Raising the Reproducibility Bar

Joseph Wonsil
University of British Columbia
Vancouver, British Columbia, Canada
jwonsil@student.ubc.ca

Rúbia Guerra
University of British Columbia
Vancouver, British Columbia, Canada
rubiarg@cs.ubc.ca

Adam Pocock
Oracle Labs
Burlington, Massachusetts, USA
adam.pocock@oracle.com

Jack Sullivan
Oracle Labs
Burlington, Massachusetts, USA
jack.t.sullivan@oracle.com

Margo Seltzer
University of British Columbia
Vancouver, British Columbia, Canada
mseltzer@cs.ubc.ca

Abstract

Current reproducibility research addresses the challenge of re-execution (executing an artifact someone else prepared), and it has chipped away at the reproducibility crisis only to reveal a comprehensibility crisis. Addressing this crisis requires that we focus on the question of why we care about reproducibility: to validate and peer-review others' work. Researchers who want to build upon published work might be able to execute the same code on the same data, but that is not the same as understanding the methodology embodied in that code.

In computational science, scientists now need both domain knowledge *and* computational knowledge to review a study well. We currently relieve them of the burden of computational knowledge by providing push-button reproducibility, but in doing so, we have inadvertently removed the "burden" of needing to understand how the analysis works. An opaque experiment that deterministically produces the exact same number across different platforms, even when executed by other users, is still an opaque experiment. Advancing science requires that a research artifact produce a roughly deterministic outcome *and* demonstrate that the computation matches the methodology and analysis appearing in its publication. We propose embracing comprehensibility as a desired outcome of reproducibility and call upon our community to explore how we can improve the comprehensibility of computational experiments. We suggest research directions using emerging technologies such as LLMs combined with existing research in provenance and virtualization to enable more scientists to generate comprehensible artifacts.

CCS Concepts

• **Software and its engineering** → *Reusability*; • **Social and professional topics** → *Software management*; • **Information systems** → Data provenance.

Keywords

reproducibility, comprehension, research artifacts

ACM REP '25, Vancouver, BC, Canada

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/10.1145/3736731.3746157

ACM Reference Format:
Joseph Wonsil, Rúbia Guerra, Adam Pocock, Jack Sullivan, and Margo Seltzer. 2025. Raising the Reproducibility Bar. In *Proceedings of ACM Conference on Reproducibility and Replicability (ACM REP '25).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3736731.3746157

1 Introduction

The reproducibility community has worked diligently to identify reproducibility barriers and produce solutions for them. Case studies and user surveys highlight both cultural and technical challenges [4, 10, 25, 26, 28]. Cultural issues have inspired policy changes at journals and conferences, such as the introduction of reproducibility badges [7]. Meanwhile, researchers are addressing technical challenges to simplify reproducible computational artifacts.

Reproducibility research is necessary to achieve the scientific standards at the heart of the scientific method. It is now time to expand the scope of computational reproducibility research. We encourage our fellow reproducibility researchers to build on top of data practices and re-execution to support knowledge transfer.

The purpose of *scientific reproducibility* is to provide to one's peers all of the tools and information necessary to scrutinize one's claims in an informed manner. An incomprehensible artifact resists scrutiny. A key property of an artifact should be its *connection to the corresponding publication.* Consider ELISA, a novel in-memory object sharing scheme for Virtual Machines (VMs) [35]. Its reproducibility artifact [34] contains a "Commentary" section in its README that bridges "the descriptions in the paper and the source code of the ELISA prototype." This bridging is critical for comprehensibility but must currently be done manually and at a burden to the authors. We advocate for a world where such artifacts are the norm without requiring significant effort.

Unfortunately, it is currently impractical for many venues to require a successfully reproduced artifact. Yet, current research shows that even just having artifacts available can lead to benefits such as higher citations [8, 23]. If artifacts are easier to review due to better reproducibility and comprehensibility, then it will be easier to restructure submission processes to include them.

We propose that the reproducibility community explore new technology, integrate lessons learned from user experience (UX) research, and revisit prior work so that the solutions we generate facilitate comprehensibility. Natural language is the root cause of many comprehensibility barriers, and recent advances in natural language processing (i.e., LLMs) have the potential to provide solutions. While LLMs are imperfect, we can benefit from careful UX

studies demonstrating how to use the technology as a supplement to thoughtful, expert-driven work grounded in critical evaluation. Additionally, UX studies focusing on the trade-off around researcher workload and a well-documented artifact can inform sustainable comprehensibility-driven workflows. Leveraging these UX studies to improve artifact comprehensibility and combining them with reproducibility tools will produce a virtuous cycle in which artifacts become more user friendly and more researchers avail themselves of tools that enable reproducibility and comprehensibility.

