Reproducibility of Code-Understanding Large Language Models: A 217-Paper Analysis of ICSE '23 & SC '24

HACKHPC@ ADMI25

Iyana Jones, Copernic Mensah, Aaliyah Lockett, Holy Agyei

High Performance Computing and Gateways 2025

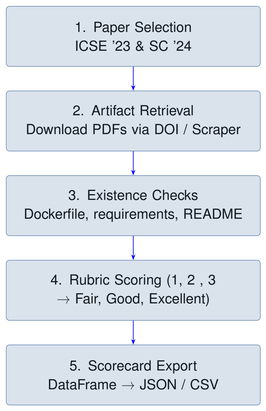
SGX3 Extend. Expand. Exemplify.

An Important Problem

Reproducibility is essential to scientific progress: without it, results cannot be validated, built upon, or trusted [6]. In machine learning and high-performance computing, complex software stacks mean that missing containerization, incomplete documentation, or unclear environment specifications can easily undermine workflow credibility and slow down collaboration.

Methodology

To evaluate the reproducibility of HPC/AI research papers in computer science and data science, a structured scorecard was used to assess the content of technical papers, the quality of the documentation and the reproducibility of the environment. This framework is informed by recent reproducibility benchmarks and HPC/AI research challenges [2].



Select papers, fetch contents, check reproducibility files, score via rubric, then export the scorecard.

We employ a detailed, nine-criterion scorecard in which each paper is assessed on a 0–4 scale (0 = not scored) and annotated with free-text notes[5].

1. Paper Availability: Is the paper open access?
2. Code & Software Availability: Is the source hosted publicly, versioned and documented?
3. Dataset Availability: Are datasets clearly shared?
4. Hardware Requirements: Are specifications unreasonable or ambigous
5. GPU Requirements: Are accelerator needs stated and justified?
6. Documentation Quality: Are README, install guides, and API docs complete?
7. Ease of Setup: Are containers like Docker provided?
8. Reproducibility of Results: Can key metrics be reproduced simply?

These 8 sub scores feed into an overall rating (Excellent, Good, Fair, Poor) alongside a summary note. By combining quantitative scores with contextual notes, our rubric delivers both a transparent metric and the qualitative insight needed to drive improvements across high performance computing and artificial intelligence research.

Results

Analyzing 217 ICSE '23 and SC '24 LLM-for-code papers on a 0–15 reproducibility scale reveals a distinctly polarized landscape: 118 papers (62%) achieved the maximum score of 15, signaling comprehensive artifact support (containers, dependency manifests, documentation, and community channels), while 16 papers (8%) scored 0, lacking any reproducibility artifacts. The remaining 55 papers (29%) scored 12, typically missing only non-essential items like CI/CD badges or discussion forums.

The mean score is 13.60, the median and upper quartile both sit at 15, and the interquartile range spans [12, 15], confirming that best practices are prevalent but not universal. To raise the baseline, we recommend shipping standard Docker templates, explicit dependency manifests, mandatory README setup/validation sections, and active issue-tracking channels.

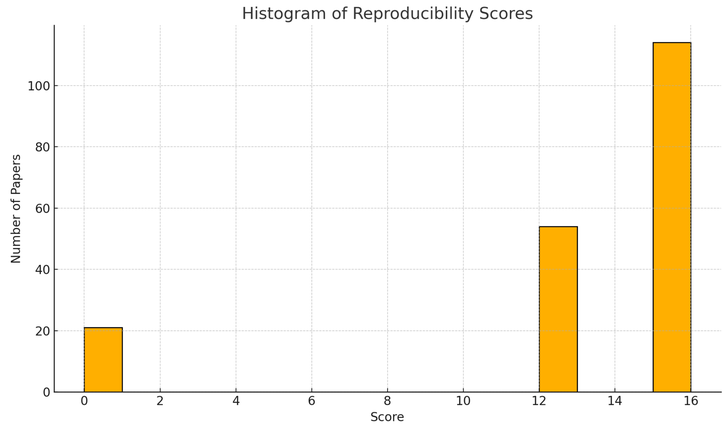


Fig. 1: Distinctive Results

Future efforts should ensure that every software release is accompanied by a fully encapsulated execution environment. Distributing a Docker or Singularity image that includes all dependencies, library versions, and operating-system configurations will prevent discrepancies between development and deployment platforms. Embedding these images in the artifact archive or container registry alongside the code repository guarantees that future users can reproduce experiments without manual environment reconstruction.

Additionally, researchers can provide machine-readable dependency specifications and automated validation workflows. A clearly documented requirements.txt or environment.yml should enumerate each library and its version, while continuous-integration scripts ought to execute the primary analysis pipeline and compare outputs against reference results. Publishing these CI configurations demonstrates that the experiments have been successfully rerun in a standardized context and provides an objective check on reproducibility.

Finally, reproducibility criteria should be formalized within peer-review and post-publication processes. Conferences and journals can adopt evaluation checklists that score container availability and dependency manifest completeness before acceptance. This transmutes reproducibility within high performance computing research and artificial intelligence a reality instead of the known 'crisis'.

Cont.

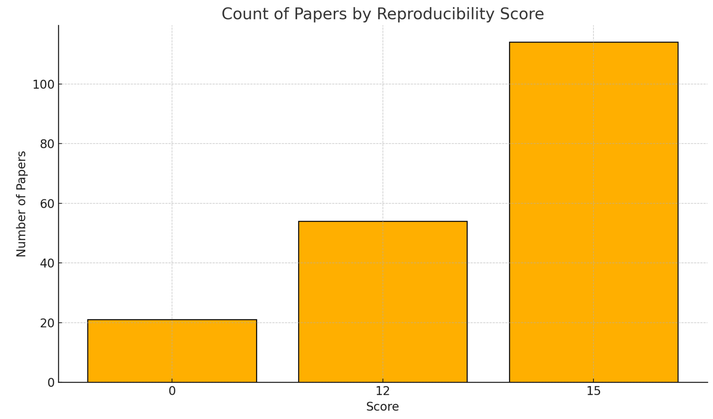


Fig. 2: Distinctive Results

Remarks

Our study demonstrates that while a majority of code understanding large language model papers[4][1] now include essential artifacts, a significant minority still omit container recipes, dependency manifests, or validation pipelines. Elevating reproducibility from an optional bonus to a baseline requirement will require both technical infrastructure (standardized containers, CI workflows) and cultural change [3] (artifact evaluation at submission, public scorecards). By publishing our portal and scorecards, we aim to provide transparency, invite corrections and updates, and track progress over time.

Acknowledgments

We gratefully acknowledge the SGX3 ADMI Hackathon organizers and mentors for their guidance and support throughout this project. None of this would have been made possible without the collaborative spirit and resources of the ADMI community.

References

[1] Proceedings of the 45th International Conference on Software Engineering (ICSE 2023). Melbourne, Australia, 2023.

[2] Benjamin A. Antunes and David R. C. Hill. "Reproducibility, Replicability and Repeatability: A Survey of Reproducible Research with a Focus on High Performance Computing". In: *Computer Science Review* (2024), p. 100655.

[3] Association for Computing Machinery (ACM). *Artifact Review and Badging – Version 1.0*. https://www.acm.org/publications/policies/artifact-review-badging. 2017.

[4] Super Computing Conference. *Reproducibility Initiative – SC23*. https://sc23.supercomputing.org/program/papers/reproducibility-initiative/. SC Conference, 2023.

[5] Google. *Google Codelabs*. https://codelabs.developers.google.com/. 2015.

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