Human–AI Collaboration in Science at Scale: A Large-scale Randomized Field Experiment

Binglu Wang[1,2,3 *], Weixin Liang[4], Jiahui Xue[1], Yuhui Zhang[4], Hancheng Cao[5], Dashun Wang[1,2,3 *], Yian Yin[6*]

[1] Kellogg School of Management, Northwestern University, Evanston, IL 60208, USA

[2] Center for Science of Science and Innovation, Northwestern University, Evanston, IL, USA

[3] Northwestern University Institute on Complex Systems, Evanston, Illinois

[4] Department of Computer Science, Stanford University, Stanford, CA 94305, USA

[5] Goizueta Business School, Emory University, Atlanta, GA 30332, USA

[6] Department of Information Science, Cornell University, Ithaca, NY 14850, USA

* Correspondence to: binglu.wang@kellogg.northwestern.edu, dashun.wang@kellogg.northwestern.edu, yian.yin@cornell.edu

Abstract:

Collaboration is the defining mode of modern science, yet its core mechanism—feedback—remains hard to observe, difficult to scale, and unequally distributed. Here we test whether large language models (LLMs) can contribute to this hidden but vital practice. In a large-scale randomized field experiment, we delivered customized LLM-generated feedback for over 34,000 *arXiv* preprints across 150 fields and 131 geographic regions. Relative to controls, authors who received feedback had a significantly higher likelihood of revising their manuscripts, corresponding to a 12.5% relative increase over the baseline revision rate. These effects nearly doubled among authors based outside English-dominant research regions, consistent with the idea that AI feedback provides the greatest benefit where access to timely critique is otherwise limited. These revisions extended beyond language polishing: treated authors in these settings were significantly more likely to strengthen theoretical soundness and improve novelty. Further heterogeneity analyses reveal a non-linear pattern by intellectual proximity to AI: effects peak at moderate distance, indicating impact is maximized where relevance is high but redundancy is low. Importantly, exposure also increased subsequent adoption of LLM tools in authors' later papers, suggesting longer-run shifts in scientific practice. Together, these results provide causal evidence that AI can participate in a central collaborative practice of science, expand access to critique for scholars working in feedback-

scarce contexts, and accelerate the diffusion of new tools, carrying broad implications for scientists, their collaborators, and the institutions and policymakers that shape the increasingly AI-integrated global research ecosystem.

1. Introduction

Collaboration has become the defining mode of modern science. Across nearly every field, teams drive the majority of discoveries, and collaborative work underpins many of the most important innovations and breakthroughs [1-3]. This growing importance also reflects the changing reality of science, as scientific challenges expand in scale and complexity [4, 5] while individual scientists increasingly narrow their focus and expertise [6] or face difficulties in branching into new terrains [7]. By combining perspectives and approaches, collaboration enables scientists to generate ideas, refine methods, and strengthen conclusions, making it indispensable for progress.

At the heart of collaboration lies feedback. As Zuckerman observed [8], science often advances by "bouncing ideas off each other." Feedback is the mechanism through which scientists challenge assumptions, surface anomalies, and sharpen contributions [9, 10]. Yet despite its central role, this essential collaborative practice is largely invisible in data, taking place through various formal and informal exchange channels that rarely leave a trace. Its hidden nature makes feedback difficult to observe, to scale, and to intervene in, despite its influence on the direction and quality of scientific work.

At the same time, access to feedback is deeply unequal. Opportunities to exchange ideas are stratified along geographic, linguistic, and institutional lines [9, 11, 12]. Elite networks provide dense, continuous flows of critique, while scholars outside English-dominant or resource-rich settings face systemic barriers to receiving timely and constructive feedback [13-15]. Feedback thus drives progress but also reproduces inequality in science, determining whose ideas are refined and amplified, and whose remain underdeveloped.

Against this backdrop, large language models (LLMs) raise a provocative question: can AI participate in this hidden but vital form of collaboration? Beyond producing text, LLMs are increasingly capable of synthesizing information and offering critique [16-19]. If they can

provide useful feedback, they might expand collaborative capacity and reduce disparities in who has access to it.

Yet despite rapid advances in LLMs, there is substantial skepticism about whether they can play a meaningful role in science. While LLMs have demonstrated measurable gains in codified domains such as coding, writing, and professional decision-making [20-25], scientific research differs in fundamental ways. It is tacit, relying on context-sensitive judgments and accumulated expertise; it is highly specialized, with each field defining its own standards of rigor and contribution; and it is credibility-dependent, where the value of feedback hinges on whether it is trusted and defensible to expert peers. And ultimately, scientific research is a venture into the unknown. These features make collaboration with AI far less straightforward. Critics argue that AI feedback is often too generic, not credible, or prone to errors, raising doubts about whether it could meaningfully influence scientific practice [26, 27]. Others worry that reliance on AI could dilute the human expertise on which discovery depends [28, 29].

Here we test this possibility in a large-scale randomized field experiment on more than 34,000 scientific manuscripts. By delivering customized AI feedback to authors across 150 scientific fields and 131 geographic regions, we evaluate whether AI can meaningfully contribute to the collaborative practice of offering feedback. We examine both short-term effects on revisions and longer-term effects on adoption, with special attention to disparities across linguistic and institutional contexts. In doing so, we provide causal evidence on whether AI can engage in the collaborative practices that underpin scientific progress.

The timing of our experiment is central to its contribution. Conducted in early 2024, our intervention occurred during the brief onset phase of LLM adoption in scientific writing [30, 31], after the tools became widely available but before they were routinely embedded in research workflows. This moment allows us to capture a clean view of how AI enters a core collaborative practice, offering a new baseline for understanding human-AI collaboration as these tools become increasingly integrated into the scientific process.

2. Experiment Design

We examine a collaboration setting that mirrors how researchers draft manuscripts and seek critique from coauthors. This process plays a central role in advancing research, and allows us

to observe how collaboration shapes subsequent outcomes. To operationalize this process, we introduced an AI "collaborator" at scale, delivering structured yet customized feedback on scientific drafts. Recent advances have shown that LLM-based systems can generate comments resembling those of peer reviewers, with quality validated by experts and even preferred to some human reviews [16, 32]. Building on these advances, our study moves beyond validation and reviewer assistance to evaluate whether AI-generated feedback can influence scientific production itself, delivering it directly to authors while their manuscripts are still in development[1].

Sample collection and randomization: We collected preprints initially posted on *arXiv* between January and June 2024, and randomized at the paper level, stratified by 150 fields (Fig. 1). The initial randomized cohort comprised 34,340 manuscripts, from which we extracted 50,454 authors with valid email addresses across 131 regions worldwide (Extended Data Figure 1). During the period between random assignment and feedback delivery, 3,320 manuscripts (9.7%) posted an updated version. Because feedback was generated based on each paper's initially posted version, it could not plausibly inform revisions that had already occurred. We thus focus our analyses on preprints that had only the first version when feedback was delivered, to align our analysis with our causal estimand (the effect of receiving feedback for initial manuscripts on subsequent revision). The overall analysis sample contains 31,020 manuscripts (treatment group: $N = 15{,}542$; control group: $N = 15{,}478$). Baseline covariates remain balanced after restriction (SI Table S1). This risk-set definition follows standard causal-inference practice, aligning time-zero eligibility with treatment assignment and ensuring that only manuscripts genuinely at risk of exposure contribute to the estimated effect.

Treatment and control conditions: Authors in the treatment group received an email (See email template in SI Section 2) containing a private link to an interactive webpage with customized AI-generated feedback on their preprint, while control authors received no intervention. Feedback was generated using a state-of-the-art LLM framework, augmented with a multi-agent pipeline that (i) conducted review analysis, (ii) suggested alternative titles, and (iii) checked grammar. We further apply in-context learning to tailor the comments toward the specific content of each preprint (see feedback examples in SI section 1.3 and prompt detail

[1] The study was reviewed and approved by the Institutional Review Board at Northwestern University (STU00220102.)

in SI Section 10.1). For security and privacy, feedback links were encrypted, and access was restricted to the corresponding author via a private webpage.

Timeline: To manage the logistics of generating and distributing feedback at scale, we divided the sample into two cohorts. The first cohort included preprints first posted on *arXiv* January-March 2024 ($N = 17{,}411$); feedback, generated on each paper's first posted version, was delivered in June 2024. The second cohort consisted of preprints first posted April-June 2024 ($N = 16{,}929$), with feedback delivered in August 2024. Authors therefore received feedback between one and five months after initial posting.

Outcome measures: We trace two major outcome measures. (1) Short-term revisions: The number of revised versions posted within a month of feedback delivery. (2) Long-term adoption: Whether authors subsequently adopted LLM tools in their own writing. To assess this, we collected all publications posted by treatment and control authors in a twelve-month window after the intervention and applied a state-of-the-art AI detection model [30] to detect LLM-generated content. Detailed descriptions of these measurements are provided in SI Section 3.

Together, this design enables us to test whether AI can meaningfully participate in the collaborative practice of offering feedback, influence revisions in the short term, and shape the adoption of new tools in the longer term.

3. Results

AI feedback and manuscript revision A key test of whether feedback matters is to see if it alters the work. In this respect, revisions offer a clear, observable signal of engagement because they require deliberate effort and a decision to act on critique. Here we assess revision behaviors by tracking the number of manuscript revisions that were updated on *arXiv* following the intervention. Using time-stamped records, we measure the number of revised versions authors submitted within a 30-day window after receiving feedback. To estimate this effect, we fit the OLS regression model:

$$Revision_i = \beta_0 + \beta_1 \times Treatment_i + \varepsilon_i,$$

where $Revision_i$ is the number of revised versions paper i has post-treatment and $Treatment_i$ denotes assignment to the AI-feedback condition. The coefficient β_1 captures the intent-to-treat effect of receiving AI-generated feedback on the revision likelihood. Across the full sample, treated authors were significantly more likely to revise their manuscripts than the controls, with the average number of revisions increasing by 0.005, corresponding to a 12.5% relative increase over the baseline revision rate (Fig. 2A, $p < 0.05$, SI Table S2). Further, this difference appears especially strong for teams affiliated with institutions outside English-dominant regions, whose number of revisions featured a 22.4% relative increase over baseline in this group (0.037 vs 0.045, $p < 0.01$, SI Table S2). See SI S5 for more details as well as additional robustness checks (SI S6SI Fig. S1).

The stronger revision response among authors affiliated with institutions outside English-dominant regions raises an important question: what exactly did they revise? One possibility is that the observed effect was primarily driven by language polishing. To test this, we developed a systematic computational pipeline to compare the original and revised versions of *arXiv* preprints in this group. We parsed all PDFs and converted them into text, segmenting each paper into semantic blocks, which are context-preserving units that enabled precise alignment of corresponding sections across versions (see SI Section 10.2 for details). Within each block, we used large language models to align sentences and classify the nature of edits, distinguishing between content updates, language refinements, typographical corrections, and formatting adjustments (methods in SI Section 4) [33]. Each stage of the pipeline was iteratively refined using human-labeled samples to calibrate model behavior. This design allowed us to detect and interpret revisions with far greater granularity than conventional text-comparison approaches, revealing not only how much authors revised but what intellectual dimensions of their work changed.

At the individual sentence level, we find no differences for language, typographical, or formatting edits. Instead, treatment authors were significantly more likely to update the content itself (Fig. 2B; SI Table S3). This suggests that revisions due to LLM feedback are not limited to language polishing but reflect more substantive changes.

To understand the nature of the content that the authors updated, we further analyzed revisions at the semantic block level, inferring revision intentions across eleven categories [16] (see prompt details in SI Section 10.4). Indeed, we find suggestive evidence that treatment authors were substantially more likely to make revisions addressing substantive issues, corresponding to a 32.4% relative increase in revisions aimed at improving theoretical soundness (Fig. 2C; SI Table S4). We also find suggestive evidence that the treated group made significantly more changes addressing novelty (63% increase). These findings again reject the hypothesis that AI feedback predominately improves minor language issues. Instead, they suggest that targeted AI feedback not only encouraged additional revisions but appeared to nudge authors toward reinforcing the intellectual core of their manuscripts.

Together, these results suggest that AI-generated feedback can meaningfully alter short-term scientific practice. Despite numerous concerns and potential shortcomings of LLMs [26-29], our results highlight substantive roles of LLMs in providing feedback [30, 34]. By prompting content revisions, particularly those tied to theory, clarity, and prior literature, AI may participate in the collaborative practice of critique rather than functioning merely as a language polishing tool.

Heterogeneity analysis The strong effect observed among authors affiliated with institutions outside English-dominant regions (Fig. 3A; SI Table S5) is consistent with the idea that AI feedback may disproportionately aid those for whom valuable feedback is otherwise limited. This further raises the question of who is most responsive to AI-generated feedback, as opportunities to exchange ideas are shaped by social and geographical factors. To this end, we first conduct geographic comparisons, restricting analysis to countries with more than 300 observations to ensure stable estimates.

Region-level patterns reveal marked asymmetries (Fig. 3B; SI Table S6). Positive significant effects appear in Israel, South Korea, and China, where treated authors revised substantially more than their counterparts in the control group (Fig. 3B). Several other countries, including Brazil, Germany, and Japan, show effects that are directionally positive but not statistically significant at conventional levels, probably due to smaller sample sizes. In contrast, the effects in the United States, Australia, the United Kingdom, and Canada are close to zero. Such region-level variation is consistent with the uneven distribution of traditional collaborative

infrastructures, suggesting that AI feedback appears to offer meaningful opportunities in regions where these channels are less accessible.

Beyond language and geography, responsiveness may also vary with intellectual positioning. Proximity to AI may equip scholars to exploit such tools [35], yet the "intellectual distance penalty" predicts that experts scrutinize ideas encroaching on their own specialization [36]. We measured proximity using *arXiv*'s subject co-classification network, computing each paper's shortest-path distance to AI (SI Fig. S3) and asking if the effect is stronger for researchers who are in further or closer proximity. The results suggest neither, revealing a non-linear pattern: authors at a moderate distance from AI were 40.9% more likely to revise in response to treatment, whereas effects for the closest and farthest fields were statistically indistinguishable from zero (Fig. 3C; SI Table S7).

This inverted-U relationship rules out simple explanations in which AI feedback is most useful either where AI expertise is highest (close proximity) or where need is greatest (far proximity). Instead, it suggests a more complex interaction between topical proximity, field norms, and how easily authors can incorporate external critique. Several mechanisms may contribute to this relationship. For example, a perception mechanism suggests that whereas insiders and distant outsiders may dismiss AI-generated comments [36], researchers at an intermediate distance may be both close enough to assess relevance yet far enough to value external input. Overall, the strongest effects at intermediate distance indicate that intellectual positioning within the scientific landscape shapes how researchers engage with AI-generated feedback.