2 The Need for Reproducibility Comprehension

We identify three specific motivations for research into reproducibility comprehension.

(1) Scientific reproducibility must ultimately support peer-review of scientific results.
(2) Today's reproducibility tools often lead to opaque science.
(3) Researchers report that ensuring reproducibility is time-intensive.

In the National Academies of Sciences, Engineering, and Medicine's report on reproducibility and replicability, one of their stated conclusions is that the "scientific enterprise depends on the ability of the scientific community to scrutinize scientific claims and to gain confidence over time in results and inferences that have stood up to repeated testing" [18]. It follows that technical efforts to support reproducibility should support the process by which scientists can *scrutinize scientific claims*. Existing research in open science, FAIR (Findability, Accessibility, Interoperability, and Reusability [31]) data repositories, virtualization, and code re-execution all support scientific scrutiny but do not achieve it.

Computational science increases the burden of understanding to include not only the domain expertise about the research question being answered, but the computational expertise to understand and use the methods. While an ecologist will likely be happy to have an R environment managed for them, they should still be able to meaningfully engage with the R scripts executing within that environment. The challenge we face is decoupling the computational expertise from the domain expertise. How do we help a reviewer relate the algorithms and methods embodied in the code to the prose in the paper? Pedagogical research recommends that learners require an "appropriate level of challenge" [15], because understanding requires cognitive effort. An activity can be either too hard *or* too easy, and when an activity is too easy, a learner "completes the goal with little to no effort and is not pushed to improve" [15]. While comprehension of an artifact is different from conventional classroom learning, we claim that the requirement for cognitive effort applies to scrutinizing research artifacts.

Push-button or automatic reproducibility shifts the goal of artifact review from comprehension and scientific scrutiny to the simpler goal of seeing a research artifact produce the results described in the paper. Consider MERIT, the fully automatic machine learning reproducibility system [33]. It works only because it is embedded in Tribuo [22], which trains the models it is reproducing. Its users need not understand the program they are reproducing, because the system is automatic. Furthermore, Tribuo itself is inflexible; it does not allow users to control the training loop, which substantially improves the provenance quality at the cost of making certain computations difficult to perform. This fundamentally enables MERIT's automatic reproducibility, and the authors admit that letting "users control the training loop" is a barrier to reproducibility. Therefore, if a researcher trains a Tribuo model for a novel purpose as part of a publication, an artifact reviewer could reproduce the results without confirming that the model architecture and parameters correspond to those described in the publication. Despite being "reproducible", that model is still opaque and resists scrutiny. MERIT is still far preferable to the alternative, as providing no assistance returns us to a time when researchers struggled to re-execute code at all resulting in the "reproducibility crisis" [4]. As we discuss in Section 5, future work could build upon tools such as MERIT, using provenance data to facilitate comprehension [6].

The additional computational knowledge required to make computational experiments more reproducible and to understand how to reproduce a computational experiment adds mental overhead that dissuades scientists. Surveys revealed that researchers feel there is not enough time for reproducible practices and that they do not have enough knowledge and training of these practices [4, 10, 25, 26, 28]. Other studies have demonstrated that even when authors are sharing artifacts a reviewer will likely required moderate to significant effort to achieve satisfactory reproducibility [28–30, 32]. We now face a balancing act: from the perspective of an author a reproducibility artifact must not be "too hard" to produce, from the perspective of a reviewer it must not reduce engagement or it makes the artifact hard to scrutinize, but if it is too challenging to reproduce, reviewers might not have sufficient time and energy. Addressing this problem requires investigating an artifact's role and defining what it means to comprehend something.

3 Cognitive Background for Comprehensibility

Given that comprehension is a process occurring in the human mind, we discuss the cognitive background involved in generating a computational experiment and the role artifacts play in scientific review. We take inspiration from other human and cognitive-focused fields, such as pedagogy, and propose a new framework for identifying the actions a reviewer can accomplish with an artifact.

When scientists produce a publication, they have not only written a paper and produced a research artifact, they have built a mental "theory" [19]. Ryle introduced this theory [24], and Naur discussed how a programming team builds this theory while creating a program [19]. He states that this theory "is understood as the knowledge a person must have in order not only to do certain things intelligently but also to explain them, to answer queries about them, to argue about them, and so forth". Crucially, the theory involves more than just how the program executes; it includes its relationship to the real world in concepts inexpressible as a program or documentation. Someone lacking this theory is unable to effectively engage with the program, or in our case, a research artifact. Supporting scientific scrutiny requires that publications with computational artifacts provide methods to facilitate reconstruction of the original author's theory in the mind of someone else. In other words, an artifact should act partially as a pedagogical tool for the scientists interacting with it, to support their comprehension process.

