NEJM AI

NEJM AI 2024; 1 (8)
DOI: 10.1056/AIoa2400196

ORIGINAL ARTICLE

Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis

Weixin Liang, M.S.,[1] Yuhui Zhang, M.S.,[1] Hancheng Cao, Ph.D.,[1] Binglu Wang, M.S.,[2] Daisy Yi Ding, M.S.,[3] Xinyu Yang, B.E.,[4] Kailas Vodrahalli, M.S.,[5] Siyu He, Ph.D.,[3] Daniel Scott Smith, Ph.D.,[6] Yian Yin, Ph.D.,[4] Daniel A. McFarland, Ph.D.,[6] and James Zou, Ph.D.[1,3,5]

Received: February 22, 2024; Revised: May 17, 2024; Accepted: May 28, 2024; Published: July 17, 2024

Abstract

BACKGROUND Expert feedback lays the foundation of rigorous research. However, the rapid growth of scholarly production challenges the conventional scientific feedback mechanisms. High-quality peer reviews are increasingly difficult to obtain.

METHODS We created an automated pipeline using Generative Pretrained Transformer 4 (GPT-4) to provide comments on scientific papers. We evaluated the quality of GPT-4's feedback through two large-scale studies. We first quantitatively compared GPT-4's generated feedback with human peer reviewers' feedback in general scientific papers from 15 *Nature* family journals (3096 papers in total) and the International Conference on Learning Representations (ICLR) machine learning conference (1709 papers). To specifically assess GPT-4's performance on biomedical papers, we also analyzed a subset of 425 health sciences papers from the *Nature* portfolio and a random sample of 666 submissions to *eLife*. Additionally, we conducted a prospective user study with 308 researchers from 110 institutions in the fields of artificial intelligence and computational biology to understand how researchers perceive feedback generated by our system on their own papers.

RESULTS The overlap in the points raised by GPT-4 and by human reviewers (average overlap of 30.85% for *Nature* journals and 39.23% for ICLR) is comparable with the overlap between two human reviewers (average overlap of 28.58% for *Nature* journals and 35.25% for ICLR). Results on *eLife* and a subset of health sciences papers as categorized by the *Nature* portfolio show similar patterns. In our prospective user study, more than half (57.4%) of the users found GPT-4–generated feedback helpful/very helpful, and 82.4% found it more beneficial than feedback from at least some human reviewers. We also identify several limitations of large language model (LLM)–generated feedback.

Mr. Liang, Mr. Zhang, and Dr. Cao contributed equally to this article.

The author affiliations are listed at the end of the article.

Dr. Zou can be contacted at jamesz@stanford.edu or at 350 Jane Stanford Way, Room 369, Stanford, CA 94305.

CONCLUSIONS Through both retrospective and prospective evaluation, we find substantial overlap between LLM and human feedback as well as positive user perceptions regarding the usefulness of LLM feedback. Although human expert review should continue to be the foundation of the scientific process, LLM feedback could benefit researchers, especially when timely expert feedback is not available and in earlier stages of manuscript preparation. (Funded by the Chan–Zuckerberg Initiative and the Stanford Interdisciplinary Graduate Fellowship.)

Introduction

Effective feedback among peer scientists not only elucidates and promotes the way that new discoveries are made, interpreted, and communicated, but it also catalyzes the emergence of new scientific paradigms by connecting individual insights, coordinating concurrent lines of thought, and stimulating constructive debates and disagreement.[1] However, the process of providing timely, comprehensive, and insightful feedback on scientific research is often laborious, resource intensive, and complex.[2] This complexity is exacerbated by the exponential growth in scholarly publications and the deepening specialization of scientific knowledge.[3,4] Traditional avenues, such as peer review and conference discussions, exhibit constraints in scalability, expertise accessibility, and promptness. For instance, it has been estimated that peer review — one of the major channels of scientific feedback — costs over 100 million researcher hours and $2.5 billion in a single year.[5] Yet, at the same time, it has been increasingly challenging to secure enough qualified reviewers who can provide high-quality feedback given the rapid growth in the number of submissions.[6-10] For example, the number of submissions to the International Conference on Learning Representations (ICLR) machine learning conference increased from 960 in 2018 to 4966 in 2023.

Whereas a shortage of high-quality feedback presents a fundamental constraint on the sustainable growth of science overall, it also becomes a source of deepening global scientific inequities. Marginalized researchers, especially those from nonelite institutions or resource-limited regions, often face disproportionate challenges in accessing valuable feedback, perpetuating a cycle of systemic scientific inequities.[11,12] Given these challenges, there is an urgent need for scalable and efficient feedback mechanisms that can enrich and streamline the scientific feedback process. Such advancements hold the promise of elevating the quality and scope of scientific research.[13,14]

Large language models (LLMs)[15-17] have opened up great potential in various applications.[18-21] Although LLMs have made remarkable strides in various domains, the promises and perils of leveraging LLMs for scientific feedback remain largely unknown. Despite recent attempts that explore the potential uses of such tools in areas such as automating paper screening,[22] error identification,[23] and checklist verification,[1,24] we lack large-scale empirical evidence on whether and how LLMs may be used to facilitate scientific feedback and augment current academic practices.

In this work, we present the first large-scale systematic analysis characterizing the potential reliability and credibility of leveraging LLMs for generating scientific feedback. Specifically, we developed a Generative Pretrained Transformer 4 (GPT-4)–based scientific feedback generation pipeline that takes the raw PDF of a paper and produces structured feedback (Fig. 1A and Supplementary Methods in the Supplementary Appendix). The system is designed to generate constructive feedback across various key aspects, mirroring the review structure of leading interdisciplinary journals[25,26] and conferences,[27-31] including significance and novelty, potential reasons for acceptance, potential reasons for rejection, and suggestions for improvement. To characterize the informativeness of GPT-4–generated feedback, we conducted both a retrospective analysis and a prospective user study.

