Humanities & Social Sciences Communications

COMMENT

https://doi.org/10.1057/s41599-025-06110-1 OPEN

Check for updates

Lessons from an AI-Sprint: a proposal for measuring human-AI cooperation in research

Leonard Wendering[1]✉, Marc Ratkovic[1], Thomas Gschwend[1], Henning Schoenenberger[2] & Niki Scaplehorn[3]

Generative artificial intelligence is transforming the way scholars draft, revise, and publish. Yet, academia lacks a systematic way to measure these shifts and risks relying on anecdotal evidence in evaluating whether AI elevates or erodes scholarly standards. This Comment draws on a pre-registered, three-day field experiment that addressed this lack of measurement by pairing twenty-two early-career researchers with and without AI tools to improve scholarly manuscripts for journal submission. However, the AI models used in the field experiment are already outdated and outperformed by more powerful reasoning models, situating the results as a snapshot in time. This Comment calls for recurring events with a similar set of evaluation criteria to combine the results in a publicly available dataset. Monitoring the quality of researcher-AI collaboration is necessary if academia wants to keep track of AI's rapid impact on research practice.

[1] University of Mannheim, Mannheim, Germany. [2] Springer Nature Ltd., Heidelberg, Germany. [3] Springer Nature Ltd., Berlin, Germany.
✉email: leonard.wendering@uni-mannheim.de

Introduction

How can the research community evaluate the impact of increasingly powerful AI models on scholarly labor? The rapid evolution of artificial intelligence presents a challenge to the academic community's ability to understand and adapt to AI's influence. In September 2024, the University of Mannheim and Springer Nature jointly organized a three-day-long "AI-Sprint" that paired early-career scholars with AI tools to evaluate their impact on scholarly writing. The participants' goal was to improve their drafts, with the opportunity to submit them to this journal. The AI-Sprint demonstrated how a partnership between a publisher and a university offers possibilities to evaluate emerging AI tools. We suggest repeating and coordinating similar events across institutions to continuously track the impact of AI on academic writing and publishing.

AI tools are already part of the daily life of universities. Their potential and promise lie in the benefits they offer to various aspects of academic research, such as assisting with writing, analysis, and discovery (Meyer et al., 2023; Salvagno et al., 2023). Indeed, AI proves beneficial in tasks relevant to researchers, such as improving writing, developing new ideas, and analyzing data (e.g., Dell'Acqua et al., 2023; Ratkovic et al., 2025). However, recent results demonstrate limitations of AI models. Social scientists tasked with replicating previous work do not perform better when paired with AI than researchers without such support (Brodeur et al., 2025).

While current research provides snapshots, AI's rapid development limits how contributions reflect AI's ability to aid researchers, requiring recurring measurements. In this Comment, we outline practical steps for replicating AI-Sprints across institutions and discuss considerations for effectively monitoring AI's evolving academic impact.

Understanding AI's scholarly impact: the Mannheim AI-Sprint

For a weekend, twenty-two social scientists were randomly assigned to either an AI-assisted group or a control group with no access to AI. The goal of the experiment was to determine whether AI assistance enhances the manuscript quality of early-career researchers. In the aftermath, both groups had the opportunity to submit their papers to a dedicated peer reviewed journal Collection, of which this Comment also forms a part.

The full results from this experiment, which are presented separately (Ratkovic et al., 2025), indicate benefits for the group using AI tools, improving the clarity and coherence of their manuscripts. The ratings by five faculty members show no differences in the remaining dimensions (depth of analysis, literature integration, methodological rigor, and originality). Analysis of open-ended responses in additionally administered self-reflection surveys of the participants using LIWC (Pennebaker et al., 2015) suggests a greater momentum in drafting for the treated group, as evidenced by an increase in vocabulary related to action and temporality. An evaluation of manuscripts by AI (Sakana AI) mirrors human ratings, with non-significant differences in clarity and coherence. Taken together, working with AI improves the manuscripts of early-career researchers.