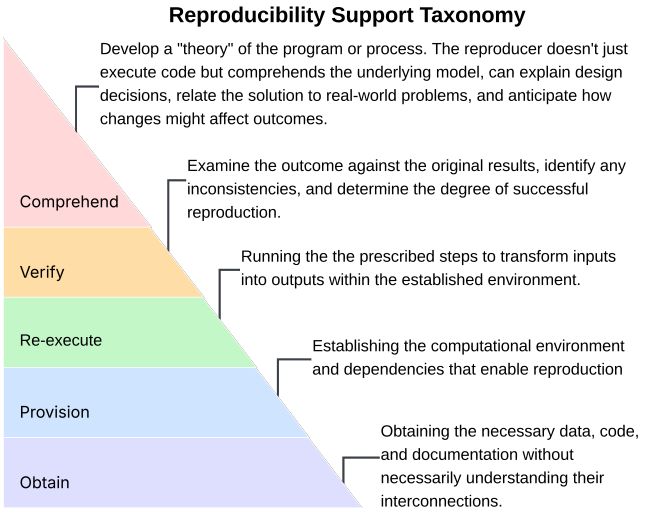

Figure 1: A "reproducibility support taxonomy" identifying each of the actions enabled and the artifacts necessary to achieve the kind of reproducibility that facilitates scientific scrutiny.

Bloom's Cognitive Taxonomy provides a categorization for the different stages of learning [5]. Note that comprehension[1] is its second stage. Bloom defined *comprehension* as a phenomenon occurring exclusively within the mind of a human being as they engage with educational material and is the next step past simply remembering facts. When someone has comprehended or understood something, they can explain it in their own words and draw logical conclusions based on the information they learned. When a scientist engages with a paper and its associated artifact, if they reach the comprehension stage they have made substantial progress towards re-creating the original authors' mental theory. Ideally, the artifact will encourage this theory building process.

An artifact's properties enable varying types of interactions that support reproducibility, and ultimately, a scientist's comprehension. We propose a "reproducibility support taxonomy" (Figure 1), inspired by Bloom's taxonomy [5], to illustrate stages of interactions that suggest how reproducibility tools can assist in creating artifacts ideal for the scientific process. Unlike Bloom's taxonomy, which describes a mental process, our taxonomy focuses on the interactions with a user that an artifact supports. The taxonomy's base, *obtain*, is the artifact's availability, as accessing it is a prerequisite to any other action. The FAIR repositories, venue availability requirements, and badging all point to this stage. The next stage, *provision*, concerns the artifact's computational environment and whether a researcher can access or recreate one sufficient to execute the artifact. Virtualization technologies and reproducibility packaging solutions assist with this stage and often the next stage: *re-execute*. For a researcher to successfully *re-execute* an artifact, the artifact should contain an unambiguous set of computational steps that are free of errors. Workflow languages, build systems, and reproducibility packages all assist with this stage. Comparing the results claimed in the paper to the artifact's output is the penultimate stage of the taxonomy, *verify*. Most tools leave this task to the user; however, noWorkflow [17], MERIT [33], and MXLP [16] all provide methods of comparing results of executions.

Most importantly, this taxonomy reveals a critical gap towards the goal of scientific scrutiny: *comprehension*. Few artifacts help a user re-build the "theory" underlying the artifact. This outcome is not entirely surprising, as it's the lower stages of the taxonomy that enable the higher stages. Only now, with the help of existing tools and techniques, can we begin to address comprehension. Boufford et al.'s LLM-generated provenance summaries [6] are, perhaps, a first step. Unlike most tools that prioritize automation over explanatory power, these summaries are narratives of execution graphs. We posit that reproducibility tools must balance automation with explanatory power to scaffold true understanding rather than just repeatable execution.