Methods

RETROSPECTIVE EVALUATION

To evaluate the quality of LLM feedback retrospectively, we systematically assessed the content overlap between human feedback given to submitted manuscripts and the LLM feedback using two large-scale datasets. The first dataset, sourced from *Nature* family journals, includes 8745 comments from human reviewers for 3096 accepted papers across 15 *Nature* family journals, including *Nature*, *Nature Biomedical Engineering*, *Nature Human Behavior*, and *Nature Communications* (Supplementary Methods and Tables S1 and S4). The second dataset comprises 6505 comments from human reviewers for 1709 papers from the ICLR, a leading venue for artificial intelligence (AI) research (Supplementary Methods and Tables S2 and S4). These two datasets complement each other. The first

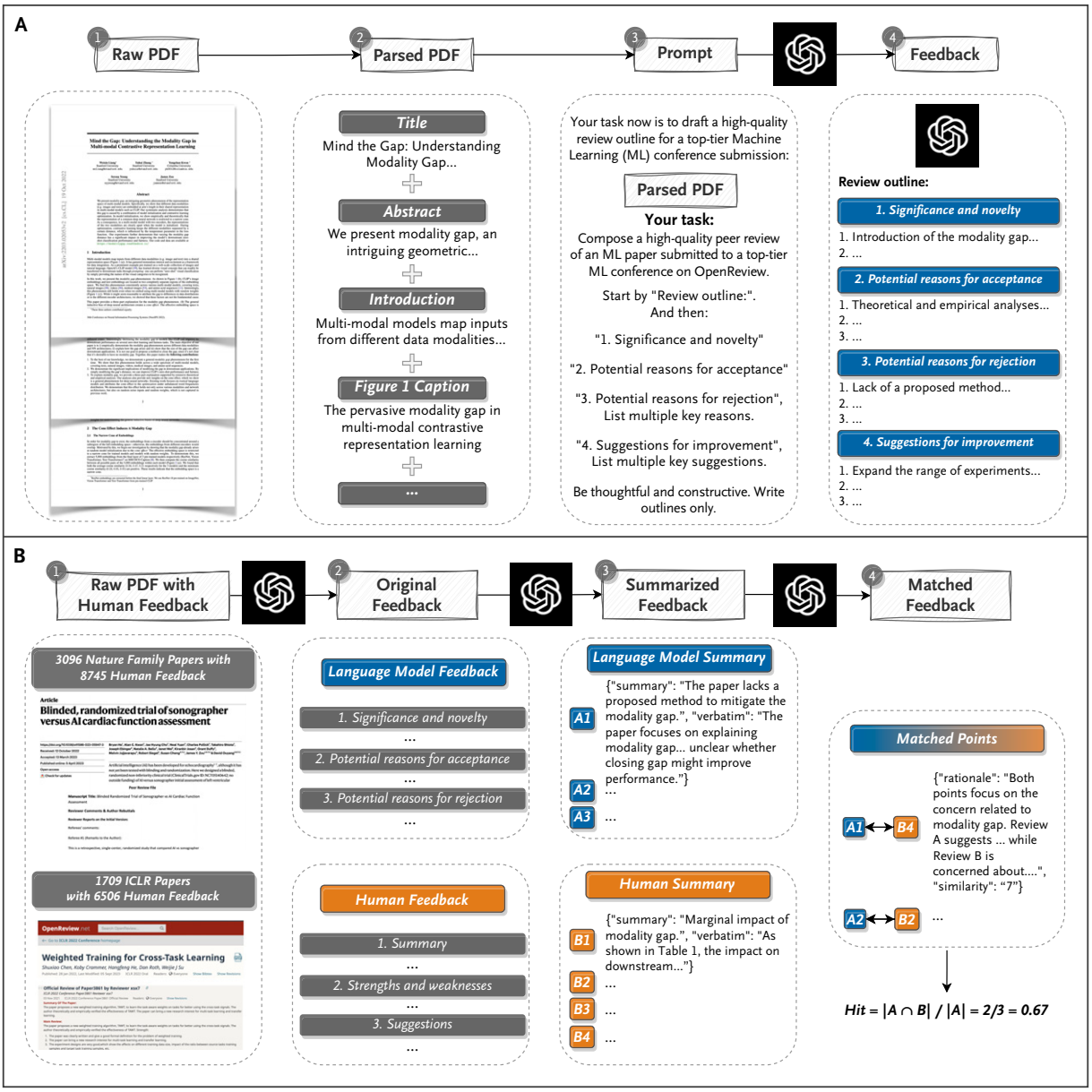

Figure 1. Characterizing the Capability of LLMs in Providing Helpful Feedback to Researchers.

Panel A shows the pipeline for generating LLM scientific feedback using Generative Pretrained Transformer 4 (GPT-4). Given a PDF, we parse and extract the paper's title, abstract, figure and table captions, and main text to construct the prompt. We then prompt GPT-4 to provide structured comments with four sections following the feedback structure of leading interdisciplinary journals and conferences: significance and novelty, potential reasons for acceptance, potential reasons for rejection, and suggestions for improvement. Panel B shows retrospective analysis of LLM feedback on 3096 *Nature* family papers and 1709 International Conference on Learning Representations (ICLR) papers. We systematically compare LLM feedback with human feedback using a two-stage comment-matching pipeline. The pipeline first performs extractive text summarization to extract the points of comments raised in LLM and human-written feedback, respectively, and then, it performs semantic text matching to match the points of shared comments between LLM feedback and human feedback. Panel C shows a prospective user study survey with 308 researchers from 110 U.S. institutions in the field of artificial intelligence (AI) and computational biology. Each researcher uploaded a paper they authored and filled out a survey on the LLM feedback generated for them. LLM denotes large language model.

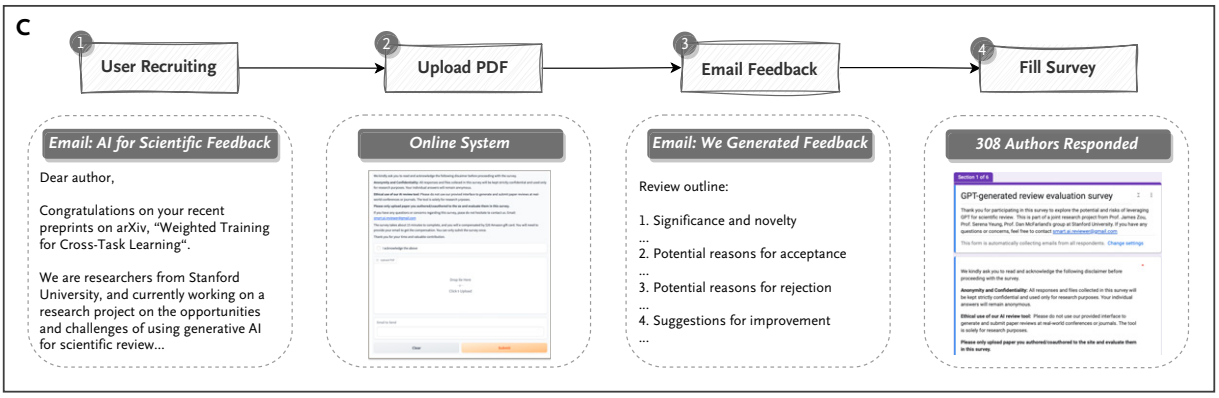

Figure 1. Continued

dataset (*Nature* portfolio journals) spans a broad range of prominent journals across various scientific disciplines and impact levels, thereby capturing both the universality and variations in human-based scientific feedback. The second dataset (ICLR) provides an in-depth perspective of scientific feedback within leading venues of a rapidly evolving field: machine learning. Importantly, this second dataset includes expert feedback on both accepted and rejected papers.

