Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Harnessing GPT for Enhanced Academic Writing: Evidence from a Field Experiment with Early-Career Researchers in the Social Sciences

Marc Ratkovic

marc.ratkovic@uni-mannheim.de

University of Mannheim

Thomas Gschwend

University of Mannheim

Leonard Wendering

University of Mannheim

Kevin Bauer

University of Mannheim and Goethe University Frankfurt

Anna-Sophia Kurella

University of Mannheim

Oliver Rittman

University of Mannheim

Melanie Sauter

University of Mannheim

Nicole Schwitter

University of Mannheim and University of Warwick

Research Article

Keywords: Generative AI, Large Language Models, Randomized Controlled Trials, Pre-Registration

Posted Date: June 20th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6937665/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Harnessing GPT for Enhanced Academic Writing: Evidence from a Field Experiment with Early-Career Researchers in the Social Sciences

Marc Ratkovic[1], Thomas Gschwend[1], Leonard Wendering[1], Kevin Bauer[1,2], Anna-Sophia Kurella[1], Oliver Rittman[1], Melanie Sauter[1], Nicole Schwitter[1, 3]
[1]University of Mannheim and [2]Goethe University Frankfurt [3] University of Warwick

Abstract

Large Language Models (LLMs) bring both risk and opportunity to scholarly communication, yet their effects on expert-level research remain untested. We conducted the first randomized controlled trial of GPT-4 in expert academic writing, implemented as a three-day hackathon (n=22 early-career social scientists) in collaboration with Springer Nature. Participants were randomized to unrestricted GPT access or a no-AI control. Manuscripts were evaluated pre- and post-intervention by expert faculty and postdoctoral researchers on six dimensions: clarity, coherence, originality, methodological rigor, depth of analysis, and literature relevance. AI assistance produced significant improvements in clarity (Δ=+0.49, p=0.009) and coherence (Δ=+0.43, p=0.036) without affecting other dimensions. Principal-components analysis confirmed AI's selective enhancement of organizational features. SakanaAI, a GPT-based reviewer, aligned broadly with human assessments. A linguistic analysis of participants' reflective journals revealed increased process-oriented language and reduced temporal references, suggesting a reallocation of cognitive resources. Although concerns about overreliance and academic integrity persist, our findings imply that delegating routine structuring tasks to AI can free scholars to focus on generating novel insights.

We thank Henning Schoenenberger, Karen Ehrhardt-Dreier, Niki Scaplehorn, and Vivien Bender from Springer Nature for their collaboration and support. We also thank all participants in the event. A version of this manuscript was presented at the University Future Festival, TU Graz, 13.5.2025.

Introduction

How will generative artificial intelligence reshape scientific practice? In recent years, Large Language Models (LLMs) like ChatGPT have transformed how we generate text, touching every component of the research pipeline. This Promethean technology offers significant opportunities for clarifying and structuring work (Salvagno et al., 2023; Meyer et al., 2023)-raising hopes for democratizing scholarly communication, especially among early-career researchers and non-native English speakers. Yet over-reliance can reduce critical thinking for everyone from undergraduates (Shirkey, 2025) to peer reviewers (Liang et al., 2024), and may even facilitate research fraud (Májovský et al., 2023). In assessing the use of AI, previous experimental work has found productivity gains through contrived situations or harmonized tasks (e.g., Dell'Acqua et al., 2023; Noy & Zhang, 2023). We know surprisingly little about how they affect real academic work. We therefore examine how AI assists early-career scholars working on their own projects, with the aim of identifying the specific dimensions along which scholarly writing and assessment may benefit or be harmed from incorporating AI.

