ChatGPT identifies gender disparities in scientific peer review

Jeroen P. H. Verharen

Department of Molecular and Cell Biology and Helen Wills Neuroscience Institute,
University of California, Berkeley, USA. Correspondence to: jeroenverharen@gmail.com

ABSTRACT

The peer review process is a critical step in ensuring the quality of scientific research. However, its subjectivity has raised concerns. To investigate this issue, I examined over 500 publicly available peer review reports from 200 published neuroscience papers in 2022-2023. OpenAI's generative artificial intelligence *ChatGPT* was used to analyze language use in these reports. It demonstrated superior performance compared to traditional lexicon- and rule-based language models. As expected, most reviews for these published papers were seen as favorable by *ChatGPT* (89.8% of reviews), and language use was mostly polite (99.8% of reviews). However, this analysis also demonstrated high levels of variability in how each reviewer scored the same paper, indicating the presence of subjectivity in the peer review process. The results further revealed that female first authors received less polite reviews than their male peers, indicating a gender bias in reviewing. In addition, published papers with a female senior author received more favorable reviews than papers with a male senior author, for which I discuss potential causes. Together, this study highlights the potential of generative artificial intelligence in performing natural language processing of specialized scientific texts. As a proof of concept, I show that *ChatGPT* can identify areas of concern in scientific peer review, underscoring the importance of transparent peer review in studying equitability in scientific publishing.

INTRODUCTION

The peer review process is a crucial step in the publication of scientific research, where manuscripts are evaluated by independent experts in the field before being accepted for publication. This process helps ensure the quality and validity of scientific research and is a cornerstone of scientific integrity. Despite its importance, concerns have been raised regarding subjectivity in this process that may affect the fairness and accuracy of evaluations[1-5]. Indeed, most journals engage in single-blind peer review, in which the reviewers have information about the authors of the paper, but not vice versa. While some studies have found evidence of disparities in peer review as a result of gender bias, the scope and methodology of these studies are often limited[6-7]. One larger study, performed by an ecology journal, found no evidence of gender bias in reviewing, but did find a bias against non-English-speaking first authors[8]. Additionally, other factors, such as the seniority and institutional affiliation of authors, may influence the evaluation process and lead to biased assessments of research quality[6]. As such, papers from more prestigious research institutions may receive better reviews[9]. It is crucial to identify potential sources of disparity in the reviewing process to maintain scientific integrity and find areas for improvement within the scientific pipeline.

Natural language processing tools have shown promise in analyzing large amounts of textual data and extracting meaningful insights from evaluations[10-12]. However, applying these tools to scientific peer review has been challenging due to the specialized construction and language use in such reports. A recent study that manually annotated language use in peer reviews has shown great potential[13], but algorithms struggled to perform well in this task[14-15]. Recent advances in generative artificial intelligence, such as OpenAI's *ChatGPT*, offer new possibilities for studying scientific peer review. These models can process vast amounts of text and provide accurate sentiment scores and language use metrics for individual sentences and documents. As such, using generative artificial intelligence to study scientific peer review may ultimately help improve the

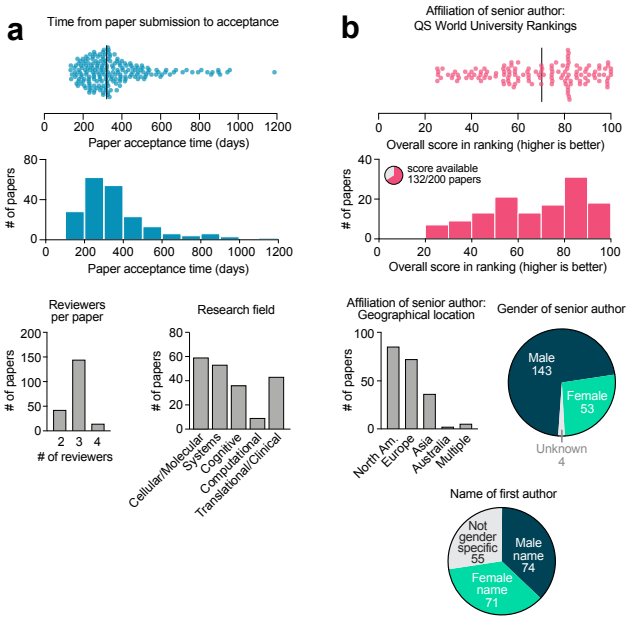


Figure 1: Characteristics of the 200 papers included in this analysis

(**a**) Paper metrics. (**b**) Author metrics. More information on how these metrics were collected and defined can be found in the **Methods** section.

overall quality and fairness of scientific publications and identify areas of concern in the way towards equitable academic research.