Together, these findings suggest that AI-generated feedback may act as a collaborative equalizer. It expands access to critique for those traditionally excluded from elite feedback networks, benefiting scholars working in regions outside English-dominant research environments and those located in geographically peripheral institutions. At the same time, receptiveness is not simply a function of material resources; intellectual distance also shapes who benefits most. The strongest effects emerge not among fields closest to AI nor among those most distant, but in between. This inverted-U pattern suggests that rather than tracking simply with expertise or need, the impact of AI feedback appears to depend on more subtle positioning within the scientific landscape, maximized not at the extremes but where relevance is high enough and redundancy low enough for AI feedback to matter.

AI feedback and future AI uses Although our intervention occurred only once, its influence may extend beyond immediate revisions. This raises an interesting question of whether exposure to AI-generated feedback leaves detectable traces in later research practice. To this end, we examined whether exposure to AI-generated feedback increased subsequent adoption of LLMs. We tracked each author's *arXiv* preprints within a twelve-month window after the intervention and analyzed abstracts with a state-of-the-art AI detection model [30], which assigns a paper-level score α estimating the fraction of text likely generated by LLM. We define adoption as high α in post-treatment papers ($\alpha_{post} > \tau$, where $\tau = 0.1$ in our primary analysis). Consistent with the main analysis, we focus on authors with minimal prior use ($\alpha_{pre} \leq \tau$). We estimated the treatment effect using an OLS regression:

$$Adoption_p = \beta_0 + \beta_1 \times Treatment_p + \varepsilon_p ,$$

where $Adoption_p$ is an indicator variable equal to 1 if researcher p adopted LLM tools in subsequent manuscript writing, and $Treatment_p$ denotes assignment to the AI-feedback condition. The coefficient β_1 captures the intent-to-treat effect of receiving AI-generated feedback on the likelihood of later LLM adoption.

We find that, among authors whose pre-treatment writing showed minimal detectable LLM involvement, those in the treatment group were 5.47% more likely to adopt LLM tools in subsequent work within twelve months (Fig. 4; SI Table S8). This effect was again driven by authors based outside English-dominant research regions, who exhibited a 9.56% increase in adoption. Robustness checks that alter the post-treatment window (three to twelve months, Fig. 4B) and vary the adption threshold τ (0 to 0.2, SI S6) yield consistent results, with authors continuously showing large and statistically significant adoption effects and the effects are consistently larger for non-English dominant group. These adoption patterns mirror the revision findings, suggesting that the same group that show an increased engagement with substantive revisions are also more likely to incorporate AI into their research workflows over the longer term. This further suggests that the effect we observe does not merely reflect transient experimentation but rather sustained integration, as measured by repeated use across multiple subsequent manuscripts.

Together, these results demonstrate that AI feedback can shape collaborative practices at scale, not only prompting immediate revisions but also accelerating the diffusion of AI into research workflows. The longer-term effects were strongest among those historically excluded from

dense feedback networks, suggesting that AI can help lower barriers to participation and act as a catalyst for broader adoption of new tools. In this way, AI functions not only as a short-term collaborator but also as an enduring influence on how scientific work is conducted.

4. Discussion

This study provides causal evidence that large language models can meaningfully participate in a central collaborative practice of science: offering feedback. In a large-scale field experiment, exposure to AI-generated critique prompted substantive revisions to manuscripts and increased subsequent adoption of AI tools. These effects were not evenly distributed. They were stronger where feedback networks are thinner, and in fields positioned at a moderate intellectual distance from AI. Taken together, these findings suggest that AI has the potential to act as a collaborative equalizer in science, expanding access to critique where traditional channels are less accessible.

Importantly, the consequences of AI feedback extended beyond immediate revisions. Authors not only incorporated more changes that strengthened theoretical soundness and novelty, but also demonstrated a greater likelihood of integrating AI tools into future writing. Qualitative responses reinforced these patterns: participants described the feedback as clear and useful, and many expressed an intention to continue using such tools. These responses are consistent with the hypothesis that the intervention may not only shape the content of individual manuscripts but also leave lasting traces in research practices.

The heterogeneity we observe provides conceptual clues about how human-AI collaboration unfolds. The inverted-U relationship by intellectual distance is particularly informative: AI feedback had the greatest impact not in fields closest to AI, nor in those farthest removed, but at an intermediate distance [37, 38]. This rules out simple explanations based solely on expertise or need. One plausible interpretation is that feedback is most influential where it is both interpretable and not redundant with existing expertise, and where it is close enough to be relevant yet distant enough to add something new. Other mechanisms may also contribute, including field norms around automated tools, baseline familiarity with LLMs, and how well generic critique maps onto disciplinary standards. Disentangling these channels is an important direction for future work, but the core implication of this finding suggests that intellectual

positioning within the landscape of AI appears to strongly shape how researchers respond to AI-generated feedback.

The timing of the study is also noteworthy. Conducted in early 2024, our experiment captures scientific practice at the onset of LLM adoption, when these tools were widely available and broadly known, but not yet fully normalized into drafting and collaboration workflows. This produced a natural mix of researchers, with some already experimenting with LLMs but many still relying on traditional feedback channels. Random assignment in our experiment balances these dispositions across treatment and control, allowing us to isolate how exogenous AI feedback affects revision and adoption in a critical regime where AI is accessible but not yet embedded in everyday work. As technology evolves [39, 40][2], adoption accelerates and baseline use rises, the absolute magnitudes of our effects may evolve, but the underlying mechanisms we document, including equity gains where feedback is scarce, substantive improvements to intellectual content, and sensitivity to intellectual positioning, are likely to remain central. In this sense, these findings are both timely and fundamental, providing a baseline for understanding how human-AI collaboration takes hold in the increasingly AI-infused global research ecosystem.

Our study also advances efforts to understand and experiment with the scientific process itself. Field experiments in science are notoriously difficult to execute or scale, with most prior trials limited to small samples or narrow domains [41-44]. By leveraging bibliometric infrastructure and automated feedback systems, we demonstrate the feasibility of running large-scale, personalized field experiments on live scientific practice. This opens new opportunities to not only observe scientific behavior retrospectively but to test interventions capable of shaping how knowledge is produced, revised, and adopted across global communities.

At the same time, our findings raise questions about the future of collaboration in an era of increasingly AI-integrated research. If AI can reliably provide critique on clarity, theory, and prior work, human collaborators may shift toward higher-order conceptual and design roles, but marginal collaborations may also be displaced. A further challenge arises from standardization: as more researchers lean on similar underlying models, science risks a form of homogenization [16], with diverse lines of critique converging toward a narrow band of

[2] Other AI feedback tools such as https://paperreview.ai/ and https://www.refine.ink/.

recommendations. This concern parallels emerging evidence that, while AI can raise individual performance, it may also reduce collective diversity unless deliberately counteracted[45, 46]. These possibilities highlight the importance of designing human-AI collaboration to preserve heterogeneity, through plural model architectures, varied objectives, and institutional norms that reward dissent and methodological diversity.

Overall, our results suggest that large language models can do more than streamline writing or surface edits. They can influence the collaborative practices that shape scientific knowledge itself. By expanding access to critique and empowering scholars historically excluded from dense feedback networks, AI holds the potential to make science more inclusive. By highlighting the conditions under which its effects are strongest, our study points to pathways for integrating AI into research in ways that preserve diversity, enhance credibility, and maximize surplus. As AI becomes more deeply woven into the research ecosystem, its role in collaboration may carry broad implications, not only for scientists and their collaborators, but also for the institutions and policymakers that shape the future of global science.

Reference:

1. Wuchty, S., B.F. Jones, and B. Uzzi, *The increasing dominance of teams in production of knowledge.* Science, 2007. **316**(5827): p. 1036-1039.
2. Jones, B.F., S. Wuchty, and B. Uzzi, *Multi-university research teams: Shifting impact, geography, and stratification in science.* science, 2008. **322**(5905): p. 1259-1262.
3. Woolley, A.W., et al., *Evidence for a collective intelligence factor in the performance of human groups.* science, 2010. **330**(6004): p. 686-688.
4. Price, D.J.D.S., *Little science, big science*. 1963: Columbia University Press.
5. Wang, D. and A.-L. Barabási, *The science of science*. 2021: Cambridge University Press.
6. Jones, B.F., *The burden of knowledge and the "death of the renaissance man": Is innovation getting harder?* The Review of Economic Studies, 2009. **76**(1): p. 283-317.
7. Hill, R., et al., *The pivot penalty in research.* Nature, 2025: p. 1-8.
8. Zuckerman, H., *Scientific elite: Nobel laureates in the United States*. 1977: Transaction Publishers.
9. Merton, R.K., *The Matthew effect in science: The reward and communication systems of science are considered.* Science, 1968. **159**(3810): p. 56-63.
10. Kuhn, T., *The structure of scientific revolutions*. 1996: Princeton University Press.
11. Bourdieu, P., *The forms of capital*, in *The sociology of economic life*. 2018, Routledge. p. 78-92.
12. Nielsen, M.W. and J.P. Andersen, *Global citation inequality is on the rise.* Proceedings of the National Academy of Sciences, 2021. **118**(7): p. e2012208118.
13. Allison, P.D., J.S. Long, and T.K. Krauze, *Cumulative advantage and inequality in science.* American Sociological Review, 1982: p. 615-625.
14. Huang, J., et al., *Historical comparison of gender inequality in scientific careers across countries and disciplines.* Proceedings of the national academy of sciences, 2020. **117**(9): p. 4609-4616.
15. Graves Jr, J.L., et al., *Inequality in science and the case for a new agenda.* Proceedings of the National Academy of Sciences, 2022. **119**(10): p. e2117831119.
16. Liang, W., et al., *Can large language models provide useful feedback on research papers? A large-scale empirical analysis.* NEJM AI, 2024. **1**(8): p. AIoa2400196.
17. Zou, J. and E.J. Topol, *The rise of agentic AI teammates in medicine.* The Lancet, 2025. **405**(10477): p. 457.
18. Swanson, K., et al., *The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies.* Nature, 2025: p. 1-3.
19. Shao, E., et al., *SciSciGPT: Advancing Human-AI Collaboration in the Science of Science.* arXiv preprint arXiv:2504.05559, 2025.
20. Noy, S. and W. Zhang, *Experimental evidence on the productivity effects of generative artificial intelligence.* Science, 2023. **381**(6654): p. 187-192.
21. Peng, S., et al., *The impact of ai on developer productivity: Evidence from github copilot.* arXiv preprint arXiv:2302.06590, 2023.
22. Dell'Acqua, F., et al., *Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality.* Harvard Business School Technology & Operations Mgt. Unit Working Paper, 2023(24-013).
23. Boussioux, L., et al., *The crowdless future? Generative AI and creative problem-solving.* Organization Science, 2024. **35**(5): p. 1589-1607.

24. Eloundou, T., et al., *GPTs are GPTs: Labor market impact potential of LLMs.* Science, 2024. **384**(6702): p. 1306-1308.
25. Brynjolfsson, E., D. Li, and L. Raymond, *Generative AI at work.* The Quarterly Journal of Economics, 2025. **140**(2): p. 889-942.
26. Si, C., D. Yang, and T. Hashimoto, *Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers.* arXiv preprint arXiv:2409.04109, 2024.
27. Gao, Y., et al., *Take caution in using LLMs as human surrogates.* Proceedings of the National Academy of Sciences, 2025. **122**(24): p. e2501660122.
28. Kosmyna, N., et al., *Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task.* arXiv preprint arXiv:2506.08872, 2025.
29. Bastani, H., et al., *Generative AI without guardrails can harm learning: Evidence from high school mathematics.* Proceedings of the National Academy of Sciences, 2025. **122**(26): p. e2422633122.
30. Liang, W., et al., *Quantifying large language model usage in scientific papers.* Nature Human Behaviour, 2025: p. 1-11.
31. Kusumegi, K., et al., *Scientific production in the era of large language models.* Science, 2025. **390**(6779): p. 1240-1243.
32. Thakkar, N., et al., *Can llm feedback enhance review quality? a randomized study of 20k reviews at iclr 2025.* arXiv preprint arXiv:2504.09737, 2025.
33. Jiang, C., W. Xu, and S. Stevens, *arXivEdits: Understanding the Human Revision Process in Scientific Writing.* In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. , 2022: p. 9420–9435.
34. Kobak, D., et al., *Delving into LLM-assisted writing in biomedical publications through excess vocabulary.* Science Advances, 2025. **11**(27): p. eadt3813.
35. Otis, N., et al., *The uneven impact of generative AI on entrepreneurial performance.* 2024.
36. Boudreau, K.J., et al., *Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science.* Management science, 2016. **62**(10): p. 2765-2783.
37. Acemoglu, D., *Technical change, inequality, and the labor market.* Journal of economic literature, 2002. **40**(1): p. 7-72.
38. Autor, D.H., F. Levy, and R.J. Murnane, *The skill content of recent technological change: An empirical exploration.* The Quarterly journal of economics, 2003. **118**(4): p. 1279-1333.
39. Jin, Y., et al., *Agentreview: Exploring peer review dynamics with llm agents.* arXiv preprint arXiv:2406.12708, 2024.
40. Taechoyotin, P., et al. *MAMORX: Multi-agent multi-modal scientific review generation with external knowledge*. in *Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*. 2024.
41. Blau, F.D., et al., *Can mentoring help female assistant professors? Interim results from a randomized trial.* American Economic Review, 2010. **100**(2): p. 348-352.
42. Boudreau, K.J. and K.R. Lakhani, *Innovation experiments: Researching technical advance, knowledge production, and the design of supporting institutions.* Innovation Policy and the Economy, 2016. **16**(1): p. 135-167.
43. Boudreau, K.J., et al., *A field experiment on search costs and the formation of scientific collaborations.* Review of Economics and Statistics, 2017. **99**(4): p. 565-576.
44. Liu, L., et al., *Data, measurement and empirical methods in the science of science.* Nature human behaviour, 2023: p. 1-13.

45. Hao, Q., et al., *AI Expands Scientists' Impact but Contracts Science's Focus.* arXiv preprint arXiv:2412.07727, 2024.
46. Doshi, A.R. and O.P. Hauser, *Generative AI enhances individual creativity but reduces the collective diversity of novel content.* Science advances, 2024. **10**(28): p. eadn5290.

