SHORT: 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 and revealed a comprehensibility crisis. Our community is uniquely positioned to address this crisis by focusing on the question of why we care about reproducibility: to validate and peer-review others' work. The introduction of computation to the scientific method added new burdens to scientists, as they now need domain knowledge *and* computational knowledge to review a study successfully. We can relieve them of that burden by providing push-button reproducibility, but we then inadvertently remove the "burden" of needing to understand how the analysis works. An opaque experiment that deterministically spits out the exact same number across different platforms, even when operated by other users, is still an opaque experiment. Ideally, a research artifact should produce a roughly deterministic outcome *and* demonstrate that the computation matches the methodologies and analyses claimed in its corresponding publication. We propose embracing comprehensibility as a necessary facet 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

• **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Keywords

Do, Not, Us, This, Code, Put, the, Correct, Terms, for, Your, Paper

Conference acronym 'XX, Woodstock, NY

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

ACM Reference Format:
Joseph Wonsil, Rúbia Guerra, Adam Pocock, Jack Sullivan, and Margo Seltzer. 2025. SHORT: Raising the Reproducibility Bar. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

In recent years, the reproducibility community has worked diligently to identify barriers and propose solutions. Case studies and user surveys have highlighted both cultural and technical challenges [1, 4, 11, 12, 14]. Cultural issues have inspired policy changes at journals and conferences, such as introducing reproducibility badges[1]. Meanwhile, developers are addressing technical challenges to simplify the creation of reproducible computational experiments.

This research is necessary for reaching scientific standards that promote trust in the scientific method; however, we believe it is the right time to expand the scope of computational reproducibility research. Many reproducibility studies highlight FAIR [16] as the principles to follow for research artifacts. Yet when researchers test the automatic reproducibility of FAIR repositories [15, 17], they run into issues as simple as being unable to determine how to execute the scripts. These concerns are not a condemnation of the FAIR principles, nor the repositories implementing them, because we would not even be able to test the reproducibility of these experiments without them! Rather, we encourage our fellow reproducibility researchers to take the next steps past accessible data and even re-execution. It is not enough for an artifact to be found, accessed, and re-executed - it must be comprehensible. We, therefore, propose a "reproducibility artifact taxonomy" inspired by Bloom's cognitive taxonomy [2] to help guide the direction of reproducibility tools towards creating artifacts ideal for the scientific process Figure 1.

Ideally, a research artifact is not only deterministic in its outcome but comprehensible in how it reached that outcome. The purpose of *scientific reproducibility* is for one's peers to scrutinize scientific claims, hence why interpretation is at the top of our taxonomy. An incomprehensible artifact, whether outright missing or plain messy, resists scrutiny. A key property of comprehensibility should be its *connection to the corresponding publication.* Consider ELISA [19], its reproducibility artifact[2] contains within its README a "Commentary" section responsible for bridging "the descriptions in the paper and the source code of the ELISA prototype." This type

[1] https://www.cos.io/initiatives/badges
[2] https://github.com/yasukata/ELISA

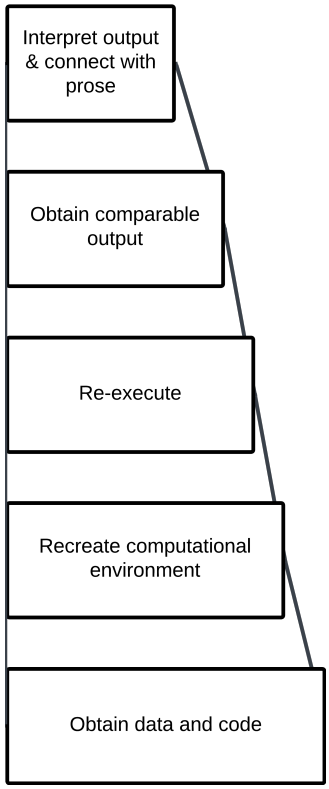

Figure 1: A "reproducibility artifact taxonomy" illustrating all the necessary steps to achieve an ideal research artifact.

of bridging is critical for comprehensibility but must currently be done manually and with great effort by the authors. We imagine a world where this type of artifact can be the norm without such significant effort.