This study had three main objectives. The first aim was to test whether the latest advances in generative artificial intelligence, such as OpenAI's *ChatGPT*, can be used to analyze language use in specialized scientific texts, such as peer reviews. The second aim was to explore subjectivity in peer review by looking at consistency in favorability across reviews for the same paper. The last aim was to test whether the identity of the authors, such as institutional affiliation and gender, affect the favorability and language use of the reviews they receive.

RESULTS

An analysis of scientific peer review

Nature Communications has engaged in transparent peer review since 2016, giving authors the option to (and since 2022 requiring authors to) publish the peer review history of their paper[16]. To explore language use in these reports, I downloaded the primary (i.e., first-round) reviews from the last 200 papers in the neuroscience field published in this journal. This yielded a total of 572 reviews from 200 papers, with publications dates ranging from August 2022 to February 2023. Additional metrics of these papers were manually collected (**Fig. 1a and 1b**), including the total time until paper acceptance, the subfield of neuroscience, the geographical location and QS World Ranking score of the senior author's institutional affiliation, the gender of the senior author, and whether the first author had a male or female name (see **Methods** for more information on classifications and a rationale for the chosen metrics). These metrics were collected to test whether they influenced the favorability and language use of the reviews that a paper received.

Sentiment analysis

To assess the sentiment and language use of each of the peer review reports, I asked OpenAI's generative artificial intelligence *ChatGPT* to extract two scores from each of the reviews (**Fig. 2a**). The first score was the *sentiment score*, and measures how favorable the review is. This metric ranges from -100 (negative) to 0 (neutral) to +100 (positive). Sentiment reflects the reviewer's opinion about the paper and is what presumably drives the decision for a paper to be accepted or rejected. The second score was the *politeness score*, which evaluates how polite a review's language is, measured on a scale from -100 (rude) to 0 (neutral) to +100 (polite). *ChatGPT* was able to extract sentiment and politeness scores for all of the 572 reviews, and usually included a reasoning of how it established the score (**Supplementary Fig. 1**).

The accuracy and consistency of the generated scores was validated in four different ways. First, for a representative sample of the reviews, I read both the review and *ChatGPT*'s reasoning of how it came to the scores (for examples see **Supplementary Fig. 1**). I established that the algorithm was able to extract the most important sentences from each of the reviews and to provide a plausible score. Second, since generative artificial intelligence can provide different answers every time it is prompted, the algorithm was asked to provide scores for each review twice. This yielded a significant correlation between the first and second iteration of scoring ($P < 0.0001$ for both sentiment and politeness scores; **Supplementary Fig. 2**); the average of the two scores was used for all subsequent analyses in this paper. Third, manipulated reviews (in which I manually re-wrote a