4 Now is the Ideal Time

The current landscape of cultural shifts towards reproducibility, existing reproducibility solutions, and emerging technologies in natural language processing makes now the ideal time to push for comprehensibility. While researchers view time and effort as barriers to pursuing proper reproducibility practices [4, 26], more journals are promoting artifact badges [7] or requiring open science practices. We also see communities such as Papers with Code [2] that encourage publicizing your code alongside your paper. These changes encourage researchers to pursue reproducibility practices, and we should promote practices that encourage proper scientific scrutiny via reproducibility comprehension. Without further guidance, current reproducibility research and approaches focusing on push-button reproducibility could converge on the opaque science we described earlier. For example, badges are helpful, but artifact evaluation committees (in the ACM, at least) consist of volunteers with unclear guidelines. In fact, the ACM states that "it is still too early to establish more specific guidelines for artifact and replicability review."[2] In practice, authors are incentivized to create the simplest artifact possible to ensure that the reviewer has an effortless experience reproducing the artifact. This approach encourages automatic reproducibility, often via virtualization and build systems. However, this process does not necessarily help validate the claims from a publication, nor does it facilitate other researchers' understanding and ability to build upon the work.

Recent advances in natural language processing using large language models (LLMs) might allow tasks that would otherwise require substantial manual effort to be achieved automatically. A significant component of comprehensibility is, of course, effective construction of natural language. LLMs open up new avenues of research that could explicitly help with comprehension and consistency between a computational experiment and its corresponding paper. These models have known weaknesses that we should recognize, but they have the potential to help.

[1] In a later revision, other researchers changed the name to "understanding" [3], but for the sake of consistency, we will continue to use comprehension in this paper.

[2] https://www.acm.org/publications/policies/artifact-review-badging - retrieved April 2, 2025

5 Proposed Research Areas

We propose that the following research areas should be explored to increase the comprehensibility of reproducibility artifacts.

To build a strong foundation for future work, we first recommend that reproducibility researchers conduct more user studies or at least explore existing research on user experience design and learning theory. Comprehensibility is a human-centric phenomenon, and any tools we generate to achieve it should, therefore, include human-centric studies. If the tools we build for scientists do not have scientists' input, we risk scientists disregarding them. Key points of interest include determining how users currently interact with and structure their workflows, identifying locations where instrumentation such as provenance collection could be placed automatically to reduce user effort for tracking experiments, and finding what features of an artifact are conducive to learning.

We suggest using LLMs as *supportive* tools during the research and writing processes. Imagine an AI assistant that provides reproducibility suggestions while scientists work. If the scientist were to add a new Python package to their script, the assistant could provide a suggestion to add it to the project's "requirements.txt" file and then suggest a Dockerfile capable of building a suitable environment. Alternatively, they can suggest connections between the otherwise disjoint code and paper, facilitating the creation of an artifact similar to ELISA [35], where sections of the code are associated with sections of the paper. A related example is DeepWiki [1] that makes it easier to understand a codebase by automatically generating documentation for a repository. However, our emphasis is on *suggestions* rather than complete generation to ensure that the scientist retains agency and is ultimately responsible for changes in a project. While OpenAI has recently demonstrated that LLMs are *capable of*, but not yet proficient at, replicating machine-learning research just by reading the papers [27], we are not suggesting that LLMs take over our replication studies in this manner. Doing so would be replacing one opaque scientific method with another.

Alternatively, LLMs can facilitate the accuracy and consistency of a paper across iterations. Suppose a researcher updates a numeric result in their paper, downgrading recall from 99% to 98%. If the researcher already had prose that referred to the previous value (e.g., "Our model achieves a recall of 98%. This result outperforms the state-of-the-art which achieves 98%."), that prose would now be incorrect. In this contrived example, they might remember to update the following sentence, but will they remember to update their forward reference in their introduction? Or their summary within their conclusion? What if out of a team of 10 authors one makes all the changes, and then another rolls back only some? LLMs have the potential to perform this type of proofreading. Then, rather than change the prose, the model will bring any inconsistent writing to the researcher's attention - not modify the paper (Proof of concept in Appendix A). This methodology ensures that even though a researcher wrote the paper with AI assistance, the author remains responsible for the contents of the paper.

Aided by new technology and the cultural shifts we described earlier, we recommend revisiting previously explored or current research areas to add comprehensibility as a key component. The executable paper [9] is a previous topic that has seen less attention recently. This method brings the code and prose closer together by hosting them in the same document thereby providing a reviewer the opportunity to engage with the code relevant to the section of the paper they are reading. An executable paper can provide push-button reproducibility, but in this instance the proximity of the result to the code that generated it increase transparency and reduce the effort a reviewer needs to verify the result. Two barriers to their current adoption include venue submission standards and difficulty preparing the code for publication. Given the adoption of reproducibility badges and open science initiatives, executable papers could see more popularity as researchers and venues look for accessible ways to promote open science. Even code preparation can be more accessible as recent research has demonstrated that code can be organized and explained without the use of language models, instead using provenance or static analysis [12, 14]. However, for those who want the LLM-powered research assistant, a language model could augment a tool that collects contextual information to assist a user with research programming, such as Burrito [11].