To assess the generalizability of our findings to biomedical and health sciences papers, we also include a random sample of 666 submissions from *eLife*, a leading open-access journal in the life sciences and biomedicine, along with their 1632 associated reviews. We also analyzed a subset of 425 papers categorized under the health science category by the *Nature* portfolio.

We developed a retrospective comment-matching pipeline to evaluate the overlap between feedback from LLMs and human reviewers (Fig. 1B, Supplementary Methods, and Figs. S9 and S16). The pipeline first performs extractive text summarization[32-35] to extract the comments from both LLM and human-written feedback (Figs. S10 and S17). It then applies semantic text matching[36-38] to identify shared comments between the two feedback sources (Figs. S10 and S18). We validated the pipeline's accuracy through human verification, yielding an F1 score of 96.8% for extraction (Table S3A and Supplementary Methods) and an F1 score of 82.4% for matching (Table S3B and Supplementary Methods).

PROSPECTIVE USER STUDY AND SURVEY

We conducted a survey study on 308 researchers from 110 institutions who opted in to receive LLM-generated scientific feedback on their own papers and were asked to evaluate its utility and performance (Fig. 1C, Supplementary Methods, and Fig. S15). Although our sampling approach is subject to biases of self-selection, the data can provide valuable insights and subjective perspectives from researchers that complement our retrospective analysis.[39,40]

Results

FINDINGS FROM RETROSPECTIVE EVALUATION

The results from the retrospective evaluation are illustrated in Figures 2 and 3.

LLM Feedback Significantly Overlaps with Human-Generated Feedback

We began by examining the overlap between LLM feedback and human feedback on *Nature* family journal data (Table S1) through the retrospective matching pipeline. More than half (57.55%) of the comments raised by GPT-4 were raised by at least one human reviewer (Fig. S1A). This suggests a considerable overlap between LLM feedback and human feedback, indicating potential accuracy and usefulness of the system. When comparing LLM feedback with comments from each individual reviewer, approximately one third (30.85%) of comments raised by GPT-4 overlapped with comments from an individual reviewer (Fig. 2A). The degree of overlap between two human reviewers was similar (28.58%) after controlling for the number of comments (Supplementary Methods and Figs. S12 to S14). As validation of the results from the retrospective matching pipeline, four human annotators independently assessed LLM feedback and human reviews for 400 *Nature* portfolio papers. They also found that the overlap between LLM's

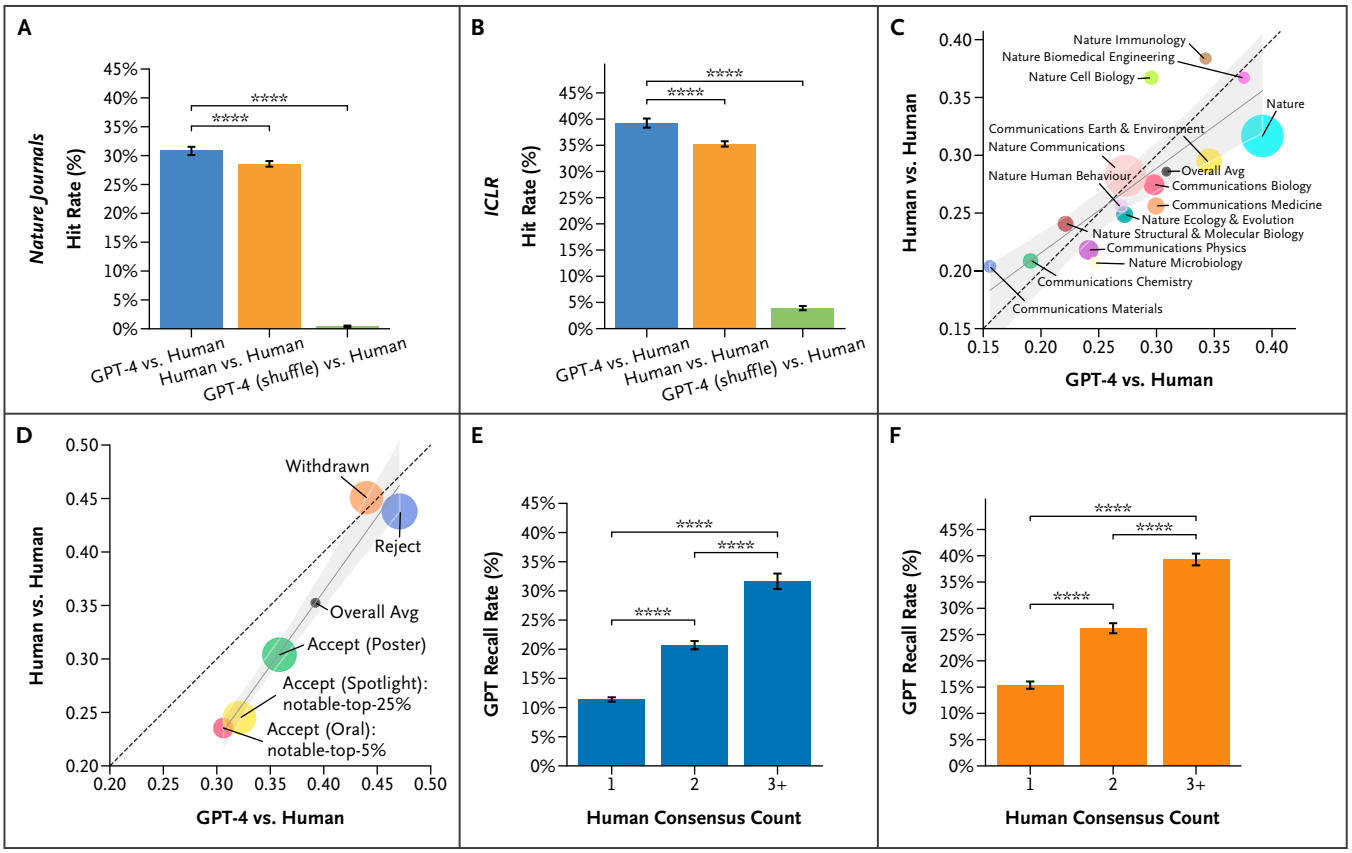

Figure 2. Retrospective Analysis of LLMs and Human Scientific Feedback.