Our contribution to these issues centers on a three-day research hackathon at the University of Mannheim (Sept. 2024), run in conjunction with Springer Nature, to assess the impact of AI on academic research. By recruiting only PhD and social scientists, our design zeroes in on experts creating expert knowledge rather than students drafting simple assignments, or routine writing or coding tasks (e.g. Cui et al., 2024; Pankiewicz & Baker, 2023). Our participants were all fluent in English, but only two were native speakers, giving us insight into the potential of AI to contribute to scholarship in non-US settings. Following a pre-registered design,[1] twenty-two participants were randomized to a treatment condition with unrestricted GPT access or a no-AI control. They drafted manuscripts for submission to a dedicated special issue in *Humanities and Social Science Communications (HSSC)* and completed open-ended, self-reflective writing prompts to track shifts in metacognition. Faculty and established postdoctoral evaluators (the last five authors) then scored pre- and post-event manuscripts on clarity, coherence,originality, methodological rigor, depth of analysis, and literature

[1]The pre-registration plan and all supplemental materials are available at https://osf.io/f4epg. Ethical approval was secured at the University of Mannheim. All participants were aware of the experiment and consented to the use of their data.

relevance. These expert assessments were subsequently compared to scores generated by SakanaAI, a GPT-based paper evaluation tool.

We make three contributions. First, our field experiment moves beyond aggregate productivity measures to identify precisely which facets of writing GPT improves. Treatment papers score significantly higher on clarity and coherence, with no change in originality, methodological rigor, or literature relevance. A principal-components analysis reveals three structural dimensions: depth and rigor, clarity and coherence, and literature relevance. Importantly, our intervention shifts only the second cluster.

Second, we probe how GPT alters writers' self-reflection. Using LIWC (Pennebaker, et al., 2015; Benoit, et al. 2018) to analyze participants' writing reflections (see, e.g., Kumar et al., 2024), we find that AI users employ more process-oriented language ("work," "results," "improve") and fewer temporal markers ("now," "time," "today"). This suggests that GPT not only enhances surface polish, but by speeding up the generation of text can also shift cognitive engagement with the writing process.

Third, building on Liang et al. (2024), we evaluate AI as a peer-review proxy. While SakanaAI's scores align broadly with human judgments, we observe systematic discrepancies in how AI weights substantive rigor versus stylistic polish. Nonetheless, the overall agreement at baseline indicates that LLMs could serve as an effective first-pass screening tool.

Our sample size of n=22 raises power concerns; we address these through blocking, permutation-adjusted inference, and strong incentive structures (Springer Nature special issue access and manuscript awards) to maximize compliance and reduce implementation variance. We ultimately conclude on a fundamentally humanistic note: used effectively, shifting some of the burden of structuring and clarifying academic argument to AI frees up space for scholars to pursue original insight derived from lived experience.

Materials and Methods

Participants and Design. We recruited 22 doctoral and postdoctoral researchers from multiple social disciplines: political science (14), sociology (2), economics (2), psychology (2), finance (1), and policy analysis (1). All participants were fluent in English, but native languages varied.[2] Recruitment occurred via professional listservs, social media channels affiliated with the University of Mannheim, and direct outreach to universities through the region. All participants provided informed consent, ethical approval was granted by the University of Mannheim Institutional Review Board (IRB).

Participants submitted preliminary drafts of ongoing manuscripts as baseline writing samples, allowing us to stratify randomization on two key criteria: (1) prior experience using generative AI tools (GPT or other LLMs) to write or edit academic drafts (yes/no), and (2) manuscript completeness (complete draft available or not). Among participants who had prior AI experience (n=15), further stratification accounted for whether they had previously published solo-authored academic work, and their disciplinary affiliation (political science or not). Additional control variables—including age, gender, specific discipline, language proficiency, number of prior academic presentations, and publication history—were recorded prior to the event. Further details are in our pre-registration plan.

Participants were divided into two equal groups (AI-assisted and Control) following stratification. To ensure that participants and organizers were blind to treatment condition through the preliminary instructional session, final assignment to treatment condition was done after the conclusion of instruction.