However, our experiment, as well as similar efforts, only provides one measurement, a snapshot of reality. Because every causal estimate is tethered to the moment it was measured, its explanatory power fades as the world changes (Munger, 2019, 2023). This is particularly relevant for the rapid development of increasingly powerful AI models. For instance, the company behind ChatGPT introduced a new generation of models capable of reasoning a week prior to the AI-Sprint (OpenAI, 2024). By employing greater computing power for answer generation and systematically evaluating diverse reasoning paths, these models outperform conventional large language models across most tasks (OpenAI, 2024). These reasoning models are now widely available and more powerful, having saturated multiple AI benchmarks (e.g., Kavukcuoglu, 2025; OpenAI, 2025). Consequently, the insights from our 2024 experiment are best interpreted as historically situated evidence rather than a timeless verdict.

We propose repeating similar events across sites and pooling the data to arrive at a continuous and updated assessment of how scholars adopt AI. Without such monitoring, researchers, universities, and publishers are forced to rely on anecdotal evidence when discussing the relevance of AI for academia. Although this might be slightly more comfortable, implementing systematic approaches is not much more difficult than avoiding them, and doing so will yield substantial benefits.

The AI-Sprint offers an example of how to replicate the same approach with modest resources. Two seminar rooms, a small pool of AI-usage credits, and basic travel stipends cover the essentials, while light refreshments and two daily exercise breaks sustain focus. Because the format is lightweight, any university can run a similar AI-Sprint.

The Mannheim AI-Sprint comprised the following steps: first, the group of accepted participants received an introduction to two AI tools. Second, randomization assigned participants to either an AI-assisted (treatment) group or a control group without access to AI. Third, both groups' objective was to advance their manuscripts toward journal submission over the weekend. The treatment group was free to use their AI access for sentence polishing, outline expansion, or drafting entire sections. They remained solely responsible for every word produced. After the AI-Sprint, all participants received access to the AI tools and could submit their work to this journal until the end of January 2025. To track how participants experienced the sprint, we surveyed them six times: once before the event, four times during the weekend, and once after the event. Faculty members evaluated the participants' manuscripts in the versions prior to the workshop and at the end of the weekend. Overall, our design captured both the objective shifts in manuscript quality and the participants' subjective experiences of AI-assisted writing.

Hosting a similar event benefits all stakeholders beyond the immediate publishing opportunity for participants. Publishers receive early detection of how AI usage affects the quality of research. Likewise, universities can identify areas in which AI will aid human work and which domains will remain human-centered. Further, they can adapt courses and test formats for undergraduates and keep regulations for ethical human-machine interactions up to date. The benefits for researchers themselves are two-fold. Researchers participating in an AI-Sprint can utilize the latest tools in focused environments to transform a draft into a publishable manuscript. The benefits for researchers extend beyond the participants in AI-Sprints. Everyone can utilize the resulting metrics to identify areas where AI support is effective and areas where a human-in-the-loop approach remains irreplaceable.

Examining the papers accepted to this Collection, our AI-Sprint demonstrates the success of this concept: three of the twenty-two participants have so far successfully passed the peer-review process and had their work published. These publications originate from various subfields in the social sciences. One contribution provides quasi-experimental evidence for the "rally around the flag effect" (Mueller, 1970), examining whether a crisis increases support for governmental actors (Muhammad and Undzėnas, 2025). Using data from the 10th wave of the European Social Survey, they find an increase in public support for the

European Union right after Russia's invasion of Ukraine. Warode (2025) introduces a model to analyze how left- and right-leaning German political candidates associate different meanings with the terms "left" and "right". By comparing the semantic embeddings of answers to open-ended survey questions from political candidates with their self-placements, Warode detects positive connotations associated with a candidate's ideology and negative connotations associated with the opposing ideology. The third contribution by Gelvez (2025) aggregates multiple machine-learning models into a super-learner (Van Der Laan et al., 2007) to predict police and military violence in Colombia and Mexico. He achieves over 92% predictive accuracy, finding that geographic factors are the most influential predictors in Colombia, whereas socioeconomic variables are the most important in Mexico. Together, these publications demonstrate that scholars employing a range of methodological approaches and research interests can effectively utilize the AI-Sprint concept to produce high-quality, peer-reviewed research.