Fundamentally, can we build these tools that make reproducibility and comprehensibility more accessible *and* help users learn better reproducibility practices? Survey respondents explain that lack of training and knowledge around good practices affect reproducibility outcomes [4, 26]. While workshops, classes, and self-driven learning can help, those solutions already existed in some form or another at the time of those surveys. Given researchers' self-reported lack of time, they might not always be able to make proactive decisions regarding their formal training. While we believe such an investment in their development is well worth it, they might only learn enough to get through their current publication before moving on to the next. Using the techniques we described in the previous paragraphs, tools that can "nudge" scientists toward producing more reproducible work might help them develop skills over time. Consider a tool that could generate a template of Krafczyk et al.'s proposed *Reproduction Package* and then provide assistance in keeping the components consistent [13].

6 Conclusion

We call upon the reproducibility community to investigate new technical solutions to old problems. Now is the time to shift our attention to previously challenging issues given the new technologies in AI, the trajectory of field towards opaque push-button reproducibility, and the growing cultural emphasis on open data and transparent research practices. Adding computation to the scientific process has created additional burdens for scientists. While current research attempts to relieve those burdens, they inadvertently add new barriers to engaging with research artifacts. We should provide solutions that help scientists mentally engage with research artifacts, as well as relieve computational burdens. We propose attention to the following areas: exploring ways scientists engage with their workflows and research artifacts via user studies, tightening the connection between publication prose and code, harnessing the capabilities of LLMs across various components of the analysis and paper writing pipeline, and combining these new areas with existing solutions to make reproducible and comprehensible artifacts that encourage engagement. Committing to these areas enable more effective artifact review, encourages building off existing work, and increases trust in the scientific method.

References

[1] 2025. DeepWiki. deepwiki.com.

[2] 2025. Papers with Code. https://portal.paperswithcode.com/.

[3] Lorin W. Anderson and David R. Krathwohl. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives.* Longman.

[4] Monya Baker. 2016. 1,500 Scientists Lift the Lid on Reproducibility. *Nature* 533, 7604 (May 2016), 452–454. doi:10.1038/533452a

[5] Benjamin S Bloom, Max D Engelhart, Edward J Furst, Walker H Hill, and David R Krathwohl. 1956. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook 1: Cognitive Domain.* Longman New York.

[6] Nichole Boufford, Joseph Wonsil, Adam Pocock, Jack Sullivan, Margo Seltzer, and Thomas Pasquier. 2024. Computational Experiment Comprehension Using Provenance Summarization. In *Proceedings of the 2nd ACM Conference on Reproducibility and Replicability (ACM REP '24)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3641525.3663617

[7] Center for OpenScience. 2025. Open Science Badges. https://www.cos.io/initiatives/badges.

[8] Eitan Frachtenberg. 2022. Research Artifacts and Citations in Computer Systems Papers. *PeerJ Computer Science* 8 (Feb. 2022), e887. doi:10.7717/peerj-cs.887

[9] Ann Gabriel and Rebecca Capone. 2011. Executable Paper Grand Challenge Workshop. *Procedia Computer Science* 4 (2011), 577–578. doi:10.1016/j.procs.2011.04.060

[10] Devarshi Ghoshal, Drew Paine, Gilberto Pastorello, Abdelrahman Elbashandy, Dan Gunter, Oluwamayowa Amusat, and Lavanya Ramakrishnan. 2020. Experiences with Reproducibility: Case Studies from Scientific Workflows. In *Proceedings of the 4th International Workshop on Practical Reproducible Evaluation of Computer Systems.* ACM, Virtual Event Sweden, 3–8. doi:10.1145/3456287.3465478

[11] Philip J. Guo and Margo Seltzer. 2012. {BURRITO}: Wrapping Your Lab Notebook in Computational Infrastructure. In *4th USENIX Workshop on the Theory and Practice of Provenance (TaPP 12)*. 7–7.