Intervention and Writing Task. The writing task involved participants working on their submitted project, either a dissertation chapter or working paper. Participants in the AI-assisted group were provided access to the commercial version of ChatGPT and allowed to use this or any AI tool as they wished, interacting with the model iteratively during drafting and editing to generate text, clarify ideas, and refine written content. Participants were specifically coached on effective prompt engineering (e.g., asking for outlines, paragraph refinement, grammar

[2] Native languages were: German (5), Spanish (3), English (2), Chinese (2), Russian (2), Turkish (2), then one each for Lithuanian, Italian, Urdu, Albanian, Arabic, and Vietnamese.

improvement), but were solely responsible for ensuring the content met scholarly standards and the assignment guidelines.

Control group participants completed identical tasks independently, without any AI assistance. Both groups worked under identical time constraints (two-hour sessions) and followed similar daily schedules, ensuring comparability in working conditions.

The experiment employed a two-stage design. Prior to the event, participants submitted a baseline manuscript draft, forming the basis for initial evaluation. After random assignment and instruction on Day 1, participants proceeded under treatment conditions for subsequent writing sessions. Final outputs were collected at the end of the event, with an optional eight-week revision window provided to both groups. Notably, Springer Nature's involvement—specifically creating a special issue of Humanities and Social Sciences Communications (HSSC)—significantly incentivized participation and compliance by enabling participants from both groups equal access to publication opportunities. The delayed submission option effectively removed short-term competitive advantages associated with GPT use, fostering equitable conditions and strengthening internal validity. To encourage participation and focus, four €200 awards were disbursed: one for best and one for most-improved manuscript in the treatment condition and in the control condition.

Evaluation Procedures: Faculty Expert Assessments. Five expert evaluators (faculty and experienced postdoctoral researchers) independently assessed each participant's writing across six key dimensions aligned with HSSC criteria: Clarity, Coherence, Depth of Analysis, Methodological Rigor, Originality, and Relevance of the Literature Review. Evaluations employed a standardized 5-point Likert scale (1 = very poor, 5 = excellent). Evaluators were blinded to participants' group assignments but were aware of whether the submission was pre-treatment (baseline) or post-treatment. Initial baseline evaluations occurred independently; subsequently, evaluators convened briefly to coordinate interpretations of the scoring rubric. We intentionally refrained from enforcing strict consensus, allowing varied academic perspectives to inform evaluations. The primary outcome measure was the change in evaluation scores (post-treatment minus pre-treatment) across dimensions.

Evaluation Procedures: AI-Based Evaluations. In parallel, manuscripts were evaluated using an AI-based reviewer via SakanaAI, a GPT-based assessment tool initially designed for machine learning conferences (e.g., NeurIPS). We adapted the SakanaAI prompts explicitly to HSSC evaluation standards, ensuring direct comparability between AI and human assessments.[3] The AI evaluations utilized identical dimension criteria and scoring scales provided to faculty reviewers. Evaluations were standardized, employing the same AI model parameters and prompt structure for every essay, with no adaptive learning between manuscripts. Like human reviewers, the AI evaluator was blinded to treatment conditions, thus providing independent assessments suitable for comparison with faculty-generated scores.

Analysis of Self-Reflective Writing. Participants provided structured reflective writing responses to open-ended prompts at scheduled intervals (after each afternoon and evening session) and during a follow-up six weeks post-event.[4] Prompt content encouraged reflective metacognition about their writing processes without explicitly referencing treatment conditions. Responses were analyzed quantitatively using Linguistic Inquiry and Word Count (LIWC, Pennebaker, et al., 2015; Benoit, et al. 2018), a validated dictionary-based text analysis tool assessing linguistic markers of psychological processes. Differences in LIWC scores between treatment and control groups were evaluated statistically to identify shifts in reflective thought processes attributable to GPT assistance.

Hypotheses. We pre-registered the following five hypotheses at the Open Science Framework. The first two are our primary hypotheses:

---

[3] The Python files for executing the analysis are included with the replication materials. They are minimally edited versions of those on SakanaAi's github https://github.com/SakanaAI/AI-Scientist; last accessed 20 June, 2025.