We propose that the reproducibility community explore new technology, revisit prior work, and integrate lessons learned from user experience (UX) research. Natural language is the root cause of many comprehensibility barriers, and recent advancements in natural language processing in the form of large language models (LLMs) have the potential to provide solutions. With LLMs, new avenues of artifact comprehensibility are achievable by supplementing prior work, such as executable papers, provenance collection, and workflow languages. We have already seen an example of LLMs supporting the comprehensibility of research artifacts by explaining system-level provenance logs [3]. We will not hesitate to note the tendency of LLMs to provide nonfactual information; that is why we are proposing careful studies of how we can use this technology as a *supplement* to more grounded work. Writing a paper will inherently require a human-in-the-loop. That human can use the LLM to identify inconsistencies in prose and irreproducible workflows rather than trust the LLM to write the prose or workflow on its own. Incorporating human input into our research will also require incorporating elements from UX design research into our tools and solutions.

Following these research threads and incorporating comprehensibility into research artifacts will raise the standards for reproducibility artifacts, and the added value will also promote the use of existing reproducibility tools.

2 The Need for Comprehensibility

We identify three specific reasons why there is a need for more research into reproducibility comprehension.

(1) Scientific reproducibility ultimately is about supporting peer-review and verification of scientific results.
(2) The tools and cultural shifts induced by recent reproducibility research could lead to opaque science.
(3) Researchers report extra overhead hinders reproducibility.

Trust in the scientific method hinges on the ability of scientists to examine and validate each other's claims successfully. 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." [10] It then follows that technical efforts to support reproducibility should support the process by which scientists can *scrutinize scientific claims*. Existing research in open science, FAIR data repositories, virtualization, and code re-execution all support this process. This research contributes towards ideal artifacts, but none reach the top of the artifact taxonomy hierarchy Figure 1. Even with each of these components, a research artifact can be utterly inscrutable without careful, methodical work.

Despite the importance of scientific peer review, the current landscape of reproducibility could be shifting the scientific community toward more opaque science. Cognitive effort is a requirement for comprehension, which is why pedagogical research recommends that learners require an "appropriate level of challenge" [9]. 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" [9]. While a research artifact is not the same type of learning as in an undergraduate classroom, we believe the requirement for cognitive effort is transferable to research artifacts.

Push-button or automatic reproducibility shifts the goal of artifact review from comprehension and scientific scrutiny to achieving the less effortful goal of seeing a research artifact produce the same results described in the paper. Consider MERIT, the fully automatic machine learning reproducibility system created by Wonsil et al.. MERIT works because it lives within the ML framework that trains the models it is reproducing. Its users are *required* to relinquish the burden of understanding what the library is doing so that the library can enforce strict restraints, enabling its efficient and complete provenance collection. In their discussion, they even note that a barrier to achieving the same reproducibility in other systems is letting "users control the training loop." Therefore, if a researcher trains a 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. Hence, despite being "reproducible", that scientific artifact is still opaque and resisted scrutiny. However,

by not providing *any* assistance, we return to the unrealistic expectations that sparked our "reproducibility crisis" [1] where scientists must understand *both* their domain and computational tooling.

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 have demonstrated that researchers feel there is not enough time for reproducible practices, nor do researchers have enough knowledge and training of these practices [1, 4, 11, 12, 14]. We now must perform a balancing act: a reproducibility artifact must not be "too easy" to reproduce or it resists scrutiny, but it cannot be too challenging to generate or reproduce else researchers might not allocate it suitable time and effort.

3 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.

Many cultural barriers to reproducibility are starting to change. While we previously mentioned that researchers view time and effort as a barrier to pursuing proper reproducibility practices [1, 12], more and more journals are promoting artifact badges[3] or requiring open science practices. We even see communities such as Papers with Code[4] that encourage publicizing your code alongside your paper. These changes incentivize researchers to pursue reproducibility practices, and we should participate in promoting 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."[5] In practice, authors are incentivized to create the simplest artifact possible to ensure that the reviewer has the least resistance possible while reproducing the artifact. This approach encourages automatic reproducibility, often via virtualization and build systems. However, as we discussed earlier, this process does not necessarily help validate the claims from their publication, nor will it facilitate other researchers' understanding and building off the work.

Recent significant advancements in natural language processing using large language models (LLMs) indicate that tasks that would otherwise require substantial manual effort are now achievable through automatic means. A significant component of comprehensibility is, of course, 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 are not the end-all solution to our woes, as they have significant weaknesses that we should address, but they have the potential to help significantly.