Panel A shows retrospective overlap analysis between feedback from the LLM versus individual human reviewers on papers submitted to *Nature* family journals. Approximately one third (30.85%) of comments raised by Generative Pre-trained Transformer 4 (GPT-4) overlap with the comments from an individual reviewer (hit rate). "GPT-4 (shuffle)" indicates feedback from GPT-4 for another randomly chosen paper from the same journal and category. As a null model, if the LLM mostly produces generic feedback applicable to many papers, then there would be little drop in the pairwise overlap between LLM feedback and the comments from each individual reviewer after the shuffling. In contrast, the hit rate drops substantially from 57.55 to 1.13% after shuffling, indicating that the LLM feedback is paper specific. Panel B shows that in the International Conference on Learning Representations (ICLR), more than one third (39.23%) of GPT-4–raised comments overlap with the comments from an individual reviewer. The shuffling experiment shows a similar result, indicating that the LLM feedback is paper specific. In Panels C and D, the overlap between LLM feedback and human feedback appears comparable with the overlap observed between two human reviewers across *Nature* family journals (Panel C; r=0.80, P<0.001) and across ICLR decision outcomes (Panel D; r=0.98, P=0.003). In Panels E and F, comments raised by multiple human reviewers are disproportionately more likely to be hit by GPT-4 on *Nature* family journals (Panel E) and ICLR (Panel F). The x axis indicates the number of reviewers raising the comment. The y axis indicates the likelihood that a human reviewer comment matches a GPT-4 comment (GPT-4 recall rate). In Panels G and H, comments presented at the beginning of a reviewer's feedback are more likely to be identified by GPT-4 on *Nature* family journals (Panel G) and ICLR (Panel H). The x axis indicates a comment's position in the sequence of comments raised by the human reviewer. Error bars represent 95% confidence intervals. LLM denotes large language model. *P<0.05; ****P<0.001.

feedback and human reviews (28% hit rate) was similar to the overlap between two human reviews (25% hit rate). This indicates that the overlap between LLM feedback and human feedback is comparable with the overlap observed between two human reviewers.

We further stratified these overlap results by academic journals (Fig. 2C). Whereas the degree of overlap between LLM feedback and human comments varied across different academic journals within the *Nature* family — from 15.58% in *Nature Communications Materials* to 39.16% in *Nature* — the overlap between LLM feedback and human feedback comments largely mirrored the overlap found between two human reviewers. The robustness of the finding further indicates that scientific feedback generated from the LLM is similar to what researchers could get from peer reviewers.

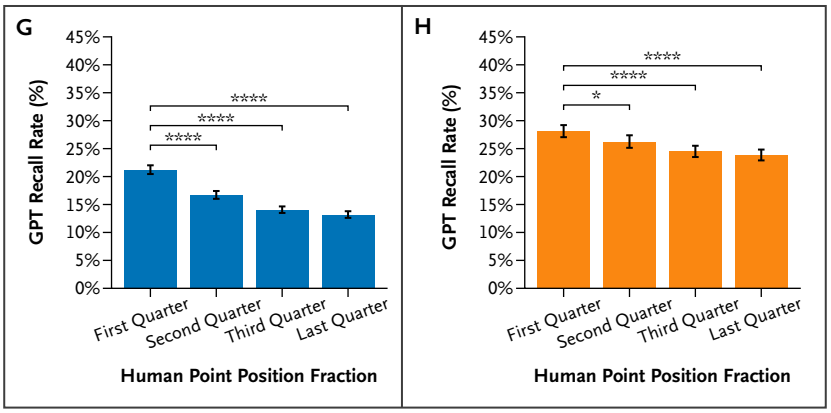

Figure 2. Continued

As additional sensitivity experiments, we found that the overlap analysis was consistent across other statistical metrics, including the Szymkiewicz–Simpson overlap coefficient, the Jaccard index, and the Sørensen–Dice coefficient (Fig. S2). For a subset of 408 *Nature* family publications, we also obtained associated Research Square preprints. The GPT-4's feedback on these preprints also significantly overlaps the reviewers' comments; 55.4% of the points raised by GPT-4 were raised by at least one human reviewer (Fig. S3).

In parallel experiments, we investigated the comment overlap between LLM feedback and human feedback on ICLR papers data (Table S2), and the results were largely similar. A majority (77.18%) of the comments raised by GPT-4 were also raised by at least one human reviewer (Fig. S1B), indicating considerable overlap between LLM feedback and human feedback. When comparing LLM feedback with comments from each individual reviewer, more than one third (39.23%) of comments raised by GPT-4 overlapped with comments from an individual reviewer (Fig. 2B). The overlap between two human reviewers was similar (35.25%) after controlling for the number of comments (Supplementary Methods and Figs. S11 and S14). We further stratified these overlap results by the decision outcomes of the papers (Fig. 2D). Similar to results from *Nature* family journals, we found that the overlap between LLM feedback and human feedback comments largely mirrored the overlap found between two human reviewers.

The results on biomedical and health sciences papers were consistent with our previous findings. Among *eLife* samples, 32.07% of the comments raised by GPT-4 overlapped with those from an individual reviewer, and the degree of overlap between two human reviewers was similar at 30.58% (Fig. S4). On the subset of health science papers from the *Nature* portfolio, 31.48% of the comments raised by GPT-4 overlapped with those from an individual reviewer, and the overlap between two human reviewers was 27.91% (Fig. S5). These additional analyses strengthen our original findings and demonstrate the potential of LLMs to provide feedback on biomedical and health sciences papers.