[4] Examples include "What are the strengths and weaknesses of your manuscript so far? What sections, or arguments, are strongest? Which need the most work?" and "What problems did you encounter in your writing today? How did you troubleshoot them? Did you have any success in dealing with them?"

	Originality	Clarity	Coherence	Depth of Analysis	Literature Review	Methodological Rigor
Estimate	0.11	0.49	0.43	0.07	0.28	0.27
t, unclustered	0.58	2.29	1.84	0.13	0.9	1.45
t, clustered	0.27	1.99	1.77	0.25	1.11	1.26
t, stratified and clustered	0.61	2.61	2.29	0.37	1.49	1.46
Westfall-Young adjusted p	0.87	0.01	0.04	0.94	0.37	0.38

Table 1: Main Effect Estimates. The rows contain the estimates on a Likert scale, then t-statistics with no clustering or stratification; clustering but without stratification; and both stratified and clustered. These rows illustrate the impact of the design on our inference. Westfall-Young adjusted p-values appear in the bottom row. Our assessors uncovered a statistically significant difference in clarity and coherence, but not in any other dimension.

H1: Participants who use GPT will produce better-structured academic papers compared to those who do not, as measured by faculty-assessed dimensions of academic quality.
H2: We predict a null effect on measures of innovation, again using the faculty-assessed dimensions.

The next two are secondary hypotheses, as they were interesting to us but not central to our claims:

H3: We expect a larger effect in productivity for papers in their early, rather than later stages. This will be projects that do not have a complete draft.
H4: GPT usage will result in higher productivity, measured by the time spent on writing and the completeness of the final submission.
H5: GPT-trained models will closely mimic faculty assessments in evaluating academic writing.

We end up not reporting results for H3 and H4, as we found, after consulting with the evaluators, that word count was a poor proxy for productivity. Some works came in as quite long documents with sketches of ideas, and were trimmed down, and a few documents came in as just a page or two or a proposal.

Statistical Analyses. We utilized mixed-effects linear regression models to capture the nested structure of our data (multiple reviewers, multiple quality dimensions, stratified participant clusters). Given our modest sample size (n=22), we adopted a

	Originality	Clarity	Coherence	Depth of Analysis	Literature Review	Methodological Rigor	Percent Explained
Dimension 1	-0.25	-0.33	-0.41	0.71	-0.09	0.37	0.56
Dimension 2	-0.83	0.47	0.29	0.1	-0.02	-0.02	0.27
Dimension 3	-0.19	-0.13	-0.27	-0.03	0.89	-0.27	0.1

Table 2: Principal Component Analysis. The first three principal components of the null distribution among faculty assessors explains 93% of the total variance. The first dimension loads on the quality of the analysis and method; the second on clarity and coherence; the third on the literature review. Our intervention manipulated the clarity and coherence dimension.

pre-registered family wise error rate of 0.1, but as we are testing across six-dimensions, we adjust for multiple comparisons using a Westfall-Young adjustment, (Westfall and Young, 1993).

Results

Overall Effect. The results indicate that GPT primarily improves structural rather than substantive aspects of scholarly writing, consistent with our pre-regiistered expectations. As shown in Table 1, GPT assistance yielded statistically significant improvements in two dimensions: clarity (estimated effect = 0.486, adjusted p = 0.009) and coherence (estimated effect = 0.427, adjusted p = 0.036). The estimated effects for clarity and coherence correspond to moderate practical gains on a five-point Likert scale, suggesting that GPT assistance notably improved readability and logical flow. In contrast, no significant effects emerged for deeper intellectual