[3]https://www.cos.io/initiatives/badges
[4]https://portal.paperswithcode.com/
[5]https://www.acm.org/publications/policies/artifact-review-badging - retrieved April 2, 2025

4 Proposed Research Areas

We propose that the following areas of research 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. 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. This concern should be especially prevalent in our proposed research areas since we want to reduce the computational load they experience, which we can only "measure" through user studies.

It will be no surprise to recommend research into LLMs right now, but we nonetheless believe there are areas of interest that LLMs are uniquely situated to assist within our domain. We specifically propose research into using LLMs as supportive tools while scientists conduct their research and write their papers. One of the simplest forms a tool in this realm can take is an AI assistant that provides reproducibility-oriented suggestions as scientists develop their computational experiments. If the scientist were to add a new Python package to their script, the assistant could provide a Copilot-style suggestion to add that package to the project's "requirements.txt" file and then suggest a Dockerfile capable of building an environment suitable for their experiment. The emphasis here is on *suggestions* to ensure that the scientist is always ultimately responsible for changes in the project. How this assistant would interact with the scientist and how it could provide these suggestions would require knowledge from the user studies described earlier. While OpenAI has recently demonstrated that LLMs are *capable of*, but not yet proficient at, replicating machine-learning research just by reading the papers [13], 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.

Language models also have the potential to increase the cohesion of a project's code and paper. Prior work already explains the importance of linking prose from a publication to the code it describes [7] to increase the transparency of the research. LLMs can process that prose and help generate proposed connections between these otherwise disjoint components of a research project, facilitating the creation of an artifact similar to the previously described ELISA [19]. We already see the beginnings of such a tool by Boufford et al. when they demonstrated that users find LLM-generated textual descriptions of workflows helpful in comprehending what they do [3].

Alternatively, LLMs can facilitate the accuracy and consistency of a paper across its iterations. For example, suppose a researcher updates a numeric result in their paper, downgrading their recall from 99% to 98%. If the researcher already had prose that referred to the previous value, it might no longer be correct! "Our model achieves a recall of 98%. This result outperforms the state-of-the-art which achieves 98%." 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 a team of 10 authors is writing this paper, and one makes all the changes, and then another

rolls back only some? LLMs have the potential to perform this type of proofreading automatically. Then, rather than change the prose, the model will bring any inconsistent writing to the researcher's attention - not modify the paper. This methodology ensures that even though a researcher wrote the paper with AI assistance, there is no risk of hallucinations tainting the text.

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 is a previous research topic that has seen less attention recently. This approach to paper writing also brings the code and publication closer but without an LLM by hosting them in the same document. 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 people 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 [6, 8]. However, for those who want the LLM-powered research assistant, a language model could easily augment a tool that collects contextual information to assist a user with research programming, such as Guo and Seltzer's Burrito.

Finally, a research question we think is fundamental: can we build these tools that make reproducibility and comprehensibility more accessible and simultaneously help research programmers learn better reproducibility practices? Survey respondents explain that lack of training and knowledge around good practices affect reproducibility outcomes [1, 12]. While workshops, classes, and self-driven learning can address this issue, those solutions already existed in some form or another during those surveys. Given the lack of time for researchers, 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 understand what they could be missing. Consider a tool that could generate a template of Krafczyk et al.'s proposed *Reproduction Package* and then provide automatic assistance in keeping the components consistent and up-to-date. User experience, pedagogical, and technical questions need answering to realize such a system.

5 Conclusion

We call upon the reproducibility community to investigate new technical solutions to old problems. It is currently the ideal time to shift our attention to previously challenging issues given the emerging new technologies in natural language processing, the trajectory the field is going in, and the cultural shift we have seen in the scientific community. Adding computation to the scientific process has created additional burdens for scientists. While current research attempts to relieve those burdens, they will inadvertently add new barriers to engaging with research artifacts. In addition to relieving computational burdens, we should provide solutions that help scientists engage with research artifacts. We propose bringing our attention to the following areas: finding ways to tighten the connection between publication prose and data analysis, 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 verification and re-use. A commitment to more comprehensible computational artifacts enables more effective artifact review, encourages building off existing work, and increases trust in the scientific method.

References

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

[2] 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.

[3] 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

[4] 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

[5] 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.

[6] 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

[7] 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

[8] 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

[9] 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.

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

[11] 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

[12] 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

[13] 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. [n. d.]. PaperBench: Evaluating AI's Ability to Replicate AI Research. ([n. d.]).

[14] 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

[15] 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

[16] 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

[17] 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

[18] 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

[19] 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