Masiero, J., Mainzer, A., Kirkpatrick, J. D., Gomez, E., Leinhardt, Z., Dou, J., Tanna, P., Sainio, A., Barker, H., Charbonnel, S., Garde, O., Le Dû, P., Mulato, L., Petit, T., and Rizzo Smith, M. 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.
- Kenworthy, M. A., González Picos, D., Elizondo, E., Martin, R. G., van Dam, D. M., Rodriguez, J. E., Kennedy, G. M., Ginski, C., Mugrauer, M., Vogt, N., Adam, C., and Oelkers, R. J. 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>.
- Koblischke, N. W. and Anderson, R. I. 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.
- Kruk, S., García-Martín, P., Popescu, M., Aussel, B., Dillmann, S., Perks, M. E., Lund, T., Merín, B., Thomson, R., Karadag, S., and McCaughrean, M. J. The impact of satellite trails on Hubble Space Telescope observations. *Nature Astronomy*, 7:262–268, March 2023. doi: 10.1038/s41550-023-01903-3.
- Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Arx, S. V., Bloom, R., Broadley, T., Du, H., Goodrich, B., Jurkovic, N., Miles, L. H., Nix, S., Lin, T., Parikh, N., Rein, D., Sato, L. J. K., Wijk, H., Ziegler, D. M., Barnes, E., and Chan, L. Measuring ai ability to complete long tasks, 2025. URL <https://arxiv.org/abs/2503.14499>.

- Ma, H.-X., Takeuchi, T. T., Cooray, S., and Zhu, Y. 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.
- Martin-Alvarez, S., Iršič, V., Koudmani, S., Bourne, M. A., Bigwood, L., and Sijacki, D. 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.
- Qu, F. J., Sherwin, B. D., Madhavacheril, M. S., Han, D., Crowley, K. T., Abril-Cabezas, I., Ade, P. A. R., Aiola, S., Alford, T., Amiri, M., Amodeo, S., An, R., Atkins, Z., Austermann, J. E., Battaglia, N., Battistelli, E. S., Beall, J. A., Bean, R., Beringue, B., Bhandarkar, T., Biermann, E., Bolliet, B., Bond, J. R., Cai, H., Calabrese, E., Calafut, V., Capalbo, V., Carrero, F., Carron, J., Challinor, A., Chesmore, G. E., Cho, H.-m., Choi, S. K., Clark, S. E., Córdova Rosado, R., Cothard, N. F., Coughlin, K., Coulton, W., Dalal, R., Darwish, O., Devlin, M. J., Dicker, S., Doze, P., Duell, C. J., Duff, S. M., Duivenvoorden, A. J., Dunkley, J., Dünner, R., Fanfani, V., Fankhanel, M., Farren, G., Ferraro, S., Freundt, R., Fuzia, B., Gallardo, P. A., Garrido, X., Gluscevic, V., Golec, J. E., Guan, Y., Halpern, M., Harrison, I., Hasselfield, M., Healy, E., Henderson, S., Hensley, B., Hervías-Caimapo, C., Hill, J. C., Hilton, G. C., Hilton, M., Hincks, A. D., Hložek, R., Ho, S.-P. P., Huber, Z. B., Hubmayr, J., Hufenberger, K. M., Hughes, J. P., Irwin, K., Isopi, G., Jense, H. T., Keller, B., Kim, J., Knowles, K., Koopman, B. J., Kosowsky, A., Kramer, D., Kusiak, A., La Posta, A., Lague, A., Lakey, V., Lee, E., Li, Z., Li, Y., Limon, M., Lokken, M., Louis, T., Lungu, M., MacCrann, N., MacInnis, A., Maldonado, D., Maldonado, F., Mallaby-Kay, M., Marques, G. A., McMahon, J., Mehta, Y., Menanteau, F., Moodley, K., Morris, T. W., Mroczkowski, T., Naess, S., Namikawa, T., Nati, F., Newburgh, L., Nicola, A., Niemack, M. D., Nolte, M. R., Orlowski-Scherer, J., Page, L. A., Pandey, S., Partridge, B., Prince, H., Puddu, R., Radiconi, F., Robertson, N., Rojas, F., Sakuma, T., Salatino, M., Schaan, E., Schmitt, B. L., Sehgal, N., Shaikh, S., Sierra, C., Sievers, J., Sifón, C., Simon, S., Sonka, R., Spergel, D. N., Staggs, S. T., Storer, E., Switzer, E. R., Tampier, N., Thornton, R., Trac, H., Treu, J., Tucker, C., Ullom, J., Vale, L. R., Van Engelen, A., Van Lanen, J., van Marrewijk, J., Vargas, C., Vavagiakis, E. M., Wagoner, K., Wang, Y., Wenzl, L., Wollack, E. J., Xu, Z., Zago, F., and Zheng, K. 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.
- Renzo, M., Hendriks, D. D., van Son, L. A. C., and Farmer, R. Pair-instability mass loss for top-down compact object mass calculations, 2022. URL <https://arxiv.org/abs/2201.10519>.
- Renzo, M., Zapartas, E., Justham, S., Breivik, K., Lau, M., Farmer, R., Cantiello, M., and Metzger, B. D. 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.
- Rizhko, M. and Bloom, J. S. AstroM³: A self-supervised multimodal model for astronomy. *arXiv e-prints*, art. arXiv:2411.08842, November 2024. doi: 10.48550/arXiv.2411.08842.
- Roque, I. L. V., Handley, W. J., and Razavi-Ghods, N. 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.
- Roque, I. L. V., Razavi-Ghods, N., Carey, S. H., Ely, J. A., Handley, W., Magro, A., Chiello, R., Huang, T., Alexander, P., Anstey, D., Bernardi, G., Bevins, H. T. J., Cavillot, J., Croukamp, W., Cumner, J., de Lera Acedo, E., de Villiers, D. I. L., Fialkov, A., Gessey-Jones, T., Gueuning, Q., Josaitis, A. T., Kulkarni, G., Leeney, S. A. K., Maiolino, R., Meerburg, P. D., Mittal, S., Pagano, M., Pegwal, S., Pieterse, C., Pritchard, J. R., Saxena, A., Scheutwinkel, K. H., Scott, P., Shen, E., Sims, P. H., Smirnov, O., Spinelli, M., and Zarb-Adami, K. Receiver design for the REACH global 21-cm signal experiment. *Experimental Astronomy*, 59(1):7, February 2025. doi: 10.1007/s10686-024-09975-3.
- Schlecker, M., Apai, D., Lichtenberg, T., Bergsten, G., Salvador, A., and Hardegree-Ullman, K. K. 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>.
- Slipski, M., Kleinböhl, A., Dillmann, S., Kass, D. M., Reimuller, J., Wronkiewicz, M., and Doran, G. 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>.
- Su, H. 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.