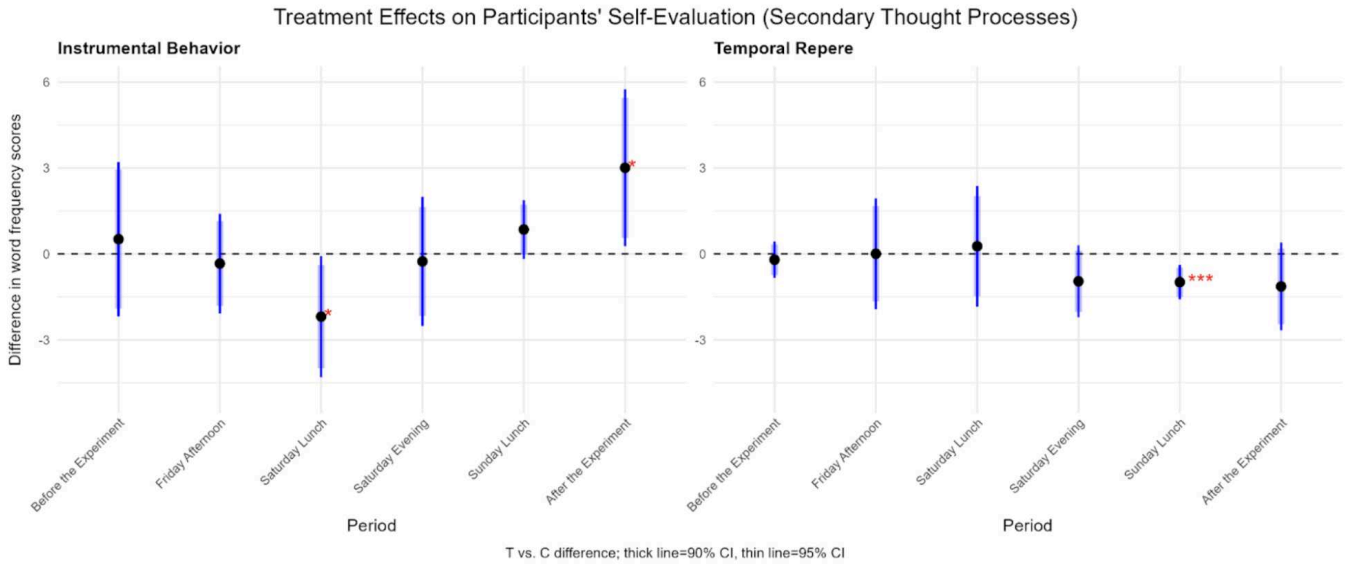


Figure 1: Self-Assessment Dimensions Uncovered by LIWC. The figure presents the results of LIWC applied to participant self-reflections before and through the event. Multiple-correction 95% intervals are presented and stars for *p<0.1, .05, 0.01*. We notice an uptick in instrumental words, like "make" and "put", where the treatment group is moving away from temporal words ("when", "now", "then").

dimensions such as originality (estimate = 0.114, p = 0.875), analytical depth (estimate = 0.068, p = 0.935), methodological rigor (estimate = 0.273, p = 0.384), or relevance of literature review (estimate = 0.277, p = 0.368).

Correlation Among Dimensions (Human Assessments). To contextualize these findings, we analyzed correlations among evaluative dimensions using principal component analysis (PCA) performed on the null distribution from our permutation tests (see Table 2). The PCA analysis indicated clear structure in the evaluation criteria. Specifically, the first principal component, accounting for approximately 56% of variance, loaded strongly on deeper evaluative criteria—primarily Depth of Analysis and Methodological Rigor—the dimensions not significantly affected by AI assistance. The second principal component, capturing about 27% of variance, aligned closely with clarity and coherence, precisely the two dimensions in which GPT yielded significant improvements. This suggests a coherent dimension of surface-level textual quality distinct from analytical depth. Literature review quality emerged independently in a third principal component (10% of the variance), underscoring a distinct but lower-weighted construct within our human assessors' judgments.

Thus, GPT assistance specifically enhances the clarity-coherence dimension of scholarly communication, a result consistent with the notion that generative AI primarily improves structural aspects rather than deeper intellectual innovation or methodological insight.