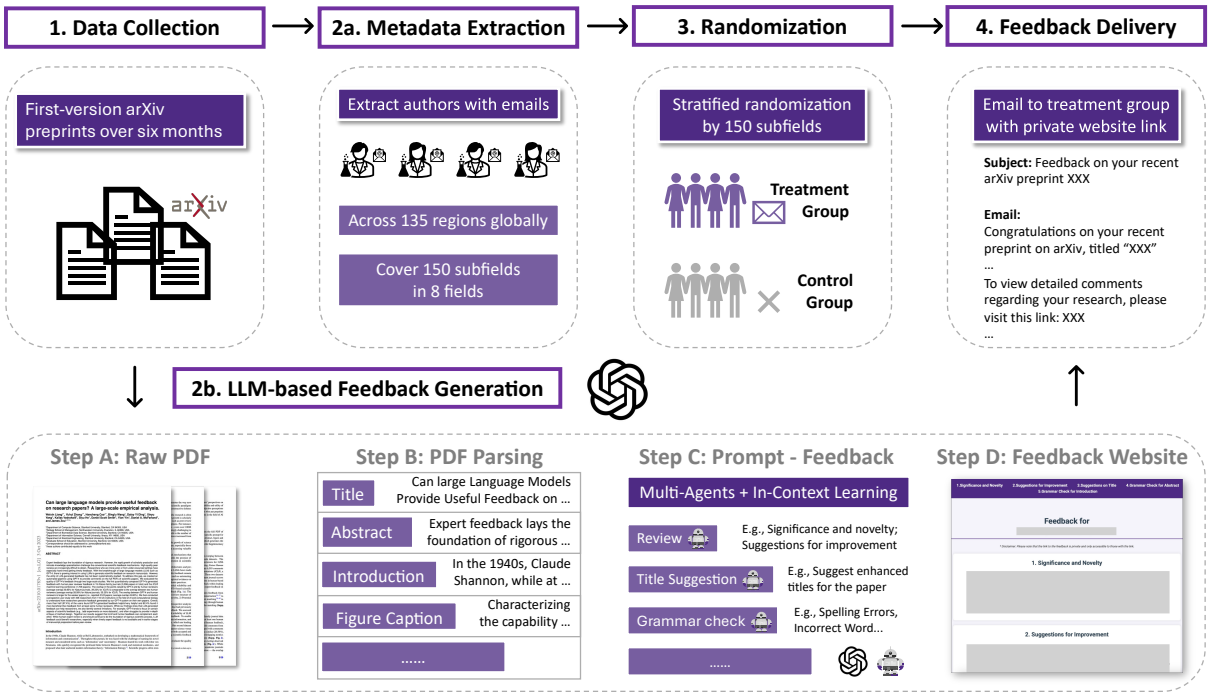


Figure 1. LLM feedback generation and experiment workflow. (Step 1) We collected all recent *arXiv* preprints (Jan–Jun 2024) for which only the first version was available. **(Step 2)** For each preprint, we extracted metadata (authors, affiliations, field, **2a**), and parsed structured text from the raw PDFs to generate customized feedback using an LLM-based system (**2b**). **(Step 3)** We implemented a stratified randomization process across 150 fields, assigning each paper to either the treatment or control group. **(Step 4)** Each author in the treatment group receives a personalized email, with a private link giving them access to the AI-generated feedback.

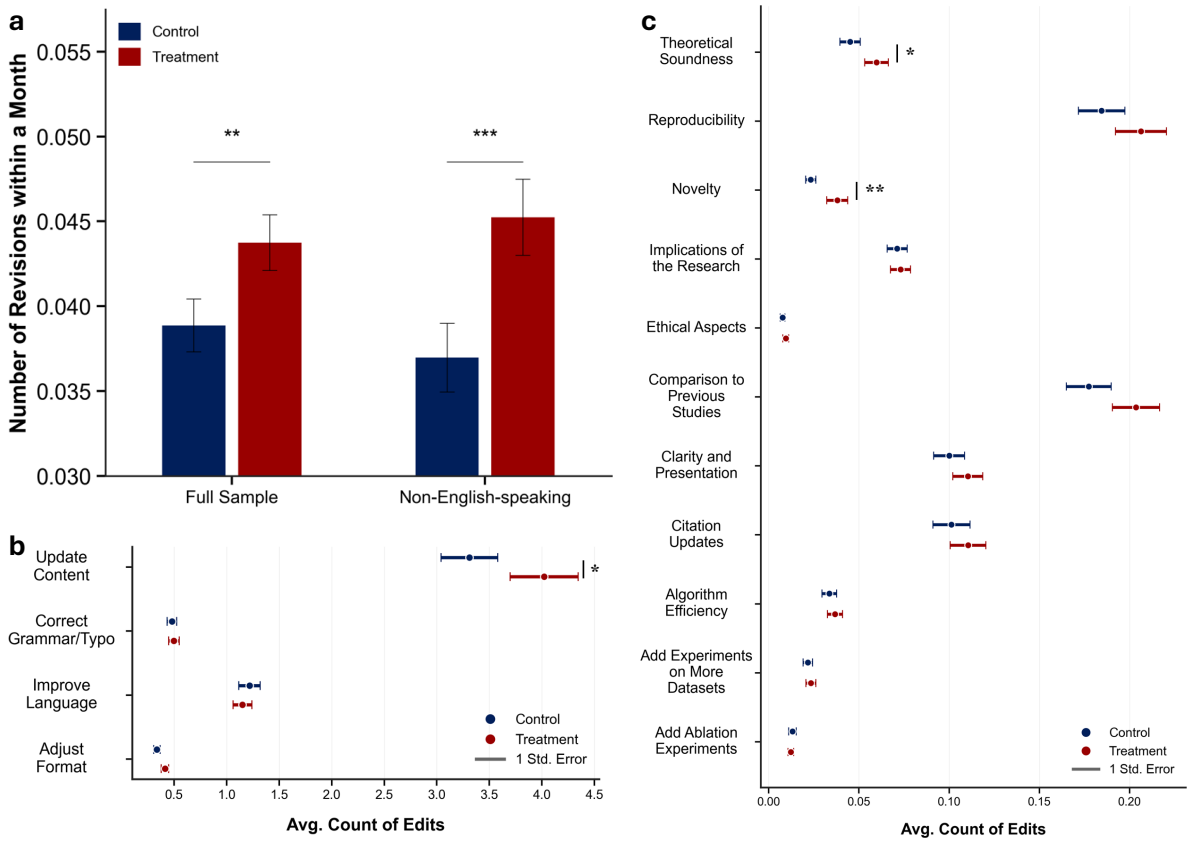


Figure 2. Effects of AI-generated feedback on short-term scientific practice. (a) Number of revisions within one month of feedback delivery. Across the full sample, authors receiving AI-generated feedback were likely to revise their manuscripts 12.5% more than controls, with effects concentrated among non-English-speaking teams, who showed a 22.4% higher number of revisions. **(b)** Sentence-level analysis of revision types among non-English group. Differences between treatment and control groups were negligible for formatting, grammar or language refinement, but treated authors were significantly more likely to update substantive content. **(c)** Block-level analysis of revision intentions among non-English group. Treated papers were more likely to incorporate changes addressing deeper intellectual concerns, including theoretical soundness and novelty. Together, these findings show that AI feedback prompts revisions that extend beyond surface-level polishing, demonstrating that LLMs can meaningfully participate in the collaborative practice of critique. Standard significance thresholds are indicated: $***$ $P < 0.01$, $**$ $P < 0.05$, $*$ $P < 0.1$.

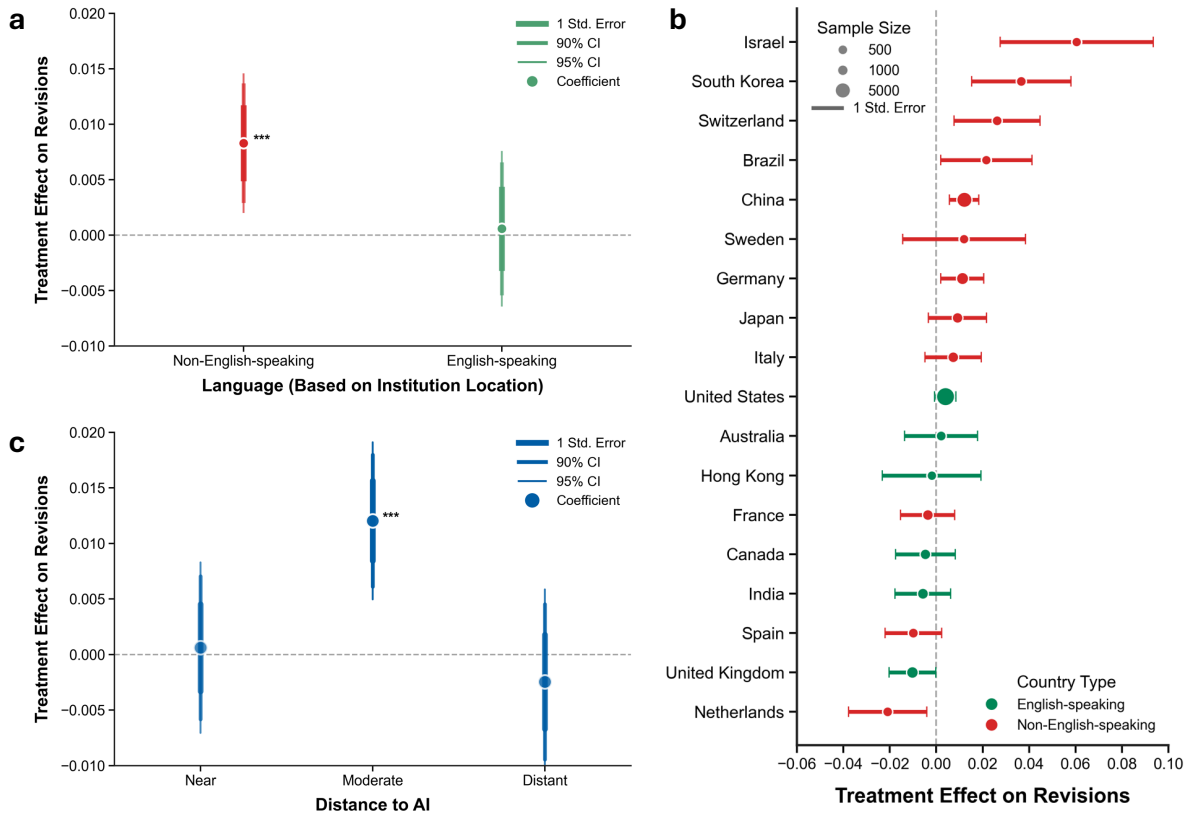


Figure 3. Heterogeneous effects of AI feedback across language, geography, and intellectual distance. (a) Treatment effects by institutional language environment. Non-English-speaking teams exhibit a strong positive response to LLM feedback, whereas English-speaking teams show no measurable effect. **(b)** Region-level heterogeneity across major research-producing nations. Positive effects are observed in Israel, South Korea, and China, while effects are close to 0 in the United States, United Kingdom, and other English-speaking countries. Point size reflects the number of preprints per region. **(c)** Treatment effects by intellectual distance to AI, measured as shortest-path distance in a topic co-appearance network. Effects follow a non-linear pattern: researchers at moderate distance are significantly more likely to revise in response to AI feedback, while those closest to AI appear resistant, and those farthest away find the feedback less relevant. These findings indicate that AI feedback expands collaborative capacity where it is most scarce, acting as a "collaborative equalizer" across linguistic and geographic divides, while also revealing that receptiveness depends on intellectual positioning within science. $*** P < 0.01$, $** P < 0.05$, $* P < 0.1$.

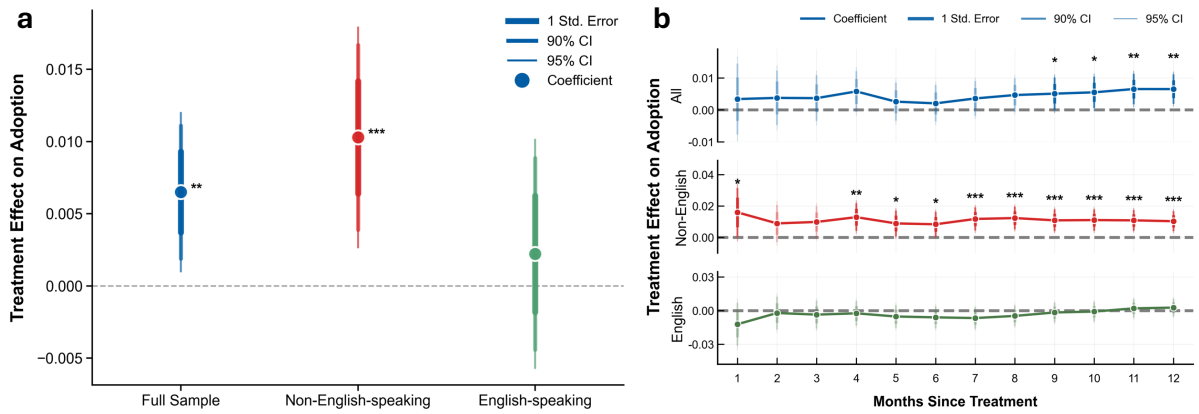
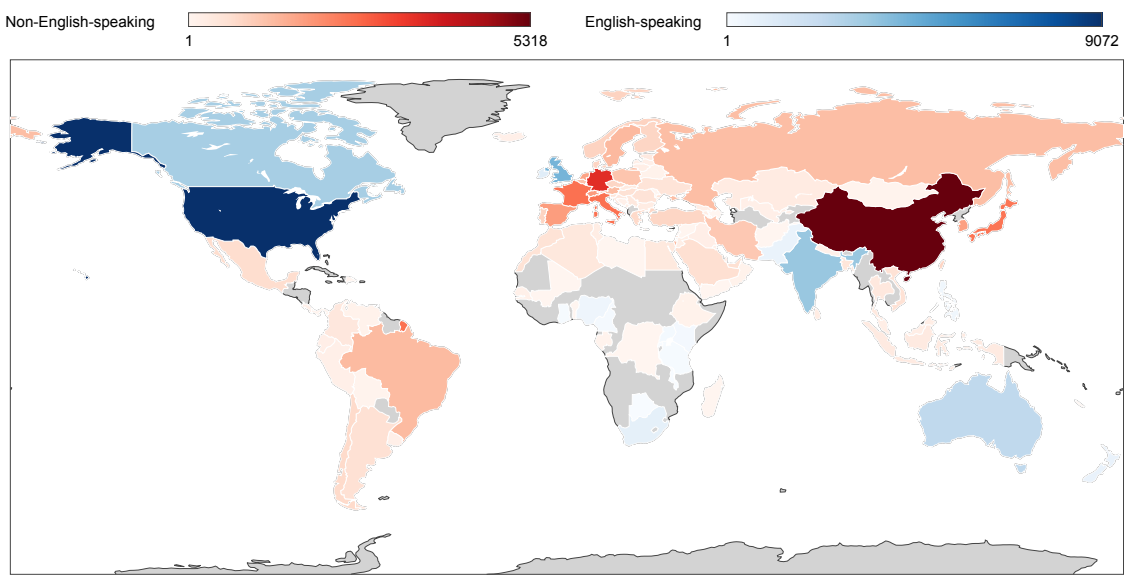


Figure 4. Treatment effect on likelihood of adoption. (a) Estimates indicate a marginal positive treatment effect on the likelihood of adoption ($\alpha_{post} > 0.1$) among authors with minimal prior use ($\alpha_{pre} \leq 0.1$). We further explore heterogeneous treatment effects by splitting our sample into two groups based on whether the authors are from English or non-English speaking countries. Authors from non-English-speaking countries in the treatment group show a significantly higher likelihood of adopting LLM tools in subsequent academic writing. Conversely, the analysis reveals no significant difference in revision rates between the treatment and control groups for English-speaking authors. **(b)** We extend the analysis in panel (a) by varying the outcome window from 1 months to 12 months. The direction of the treatment effect is consistent across these alternative time windows. $***$ $P<0.01$, $**$ $P<0.05$, $*$ $P<0.1$.