[12] Jingmei Hu, Jiwon Joung, Maia Jacobs, Krzysztof Z. Gajos, and Margo I. Seltzer. 2020. Improving Data Scientist Efficiency with Provenance. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering.* ACM, Seoul South Korea, 1086–1097. doi:10.1145/3377811.3380366

[13] M. S. Krafczyk, A. Shi, A. Bhaskar, D. Marinov, and V. Stodden. 2021. Learning from Reproducing Computational Results: Introducing Three Principles and the Reproduction Package. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379, 2197 (March 2021), 20200069. doi:10.1098/rsta.2020.0069

[14] Barbara Lerner, Emery Boose, Orenna Brand, Aaron M. Ellison, Elizabeth Fong, Matthew Lau, Khanh Ngo, Thomas Pasquier, Luis A. Perez, Margo Seltzer, Rose Sheehan, and Joseph Wonsil. 2023. Making Provenance Work for You. *The R Journal* 14, 4 (Feb. 2023), 141–159. doi:10.32614/RJ-2023-003

[15] Marsha C. Lovett, Michael W. Bridges, Michele DiPietro, Susan A. Ambrose, and Marie K. Norman. 2023. *How Learning Works: Eight Research-Based Principles for Smart Teaching.* John Wiley & Sons, Incorporated, Newark, UNITED STATES.

[16] Alexandre Zouaoui Michael Arbel. 2024. MLXP: A Framework for Conducting Replicable Machine Learning eXperiments in Python. arXiv preprint arXiv:2402.13831.

[17] Leonardo Murta, Vanessa Braganholo, Fernando Chirigati, David Koop, and Juliana Freire. 2015. noWorkflow: Capturing and Analyzing Provenance of Scripts. In *International Provenance and Annotation Workshop (Lecture Notes in Computer Science)*, Bertram Ludäscher and Beth Plale (Eds.). Springer International Publishing, Cham, 71–83. doi:10.1007/978-3-319-16462-5_6

[18] National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and Replicability in Science.* National Academies Press (US), Washington (DC). doi:10.17226/25303

[19] Peter Naur. 1985. Programming as Theory Building. *Microprocessing and Microprogramming* 15, 5 (May 1985), 253–261. doi:10.1016/0165-6074(85)90032-8

[20] Rochana R. Obadage, Sarah M. Rajtmajer, and Jian Wu. 2024. SHORT: Can Citations Tell Us about a Paper's Reproducibility? A Case Study of Machine Learning Papers. In *Proceedings of the 2nd ACM Conference on Reproducibility and Replicability.* ACM, Rennes France, 96–100. doi:10.1145/3641525.3663628

[21] Perplexity. 2023. Perplexity.Ai (AI Chatbot) [Large Language Model]. https://www.perplexity.ai/.

[22] Adam Pocock. 2021. Tribuo: Machine Learning with Provenance in Java. *arXiv preprint arXiv:2110.03022* (2021). arXiv:2110.03022

[23] Edward Raff. 2023. Does the Market of Citations Reward Reproducible Work?. In *Proceedings of the 2023 ACM Conference on Reproducibility and Replicability.* ACM, Santa Cruz CA USA, 89–96. doi:10.1145/3589806.3600041

[24] Gilbert Ryle. 1949. *The Concept of Mind.* Hutchinson London.

[25] Evanthia Kaimaklioti Samota and Robert P. Davey. 2021. Knowledge and Attitudes Among Life Scientists Toward Reproducibility Within Journal Articles: A Research Survey. *Frontiers in Research Metrics and Analytics* 6 (June 2021). doi:10.3389/frma.2021.678554

[26] Sheeba Samuel and Birgitta König-Ries. 2021. Understanding Experiments and Research Practices for Reproducibility: An Exploratory Study. *PeerJ* 9 (April 2021), e11140. doi:10.7717/peerj.11140

[27] Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Chan Jun Shern, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. 2025. PaperBench: Evaluating AI's Ability to Replicate AI Research. (April 2025).

[28] Victoria Stodden, Matthew S. Krafczyk, and Adhithya Bhaskar. 2018. Enabling the Verification of Computational Results: An Empirical Evaluation of Computational Reproducibility. In *Proceedings of the First International Workshop on Practical Reproducible Evaluation of Computer Systems.* ACM, Tempe AZ USA, 1–5. doi:10.1145/3214239.3214242

[29] Victoria Stodden, Jennifer Seiler, and Zhaokun Ma. 2018. An Empirical Analysis of Journal Policy Effectiveness for Computational Reproducibility. *Proceedings of the National Academy of Sciences* 115, 11 (March 2018), 2584–2589. doi:10.1073/pnas.1708290115

[30] Ana Trisovic, Matthew K. Lau, Thomas Pasquier, and Mercè Crosas. 2022. A Large-Scale Study on Research Code Quality and Execution. *Scientific Data* 9, 1 (Feb. 2022), 60. doi:10.1038/s41597-022-01143-6

[31] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3, 1 (March 2016), 1–9. doi:10.1038/sdata.2016.18

[32] Joseph Wonsil, Nichole Boufford, Prakhar Agrawal, Christopher Chen, Tianhang Cui, Akash Sivaram, and Margo Seltzer. 2023. Reproducibility as a Service. *Software: Practice and Experience* 53, 7 (2023), 1543–1571. doi:10.1002/spe.3202

[33] Joseph Wonsil, Jack Sullivan, Margo Seltzer, and Adam Pocock. 2023. Integrated Reproducibility with Self-describing Machine Learning Models. In *Proceedings of the 2023 ACM Conference on Reproducibility and Replicability (ACM REP '23)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3589806.3600039

[34] Kenichi Yasukata. 2024. Yasukata/ELISA. https://github.com/yasukata/ELISA.