Effect on Participant Self-Reflections. We further examined whether GPT usage affected participants' reflective cognitive processes during writing, employing linguistic analysis (LIWC) on open-ended reflective prompts given after each writing session. As depicted in Figure 1, GPT assistance significantly influenced the frequency of participants' linguistic expressions related to instrumental behavior and temporal references (all confidence intervals at theadjusted for multiple comparisons). Specifically, at the final reflective period (Sunday lunch), treated participants showed a marked and statistically significant ($p < 0.001$) increase in instrumental language (e.g., action-oriented verbs such as "make," "find," "work") and a significant decrease ($p < 0.001$) in temporal references (e.g., "when," "now," "then"). These descriptive results, as they are not pre-registered, do suggest GPT led participants toward a more practically-focused and less narrative-oriented reflective stance by the end of the experiment.

Comparing AI and Human Assessors. Lastly, we systematically compared the AI-based evaluator (SakanaAI) to human faculty assessors. Figure 2 illustrates the differences in estimated treatment effects on raw evaluation scores between AI and human reviewers for each dimension. Although AI scores generally aligned with human ratings, subtle systematic discrepancies emerged. In particular, clarity and coherence showed the largest differences between human and AI assessments, albeit not statistically significant due to wide, unadjusted confidence intervals. This suggests a trend whereby human evaluators assess the changes in coherence and clarity differently than the AI evaluator, but originality and innovation similarly.

We find a similar pattern when looking at the difference between human- and AI-evaluations for each participant on the pre-event, baseline manuscripts. The lower panel of Figure 2 presents the t-statistics comparing human and AI evaluators across all twenty-two participants. Aggregate differences between human and AI evaluations on pre-event submissions appear to be similar, with no