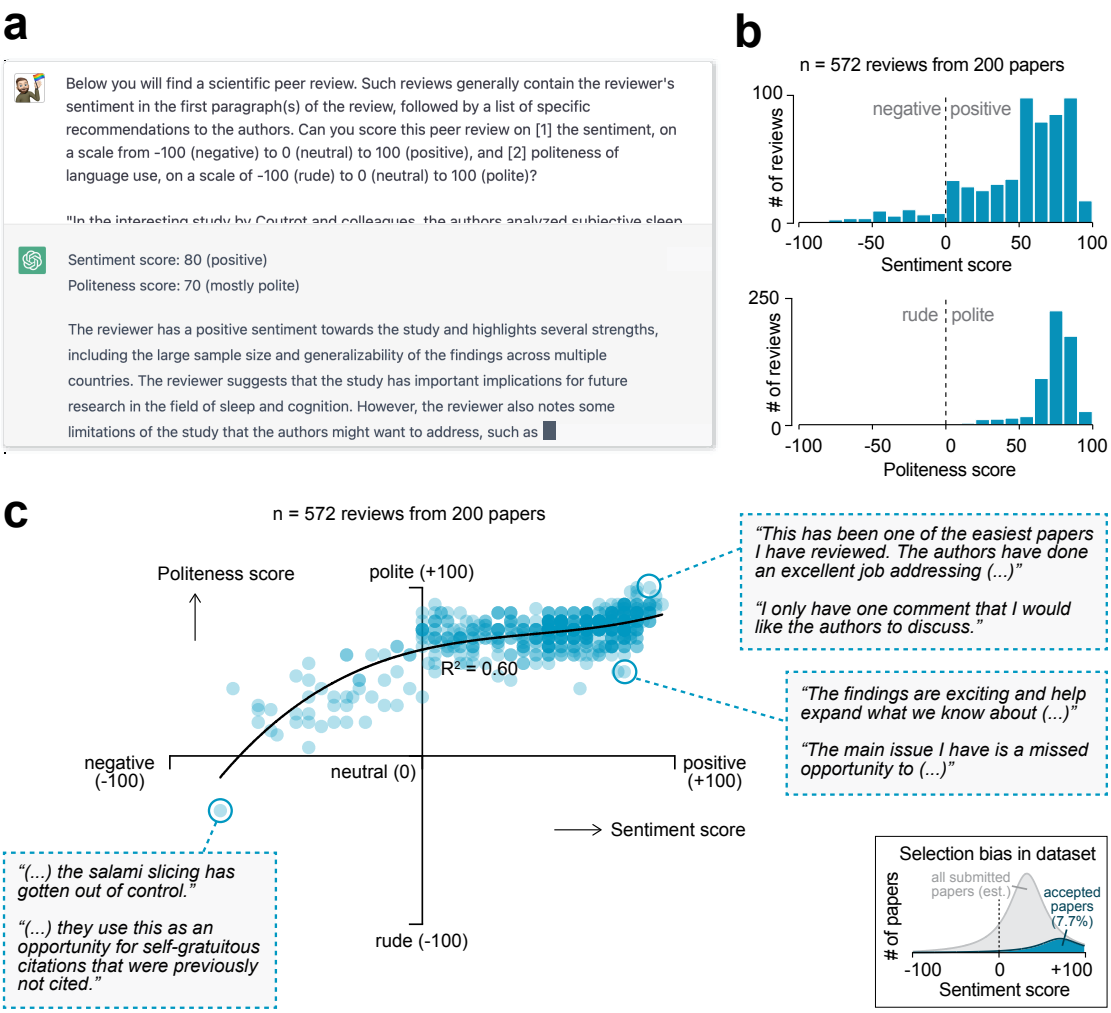


Figure 2: Sentiment analysis on peer review reports using generative artificial intelligence
(**a**) OpenAI's generative artificial intelligence model *ChatGPT* was used to extract a sentiment and politeness sores for each of the 572 first-round reviews. Shown is an example query and *ChatGPT*'s answer.
(**b**) Histograms showing the distribution in sentiment (top) and politeness (bottom) scores for all reviews.
(**c**) Scatter plot showing the relation between sentiment and politeness scores for the reviews (60% variance explained in third-degree polynomial). Insets show excerpts from selected peer reviews. Inset in the bottom right corner is a visual depiction of the expected selection bias in this dataset, as only papers accepted for publication were included in this analysis (gray area represents full pool of published and unpublished papers; not to scale).

'neutral' review in a more rude, polite, negative or positive manner) were input into *ChatGPT*, which confirmed that this changed the review's politeness and sentiment scores, respectively (**Supplementary Fig. 3**). Finally, for a subset of reviews, *ChatGPT*'s scores were compared to that of seven human scorers that were blinded to the algorithm's scores (**Supplementary Fig. 4**). Interestingly, there was high variability across human scorers, but their average score had a high correlation to that of *ChatGPT* (linear regression for sentiment score: R^2 = 0.91, P = 0.0010; for politeness score: R^2 = 0.70, P = 0.018). Importantly, *ChatGPT* was superior to the lexicon- and rule-based algorithms *TextBlob*[17] and *VADER*[18] in scoring a review's sentiment; both these algorithms did not not significantly predict the average human-scored sentiment (*TextBlob*: R^2 = 0.13, P = 0.42; *VADER*: R^2 = 0.07, P = 0.56). Together, these validations indicate that *ChatGPT* can accurately score the sentiment and politeness of scientific peer reviews, and does so better than other available tools.