Monitoring AI's impact. If other institutions repeat similar AI-Sprints with a shared evaluation rubric, snapshots from Mannheim, Melbourne, or Mexico City can be merged into one dataset reflecting how scholars harness evolving AI models. The ambition to systematically track AI's evolving impact through replication of experiments finds a precedent in the Metaketa Initiative. This initiative supports coordinated field experiments to overcome issues such as selective reporting or heterogeneous designs (Dunning et al., 2019). To do so, researchers agree to adopt common research questions and harmonize measurements (Dunning et al., 2019), a coordinated effort that hinges on a set of design choices.

To motivate scholars to participate in AI-Sprints, the task for participating scholars must remain authentic, such as advancing to a submission-ready research manuscript. AI can contribute to different aspects of the research process. Therefore, it is not necessary to narrow the focus of a sprint to tackle just one of these and future AI-Sprints may test quite different ways for AI to support article writing. Further, possibilities for publishing work in relevant journals appear to be a viable incentive.

Effective measurement demands flexible monitoring that accommodates disciplinary priorities and varied epistemologies. While this necessitates field-specific adjustments in evaluation, the core underlying question of whether researchers can use AI to produce work ready for publication more efficiently serves as a common baseline. Physicists might ensure precision in complex model descriptions and data presentation, while economists could verify the appropriate use of statistical methods and the interpretation of their results. Scholars in the humanities might focus more on AI's influence on theoretical framing and quality of argumentation. Combining the field-specific requirements while keeping central evaluations of manuscripts consistent, various subfields can contribute data. Together, this would result in insights relevant to specific fields while contributing to a broader understanding of AI's impact.

The monitoring should also reflect how AI affects researchers differently based on their proficiency in their native language, as English dominates academic publishing. Non-native English speakers face more linguistic hurdles (Amano et al., 2023; Clavero, 2010). With AI tools becoming increasingly capable of refining language, this may enable non-native speakers to articulate and refine their core scientific ideas more easily, thereby lessening the cognitive load of writing in a foreign language (Berdejo-Espinola and Amano, 2023). Criteria such as clarity and coherence of the writing can track how AI assistance influences the effective communication of complex research insights for both native and non-native speakers.

Furthermore, the storage of Sprint data should enable anyone to identify long-term patterns, work with the manuscripts without compromising author anonymity, and apply new scoring methods in the future. This could be achieved by organizations jointly running an archive, assigning each sprint output a permanent identifier, and publishing dashboards that show how the results change over time. This strategy aligns with the FAIR Guiding Principles for scientific data management, which emphasize that research outputs should be findable, accessible, interoperable, and reusable (Wilkinson et al., 2016).

Some might worry that proposed AI-Sprints trivialize the process of academic writing, collapsing research into a race, and encouraging authors to optimize for surface polish over sustained thought. These concerns continue to demonstrate the need for these sprints: Will benefits of human-researcher interaction remain limited to superficial dimensions of academic work, even as these models continue to improve? Are there necessary conditions in the collaboration to improve the depth of arguments and the originality of the work? Let us look at it differently: Just as citation indices quantified influence and plagiarism detectors formalized originality checks, a growing record of AI-assisted drafts can anchor debates on AI in data. We encourage other scholars and institutions to join the discussion about design choices and possible limitations of continuously monitoring the AI-researcher duet.

Equally important to measuring AI's impact is the transparent disclosure of its use in the research process. As AI tools become more deeply embedded in scholarly workflows—from idea generation to manuscript polishing—the academic community must establish clear and consistent standards for reporting AI involvement. Without such transparency, the integrity of peer review and the attribution of intellectual labor may be compromised. Researchers, reviewers, and readers alike benefit from knowing whether and how AI contributed to a given work. This is especially relevant as AI tools increasingly influence not just language but also structure, argumentation, and even data interpretation.