Extended Data Figure 1. Global distribution of experimental sample across English- and non-English-speaking regions. The map illustrates the geographical scope of the randomized field experiment. Blue indicates English-speaking countries, while red denotes non-English-speaking regions. Each country is shaded according to the number of manuscripts included in the study, highlighting the wide heterogeneity of participation across the global research ecosystem.

Supplementary Information

Human–AI Collaboration in Science at Scale: A Large-scale Randomized Field Experiment

Binglu Wang[1,2,3*], Weixin Liang[4], Jiahui Xue[1], Yuhui Zhang[4], Hancheng Cao[5], Dashun Wang[1,2,3 *], Yian Yin[6*]

[1] Kellogg School of Management, Northwestern University, Evanston, IL 60208, USA

[2] Center for Science of Science and Innovation, Northwestern University, Evanston, IL, USA

[3] Northwestern University Institute on Complex Systems, Evanston, Illinois

[4] Department of Computer Science, Stanford University, Stanford, CA 94305, USA

[5] Goizueta Business School, Emory University, Atlanta, GA 30332, USA

[6] Department of Information Science, Cornell University, Ithaca, NY 14850, USA

Table of Contents

1. Sample and Experiment Setup 3

1.1 Sample Construction 3

1.2 Experiment Setup 3

1.3 Feedback Example 4

2. Email Template 7

3. Text Analysis on Email Response 8

4. Content analysis 9

4.1 Initial Processing 9

4.2 Semantic Block Alignment 9

4.3 Content Update Intention Detection 9

4.4 Sentence-Level Edit Extraction 10

4.5 Average Percentage of Edit Types and Intentions 10

5. Measurement of Variables 12

5.1. Number of Revisions 12

5.2 Region Data Collection and English-Speaking Region 12

5.3 Distance to AI 12

5.4. Detection of LLM usage 13

6. Robustness Check 15

7. Potential Limitations 16

8. Figure Appendix 17

9. Table Appendix 25

10. Prompt Appendix 34

10.1 Feedback Generation 34

10.2 Block-Level Alignment Extraction 36

10.3 Sentence-Level Edit Extraction 37

10.4 Content Update Intention Detection 38

11. Reference 39

1. Sample and Experiment Setup

1.1 Sample Construction

In this study, we focused on recent preprints from *arXiv*, one of the largest preprint servers covering the physical, mathematical, and computational sciences. We obtained paper-level metadata from a publicly available data dump (Cornell *arXiv* Dataset[1]) and restricted our sample to manuscripts published between January 2024 and June 2024.

For each entry, we downloaded and parsed the corresponding PDF file (using Cermine [1]) to extract author contact information, including names, affiliations, and email addresses. To ensure data quality, we excluded papers if (1) the PDF could not be successfully downloaded or parsed, or (2) the extracted emails contained formatting errors. Importantly, the mapping between authors and papers is not one-to-one. To avoid oversampling, for authors with multiple submissions, we retained only their most recent paper in our sampling period, and for papers with more than two email addresses listed, we only kept the first and last.

We processed the first batch of data (Jan – Mar 2024) in June, yielding a cleaned sample of 17,411 papers. We then repeated the same procedure for papers published between April and June 2024 in August,[2] obtaining our second batch of data (an additional 16,929 papers).

Finally, for each round, we conducted stratified randomization based on the primary subject domain assigned by *arXiv*. In total, our final dataset includes 34,340 papers with valid PDF files, email addresses, and author information.

1.2 Experiment Setup

Feedback generation: We built on a pipeline in Liang et al. (2024) [2] and developed an automated system of scientific feedback generation based on GPT-4o. It begins by parsing the entire manuscript (full-text PDF) using a machine-learning-based PDF parser (ScienceBeam [3]) into structured components including title, abstract, main body text, and figure and table captions.

[1] https://www.kaggle.com/datasets/Cornell-University/arxiv/data

[2] For this batch, we also excluded email addresses that had already appeared in the January–March dataset to avoid duplication.

Based on the parsed content, we construct a paper-specific prompt by concatenating our instructional template with the extracted information. We then feed this prompt into GPT-4o API, which generates feedback in a single pass. To ensure clarity and consistency, we instructed GPT-4o to produce a structured feedback outline. Drawing on reviewer report formats from major machine learning conferences and *Nature* family journals, our prompt guided the model to generate review-style comments (see Section 9.1 for prompt details). In addition to review-style feedback, we also provided other forms of author-facing support including title suggestions and grammar checks. See feedback examples in Section 1.3.

Website setup: For each manuscript, we created a dedicated webpage using the generated feedback (see Website Screenshot in Figure S8). The site features five sections: (1) *Significance and Novelty*, (2) *Suggestions for Improvement*, (3) *Title Suggestions*, (4) *Grammar Check for Abstract*, and (5) *Grammar Check for Introduction*. To protect security and privacy, each web link was encoded using Base32, and access was restricted to the corresponding author via a private, non-indexed website. The website was designed to facilitate meaningful engagement by presenting the feedback in a structured, interactive format, making it easier for authors to navigate and incorporate the feedback.

Intervention email: Authors in the treated group received a standardized email containing the private web link, which enabled authors to access the generated feedback. The email informed recipients that the comments were generated with the assistance of advanced AI models. The complete email template is provided in SI Section 2.

1.3 Feedback Example

Treatment group received feedback organized into five sections, each generated through distinct pipelines. Examples of each section are provided below.

Significance and Novelty: This section summarizes the paper's key contributions and the importance of the study.

Example: *"The study addresses a critical and often overlooked aspect of recommendation systems: the integration of diversity throughout all stages of the recommendation pipeline, rather than just during reranking. It introduces the XXX, which leverages knowledge distillation and contrastive context encoding to model both positive and negative contexts, enhancing both efficiency and diversity in recommendations. Evaluates its performance using*

both offline experiments on industrial datasets and an online A/B test, providing a comprehensive analysis of the model's efficacy. Proposes an innovative method to handle the quadratic complexity of existing diversification methods, making it suitable for preranking and ranking stages."

Suggestions for improvement: This section provides structured, review-style suggestions on how the paper could be improved.

Example: *"Clarification of Mathematical Notation and Algorithms: The paper provides detailed mathematical descriptions of the proposed models and algorithms. However, some parts, especially within sections like Differentiable Sampling with GumbelTop reparametrization, can benefit from additional clarification or illustrative examples. While the equations are precise, succinct summaries of what each equation aims to achieve or any assumptions would make it accessible to a broader audience."*

"Ethical Considerations and Bias Assessment: Discussing the ethical implications of SVS, such as potential misuse or biases in gender and vocal range synthesis, can add depth to the paper. Addressing how the model handles diverse languages and cultures could also enhance its robustness and inclusivity."

"Comparison with Broader Set of Baselines: Although the paper compares CDM against several established baselines, expanding this comparison to include more recent and varied models, such as those leveraging deep reinforcement learning or graphbased approaches, would strengthen the claims about its performance and generalizability."

Suggestions on Title: We generated title suggestions by prompting GPT to retrieve semantically similar titles from our high-quality-paper database of and to generate five titles most aligned with the paper's focus.

Example: *"Recommended Titles:*

Contextual Distillation for Diverse and Efficient Industrial Recommendations

Reasoning: Emphasizes the practical application in industrial settings and highlights both the diversity and efficiency aspects"

Grammar Check for Abstract: We conducted automatic grammar and usage checks for the abstract, identifying and correcting specific errors.

Example: *"Original Sentence: as shown by metrics, provide strong superiority for the effectiveness of CDM.*

ErrorType: Comma Splice

Explanation: Comma splice in the sentence 'as shown by metrics, provide strong'

Recommended Fragment: as shown by metrics, providing strong evidence for the effectiveness of CDM."

"Original Sentence: The diversity of recommendation is equally crucial as accuracy in improving user experience.

ErrorType: Incorrect Word Usage

Explanation: Misuse of 'equally crucial as' instead of 'equally crucial to'

Recommended Fragment: The diversity of recommendation is equally crucial to accuracy in improving user experience."

Grammar Check for Introduction: A similar procedure was applied to the Introduction section, where grammatical and stylistic issues were detected.

2. Email Template

Subject: Feedback on your recent arXiv preprint [arXiv ID]

Dear [Author_First_Name],

Congratulations on your recent preprint on arXiv, titled "[Title]." We are grateful for your hard work and dedication to the field, and we value your contributions!

We are part of a team from Northwestern University, Stanford University, and Cornell University, committed to providing research feedback to scholars with the assistance of advance AI models. We have followed your work closely and, upon a thorough examination, generate the suggestions below. These suggestions cover various aspects of your work, including the writing style, research design, and title. We hope they offer you fresh insights that may enhance the depth and impact of your research.

To view detailed comments regarding your research, please visit this link [Insert a link].

Once again, congratulations on your achievement. We are certain that your work will have an impact on the future of your field and will inspire fellow researchers worldwide.

Should you have any questions, concerns, or suggestions, please do not hesitate to reach out to us at XXX. Your insights would be invaluable in helping us better support the community's research development.

Best regards,
Feedback Team

Disclaimer: The information contained in this email is intended solely for the recipient(s) named above. The content provided should not be considered to be professional or definitive advice.

3. Text Analysis on Email Response

We received a total of 78 direct responses to the intervention emails and conducted a semantic analysis of these responses.

First, the overall tone of the responses was positive, with 14% explicitly expressing negative sentiment toward the LLM-generated feedback. Several authors remarked that the feedback closely resembled comments they had received through formal peer review processes, suggesting that they perceived it as both credible and useful. For example, one author wrote, "I feel the feedback is quite informative and similar to some comments we received during peer review." Another reported, "The comments aligned surprisingly closely with our first-round peer review. It would be great if we could use AI to speed up that process." Second, several responses indicated that authors revised their manuscripts in direct response to the feedback we provided. For instance, one author noted, "This is to thank you profusely for the very helpful comments on my paper. The revised version just appeared on *arXiv*." Third, some authors expressed an interest in using similar feedback tools in the future. One wrote, "Can I share a copy of my next paper with you for feedback before I submit it for potential publication?"

We further coded the institutional affiliation of each responding author along two dimensions: (1) whether the institution is located in an English-speaking region, defined by whether the official language in this region include English, and (2) whether it ranks within the top 100 globally, based on the Times Higher Education World University Rankings 2025.[3] Our qualitative analysis (Fig. S4) of the comments received suggests that scholars from non-English speaking countries or lower-prestige institutions perceived the feedback as more constructive and valuable than their counterparts from English-speaking countries or high-prestige institutions. This disparity suggests that the use of LLMs for scientific feedback has particularly significant benefits for scholars from marginalized groups and those in resource-scarce environments, ultimately making scientific discourse more accessible to a wider and more diverse audience.

[3] World University Rankings 2025, *Times Higher Education*, https://www.timeshighereducation.com/world-university-rankings/latest/world-ranking.

4. Content analysis

4.1 Initial Processing

We conducted content analysis of 12,133 manuscripts from non-English-speaking team, consisting of 6,087 treatment manuscripts and 6,046 control manuscripts. With this subgroup, 795 manuscripts posted a revised version within 30 days of treatment (419 in the treatment group and 376 in the control group), which enables us to examine whether language factors shape the revision process and the nature of reviewer feedback.

We analyzed the parsed XML of each manuscript, which transformed the PDFs into structured, machine-readable strings. See Fig. S2 for the analysis pipeline. From the full XML string of each paper, we focused on the main textual content, explicitly excluding ancillary elements such as figures, tables, and their captions, in order to focus the analysis on the core prose. This was achieved using a custom parser that processed all <sec> (section) elements while skipping any section explicitly titled "Figures." The textual content within each section was recursively extracted to preserve the original structural order, including paragraph text and in-text citations.

4.2 Semantic Block Alignment

To enable fine-grained comparisons between the original and revised versions of each paper, we developed a two-stage process to segment and align corresponding "semantic blocks"—i.e., high-level components such as Introduction, Methods, Results, and Discussion.

We first employed a GPT-4.1 mini-based procedure to compare different versions of the sample paper at a high level. The model was prompted to segment both the original and revised versions into logically matched conceptual blocks, assigning each pair a consistent block_id and title. To minimize token usage, the model returned only the first and last sentence of each block in both versions, rather than the full content. In the second stage, we used these boundary sentences as anchors to locate and extract the corresponding full content from the original and revised texts, using another carefully designed GPT-4.1 mini prompt that accommodates minor variations (if any) on anchor sentences. See SI Section 10.2 for prompt details.

4.3 Content Update Intention Detection

To isolate and analyze substantive scholarly changes, we filtered for semantic blocks containing at least one sentence-level edit classified as Update Content in the previous step.

For each such block, we performed a higher-level intention analysis to categorize the purpose of the content update.

We employed a taxonomy of 11 revision categories introduced by Liang et al. (2024) [2] in their study of LLM-generated feedback on scientific writing. A GPT model was prompted with each block's title, original content, and revised content, and instructed to select all applicable categories from the taxonomy that described the revision, ranking them in order of relevance. The prompt emphasized a focus on meaningful content-level changes and instructed the model to use both section titles and the full semantic context of the text to ensure accurate classification of scholarly intent. See SI Section 10.4 for prompt details.