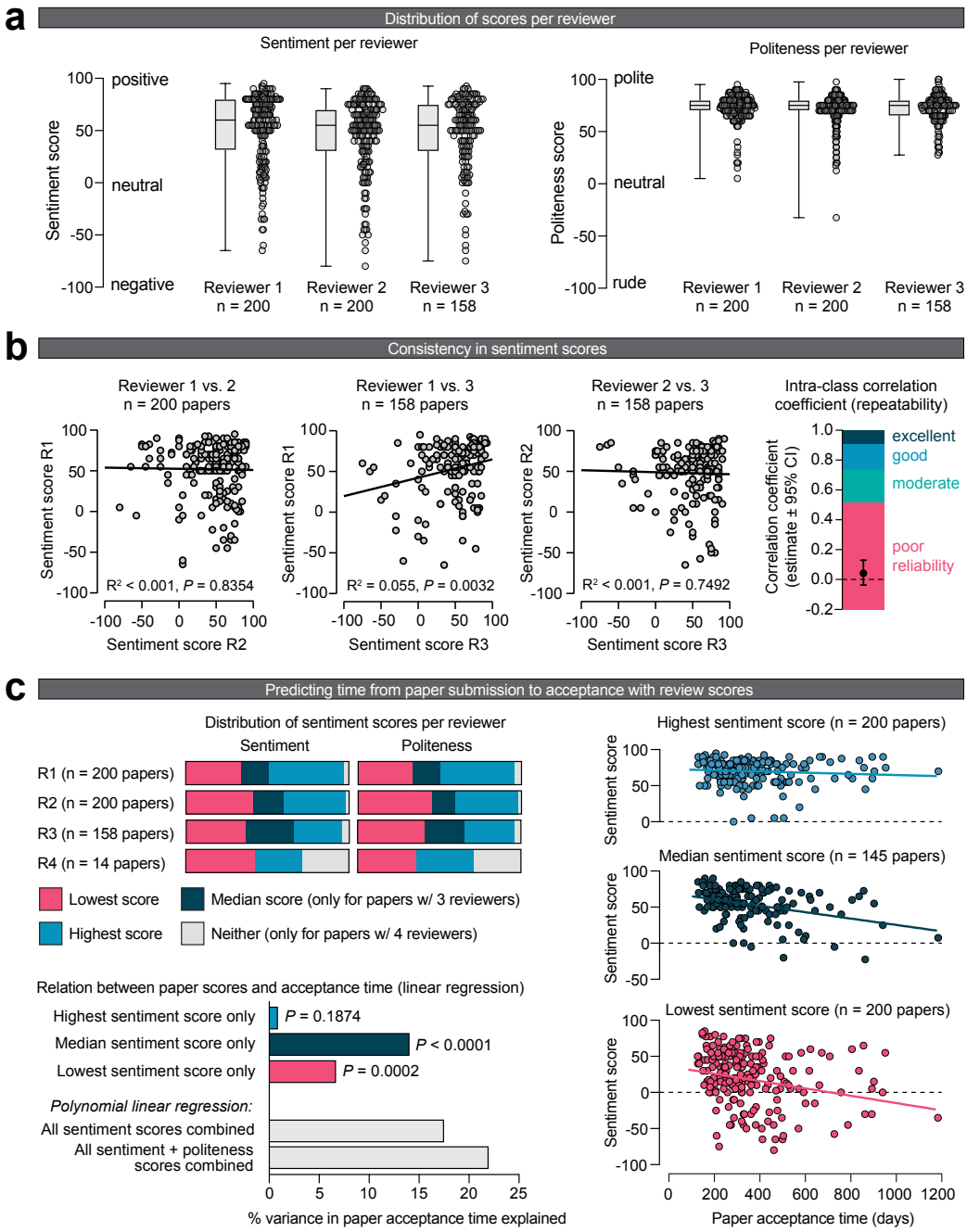


Figure 3: Consistency across reviews

(**a**) Sentiment (left) and politeness (right) scores for each of the 3 reviewers. The lower sample size for reviewer 3 is because 42 papers received only 2 reviews. No significant effects were observed of reviewer number on sentiment (mixed effects model, $F(1.929, 343.3) = 1.564$, $P = 0.2116$) and politeness scores (mixed effects model, $F(1.862, 331.4) = 1.638$, $P = 0.1977$).