In addition, because the ICLR dataset includes both accepted and rejected papers, we conducted stratification analysis and found a correlation between worse acceptance decisions and larger overlap in ICLR papers. Specifically, papers accepted with oral presentations (representing the top 5% of accepted papers) have an average overlap of 30.63% between LLM feedback and human feedback comments. The average overlap increases to 32.12% for papers accepted with a spotlight presentation (the top 25% of accepted papers), and rejected papers bear the highest average overlap at 47.09%. A similar trend was observed in the overlap between two human reviewers: 23.54% for papers accepted with oral presentations (the top 5% of accepted papers), 24.52% for papers accepted with spotlight presentations (the top 25% of accepted papers), and 43.80% for rejected papers. This suggests that rejected papers may have more apparent issues or flaws that both human reviewers and LLMs can consistently identify.

Because our primary analyses focus on GPT-4, we wanted to assess how consistent GPT-4 feedback is over time. Accordingly, we evaluated the March 2023 and June 2023 checkpoints of GPT-4 on *Nature* family and ICLR papers and found that the two GPT-4 checkpoints produced very

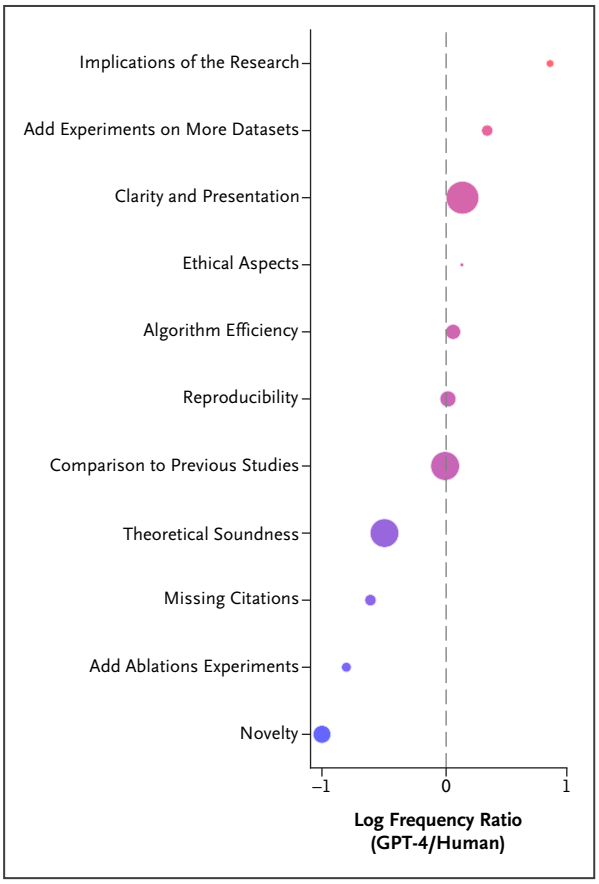

Figure 3. LLM-Based Feedback Emphasizes Certain Aspects More Than Humans.

The LLM comments on the implications of research 7.27 times more frequently than human reviewers. Conversely, the LLM is 10.69 times less likely to comment on novelty compared with human reviewers. Although both the LLM and humans often suggest additional experiments, their focuses differ. Human reviewers are 6.71 times more likely than the LLM to request additional ablation experiments, whereas the LLM is 2.19 times more likely than humans to request experiments on more datasets. Circle size indicates the prevalence of each aspect in human feedback. LLM denotes large language model.

consistent feedback (Fig. S6). We also evaluated the ability of two state-of-the-art open-source LLMs, Llama 2 (70 billion parameters) and Falcon (40 billion parameters), to provide feedback. Both open-source LLMs perform significantly worse than GPT-4 (Fig. S6).

LLMs Generate Paper-Specific Feedback

Is it possible that an LLM merely generates generic feedback applicable to multiple papers? A potential null model is that an LLM mostly produces generic feedback applicable to many papers. To test this hypothesis, we performed a shuffling experiment aimed at verifying the specificity and relevance of LLM-generated feedback. For each paper in the *Nature* family journal data, the LLM feedback was shuffled for papers from the same journal and within the same *Nature* category (Supplemental Methods). If the LLM was producing only generic feedback, we would observe no decrease in the pairwise overlap between shuffled LLM feedback and human feedback. In contrast, the pairwise overlap significantly decreased from 30.85 to 0.43% after shuffling (Fig. 2A). A similar drop from 39.23 to 3.91% was observed for ICLR (Fig. 2B). These results suggest that LLM feedback is paper specific.

Alignment Between LLMs and Humans on Major Comments

What characteristics do LLMs' comments exhibit? What are the distinctive features of the human comments that align with comments generated by LLMs? Here, we evaluate the unique characteristics of comments generated by LLMs. Our analysis revealed that comments identified by multiple human reviewers are more likely to be echoed by LLMs. For instance, in the *Nature* family journal data (Fig. 2E), a comment raised by a single human reviewer had an 11.39% chance of being identified by LLMs. This probability increased to 20.67% for comments raised by two reviewers and further to 31.67% for comments raised by three or more reviewers. A similar trend was observed in the ICLR data (Fig. 2F), where the likelihood of LLMs identifying a comment increased from 15.39% for a single reviewer to 26.21% for two reviewers and 39.33% for three or more reviewers. These findings suggest that LLMs are more likely to identify common issues or flaws that are consistently recognized by multiple human reviewers compared with specific comments raised by a single reviewer. This alignment of the LLM with human perspectives indicates its ability to identify what is generally considered as major or significant issues.

We further examined the likelihood of LLM comments overlapping with human feedback based on their position in the sequence because earlier comments in human feedback (e.g., "concern 1") may represent more significant issues. To this end, we divided each human reviewer's comment sequence into four quarters within the *Nature* journal data (Fig. 2G). Our findings suggest that comments raised in the first quarter of the review text are most likely (21.23%) to overlap with LLM comments, with subsequent

quarters revealing decreasing likelihoods (16.74% for the second quarter). Similar trends were observed in the ICLR papers data, where earlier comments in the sequence showed a higher probability of overlap with LLM comments (Fig. 2H). These findings further support that the LLM tends to align with human perspectives on what are generally considered as major or significant issues.

LLM Feedback Emphasizes Certain Aspects More Than Humans

We next analyzed whether certain aspects of feedback are more/less likely to be raised by the LLM and human reviewers. We focus on ICLR for this analysis because it is more homogeneous than *Nature* family journals, making it easier to categorize the main aspects of review. Drawing on existing research in the peer review literature within the machine learning domain,[41-44] we developed a schema comprising 11 distinct aspects of comments. We then performed human annotation on a randomly sampled subset (Supplementary Methods and Tables S5 to S7).