4.4 Sentence-Level Edit Extraction

For each pair of aligned semantic blocks, we conducted a sentence-level edit analysis. When only one block was present (i.e., either the original or the revised block is empty), we applied the block sentence-splitting prompt (see SI Section 10.3). In such cases, all sentences from the single available block were assigned edits of type Delete (if only the original existed) or Insert (if only the revised existed), and their intention was labeled as Update Content. This ensured that insertions and deletions were captured explicitly rather than treated as empty edits. When both original and revised blocks were non-empty, we instructed the GPT model: (1) Align sentences between versions. (2) If there are edits, extract token-level edits (Insert, Delete, Substitute, Reorder) for modified sentences. (3) Classify the primary intention of each edit using a predefined taxonomy (e.g., Improve Language, Correct Grammar/Typo, Update Content, Adjust Format) [4]. To avoid misclassification, we included explicit instructions to disregard superficial formatting changes such as line-break hyphenations. See SI Section 10.3 for detailed prompt design.

4.5 Average Percentage of Edit Types and Intentions

We calculated the average percentage of edit types and intentions. To quantify the distribution of different edit types for all sentences, we proceeded in two steps. (1) Edit count per paper (normalized). For each edit category c and each paper i, we computed:

$$x_{i,c} = \frac{number\ of\ edits\ in\ category\ c\ for\ paper\ i}{total\ number\ of\ sentence\ pairs\ for\ paper\ i}$$

If a paper has no edits in this category, then $x_{ic} = 0$. Here, the denominator is defined as the total number of sentence pairs for paper i, which includes not only aligned sentences but also sentences that appear only in the original or only in the revised version. This ensures that inserted and deleted sentences are properly represented.

(2) Group-level mean percentage. For each group (Treatment / Control), we then calculated the mean percentage across all papers:

$$x_c^{(group)} = \frac{1}{N} \sum_{i=1}^{N} x_{i,c}$$

Where N is the total number of papers in that group (6,046 for treatment, 6,087 for control). Thus, the reported average percentage of edits for each type represents the mean fraction of that edit type within a paper, averaged across all papers in the group.

We then quantified the distribution of different edit intentions at the block level, and we proceeded with two similar steps. (1) Update content intention count per paper (normalized). For each edit category c and each paper i, we computed:

$$x_{i,c} = \frac{number\ of\ intentions\ in\ category\ c\ for\ paper\ i}{total\ number\ of\ block\ pairs\ for\ paper\ i}$$

If a paper has no intentions in this category, then $x_{ic} = 0$. Here, the denominator is defined as the total number of block pairs for paper i, which includes not only aligned sentences but also blocks that appear only in the original or only in the revised version. (2) Group-level mean percentage. For each group (Treatment / Control), we then calculated the mean percentage across all papers. Thus, the reported average percentage of block-level update content intentions for each type represents the mean fraction of that intention within a paper, averaged across all papers in the group.

5. Measurement of Variables

5.1 Number of Revisions

We collected revision data by downloading an updated version of Cornell ArXiv Dataset from Kaggle. For each paper in our sample, we checked the number of available revisions recorded in the metadata to determine whether and how many times a manuscript had been updated.

5.2 Region Data Collection and English-Speaking Region

Region data collection: To identify author region affiliations, we extracted affiliation and email domain region codes for all 50,454 authors from the PDF files at the time of data collection. Among these, 3,105 entries lacked identifiable region information. To address this, we applied Grobid to perform structured information extraction on the corresponding PDF files for these "unknown region" entries. In parallel, we conducted a manual review of each PDF to supplement and validate the region data. For multinational organizations without an explicit region label, we assigned the location of the organization's headquarters as the author's region of affiliation (e.g., Google is located in the United States, and Huawei Technologies is located in China). In total, we successfully labeled 50,328 individual affiliations with region information. The final dataset contains 126 entries (<0.3%) for which the region could not be determined, which we label as "unknown."

English speaking region: To classify whether an author was affiliated with an English-speaking region, we checked whether the official languages of the country included English. Authors were labeled as from an English-speaking region if their country of affiliation belonged to the following list:

"Australia, Canada, Ireland, Malta, New Zealand, Singapore, South Africa, United Kingdom, United States, India, Pakistan, Nigeria, Philippines, Kenya, Uganda, Tanzania, Ghana, Botswana, Mauritius, Rwanda, Malawi, Trinidad and Tobago, Saint Kitts and Nevis, Hong Kong, British Indian Ocean Territory, Ascension Island, Cocos (Keeling) Islands, Tokelau, Cameroon."

This classification allowed us to assess the role of language background in shaping engagement with and response to the AI-generated feedback.

5.3 Distance to AI

To quantify the topical distance between scientific fields and artificial intelligence (AI), we constructed a field-level semantic distance metric based on co-occurrence patterns in *arXiv* submissions. This metric captures how conceptually proximate a given field is to AI, operationalized as proximity to the cs.AI category on *arXiv*. See Fig. S3 for details.

Data preparation: We collected 16,492 *arXiv* preprints submitted between January and March 2024. Each paper is associated with one or more category labels (e.g., cs.LG, stat.ML). These category labels serve as the basis for constructing a semantic co-occurrence network.

Co-occurrence network construction: We built an undirected field co-occurrence network using category metadata. For each pair of *arXiv* categories, we counted the number of papers in which both categories co-occurred. To reflect the strength of association between categories, we computed the Pointwise Mutual Information (PMI) for each pair and defined the edge weight as the inverse of PMI:

$$Weight_{i,j} = \frac{1}{PMI_{i,j}}$$

This formulation ensures that more semantically similar fields (with high PMI) are connected by shorter, lower-cost edges in the network.

Semantic distance calculation: Using the resulting PMI-weighted network, we computed the shortest-path distance from the cs.AI category to every other category using Dijkstra's algorithm. These distances represent the minimal semantic "path length" from AI to a target field, capturing conceptual proximity in the embedding space formed by field co-occurrence.

5.4 Detection of LLM usage

To estimate the proportion of text in *arXiv* submissions that may have been generated or substantially modified by LLMs, we adopted the distributional quantification framework proposed by Liang et al. (2024) [5]. This approach treats each document as sampled from a mixture of two distributions: human-written text (denoted P) and LLM-generated text (denoted Q). The objective is to infer the mixture weight α, representing the share of LLM usage in the target corpus.

Reference Distributions: Following Liang et al., we constructed reference corpora of human-written and LLM-generated texts to estimate token-level distributions $P(t)$ and $Q(t)$, for a

chosen list of tokens t. The estimation relies on the presence or absence of selected vocabulary tokens (primarily adjectives), which were shown to provide stable signals distinguishing human from LLM outputs.

Maximum Likelihood Estimation (MLE): Given a target document x, its likelihood is modeled as:

$$\mathcal{D}_\alpha(x) = (1-\alpha)P(x) + \alpha Q(x)$$

For a corpus of n documents, the overall log-likelihood is:

$$\alpha_{MLE} = \underset{\alpha\in[0,1]}{argmax} \sum_{i=1}^{n} log\left((1-\alpha)P(x_i) + \alpha Q(x_i)\right)$$

We computed the MLE of α for each corpus or sub-corpus of interest. We then used bootstrap resampling to construct confidence intervals, following the original method.

Modifications: We used a unified reference corpus for *arXiv* categories outside the core STEM fields. For large categories with sufficient sample sizes—namely physics, mathematics, statistics, computer science, and electrical engineering and systems science (EESS)—we trained field-specific reference distributions. For smaller categories outside these core fields, we pooled all available abstracts into a single unified reference corpus. This approach ensured more stable estimation of the human and AI distributions (P, Q) in fields with limited data.

6. Robustness Check

For the main revision outcome, we conducted a series of robustness checks by varying the observation window. We extended the short-term revision analysis shown in Figure 3A by adjusting the outcome window from 4 to 9 weeks (Figure S1). The treatment effect remains consistent in direction across these alternative time windows, with statistically significant effects particularly among non–English-speaking groups.

Similarly, we examined the robustness of the long-term adoption analysis shown in Figure 4 by varying the α detection threshold. Following prior research on α detection algorithms [5], we used a 10% transition as the main specification in the text. To test robustness, we varied the threshold to 0% (Figure S5), 5% (Figure S6), and 20% (Figure S7). The results remain stable, exhibiting consistent patterns and directions across all specifications.

7. Potential Limitations

While our pipeline and analysis follow the widely adopted practice in the field, they are not free of limitations. First, large language model (LLM)-based scientific feedback generation is rapidly evolving. Our pipeline—adapted from Liang et al. (2024)—was built on a non-reasoning model primarily tuned to computational science domains. As such, it may produce somewhat generic comments, particularly for disciplines outside computer science. This limitation likely introduces a conservative bias in our estimates: if more advanced models or domain-specific prompting were used to generate higher-quality, more tailored feedback, the observed treatment effects could plausibly be even larger. Related, we cannot rule out the possibility that some authors (in both treatment and control groups) independently used other LLM-based tools for scientific feedback and improvement, which may also attenuate the estimated effects here.

Second, there is a one- to five-month lag between the time a manuscript is posted on *arXiv* and the delivery of feedback to authors. This delay stems from the logistical challenges of parsing PDFs and generating feedback at scale. In practice, this means that feedback may not reach authors when it is most useful. Future work could explore more timely feedback delivery, which would likely enhance engagement and amplify the treatment effect—again implying that our estimates are conservatively biased.

Third, our analysis considers the intent-to-treat population—i.e., all authors who were sent feedback—rather than restricting attention to those who actually read or engaged with it. A substantial proportion of recipients may not have opened the email or visited the feedback website, introducing non-compliance into the treatment group. This likely attenuates the observed effects; had all treated authors engaged with the feedback, the impact on revision behavior and downstream outcomes could have been even stronger.

8. Figure Appendix

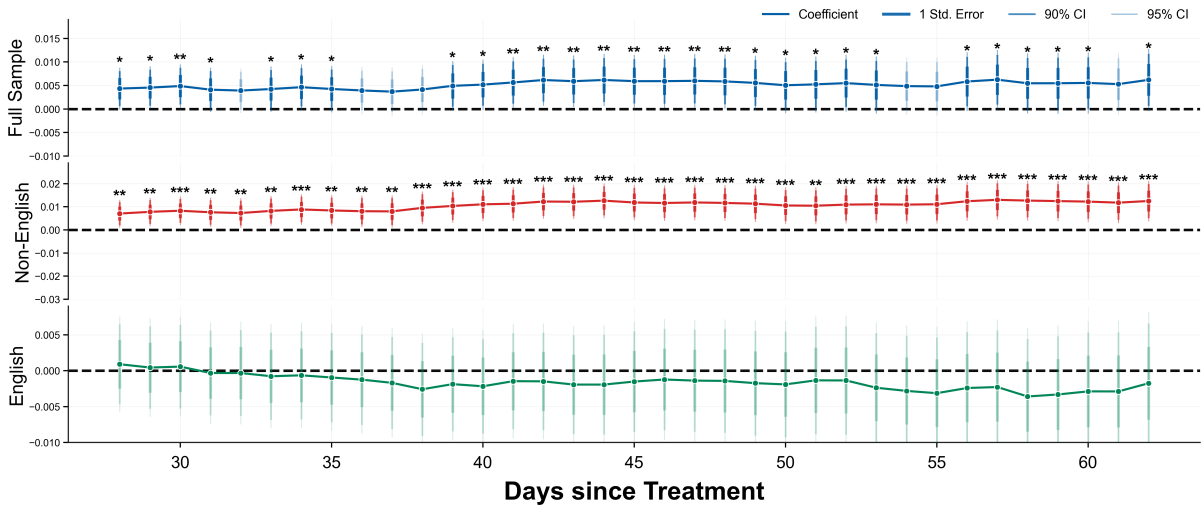


Figure S1. Robustness check of treatment effect on number of revisions. We extended the analysis in Fig. 3a by varying the outcome window from 4 weeks to 9 weeks. The treatment effect remains in the same direction across these alternative time windows.

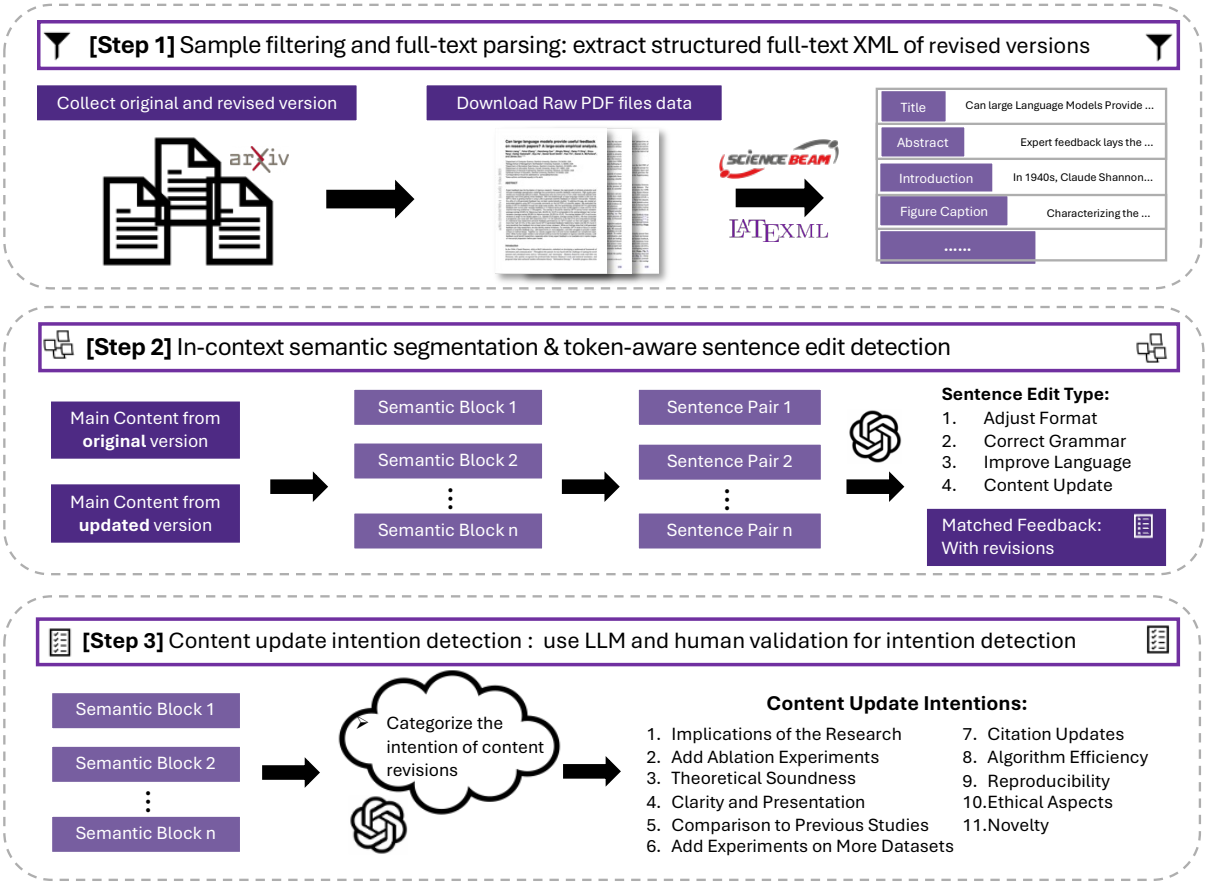


Figure S2. LLM-based edits extraction pipeline. The pipeline consists of three main steps. **(1)** Sample filtering and full-text parsing, where revised papers are identified, raw PDF files are downloaded, and structured XML is extracted for analysis. **(2)** In-context semantic segmentation and token-aware sentence edit detection, where main content from original and updated versions is segmented into semantic blocks, paired at the sentence level, and categorized by edit type (format adjustment, grammar correction, language improvement, or content update). **(3)** Content update intention detection, where LLM models and human validation are used to classify the purpose of each substantive revision into one of 11 predefined categories (e.g., implications of research, theoretical soundness, clarity and presentation, novelty).