(**b**) Correlations showing low consistency of sentiment scores across reviews for the same paper. The sentiment scores between reviewers 1 and 3 (middle panel) is the only comparison the reached statistical significance ($P = 0.0032$), albeit with a low amount of variance explained (5.5%). The intra-class correlation coefficient (ICC) measures how similar the review scores are for one paper, without the need to split review up into pairs. An ICC < 0.5 generally indicates poor reliability (i.e., repeatability)[19].

(**c**) Linear regression indicating the relation between a paper's sentiment scores and the time between paper submission and acceptance. For this analysis, reviews were first split into a paper's lowest, median (only for papers with an odd number of reviews) and highest sentiment score. The lowest and median sentiment score of a paper significantly predicted a paper's review time, but its highest sentiment score did not. Note that the relation between politeness scores and review time were not individually tested given the high correlation between sentiment and politeness, thus having a high chance of finding spurious correlations. The metric '% variance in paper acceptance time explained' denotes the R^2 value of the linear regression.

The majority of the 572 peer reviews (89.9%) were of positive sentiment; 7.9% were negative; 2.3% were neutral (i.e., a sentiment score of 0) (**Fig. 2b**). 99.8% of reviews were deemed polite by the algorithm (i.e., a positive politeness score), only 1 review was scored as rude (i.e., a negative politeness score; **Fig. 2c**, bottom left inset). A regression analysis indicated a strong relation between the reviews' sentiment and politeness scores (60% of variance explained in a third-degree polynomial regression) (**Fig. 2c**). Thus, the more positive a review, the more polite the reviewer's language generally is. It is important to note here that the papers included in this analysis were ultimately accepted for publication in *Nature Communications*, which has a low acceptance rate of 7.7%. As a result of this selection, there will be an over-representation of positive scores in this analysis (**Fig. 2c,** bottom right inset).

Consistency across reviewers

If a research paper meets certain objective standards of quality, one can reasonably expect that reviewers evaluating that paper would share a common view on its overall sentiment. To investigate if this is the case, I analyzed the consistency across review scores for the same paper (**Fig. 3**). As expected, the overall distribution of sentiment and politeness scores did not differ between the first three reviewers (**Fig. 3a**). Interestingly, a linear regression analysis of sentiment scores across reviewers indicated very low, if any, correlation between the sentiment scores of reviews for the same paper (**Fig. 3b**). The maximum variance explained in sentiment scores between reviewers was 5.5% (between reviewer 1 and 3; the only comparison that reached statistical significance). I also calculated the intra-class correlation coefficient[19] between the different reviewers, which demonstrated poor inter-reviewer reliability of scoring (ICC = 0.055, 95% confidence interval of -0.025 – 0.144). These results indicate high levels of disagreement between the reviewers' favorability of a paper, suggesting that the peer review process is subjective.

I then looked at the relation between a paper's review scores and its acceptance time (i.e., the time from paper submission to acceptance). For this analysis, review scores were first classified as the lowest, median (only for papers with an odd number of reviewers), or highest for a paper (**Fig. 3c**). A linear regression analysis indicated that the median sentiment score was the best predictor of a paper's review time (R^2 = 0.0670 , P = 0.0002), followed by the lowest sentiment score (R^2 = 0.1404, P < 0.0001) (**Fig. 3c**, bottom left panels). Interestingly, a paper's highest sentiment score did not significantly predict review time (R^2 = 0.0088 , P = 0.1874).

Exploring disparities in peer review

To explore potential sources of disparities in scientific publishing, I correlated the review scores, pooled across all papers, with the different paper and author metrics that were collected earlier (**Fig. 1b**). No significant effects were observed between sentiment and politeness scores across the different subfields of neuroscience (**Fig. 4a**). With respect to the institutional affiliation of the senior author, no effects were observed between the scores and the continent in which the senior author was based (**Fig. 4b**). Additionally, no correlation was observed between the institute's score on the QS World ranking and the paper's sentiment and politeness scores (**Fig. 4c**).