Figure 3 presents the relative frequency of each of the 11 aspects of comments raised by humans and LLM. The LLM comments on the implications of research 7.27 times more frequently than humans do. Conversely, the LLM is 10.69 times less likely to comment on novelty than humans are. While both LLMs and humans often suggest additional experiments, their focuses differ. Humans are 6.71 times more likely than the LLM to request more ablation experiments, whereas the LLM is 2.19 times more likely than humans to request experiments on more datasets. These findings suggest that the emphasis put on certain aspects of comments varies between LLMs and human reviewers. This variation highlights the potential advantages that a human–AI collaboration could provide. Rather than having LLM fully automate the scientific feedback process, humans can raise important points that the LLM may overlook. Similarly, the LLM could supplement human feedback by providing more comprehensive comments.

FINDINGS FROM THE PROSPECTIVE USER STUDY AND SURVEY

The results from the user study are illustrated in Figure 4.

Survey Responses Align with Retrospective Evaluations

Our user study provides additional evidence that is largely consistent with retrospective evaluations. First, the user study survey results corroborate the findings from the retrospective evaluation on significant overlaps between LLM feedback and human feedback. More than 70% of participants think that there is at least "partial alignment" between LLM feedback and what they think/would expect on the significant points and issues with their paper, and 35% of participants think that the alignment is considerable or substantial (Fig. 4B). Second, the survey study further corroborates the findings from the automated evaluation on the ability of the language model to generate nongeneric feedback; 32.9% of participants think that our system-generated feedback is "less specific than many but more specific than some peer reviewers," and 17.3 and 14% think that it is "about as specific as peer reviewers" or "more specific than many peer reviewers," further corroborating that LLMs can generate nongeneric reviews (Fig. 4D).

Researchers Find LLM Feedback Helpful

Participants were also surveyed about the extent to which they found the LLM feedback helpful in improving their work or understanding of a subject. The majority responded positively, with over 50.3% considering the feedback to be helpful and 7.1% considering it to be very helpful (Fig. 4A). When participants were asked to compare the LLM feedback with human feedback, 17.5% of participants considered it to be inferior to human feedback, and 41.9% considered it to be less helpful than many but more helpful than some human feedback. Additionally, 20.1% considered it to be about the same level of helpfulness as human feedback, and 20.4% considered it to be even more helpful than human feedback (Fig. 4C). Our evaluation further revealed that the perceptions of alignment and helpfulness were consistent across various demographic groups (Figs. S7 and S8).

In line with the helpfulness of the system, 50.5% of survey participants further expressed their willingness to reuse the system (Fig. 4G). The participants expressed optimism about the potential improvements that continued use of the system could bring to the traditional human feedback process (Fig. 4E and 4F). They believe that the LLM technology can further refine the quality of reviews and possibly introduce new capabilities. Interestingly, the evaluation also revealed that participants believe that authors are more likely to benefit from LLM-based feedback than other stakeholders, such as reviewers, and area chairs (Fig. 4H). Many participants envisioned a timely feedback tool for authors to receive comments on their papers in a timely manner. For example, one participant wrote, "The review took five minutes and was of a reasonably high quality. This can tremendously help authors to receive a fast turnaround feedback and help in polishing their

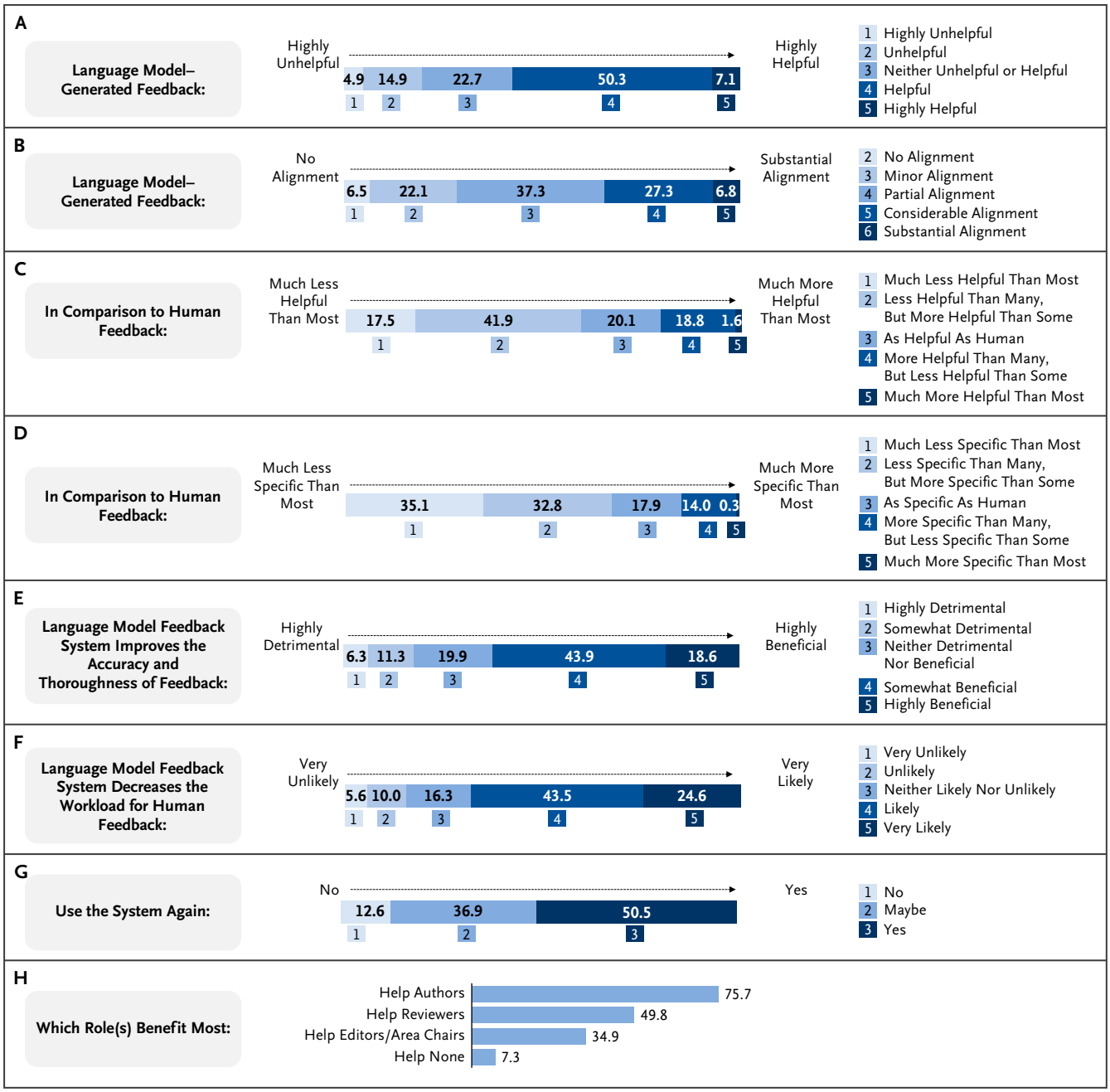

Figure 4. Human Study of LLMs and Human Review Feedback (n=308).