Calls for unified disclosure standards, such as those championed by the STM Association, highlight the urgency of this issue (STM Association Task and Finish Group, 2025). Aligning AI-Sprint protocols with these emerging frameworks would ensure that manuscripts reflect not only the quality of human-AI collaboration but also the ethical standards of academic publishing. Transparent labeling of AI contributions—whether in acknowledgments, metadata, or dedicated disclosure sections—can help distinguish between human insight and machine assistance. This clarity is essential for future research on the evolving human-AI dynamic and for maintaining trust in scholarly communication.

Conclusion

The rapid and unceasing evolution of artificial intelligence presents a challenge to the academic community's ability to understand and adapt to its influence on scholarly work. This Comment argues that isolated or infrequent assessments are insufficient. Instead, continuously updated monitoring, created through a network of recurring, harmonized "AI-Sprints," would benefit scholars, publishers, and other academic institutions.

Realizing such an ambitious yet necessary initiative requires a collaborative effort. To catalyze this effort, the academic community should consider several practical next steps: other institutions could pilot their AI-Sprints, adapting to their local contexts and disciplinary needs. Furthermore, monitoring multiple AI-Sprints will empower academia to not only react to technological advancements but also shape its future relationship

with AI, ensuring that these tools augment scholarly inquiry ethically. By institutionalizing AI-Sprints, academia can adapt its understanding of and decide upon what counts as authorship, originality, and rigor in an era of synthetic eloquence.

Data availability

No datasets were generated or analysed during the current study.

Received: 1 September 2025; Accepted: 15 October 2025;
Published online: 11 November 2025

References

Amano T, Ramírez-Castañeda V, Berdejo-Espinola V, Borokini I, Chowdhury S, Golivets M, González-Trujillo JD, Montaño-Centellas F, Paudel K, White RL, Veríssimo D (2023) The manifold costs of being a non-native English speaker in science. PLOS Biol 21(7):e3002184. https://doi.org/10.1371/journal.pbio.3002184

Berdejo-Espinola V, Amano T (2023) AI tools can improve equity in science. Science 379(6636):991. https://doi.org/10.1126/science.adg9714

Brodeur A, Valenta D, Marcoci A, Aparicio JP, Mikola D, Barbarioli B, Alexander R, Deer L, Stafford T, Vilhuber L, Bensch G (2025) Comparing human-only, AI-assisted, and AI-led teams on assessing research reproducibility in quantitative social science (I4R Discussion Paper Series 195). The Institute for Replication (I4R). https://ideas.repec.org/p/zbw/i4rdps/195.html. Accessed 25 Aug 2025

Clavero M (2010) Awkward wording. Rephrase": Linguistic injustice in ecological journals. Trends Ecol Evol 25(10):552–553. https://doi.org/10.1016/j.tree.2010.07.001

Dell'Acqua F, McFowland E, Mollick ER, Lifshitz-Assaf H, Kellogg K, Rajendran S, Krayer L, Candelon F, Lakhani KR (2023) Navigating the jagged technological frontier: field experimental evidence of the effects of AI on knowledge worker productivity and quality. SSRN Electron J https://doi.org/10.2139/ssrn.4573321

Dunning T, Grossman G, Humphreys M, Hyde SD, McIntosh C, Nellis G (2019) The Metaketa Initiative. In: Dunning T, Grossman G, Humphreys M, Hyde SD, McIntosh C, Nellis G (eds) Information, accountability, and cumulative learning, 1st edn. Cambridge University Press, Cambridge, pp 16–49. https://doi.org/10.1017/9781108381390.003

Gelvez JD (2025) Predicting police and military violence: evidence from Colombia and Mexico using machine learning models. Human Soc Sci Commun 12(1):765. https://doi.org/10.1057/s41599-025-04967-w