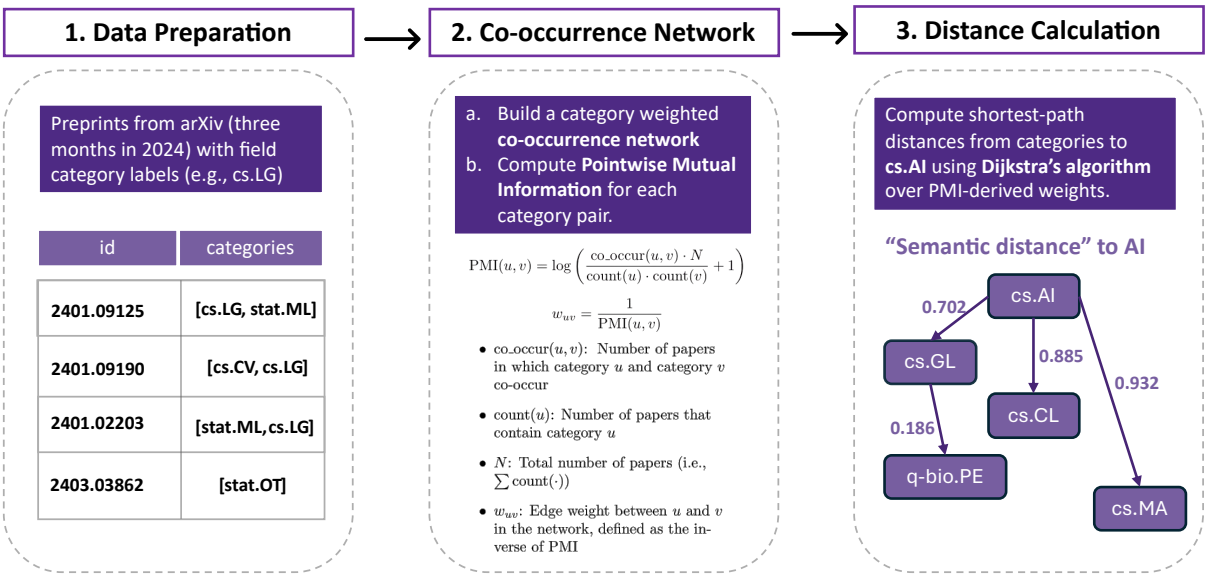


Figure S3. AI distance calculation. The procedure involved three steps: **(1)** Data preparation – We extract 16,492 preprints from *arXiv* with their field category labels (e.g., cs.LG). **(2)** Co-occurrence network construction – We build a weighted category co-occurrence network and compute Pointwise Mutual Information (PMI) for each category pair. Edge weights are defined as the inverse of PMI. **(3)** Distance calculation – We compute the shortest-path distance from each category to cs.AI using Dijkstra's algorithm over the PMI-derived network. The resulting "semantic distance" to AI reflects topical proximity, with smaller values indicating closer alignment to AI.

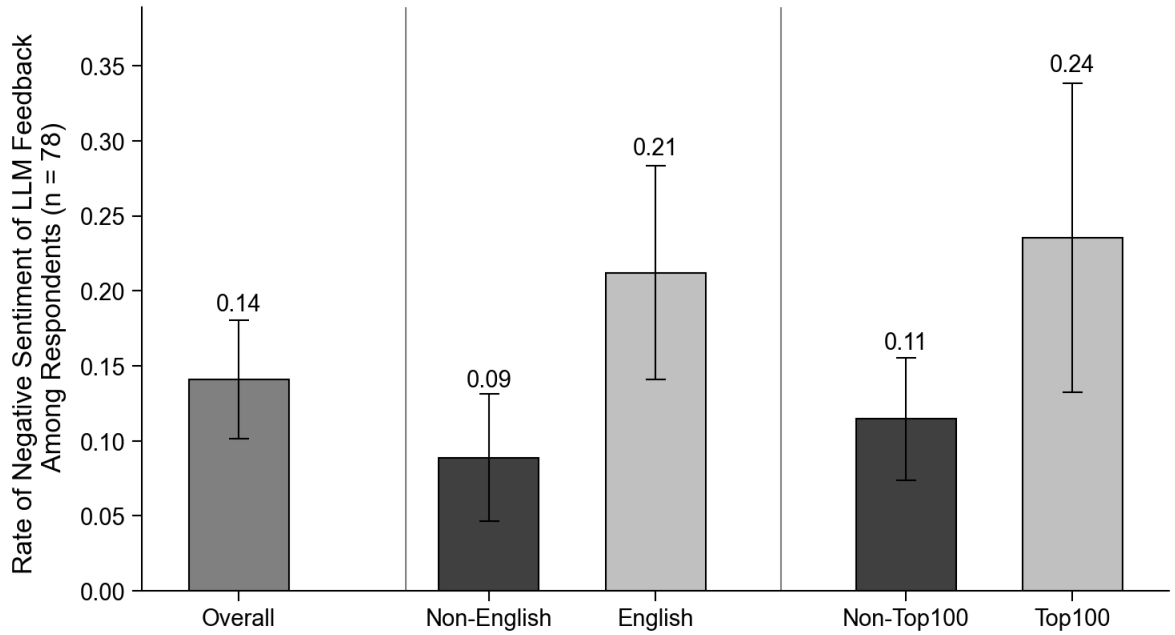


Figure S4. Rate of negative sentiment toward LLM-generated feedback by region and institutional prestige. We coded the institutional affiliation of each respondent (n = 78) along two dimensions: (1) whether the institution is located in an English-speaking region, and (2) whether it ranks within the top 100 globally, based on the Times Higher Education World University Rankings 2025. The proportion of respondents expressing negative sentiment toward the LLM-generated feedback was substantially higher among scholars affiliated with English-speaking (21%) and top 100 institutions (24%), compared to those from non-English-speaking (9%) and non-top 100 institutions (11%). This pattern suggests that researchers from non-English or lower-prestige institutions perceived the feedback as more helpful or constructive—highlighting the potential of LLM-based scientific feedback to reduce inequalities and enhance access for scholars in under-resourced or marginalized contexts.

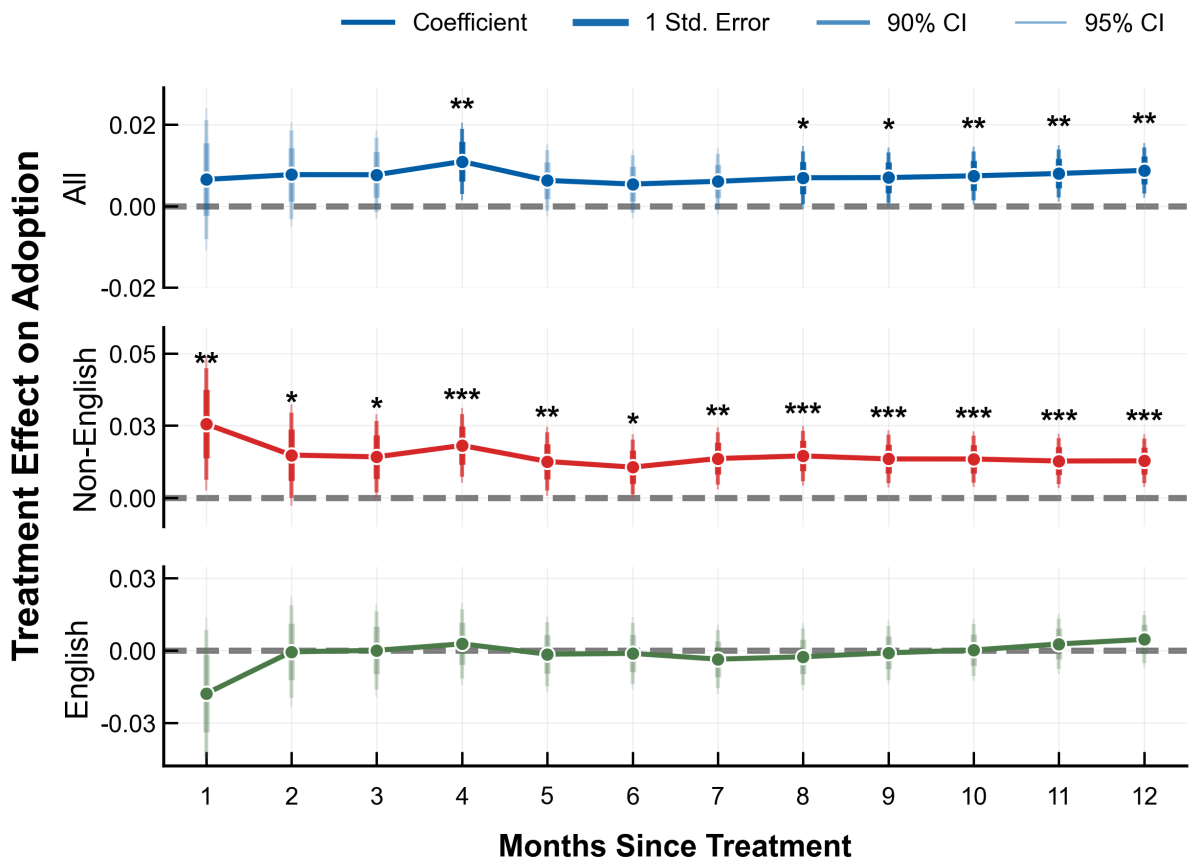


Figure S5. Robustness of adoption results to alternative α threshold (adoption defined as $\alpha_{post} > 0$).

Adoption is redefined as any detectable LLM use in post-treatment publications ($\alpha_{post} > 0$) among authors with minimal prior use ($\alpha_{pre} = 0$). Points show OSL regression estimates of the treatment effect on adoption (overall and by language subgroup); whiskers denote 95% CIs. The pattern mirrors the main specification: a positive overall effect driven by non-English authors, with no detectable change for English-speaking authors.

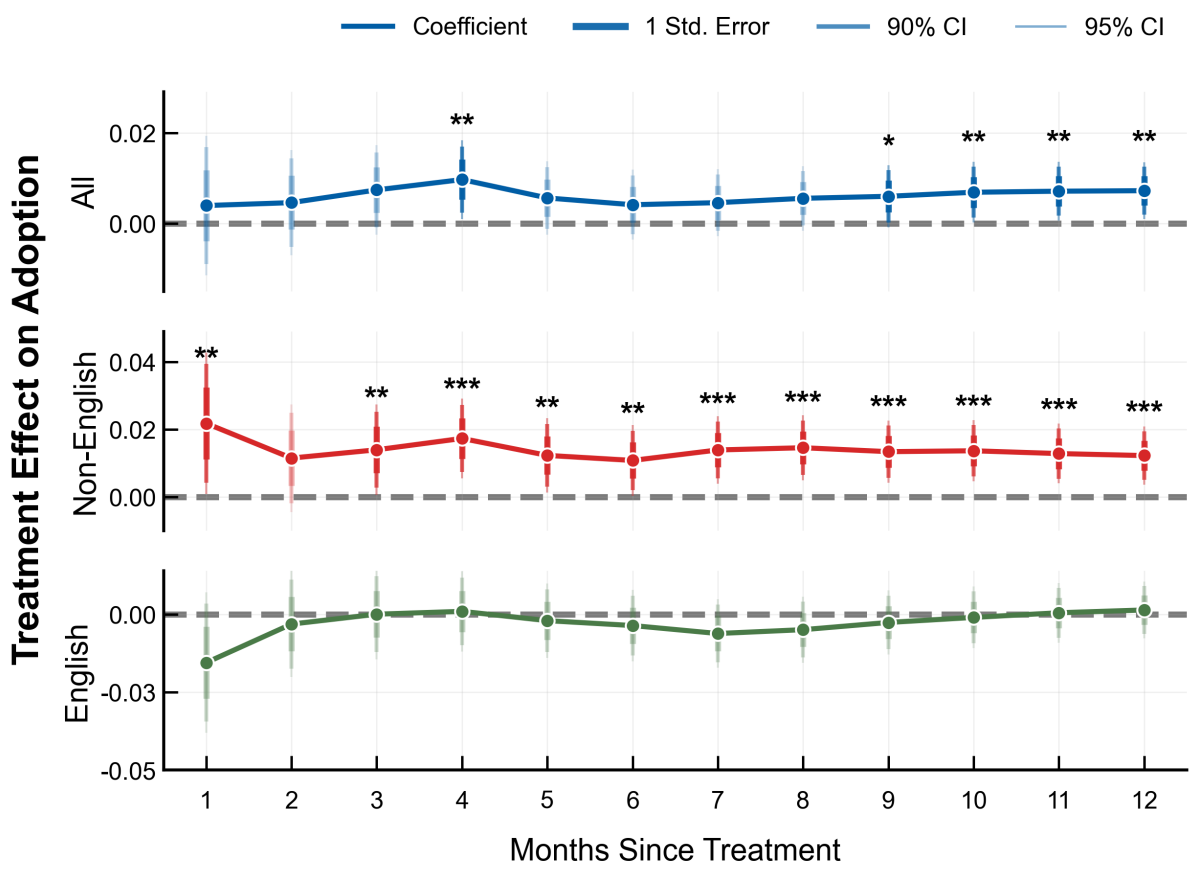


Figure S6. Robustness of adoption results to alternative α threshold (adoption defined as $\alpha_{post} > 0.05$).