Panels A and B show that LLM-generated feedback is generally helpful and has substantial overlaps with actual feedback from human reviewers. In Panels C and D, compared with human feedback, LLM feedback is slightly less helpful and less specific. Panels E and F show that users generally believe that the LLM feedback system can improve the accuracy and thoroughness of reviews and reduce the workload of reviewers. Panel G shows that most users intend to use or potentially use the LLM feedback system again. Panel H shows that users believe that the LLM feedback system mostly helps authors, followed by reviewers and editors/area chairs. Numbers are percentages. LLM denotes large language model.

submissions." Another participant wrote, "After writing a paper or a review, GPT could help me gain another perspective to re-check the paper."

Strengths and Weaknesses of LLM Feedback

Beyond generating feedback that aligns with that of humans, our results also suggest that LLMs could potentially generate useful feedback that has not been mentioned by humans. For example, 65.3% of participants think at least to some extent that LLM feedback offers perspectives that have been overlooked or underemphasized by humans. Study participants also discussed limitations of the current system. The most important limitation is its ability to generate specific and actionable feedback. Detailed results are in Supplementary Results.

Discussion

Our findings highlight the potential value and utility of leveraging LLM feedback as a valuable resource for authors seeking constructive feedback and suggestions for enhancing their manuscripts. This could be especially helpful for researchers who lack access to timely quality feedback mechanisms: for example, researchers from traditionally underprivileged regions who may not have resources to access conferences or even peer review (their works are much more likely than those of "mainstream" researchers to get desk rejected by journals and thus, seldom go through the peer review process[12]). For others, the framework could be used as a mechanism for authors to self-check and improve their work in a timely manner.

Despite the potential of LLMs to provide timely and helpful scientific feedback, it is important to note that expert human feedback will still be the cornerstone of rigorous scientific evaluation. As demonstrated in our findings, our analysis reveals limitations of the framework; for example, LLM is biased toward certain aspects of scientific feedback (e.g., "add experiments on more datasets") and sometimes feels "generic" to the authors (participants also indicate that, quite often, human reviewers are "generic"). Although it is comparable with and even better than feedback from some human reviewers, the current LLM feedback cannot substitute for specific and thoughtful human feedback by domain experts.

It is also important to note the potential misuse of LLMs for scientific feedback. We argue that LLM feedback should be primarily used by researchers to identify areas of improvement in their manuscripts before official submission. It is important that expert human reviewers deeply engage with the manuscripts and provide independent assessment without relying on LLM feedback. Automatically generating reviews without thoroughly reading the manuscript would undermine the rigorous evaluation process that forms the bedrock of scientific progress. More broadly, our study contributes to the recent discussions on the impacts of LLMs and generative AI on existing work practices. Researchers have discussed the potential of LLMs to improve productivity,[45,46] improve creativity,[47] and facilitate scientific discovery.[48] We envision that LLMs and generative AI, if deployed responsibly, could also potentially bring a paradigm change to how researchers conduct research, collaborate, and provide evaluations, influencing the way that science and technology advance.

Our study has several limitations. Our results are based on zero-shot learning from the GPT-4 model without additional fine-tuning. The retrospective evaluation is restricted to the top English-language venues and may not represent wider scientific literature. Our user study participants, mainly United States–based researchers, were self-selected, which might skew representativeness. We systematically discuss these limitations and future work in Supplementary Discussion.

Disclosures

Supported by the Stanford Interdisciplinary Graduate Fellowship (to Dr. Cao) and grants from the Chan-Zuckerberg Initiative (to Dr. Zou).

Author disclosures and other supplementary materials are available at ai.nejm.org.

We thank S. Eyuboglu, M. Yuksekgonul, D. Jurafsky, and M. Bernstein for their guidance and helpful discussions.

Author Affiliations

[1] Department of Computer Science, Stanford University, Stanford, CA
[2] Kellogg School of Management, Northwestern University, Evanston, IL
[3] Department of Biomedical Data Science, Stanford University, Stanford, CA
[4] Department of Information Science, Cornell University, Ithaca, NY
[5] Department of Electrical Engineering, Stanford University, Stanford, CA
[6] Graduate School of Education, Stanford University, Stanford, CA

References

1. Kuhn TS. The structure of scientific revolutions. Chicago: University of Chicago Press, 1962:90.
2. Horbach SP, Halffman W. The changing forms and expectations of peer review. Res Integr Peer Rev 2018;3:8. DOI: 10.1186/s41073-018-0051-5.

3. Price DJDS. Little science, big science. New York: Columbia University Press, 1963.

4. Jones BF. The burden of knowledge and the "death of the renaissance man": is innovation getting harder? Rev Econ Stud 2009;76:283-317. DOI: 10.1111/j.1467-937X.2008.00531.x.

5. Aczel B, Szaszi B, Holcombe AO. A billion-dollar donation: estimating the cost of researchers' time spent on peer review. Res Integr Peer Rev 2021;6:14. DOI: 10.1186/s41073-021-00118-2.

6. Alberts B, Hanson B, Kelner KL. Reviewing peer review. Science 2008;321:5885. DOI: 10.1126/science.1162115.

7. Björk B-C, Solomon D. The publishing delay in scholarly peer-reviewed journals. J Informetrics 2013;7:914-923. DOI: 10.1016/j.joi.2013.09.001.

8. Lee CJ, Sugimoto CR, Zhang G, Cronin B. Bias in peer review. J Am Soc Inf Sci Technol 2013;64:2-17. DOI: 10.1002/asi.22784.