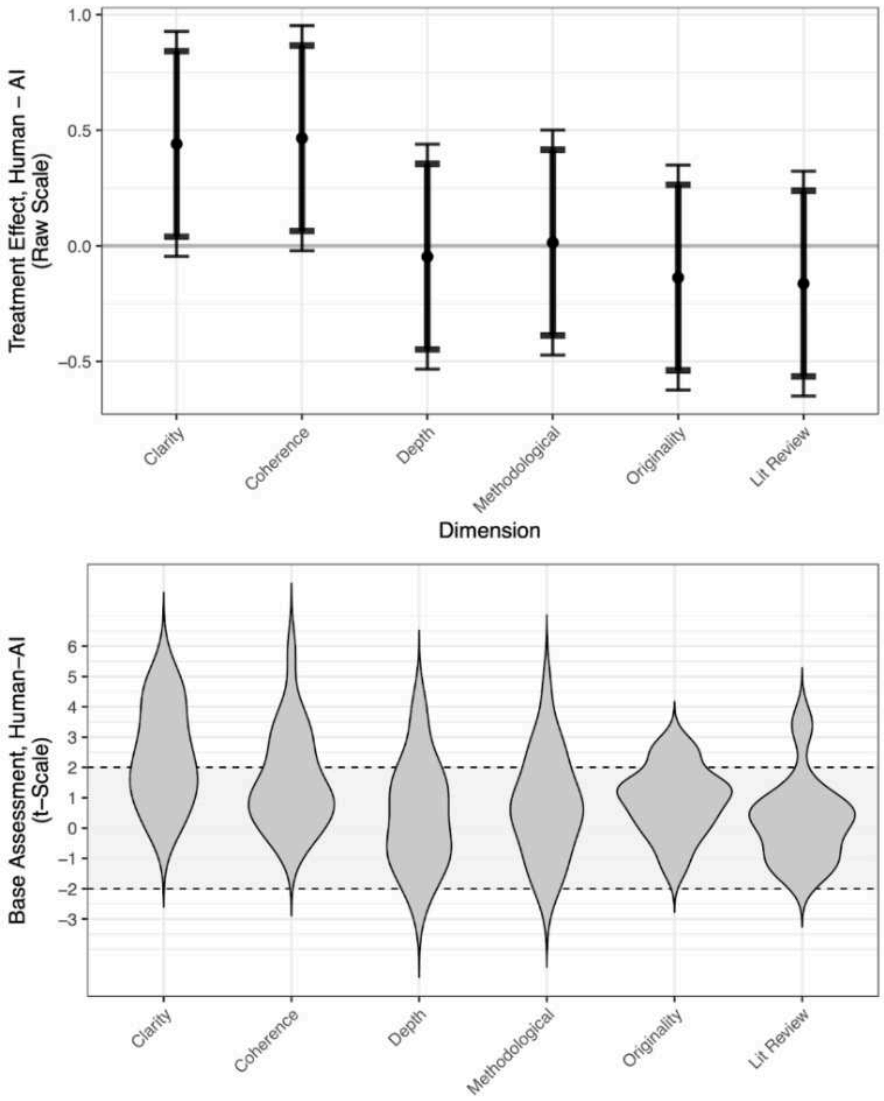


Figure 2: Comparing Human and AI Evaluations. The top figure compares human and AI estimated treatment effects, by dimensions. Thick/thin confidence intervals are 90%/95% and not adjusted for multiple comparisons. Results humans and AI evaluations are reasonably similar, though humans are more likely to provide higher clarity and coherence scores to GPT-generated content than GPT. Below is a violin plot of the participant-level t-statistics for the difference between human and AI assessments at baseline, prior to treatment. The results show broad agreement between human and AI, though again humans are a bit more likely to give higher scores than AI on clarity and coherence, even on content generated before the event.

systematic average differences, a result further substantiated by Liang, et al. (2024,

2024b). Especially for early work, AI offers a useful means of evaluation and feedback.

Discussion

Our randomized field experiment informs how generative AI can influence the production and evaluations of scholarly research. Careful, expert human evaluation allows us to assess the dimensions of impact: AI improves clarity and coherence—dimensions primarily related to textual organization and rhetorical precision—while showing negligible effects on the deeper intellectual dimensions such as analytical rigor and originality. Principal component analysis confirmed the distinct clustering of these dimensions, allowing for the precise nature of AI's influence (Dell'Acqua et al., 2023; Noy & Zhang, 2023). Substantiating earlier discussion and speculation (e.g., Salvagno et al., 2023), we find that AI can both bring benefit and avoid harm in scholarly work.

We find descriptive evidence that AI changes how we think about and interact with our writing. Participants using GPT were significantly more likely to employ instrumental and process-oriented language—words emphasizing practical actions and outcomes—and less likely to report thinking about time. These linguistic shifts imply a subtle cognitive reallocation: by alleviating the demands associated with articulating textual structure, generative AI may enable writers to redirect attention toward more substantive aspects of their research (Hou et al., 2024; Bauer et al., 2024).

Comparing evaluations between human experts suggest a distinct role for AI in the editing and review process. At a surface and aggregate level, we find agreement between the two sets of reviews, echoing Liang et al. (2024, 2024b). The baseline alignment between AI and human assessments suggest that AI can serve useful for initial manuscript triage, thereby reserving scarce expert time for deeper substantive evaluation.

Our smaller sample size allowed us to leverage our expertise and maximize internal validity, but also resulted in limiting the scope of our findings. Our participants were primarily non-native English speakers, from a set of cognate social sciences, and largely from central Europe. In creating an intensive writing experience, and

trying to maximize output, we observed outcomes in a research sprint setting that is not the norm of scholarship in our fields. Our measurements, dimensions specified by our partner journal as well as the dictionary-based LIWC, are coarse and subjective. Future research can address these constraints through similar studies with different cohorts, in different areas of the world, and with finer cognitive and even physiological measurements.