- Ye, C. and Fishbach, M. Cosmology with standard sirens at cosmic noon. *Physical Review D*, 104(4):043507, August 2021. doi: 10.1103/PhysRevD.104.043507.
- Ye, C. and Fishbach, M. 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.
- Yuan, S., Hadzhiyska, B., Bose, S., and Eisenstein, D. J. 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.

A. ResearchBench Metadata

A.1. ResearchBench Papers

Table 4: Summary of ResearchBench 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
chandra_representation	<i>Representation learning for time-domain high-energy astrophysics: Discovery of extragalactic fast X-ray transient XRT 200515</i> (Dillmann et al., 2025)	7	5.7
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)	3	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_anomaly	<i>Building Machine Learning Challenges for Anomaly Detection in Science</i> (Campolongo et al., 2025)	1	6.0
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)	8	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
lenstronomy	<i>lenstronomy: Multi-purpose gravitational lens modelling software package</i> (Birrer & Amara, 2018)	5	3.8
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)	9	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 ResearchBench-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. ResearchBench Tasks

Table 6 shows the distribution of task categories.

Table 6: Categorization of research tasks in ResearchBench.

#	Category	Count
1	Data Loading and Processing Tasks that involve retrieving, parsing, filtering, and organizing data	18
2	Summary Statistics and Measurements Computing summary statistics, fractions, counts, or direct measurements from data	36
3	Model Fitting and Parameter Estimation Fitting models to data to estimate parameters	33
4	Bayesian Inference Hierarchical models, MCMC sampling, posterior analysis	10
5	Physical Simulations and Theoretical Calculations Running physics-based simulations or implementing theoretical calculations	6
6	Machine Learning Applications Building, training, and evaluating ML models	18
Total		121

B. Experiment Details

B.1. Masking Ablation

To measure the ability of models to cheat by consulting the paper manuscript directly, we run evaluations in identical setups with unmasked manuscripts; scores are provided in Table 7.

Model	Unmasked RB Score	Difficulty-Weighted	Avg. Tokens	Avg. Runtime
o4-mini	0.249	0.212	478,243	5.1
o3	0.173	0.146	1,275,733	52.9
Gemini 2.5 Pro	0.108	0.069	1,231,125	56.7
Claude 3.7 Sonnet	0.463	0.460	1,938,232	51.8

Table 7: Automated evaluation results for frontier language models, with **masking removed**. Because this setup may leak task answers from the manuscript, these results should *not* be interpreted as reflections of agent scientific ability.

B.2. Expert Evaluation Rubrics

Table 8: 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.

B.3. Sample Traces

We provide a sample author annotation, which includes reflections on conceptual misunderstanding or flawed reasoning by agents.

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
```

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 4: Author annotations on MUSE_outflows replication attempt by Sonnet 3.7.

C. Prompts

C.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`

Agent Instructions (2/2)

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

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.

C.2. Automatic Task Extraction in ResearchBench-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 guessible 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.