9. Kovanis M, Porcher R, Ravaud P, Trinquart L. The global burden of journal peer review in the biomedical literature: strong imbalance in the collective enterprise. PLoS One 2016;11:e0166387. DOI: 10.1371/journal.pone.0166387.

10. Shah NB. Challenges, experiments, and computational solutions in peer review. Commun ACM 2022;65:76-87. DOI: 10.1145/3528086.

11. Bourdieu P. Cultural reproduction and social reproduction. In: Brown R, ed. Knowledge, education, and cultural change. New York: Routledge, 2018:71-112.

12. Merton RK. The Matthew effect in science. The reward and communication systems of science are considered. Science 1968;159:56-63. DOI: 10.1126/science.159.3810.56.

13. Chu JSG, Evans JA. Slowed canonical progress in large fields of science. Proc Natl Acad Sci USA 2021;118:e2021636118. DOI: 10.1073/pnas.2021636118.

14. Bloom N, Jones CI, Van Reenen J, Webb M. Are ideas getting harder to find? Am Econ Rev 2020;110:1104-1144. DOI: 10.1257/aer.20180338.

15. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Adv Neural Inf Process Syst 2020;33:1877-1901.

16. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst 2022;35:27730-27744.

17. OpenAI. GPT-4 technical report. March 15, 2023 (https://arxiv.org/abs/2303.08774). Preprint.

18. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183:589-596. DOI: 10.1001/jamainternmed.2023.1838.

19. Lee M, Srivastava M, Hardy A, et al. Evaluating human-language model interaction. December 19, 2022 (https://arxiv.org/abs/2212.09746). Preprint.

20. Kung, TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit Health 2023;2:e0000198. DOI: 10.1371/journal.pdig.0000198.

21. Terwiesch C. Would Chat GPT3 get a Wharton MBA? A prediction based on its performance in the operations management course. Philadelphia: Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania, 2023.

22. Schulz R, Barnett A, Bernard R, et al. Is the future of peer review automated? BMC Res Notes 2022;15:203. DOI: 10.1186/s13104-022-06080-6.

23. Liu R, Shah NB. ReviewerGPT? An exploratory study on using large language models for paper reviewing. June 1, 2023 (https://arxiv.org/abs/2306.00622). Preprint.

24. Robertson Z. GPT4 is slightly helpful for peer-review assistance: a pilot study. June 16, 2023 (https://arxiv.org/abs/2307.05492). Preprint.

25. Nature. How to write a report (https://www.nature.com/nature/for-referees/how-to-write-a-report).

26. Nature Communications. Writing your report. 2024 (https://www.nature.com/ncomms/for-reviewers/writing-your-report).

27. Rogers A, Augenstein I. How to review for ACL rolling review. February 11, 2021 (https://aclrollingreview.org/reviewertutorial).

28. Association for Computational Linguistics. ACL'23 peer review policies. 2023 (https://2023.aclweb.org/blog/review-acl23/).

29. Association for Computational Linguistics. ACL-IJCNLP 2021 instructions for reviewers. 2021 (https://2021.aclweb.org/blog/instructions-for-reviewers/).

30. International Conference on Machine Learning. ICML 2023 reviewer tutorial. 2023 (https://icml.cc/Conferences/2023/ReviewerTutorial).

31. Nicholas KA, Gordon WS. A quick guide to writing a solid peer review. Eos (Wash DC) 2011;92:233-234. DOI: 10.1029/2011EO280001.

32. Luhn HP. The automatic creation of literature abstracts. IBM J Res Develop 1958;2:159-165. DOI: 10.1147/rd.22.0159.

33. Edmundson HP. New methods in automatic extracting. J Assoc Comput Mach 1969;16:264-285. DOI: 10.1145/321510.321519.

34. Mihalcea R, Tarau P. Textrank: bringing order into text. In: Lin D, Wu D, eds. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Kerrville, TX: Association for Computational Liguistics, 2004:404-411.

35. Erkan G, Radev DR. Lexrank: graph-based lexical centrality as salience in text summarization. J Artif Intell Res 2004;22:457-479. DOI: 10.1613/jair.1523.

36. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. J Am Soc Inf Sci 1990;41:391-407. DOI: 10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-ASI1%3E3.0.CO;2-9.

37. Socher R, Huang E, Pennin J, Manning CD, Ng A. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. Adv Neural Inf Process Syst 2011;24:801-809.

38. Bowman SR, Angeli G, Potts C, Manning CD. A large annotated corpus for learning natural language inference. August 21, 2015 (https://arxiv.org/abs/1508.05326). Preprint.

39. Meyer BD, Mok WK, Sullivan JX. Household surveys in crisis. J Econ Perspect 2015;29:199-226. DOI: 10.1257/jep.29.4.199.

40. Ross MB, Glennon BM, Murciano-Goroff R, Berkes EG, Weinberg BA, Lane JI. Women are credited less in science than men. Nature 2022;608:135-145. DOI: 10.1038/s41586-022-04966-w.

41. Birhane A, Kalluri P, Card D, Agnew W, Dotan R, Bao, M. The values encoded in machine learning research. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22. New York: Association for Computing Machinery, 2022:173-184.

42. Smith JJ, Amershi S, Barocas S, Wallach H, Wortman Vaughan J. Real ML: recognizing, exploring, and articulating limitations of machine learning research. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. New York: Association for Computing Machinery, 2022:587-597.

43. Koch B, Denton E, Hanna A, Foster JG. Reduced, reused and recycled: the life of a dataset in machine learning research. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. 2021 (https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/3b8a614226a953a8cd9526fca6fe9ba5-Paper-round2.pdf).

44. Scheuerman MK, Hanna A, Denton E. Do datasets have politics? Disciplinary values in computer vision dataset development. Proc ACM Human-Computer Interact 2021;5:1-37. DOI: 10.1145/3476058.

45. Noy S, Zhang W. Experimental evidence on the productivity effects of generative artificial intelligence. Science 2023;381:187-192.

46. Peng S, Kalliamvakou E, Cihon P, Demirer M. The impact of AI on developer productivity: evidence from GitHub Copilot. February 13, 2023 (https://arxiv.org/abs/2302.06590). Preprint.

47. Epstein Z, Hertzmann A, Akten M, et al. Art and the science of generative AI. Science 2023;380:1110-1111. DOI: 10.1126/science.adh4451.

48. Wang H, Fu T, Du Y, et al. Scientific discovery in the age of artificial intelligence. Nature 2023;620:47-60. DOI: 10.1038/s41586-023-06221-2.