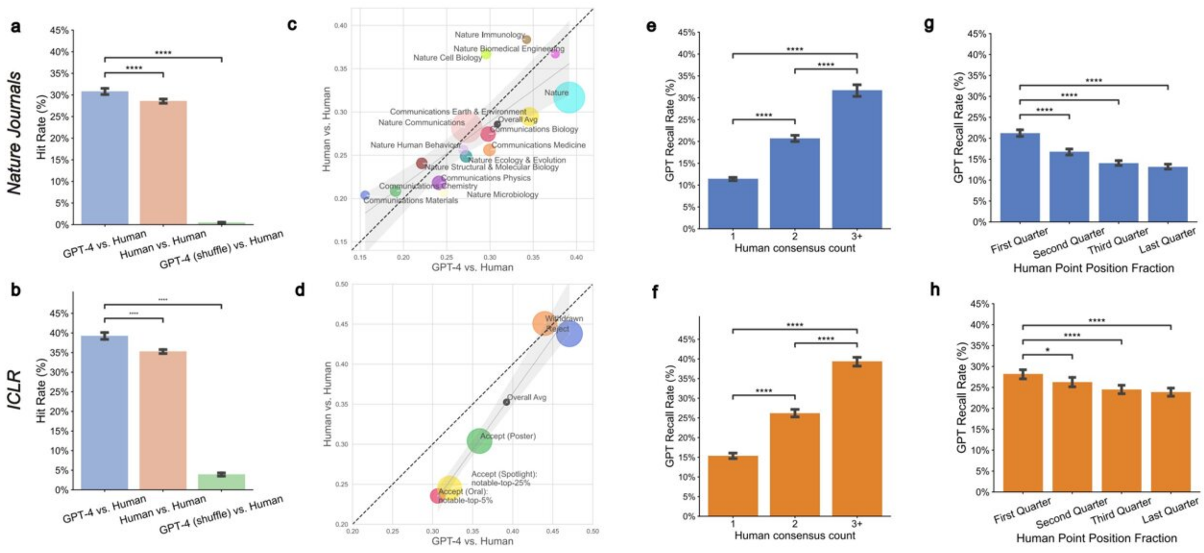
Large language models prove helpful in peer-review process

October 19 2023, by Peter Grad



Retrospective analysis of LLM and human scientific feedback. a, 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 GPT-4 raised comments 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 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. b, 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. c-d, The overlap between LLM feedback and human feedback appears comparable to the overlap observed between two human reviewers across Nature family journals (c) (r = 0.80, P = 3.69×10^{-4}) and across ICLR decision outcomes (d) (r = 0.98, P = 3.28×10^{-3}). e-f, Comments raised by multiple human reviewers are disproportionately more likely to be hit by GPT-4 on Nature Family Journals (e) and ICLR (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). g-h, Comments presented at the beginning of a reviewer's feedback are more likely to be identified by GPT-4 on Nature Family Journals (g) and ICLR (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. *P arXiv (2023). DOI: 10.48550/arxiv.2310.01783

In an era plagued by malevolent sources flooding the internet with misrepresentations, distortions, manipulated imagery and flat-out lies, it should come as some comfort that in at least one arena there is an honor system set up to ensure honesty and integrity: the peer-review process for scholarly publications.

When submitting articles on research they have done, scientists, doctors, specialists in countless fields of expertise routinely submit their work to publications that in turn recruit experts in the same field to closely review their papers.

They check for accuracy, accountability and quality. If the paper fails to meets a publication's high standards, it is returned with recommended adjustments or rejected. When a paper passes what often is robust, challenging review, it is ready for publication.

As Pulitzer Prize-winning Washington Post journalist Chris Mooney put it, "Even if individual researchers are prone to falling in love with their own theories, the broader process of peer review and institutionalized skepticism are designed to ensure that, eventually, the best ideas prevail."

Peer review has been around a long time. The *Philosophical Transactions of the Royal Society* established a formal procedure for acceptance of articles back in the 17th century, and is believed to be the first to adopt what came to be known as peer review.

It is estimated there are 5.14 million peer-reviewed articles published annually, with more than 100 million hours devoted to those reviews.

Against that backdrop, researchers at Stanford University explored how LLMs might contribute to the review process.

Citing the lengthy wait time for review (an average of four months), cost ($2.5 billion annually), and problems securing qualified reviewers who work for no pay, the researchers said assistance from LLMs could prove highly beneficial for publications and authors.

"High-quality peer reviews are increasingly difficult to obtain," said Weixin Liang, an author of the paper, "Can large language models provide useful feedback on research papers? A large-scale empirical analysis," published on the preprint server *arXiv* Oct 3. "Researchers who are more junior or from under-resourced settings have especially hard times getting timely feedback."

They tested their theory by comparing reviewer feedback on several thousand papers from *Nature* journals and the International Conference on Learning Representations machine-learning conference with GPT-4 generated reviews. They found between 31% and 39% overlap in points raised by human and machine generated reviews. On weaker submissions (articles that were rejected), GPT-4 performed even better, overlapping with human scorers 44% of the time.

The researchers also contacted the authors of those papers and found that more than half described GPT-4 commentary as helpful or very helpful. And 80% of authors said LLM feedback was more helpful than "at least some" human reviewers.

"Together our results suggest that LLM and human feedback can complement each other," Liang said. He said that such reviews can be particularly helpful in guiding authors whose papers need substantial revisions.

"Indeed, by raising these concerns earlier in the scientific process before review, these papers and the science they report may be improved," Liang said.

One author whose article was reviewed noted GPT-4 raised points that human reviewers overlooked. "The GPT-generated review suggested me to do visualization to make a more concrete case for interpretability. It also asked to address data privacy issues. Both are important, and human reviewers missed this point," the author said.

The report cautioned, however, that LLMs are not a substitute for human oversight. They cited some limitations, such as reviews that were too vague, failure to provide "specific technical areas of improvement," and in some instances lack of "in-depth critique of model architecture and design."

"It is important to note that expert human feedback will still be the cornerstone of rigorous scientific evaluation," Liang said. "While comparable and even better than some reviewers, the current LLM feedback cannot substitute specific and thoughtful human feedback by domain experts."

The work is published on the *arXiv* preprint server.

More information: Weixin Liang et al, Can large language models provide useful feedback on research papers? A large-scale empirical analysis, *arXiv* (2023). DOI: 10.48550/arxiv.2310.01783

GitHub: github.com/Weixin-Liang/LLM-scientific-feedback

Citation: Large language models prove helpful in peer-review process (2023, October 19) retrieved 31 May 2024 from https://phys.org/news/2023-10-large-language-peer-review.html