[35] Kenichi Yasukata, Hajime Tazaki, and Pierre-Louis Aublin. 2023. Exit-Less, Isolated, and Shared Access for Virtual Machines. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3 (ASPLOS 2023)*. Association for Computing Machinery, New York, NY, USA, 224–237. doi:10.1145/3582016.3582042

A LLM Proof-of-concept

This appendix presents a proof of concept demonstrating how large language models (LLMs) can enhance the comprehensibility of scientific papers. We illustrate this through two scenarios where we use Perplexity AI's Sonar [21] model to respond to prompts involving text from a paper. These scenarios are intended purely as a proof of concept, with a limited set of data as the paper section we chose is not particularly long. We chose a smaller text to demonstrate model capability, and ensure that those reviewing these scenarios would not need to read an entire paper to understand the output. Further studies should be done to determine ideal prompts, integration into a workflow, and efficacy on larger bodies of text.

A.1 Experimental Setup

We examine the AI's ability to assess the robustness of claims when experimental results change. The AI extracts key claims and then analyzes their dependence on the reported data, determining whether claims remain logically valid, need qualification, or require retraction. The model produced outputs in a tabular format, and we have preserved the content but modified the table formatting to be interpretable by LaTeX.

Our two scenarios are as follows:

(1) Two zero-shot prompts where the AI examines an original passage and a modified passage independently.
(2) One zero-shot prompt where the AI has access to both an original and modified passage and must explain the impact of the change.

A.1.1 Input Text. We used the entirety of Section 6 (Discussion and Conclusion) from Obadage et al.'s paper *Can citations tell us about a paper's reproducibility? A case study of machine learning papers* [20]. We chose a sentence that reports a numerical result and for each of the scenarios we use the original passage or a version we modified:

Original passage:

> "We trained two sentiment analysis models and achieved F1-scores of 0.7–0.86."

Modified passage:

> "We trained two sentiment analysis models and achieved F1-scores of **0.2–0.46**."

We perform this modification to make the results "worse" to evaluate how the model will respond to a change that weakens the paper's claims.

A.2 Scenario 1

We performed two zero-shot prompts with the same context and same instructions to analyze the provided section of the paper. In the first prompt, we kept the section as it was in the paper. In the second prompt, we use the modified passage. These outputs are found in Table 1 and Table 2. Each row is a claim extracted from the section by the model and, despite being performed across two zero-shot prompts, both tables contain the same claims in each row—even if the wording varies slightly.

A.2.1 Prompt.

Context. You are an AI assistant tasked with analyzing scientific papers. Your goal is to verify whether the claims made in the prose of a paper remain valid if the experimental results or numerical outcomes were to change. This is to support reproducibility and comprehensibility by assessing the logical and conceptual stability of claims independent of specific results.

Instructions. Given the prose of a scientific paper (introduction, discussion, conclusion, or claim statements) and a set of experimental results, perform the following:

(1) Identify and extract key claims or conclusions explicitly stated in the prose.
(2) Determine how strongly each claim depends on the specific reported results.
(3) Assess if the claim holds true. For example:
    - Is the claim logically or theoretically supported beyond the specific results?
    - Does the claim need to be qualified or retracted?
    - Are there any assumptions or conditions that must hold for the claim to be valid?
(4) For each claim, output a clear explanation of whether it is robust or fragile with respect to results, including reasoning and any relevant caveats.

A.3 Scenario 2

In the second scenario, we performed a zero-shot prompt with the same context as previously, but modified the instructions to explicitly compare the modified ("recompiled") version of the paper, contrasting it to the original passage. We provided both passages as part of the same prompt. The output can be found in Table 3.

A.3.1 Prompt.

Context. You are an AI assistant tasked with analyzing scientific papers. Your goal is to verify whether the claims made in the prose of a paper remain valid if the experimental results or numerical outcomes were to change. This is to support reproducibility and comprehensibility by assessing the logical and conceptual stability of claims independent of specific results.

Instructions. We have just re-compiled our paper. Confirm that any of the results we reference are consistent with how we describe them in the prose and that they support the arguments we are making.