Adoption is redefined as any detectable LLM use in post-treatment publications ($\alpha_{post} > 0.05$) among authors with minimal prior use ($\alpha_{pre} \leq 0.05$). Points show OSL regression estimates of the treatment effect on adoption (overall and by language subgroup); whiskers denote 95% CIs. The pattern mirrors the main specification: a positive overall effect driven by non-English authors, with no detectable change for English-speaking authors.

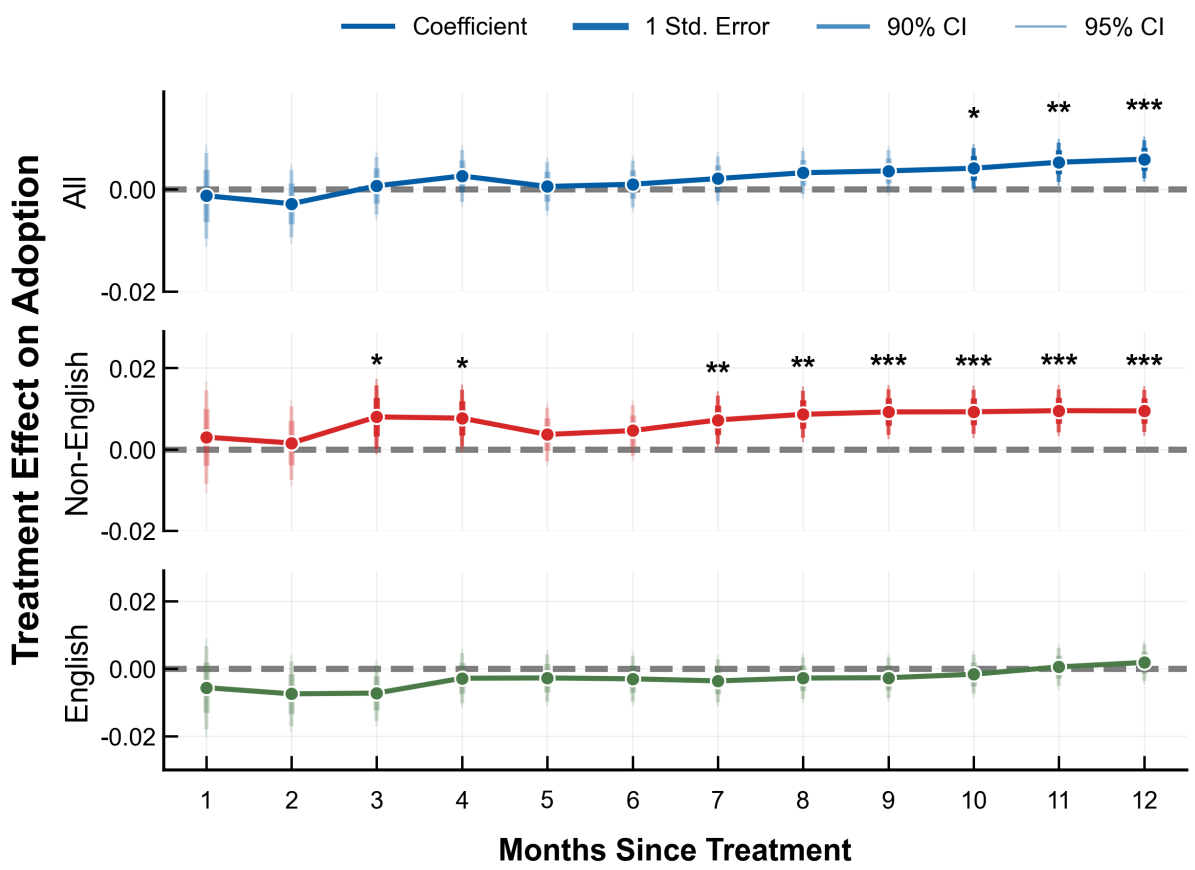


Figure S7. Robustness of adoption results to alternative α threshold (adoption defined as $\alpha_{post} > 0.2$).

Adoption is redefined as any detectable LLM use in post-treatment publications ($\alpha_{post} > 0.2$) among authors with minimal prior use ($\alpha_{pre} \leq 0.2$). Points show OSL regression estimates of the treatment effect on adoption (overall and by language subgroup); whiskers denote 95% CIs. The pattern mirrors the main specification: a positive overall effect driven by non-English authors, with no detectable change for English-speaking authors.

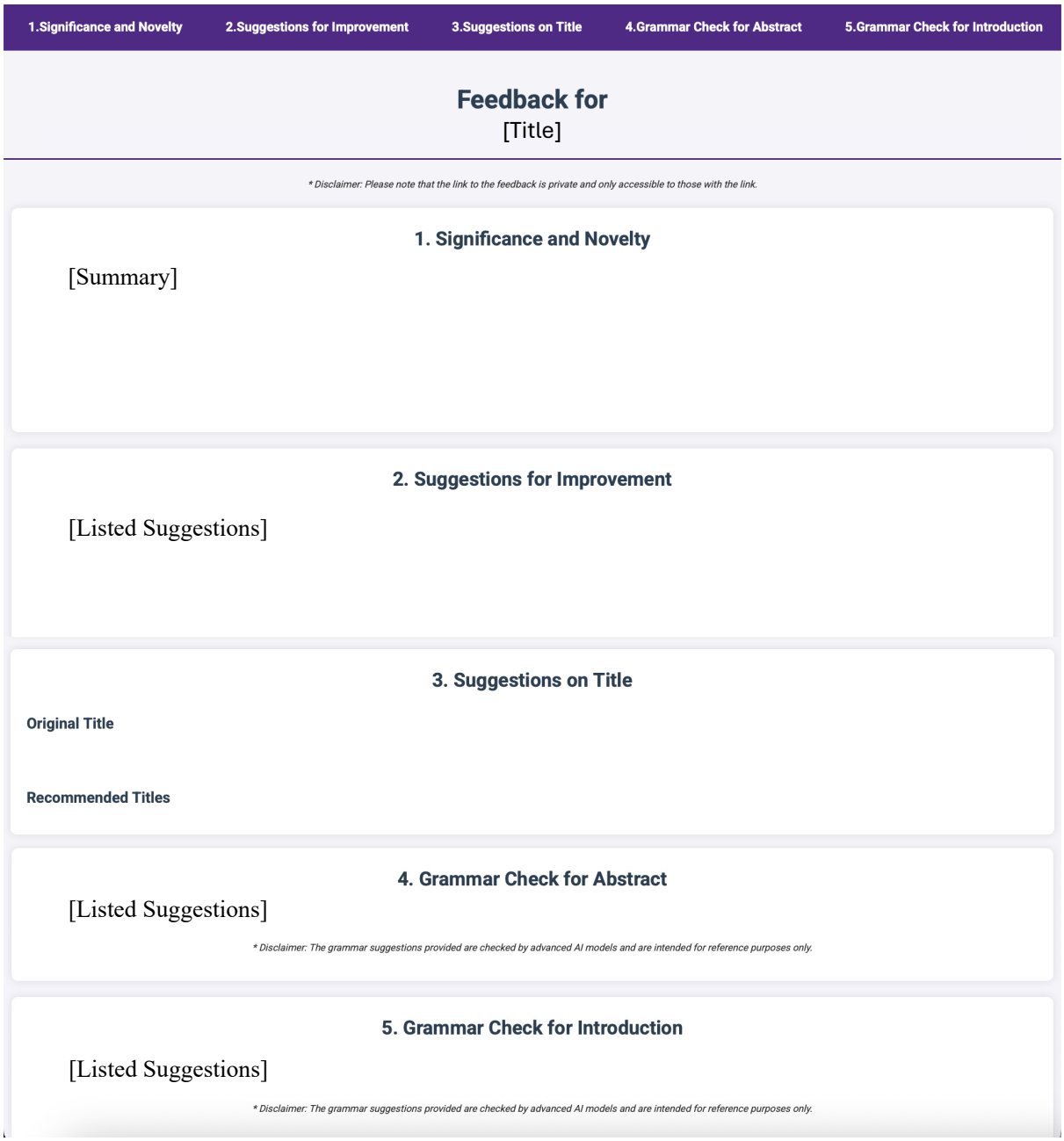


Figure S8. Example of the feedback website interface.

Screenshot of the web interface through which authors in the treatment group accessed their customized AI-generated feedback. The website presents structured comments across five sections: (1) Significance and Novelty, (2) Suggestions for Improvement, (3) Suggestions on Title, (4) Grammar Check for Abstract, and (5) Grammar Check for Introduction. Each section contains detailed suggestions automatically generated by large language models. The interface was private with personalized links provided only to authors via email.

9. Table Appendix

Manuscript-Level Feature	Control Mean	Treatment Mean	Difference	P-value
Team Size	4.657 (5.683)	4.665 (5.047)	0.008	0.897
Distance to AI	1.470 (0.393)	1.469 (0.393)	-0.001	0.882
Reference Count	39.854 (32.353)	39.869 (33.007)	0.015	0.968
English (1=Yes)	0.438 (0.496)	0.445 (0.497)	0.007	0.190
Institution Prestige (1=Yes)	0.037 (0.188)	0.038 (0.191)	0.001	0.582
α (LLM-use probability)	0.197 (0.235)	0.198 (0.236)	0.001	0.763

Table S1. Balance check between treatment (N = 15,542) and control (N = 15,478) groups for manuscript-level features. This table reports baseline characteristics of papers randomly assigned to the treatment or control groups. Means are shown with standard deviations in parentheses. Differences between groups are estimated using two-sided t-tests. No statistically significant differences are observed across key covariates. Means and standard deviations are shown for each group, with the final column reporting two-sample t-statistics. Standard significance thresholds are indicated: $***$ $p < 0.01$, $**$ $p < 0.05$, $*$ $p < 0.1$.

DV: Number of Revisions in a Month	**Full Sample**	**Non-English-Speaking**
Treatment Effect	**0.005****	**0.008*****
	(0.002)	(0.003)
Treat Time FE	Yes	Yes
Observations	31,020	17,265
Pseudo R^2	0.000	0.000

Table S2. Treatment effects on number of revisions. Estimates are from linear regression models with treatment time fixed effects. Columns 1–2 present results for the full sample and non-English-speaking groups. Treatment increases the number of revisions by 12.5% for the full sample, and 22.4% for non-English groups. Coefficients are reported with standard errors in parentheses. Standard significance thresholds are indicated: $*** p < 0.01$, $** p < 0.05$, $* p < 0.1$.

Category	Control Mean	Treatment Mean	t-stat	p-value
Adjust Format	0.336	0.413	1.631	0.1028
Improve Language	1.218	1.150	-0.500	0.6171
Correct Grammar/Typo	0.479	0.498	0.273	0.7845
Update Content	3.310	4.021	1.686	**0.0919***

Table S3. Average percentage of revised sentence in each edit types between treatment and control groups (non-English-speaking group). This table reports the percentage of sentence-level revisions within one month after treatment for the non-English-speaking group. Means for each revision category are shown separately for treatment and control groups, followed by difference tests (t-statistics and p-values). Standard significance thresholds are indicated: $***\ p < 0.01$, $**\ p < 0.05$, $*\ p < 0.1$.

Category	Control Mean	Treatment Mean	t-stat	p-value
Add Ablation Experiments	0.013	0.012	-0.355	0.7229
Add Experiments on More Datasets	0.022	0.023	0.438	0.6617
Algorithm Efficiency	0.034	0.037	0.530	0.5960
Citation Updates	0.101	0.110	0.647	0.5174
Clarity and Presentation	0.100	0.110	0.867	0.3862
Comparison to Previous Studies	0.177	0.204	1.451	0.1469
Ethical Aspects	0.008	0.009	0.863	0.3880
Implications of the Research	0.071	0.073	0.240	0.8102
Novelty	0.023	0.038	2.278	**0.0227****
Reproducibility	0.184	0.206	1.145	0.2522
Theoretical Soundness	0.045	0.060	1.690	**0.0911***

Table S4. Average percentage of revised blocks in each edit intentions between treatment and control groups (non-English-speaking group). This table reports the likelihood of block-level revision intentions within one month after treatment for the non-English-speaking group. Means for treatment and control groups are shown separately, followed by the difference test (t-statistics and p-values). Standard significance thresholds are indicated: $***$ $p < 0.01$, $**$ $p < 0.05$, $*$ $p < 0.1$.

DV: Number of Revisions in a Month	**Non-English-Speaking**	**English-Speaking**
Treatment Effect	**0.008*** **	0.001
	(0.003)	(0.004)
Treat Time FE	Yes	Yes
Observations	17,265	13755
Pseudo R^2	0.000	0.000

Table S5. Treatment effects on number of revisions by language. Estimates are from linear regression models with treatment time fixed effects. Columns 1–3 present results for the non-English-speaking group, English-speaking group, respectively. Treatment increases the number of revisions by 22.4% for non-English-speaking authors but has no significant effect for English-speaking authors. Coefficients are reported with standard errors in parentheses. Standard significance thresholds are indicated: $*** p < 0.01$, $** p < 0.05$, $* p < 0.1$.

Region	Coefficient	Sample Size	p-value	English-speaking
Israel	**0.0604* (0.0330)**	302	0.0677	No
South Korea	**0.0366* (0.0213)**	532	0.0864	No
Switzerland	0.0262 (0.0185)	551	0.1588	No
Brazil	0.0215 (0.0196)	302	0.2737	No
China	**0.0120* (0.0063)**	4,841	0.0566	No
Sweden	0.0120 (0.0265)	315	0.6512	No
Germany	0.0112 (0.0092)	2,157	0.2238	No
Japan	0.0092 (0.0125)	1,026	0.4637	No
Italy	0.0073 (0.0121)	1,112	0.5456	No
United States	0.0039 (0.0046)	8,125	0.3943	Yes
Australia	0.0021 (0.0157)	590	0.895	Yes
Hong Kong	−0.0019 (0.0212)	346	0.9271	Yes
France	−0.0037 (0.0117)	1,118	0.7507	No
Canada	−0.0047 (0.0128)	993	0.7157	Yes
India	−0.0058 (0.0120)	1,256	0.6297	Yes
Spain	−0.0098 (0.0122)	582	0.4219	No
United Kingdom	−0.0103 (0.0101)	1,811	0.3089	Yes
Netherlands	−0.0209 (0.0168)	510	0.2144	No

Table S6. Treatment effects on number of revisions by region. Region-level linear regression estimates of the effect of LLM-generated feedback on the probability that a preprint is revised

within 30 days. Estimates indicate larger positive effects in several non-English-speaking contexts (e.g., South Korea, Israel), with smaller or null effects in English-speaking settings. Coefficients are log-odds with standard errors. Coefficients are reported with standard errors in parentheses. Standard significance thresholds are indicated: $***$ $p < 0.01$, $**$ $p < 0.05$, $*$ $p < 0.1$.