Conclusion

Somewhat ironically, centering our attention on AI quickly forced us to reconsider and reemphasize the human element of our interactions, evaluations, and research. We have shown that generative AI can enhance textual organization, improving clarity and coherence, without diminishing analytical rigor or originality. A linguistic analysis revealed that AI assistance shifts reflective writing away from temporal markers toward process-oriented language. The broad but imperfect alignment between human expert and AI evaluations suggests that AI can serve as an informative screening tool, helping to filter out lower-quality manuscripts and freeing researchers' time for deeper, more nuanced evaluations. Looking forward, AI relieves scholars of routine reading and writing tasks, creating a newly-available reservoir of cognitive bandwidth. The question for us as scholars is what we do with that free time. Our answer to that question, at a personal, institutional, and disciplinary level, is exactly the answer to what it means to be a scholar in the age of AI. As always, the choice is ours.

Works Cited

Bauer, Kevin and Jussupow, Ekaterina and Heigl, Rebecca and Vogt, Benjamin and Hinz, Oliver. (November 30, 2024). All just in your head? Unraveling the side effects of generative AI disclosure in creative tasks. Available at http://dx.doi.org/10.2139/ssrn.4782554

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). *quanteda: An R package for the quantitative analysis of textual data*. *Journal of Open Source Software*, 3(30), 774. https://doi.org/10.21105/joss.00774

Cui, Z., Demirer, M., Jaffe, S., Musolff, L., Peng, S. & Salz, T. (2024). The Effects of Generative AI on High Skilled Work: Evidence from Three Field Experiments with Software Developers. doi: 10.2139/ssrn.4945566

Dell'Acqua, F., McFowland, E., Mollick, E., et al. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Harvard Business School.

Hou, Jinghui and Wang, Lei and Wang, Gang and Wang, Harry and Yang, Shuai, The Double-Edged Roles of Generative AI in the Creative Process: Experiments on Design Work (November 30, 2024). Available at SSRN: https://ssrn.com/abstract=4739471

Kumar, H., Xiao, R., Lawson, B., Musabirov, I., Shi, J., Wang, X., Luo, H., Williams, J. J., Rafferty, A. N., Stamper, J., & Liut, M. (2024). Supporting self-reflection at scale with large language models: Insights from randomized field experiments in classrooms. In Proceedings of the Eleventh ACM Conference on Learning @ Scale (pp. 86–97). Association for Computing Machinery. https://doi.org/10.1145/3657604.3662042

Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., McFarland, D. A., & Zou, J. Y. (2024). Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews. *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, PMLR 235. arXiv preprint arXiv:2403.07183. doi: 10.48550/arXiv.2403.07183

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., Vodrahalalli, K., He, S., Smith, D. S., Yin, Y., McFarland, D. A., & Zou, J. (2024b). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. NEJM AI, 1(8).

Májovský, M., Černý, M., Kasal, M., Komarc, M., & Netuka, D. (2023). Artificial Intelligence Can Generate Fraudulent but Authentic-Looking Scientific Medical Articles: Pandora's Box Has Been Opened. *Journal of Medical Internet Research*, 25. doi: 10.2196/46924

Meyer, J.G., Urbanowicz, R.J., Martin, P.C.N. et al. (2023) ChatGPT and large language models in academia: opportunities and challenges. BioData Mining 16, 2. https://doi.org/10.1186/s13040-023-00339-9

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Science, 381(6654), 187–192. https://doi.org/10.1126/science.adh2586

Pankiewicz, M., & Baker, R. S. (2023). Large language models (GPT) for automating feedback on programming assignments. In J. L. Shih et al. (Eds.), Proceedings of the 31st International Conference on Computers in Education. Asia-Pacific Society for Computers in Education.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. University of Texas at Austin.

Salvagno, M., Taccone, F. S., & Gerli, A. G. (2023). Can artificial intelligence help for scientific writing? *Critical Care*, 27:75. doi: 10.1186/s13054-023-04380-2

Shirky, C. (2025). Is AI Enhancing Education or Replacing It? *The Chronicle of Higher Education*, April 29, 2025. Retrieved from https://www.chronicle.com

Westfall, P. H., & Young, S. S. (1993). Resampling-based multiple testing: Examples and methods for p-value adjustment. John Wiley & Sons.