(1) Examine or identify what has changed between the two versions.
(2) Identify the claims that those changes affect.
(3) Assess claim robustness in light of the modified results.

A.4 Discussion

We determined that LLMs have the *potential* to provide assistance with ensuring the claims made in a paper are consistent with the results reported, and provide explanations for its reasoning when they do not.

Scenario 1. Even though the model has no access to the original results, it can identify that in the modified passage the "correlation between sentiment and reproducibility scores" is poor, and

downgrades its assessment from "Moderately robust" (Table 1) to "Fragile to some extent" (Table 2).

Scenario 2. We note that the model effectively highlighted changes between the original and modified versions, and offered a concise summary of how these changes affected the stated claims (Table 3).

We believe this functionality may be useful for authors or reviewers who need to scan for inconsistencies between reported results and their interpretations in the prose, especially given the iterative nature of computational experiments. Finally, we note that this evaluation represents a proof-of-concept based on a single, simple example. As such, it might not generalize across all types of result modifications or claim structures. Further investigation is warranted to assess the robustness and limitations of this approach across a broader set of cases.

Claim	Robustness	Explanation
Correlation between sentiment and reproducibility scores	Moderately robust	Improved model accuracy supports meaningful correlations; still preliminary and dataset-limited.
Potential to estimate reproducibility statistically	Fragile but promising	Conditional on further verification; improved results increase confidence in potential.
Citation contexts signal reproducibility efforts	Moderately robust	Supported by data and plausible assumptions; interpretation depends on citation behavior assumptions.
Citation sentiment should not replace experiments	Robust	Theoretical caution independent of results; sound methodological advice.
Limitations and future work	Robust	Methodological transparency and future plans are valid regardless of results.

Table 1: Scenario 1 - Original Passage

Claim	Assessment of Robustness	Explanation
Correlation between sentiment and reproducibility scores	Fragile to some extent	The correlation depends on the quality and size of the dataset and model performance. With low F1-scores and limited data, the strength of correlation could vary with more data or better models. The claim should be qualified as preliminary and exploratory.
Potential to estimate reproducibility statistically from citation contexts	Fragile, conditional	This claim is explicitly conditional on verification with larger datasets. It is a hypothesis supported by current trends but requires further validation before being considered robust.
Citation contexts contain signals about reproducibility efforts	Moderately robust	Conceptually plausible and supported by observed correlations, but interpretation depends on assumptions about citation behavior and sentiment meaning. It holds if the assumption that sentiment reflects reproducibility efforts is valid.
Citation context sentiment should not replace direct experiments	Robust	This is a theoretical and methodological caution that stands independently of results, emphasizing that sentiment analysis is a surrogate, not a replacement.
Limitations and future work	Robust	These are methodological acknowledgments and plans, inherently independent of results and valid as stated.

Table 2: Scenario 1 - Modified Passage

Aspect Examined	Change Identified	Claims Affected	Assessment of Claim Robustness
Sentiment analysis model F1-scores	Original: 0.7–0.86 Recompiled: 0.2–0.36	Claims about model performance and reliability of sentiment analysis results	The substantial drop in F1-scores indicates much weaker model performance than originally reported. This weakens confidence in the sentiment signals extracted, potentially undermining claims about reliably detecting reproducibility-related sentiments from citation contexts. The claim that models "exhibited" correlations still holds, but the strength and reliability of those correlations are now questionable.
Correlation trends between sentiment and reproducibility scores	No change reported in correlation direction or qualitative description	Claims about increasing positive sentiment and decreasing negative sentiment with rs_score	The qualitative trends remain consistent, supporting the conceptual claim that citation sentiments relate to reproducibility scores. However, given the low F1-scores, these trends may be noisier or less robust than initially implied. The claim is still plausible but should be presented with stronger caution regarding reliability.
Overall interpretation of findings	No change in interpretation text	Claims about potential to statistically estimate reproducibility from citation contexts and surrogate use	The lowered model performance weakens the practical feasibility of using citation sentiment as a reproducibility proxy. The cautious language ("If verified…", "may be useful") remains appropriate and necessary. The claim remains conceptually valid but is now more tentative and dependent on future improvements and larger datasets.
Limitations and future work	No change in limitations or future work description	Claims about limitations (training data size, selection bias) and future plans	These claims remain fully consistent and appropriate. The low F1-scores reinforce the stated limitation of small training data, underscoring the need for expanded labeling and dataset collection. No adjustments needed here.

Table 3: Scenario 2 - analysis of changes between original and recompiled paper versions and their impact on claims.