DV: Number of Revisions in a Month	**Low Distance**	**Medium Distance**	**High Distance**
Treatment Effect	0.002	**0.012*** **	0.001
	(0.004)	(0.004)	(0.004)
Treat Time FE	Yes	Yes	Yes
Observations	10497	10197	10326
Pseudo R^2	0.000	0.002	0.001

Table S7. Treatment effects on number of revisions by distance tertile to AI. Linear regression estimates are reported for each distance tertile (low, medium, high). Distance tertiles are based on the distribution of topical distance to *cs.AI* in the *arXiv* subject co-appearance network. The treatment effect is largest and statistically significant for medium-distance papers (0.012***), with no significant effect for low- or high-distance papers. Coefficients are reported with standard errors in parentheses. Standard significance thresholds are indicated: $***$ $p < 0.01$, $**$ $p < 0.05$, $*$ $p < 0.1$.

DV: Adoption in Twelve-Month (Binary)	**Full Sample**	**Non-English-Speaking**	**English-Speaking**
Treatment Effect	**0.0072**** (0.0032)	**0.0122***** (0.0044)	0.0014 (0.0047)
Treat Time FE	Yes	Yes	Yes
Observations	9,368	5,021	4,347
Pseudo R^2	0.001	0.002	0.000

Table S8. Treatment effects on LLM adoption. Estimates are from logistic regression models with treatment time fixed effects. Columns 1–3 report results for the full sample, non-English-speaking authors, and English-speaking authors, respectively. The full sample includes authors who have no detectable LLM use in pre-treatment writing ($\alpha \leq 0.05$) and have at least one preprint on *arXiv* post-intervention. Treatment increases the likelihood of AI adoption by 9.56 percentage points for non-English-speaking authors, while no significant effect is observed for English-speaking authors. Coefficients are reported with standard errors in parentheses. Standard significance thresholds are indicated: $***$ $p < 0.01$, $**$ $p < 0.05$, $*$ $p < 0.1$.

10. Prompt Appendix

10.1 Feedback Generation

```
review_prompt = {

"system": """Your task now is to draft a high-quality review outline for a top-tier publication venue in their field.
======
Your task:
Compose a high-quality peer review of a high-quality review outline for a top-tier publication venue in their field.
Start by "Review outline:".
And then:
"1. Significance and novelty", in bullet points.
"2. Suggestions for improvement", List multiple key critical suggestions. Always provide the context and summarize the issue before you get deep into your answer, but always start by acknowledging the good moves made by the authors first. Be as specific and detailed as possible.
Strive to provide insightful, non-obvious recommendations that can genuinely contribute to enhancing the paper. Think critically about the claims in the text and how its elements can be improved within the scope of the provided excerpt.

"""
}

title_suggestion_prompt = {
"system": """**Your Task**: You are provided with the content and title of a target academic paper in the field of artificial intelligence. Your objective is to suggest a set of five enhanced titles for this paper, aiming to amplify its academic impact while respecting the essence and intent of the original title. Use a dataset of titles from NeurIPS conference papers as your reference.

Begin by analyzing the paper's content and original title. Then, identify the top 5 NeurIPS paper titles that are most similar in theme, methodology, or findings, assigning a similarity score from 1 to 5 (5 being the most similar) to these titles. For each reference title, provide brief reasoning explaining its relevance or inspiration in relation to the target paper.

Based on these reference titles and their reasonings, propose five improved titles for the target paper. Accompany each proposed title with a rationale focusing on how it could elevate the paper's academic impact, considering aspects like clarity, engagement, and relevance.

Your output should be in JSON format, following this schema:
```
{
  "original_title": "string",   // The original title of the given paper
  "reference_titles": [// List of top 5 similar titles from NeurIPS
    {
      "title": "string",   // Title of the reference paper
      "similarity_score": "int",   // Score (1-5) for similarity to the original paper
      "reasoning": "string"   // Reasoning for its relevance (1 sentence) and inspiration on improving our paper title (2-3 sentences)
    },
    // ... 4 more reference titles with reasonings
  ],
  "improved_titles": [// List of 5 proposed improved titles
    {
      "title": "string",   // Proposed improved title
      "reasoning": "string"   // Reasoning for this title's enhancement potential
    },
    // ... 4 more improved titles with reasonings
  ]
}
```
Ensure that each proposed title and its accompanying reasoning reflect careful consideration of how to enhance the paper's appeal and relevance within the academic community, while staying true to the original title's core meaning.

Improve title clarity, engagement, and relevance.
Output JSON only.
""",
    "user": """Reference NeurIPS papers
```
{List of high-quality titles and associated keywords}
```
Target paper
```
{}
{}
""",
}
```
```

```
grammar_suggestion_prompt = {
    "system": """Your task is to proofread the provided sentence for specific grammatical accuracies, focusing on errors that commonly affect the clarity and
    professionalism of the text. Note that errors related to PDF parsing/OCR issues, such as incorrect line breaks or misinterpreted characters, should not be
    reported. Concentrate on identifying and correcting the following types of errors:

- **Spelling Errors:** Correct any misspellings.
- **Subject-Verb Agreement Errors:** Fix errors where the subject and verb do not agree.
- **Incorrect Word Usage:** Amend instances of incorrect word usage.
- **Run-On Sentences and Comma Splices:** Address improper connections between sentences.
- **Misplaced or Dangling Modifiers:** Correct misplaced or dangling modifiers.
- **Sentence Fragment:** Ensure sentences are complete and convey a complete thought.

For each identified error, provide a detailed correction in the following JSON Lines format, where each line represents a separate grammatical error and its
correction:

```jsonl
{"ErrorType": "Type of error", "Description": "Brief description of the error", "OriginalSentence": "The original sentence containing the error",
"CorrectedFragment": "The corrected sentence fragment", "Rating": "<1-10 Error Severity Levels>"}
{"ErrorType": "Type of error", "Description": "Brief description of the error", "OriginalSentence": "The original sentence containing the error",
"CorrectedFragment": "The corrected sentence fragment", "Rating": "<1-10 Error Severity Levels>"}
...
```

If no specific errors are present, it is acceptable to report that no errors were found.

Ensure minimal distortion to the original content while making the necessary grammatical corrections, maintaining the original meaning and style of the sentence
as much as possible.

Output JSONL only.
"""
}
```

10.2 Block-Level Alignment Extraction

```
block_identification_prompt = {
"""
You are a scientific editing analyst.

You are given two versions of the same paper: an original version and a revised version.

Please segment both versions **into matching semantic blocks**, such that:
- The same conceptual content (e.g., introduction, methods, experiments, conclusions) in both versions receives the **same block_id and block title**.
- If a block is **newly added** or **deleted**, it still gets its own block_id, and one side's content will be marked as missing.

For each block, return the following structure:

[
  {{
    "block_id": 1,
    "title": "Introduction",
    "original_first_sentence": "...", #original_block
    "original_last_sentence": "...",
    "revised_first_sentence": "...",  #revised_block
    "revised_last_sentence": "..."
  }},
  ...
]

---

Original Version:
{edit_before}

---

Revised Version:
{edit_after}

---

Notes:
- Do not return full paragraph content.
- Each sentence should be complete and come from the actual content.
- If a block does not exist in one version, leave that side's sentence fields as empty strings.
- Segment conservatively to keep blocks meaningful but not too short.

Now analyze and return the JSON list of semantic blocks comparing the two versions.
"""
}
```

```
full_block_extraction_prompt = {
"""
You are an intelligent assistant skilled at extracting structured content from raw text.

Your task: extract the complete block content from the provided full text, using the given first and last sentences as boundaries.

Full text content:
{full_text}

First sentence:
{first_sentence}

Last sentence:
{last_sentence}

Instructions:
1. Normally, extract everything from the **first sentence** through to the **last sentence** (inclusive).
2. The given first and last sentences may not be exactly identical to the text in the full content.
   - In this case, find the **most similar matching sentences** in the full text.
   - Ensure that the first sentence appears **before** the last sentence in the extracted content.
3. If the **first and last sentences are identical (or nearly identical) **, this usually means they mark a **block title** rather than regular content.
   - In this case, you must return the **entire block content corresponding to that title** (not just the title line).
4. If you cannot find the sentences in the text, return an empty string.

Please output only the extracted block content without extra commentary.
"""
}
```

10.3 Sentence-Level Edit Extraction

```
Sentence_split_prompt = {
"""
You are an expert in scientific writing sentence segmentation.

Your task is to split the following text into individual sentences.

Rules:
- Split the text at natural sentence boundaries
- Each sentence should be a complete thought
- Preserve the original text exactly - do not modify, correct, or rephrase anything
- Return ONLY a valid Python list of strings, with each string being one sentence
- Do not include any explanations or commentary

Text to split:
{text.strip()}

Output format example:
["This is the first sentence.", "This is the second sentence.", "This is the third sentence."]
"""
}
```

```
Sentence_edit_prompt = {
"""
You are an expert in scientific writing revision analysis.
You will be given an original version and a revised version of a paragraph. Your task is to perform sentence-by-sentence alignment between the two versions.
   ---
   **Instructions:**
   1. Split both versions into individual sentences
   2. Align sentences one-to-one as much as possible
   3. For each aligned pair, determine if there are any edits:

   **Alignment Rules:**
   - **Identical sentences**: If sentences are exactly the same, set `"edits": []`
   - **Modified sentences**: If a sentence exists in both versions but has changes, align them and extract the specific edits
   - **Inserted sentences**: If a sentence appears only in the revised version, set `"sentence-1": ""`
   - **Deleted sentences**: If a sentence appears only in the original version, set `"sentence-2": ""`

   **Edit Types:**
   - **Insert**: Adding new text
   - **Delete**: Removing text
   - **Substitute**: Replacing text with different text
   - **Reorder**: Same content, different word order

   **Edit Intentions:**
   - **Improve Language**: Enhancing clarity, style, or readability
   - **Correct Grammar/Typo**: Fixing grammatical errors or typos
   - **Update Content**: Adding, removing, or changing factual information
   - **Adjust Format**: Modifying formatting, punctuation, or structure

   **Important Notes:**
   - Ignore line-break hyphenation (e.g., "multi- field" should be treated as "multifield")
   - Keep genuine hyphens (e.g., "non-zero", "field-dependent")
   - Every sentence from both versions must appear in the output
   - Output must be a valid Python list of dictionaries
   ---
   **Output Format:**
   [
       {{
           "sentence-1": "<original sentence or empty string if inserted>",
           "sentence-2": "<revised sentence or empty string if deleted>",
           "edits": [
               {{
                   "type": "<Insert|Delete|Substitute|Reorder>",
                   "intention": "<Improve Language|Correct Grammar/Typo|Update Content|Adjust Format>",
                   "original_phrase": "<phrase from sentence-1 or empty if inserted>",
                   "revised_phrase": "<phrase from sentence-2 or empty if deleted>"
               }}
           ]
       }}
   ]

   ---
   Original Version:
   {original}
   ---
   Revised Version:
   {revised}
   ---
   Please output the result as a valid Python list of dictionaries representing the sentence alignments and edits.
"""
}
```

10.4 Content Update Intention Detection

```
Intention_detection_prompt = {
"""
You are an expert in analyzing revisions in academic writing.

Your task is to categorize the content revisions made in academic manuscripts. For each revision, I will provide:

1. **Title of the Block** – The section title (e.g., "Introduction", "Method", etc.).
2. **Original Semantic Block** – The original version of the content (may be empty).
3. **Revised Semantic Block** – The updated version after the revision.

If the **original block is empty**, treat the revised block as a **newly added section**. Focus your analysis exclusively on **content-level changes**, not on grammar, style, or tone. Take into account both the **section title** and the **full meaning** of the revised block. Choose **all applicable categories** from the list below that best describe the revised content, and **rank them in order of importance**, placing the most central category first.

**Valid Categories and Definitions:**

1. **Implications of the Research** – Emphasizes the broader impact, future applications, or significance of the findings.
2. **Add Experiments on More Datasets** – Adds experiments using additional datasets to improve generalizability or robustness.
3. **Ethical Aspects** – Discusses or modifies ethical issues (e.g., privacy, fairness, consent).
4. **Algorithm Efficiency** – Improves or suggests improvements in performance, scalability, or computation.
5. **Reproducibility** – Adds sufficient methodological detail to support replication by others. This includes clarifying experimental procedures, data preprocessing steps, model configurations, hyperparameters, training protocols, or linking to datasets or code repositories.
6. **Comparison to Previous Studies** – Compares the current work with prior research in terms of **performance, methodology, assumptions, or results**, highlighting improvements, differences, or consistency with baseline models or related methods.
7. **Theoretical Soundness** – Strengthens or changes the theoretical model, assumptions, or framework.
8. **Citation Updates** – Adds, removes, or modifies references, often seen in literature review.
9. **Add Ablation Experiments** – Introduces or discusses controlled experiments where specific components (e.g., model modules, input features, optimization strategies) are selectively modified—such as removed, disabled, simplified, or altered—to assess their individual contribution to model performance.
10. **Novelty** – Highlights or adds new, original ideas or contributions.
11. **Clarity and Presentation** – Purely stylistic or grammatical improvements, without content changes.

**Output Format:**

Return only the **category names** (e.g., `Implications of the Research, Novelty`). Separate multiple categories with commas. Do **not** include explanations or extra text.

Title of the Block: {title}
Original Semantic Block: {original_block}
Revised Semantic Block: {revised_block}
"""
}
```

11. Reference

1. Tkaczyk, D., et al. *Cermine--automatic extraction of metadata and references from scientific literature*. in *2014 11th IAPR international workshop on document analysis systems*. 2014. IEEE.
2. Liang, W., et al., *Can large language models provide useful feedback on research papers? A large-scale empirical analysis.* NEJM AI, 2024. **1**(8): p. AIoa2400196.
3. Ecer, D. and G. Maciocci, *ScienceBeam-using computer vision to extract PDF data.* Elife [Internet], 2017. **4**.
4. Jiang, C., W. Xu, and S. Stevens, *arXivEdits: Understanding the Human Revision Process in Scientific Writing.* In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. , 2022: p. 9420–9435.
5. Liang, W., et al., *Quantifying large language model usage in scientific papers.* Nature Human Behaviour, 2025: p. 1-11.