Kavukcuoglu K (2025) Gemini 2.5: Our most intelligent AI model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#gemini-2-5-thinking. Accessed 25 Aug 2025

Meyer JG, Urbanowicz RJ, Martin PCN, O'Connor K, Li R, Peng PC, Bright TJ, Tatonetti N, Won KJ, Gonzalez-Hernandez G, Moore JH (2023) ChatGPT and large language models in academia: opportunities and challenges. BioData Min. 16:20. https://doi.org/10.1186/s13040-023-00339-9

Mueller JE (1970) Presidential popularity from Truman to Johnson. Am Polit Sci Rev 64(1):18–34. https://doi.org/10.2307/1955610

Muhammad M, Undzėnas D (2025) Entangled fates—the rally effect around Europe due to Russia's war of aggression against Ukraine. Human Soc Sci Commun 12(1):915. https://doi.org/10.1057/s41599-025-05138-7

Munger K (2019) The limited value of non-replicable field experiments in contexts with low temporal validity. Soc Media + Soc 5(3):2056305119859294. https://doi.org/10.1177/2056305119859294

Munger K (2023) Temporal validity as meta-science. Res Polit 10(3):20531680231187271. https://doi.org/10.1177/20531680231187271

OpenAI (2024) Introducing OpenAI o1-preview. https://openai.com/index/introducing-openai-o1-preview/. Accessed 25 Aug 2025

OpenAI (2025) Introducing OpenAI o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/. Accessed 25 Aug 2025

Pennebaker JW, Boyd RL, Jordan K, Blackburn K (2015) The development and psychometric properties of LIWC2015. University of Texas at Austin, Austin, TX

Ratkovic M, Gschwend T, Wendering L, Bauer K, Kurella AS, Rittman O, Sauter M, Schwitter N (2025) Harnessing GPT for enhanced academic writing: evidence from a field experiment with early-career researchers in the social sciences. Res Square. https://doi.org/10.21203/rs.3.rs-6937665/v1

Salvagno M, Taccone FS, Gerli AG (2023) Can artificial intelligence help for scientific writing? Crit Care 27(1):75. https://doi.org/10.1186/s13054-023-04380-2

STM Association Task & Finish Group (2025) Recommendations for a classification of AI use in academic manuscript preparation. https://s3.eu-west-2.amazonaws.com/stm.offloadmedia/wp-content/uploads/2025/04/07023544/STM_AI_Classification_Report_April2025.pdf. Accessed 25 Aug 2025

Van Der Laan MJ, Polley EC, Hubbard AE (2007) Super learner. Stat Appl Genet Mol Biol 6(1). https://doi.org/10.2202/1544-6115.1309

Warode L (2025) Mapping left-right associations: a framework using open-ended survey responses and political positions. Human Soc Sci Commun 12(1):1318. https://doi.org/10.1057/s41599-025-05679-x

Wilkinson MD, Dumontier M, Aalbersberg IJ et al. (2016) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3:160018. https://doi.org/10.1038/sdata.2016.18

Acknowledgements

Funding was provided by Springer Nature; the Chair of Political Science, Quantitative Methods in the Social Sciences, University of Mannheim; and the Chair of Political Science, Social Data Science, University of Mannheim.

Author contributions

LW, MR, and TG conceptualized the comment, all authors wrote and reviewed the manuscript.

Competing interests

Henning Schoenenberger and Niki Scaplehorn are employees of Springer Nature. They have no participation in any editorial processes. All other authors declare no competing interests.

Ethical approval

Not applicable for this Comment, as it reports no new research involving human participants or personal data. The field experiment referred to in the cited preprint was approved by the University of Mannheim Ethics Committee in accordance with the Statute of the Ethics Committee of the University of Mannheim (15 Dec 2016; amended 26 May 2021).

Informed consent

Consent not applicable to this Comment. For the field experiment described in the cited preprint, all participants provided informed consent.

Additional information

Correspondence and requests for materials should be addressed to Leonard Wendering.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025