Finally, I looked at how the gender of the first and senior authors may affect a paper's review scores. First authors with a female name received significantly more impolite reviews, but no effect was observed on sentiment (**Fig. 4d**). To study whether these more impolite reviews for female first authors were due to an overall lower politeness score, or due to one or some of the reviewers being more impolite, I split the reviews for each paper by its lowest/ median/highest politeness score. I observed that the lower politeness scores for first authors with a female name was driven by significantly lower low and median scores (**Fig. 4d**, bottom panel). Thus, the least polite reviews a paper received were even more impolite for papers with a female first author. Conversely, female senior authors received significantly higher sentiment scores, indicating more favorable reviews, but these reviews did not differ in terms of politeness (**Fig. 4e**). An analysis of reviews split by lowest/median/highest sentiment score indicated that the reviewer that gave the most favorable review to female senior authors did so with a significantly higher score (**Fig. 4e**, bottom panel). No interactions on scores were observed between the genders of the first and senior authors (**Supplementary Fig. 5**).

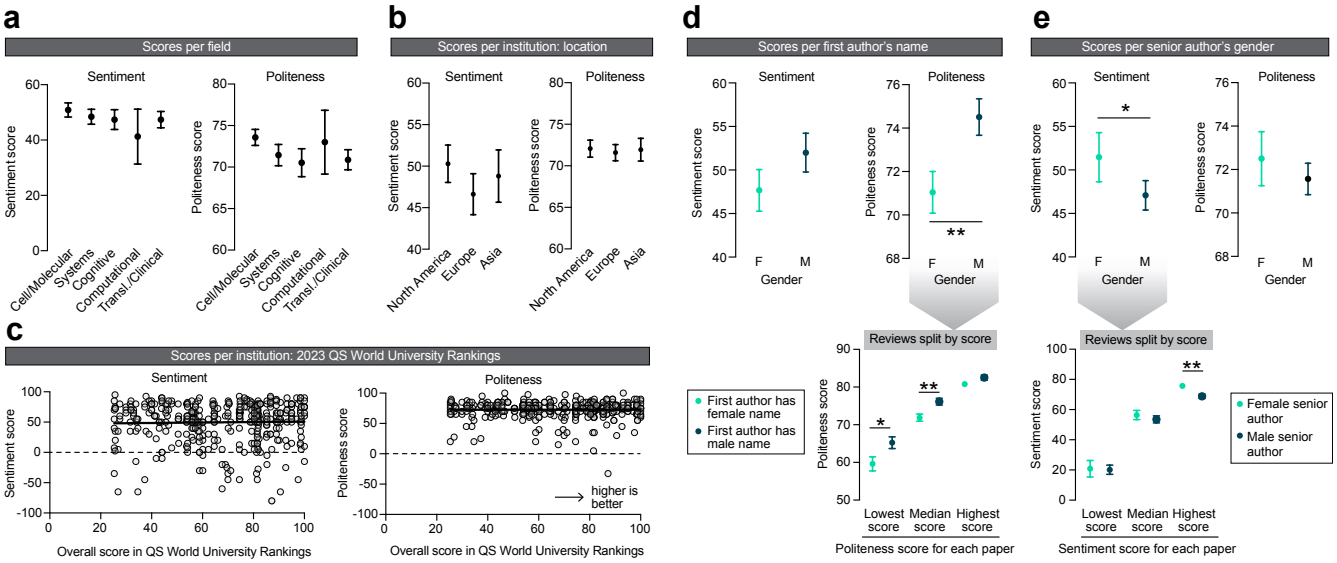


Figure 4: Exploring disparities in peer review

(**a**) Effects of the subfield of neuroscience on sentiment (left) and politeness (right) scores. No effects were observed on sentiment (Kruskal-Wallis ANOVA, $H = 2.380$, $P = 0.6663$) or politeness (Kruskal-Wallis ANOVA, $H = 8.211$, $P = 0.0842$). $n = 178, 149, 100, 20, 125$ reviews per subfield.

(**b**) Effects of geographical location of the senior author on sentiment (left) and politeness (right) scores. No effects were observed on sentiment (Kruskal-Wallis ANOVA, $H = 1.856$, $P = 0.3953$) or politeness (Kruskal-Wallis ANOVA, $H = 0.5890$, $P = 0.7449$). $n = 239, 208, 103$ reviews per continent.

(**c**) Effects of QS World Ranking score of the senior author' institutional affiliation on sentiment (left) and politeness (right) scores. No effects were observed on sentiment (Linear regression, $R^2 = 0.0006$, $P = 0.6351$) or politeness (Linear regression, $R^2 < 0.0001$, $P = 0.9804$). $n = 430$ reviews.

(**d**) Effects of the first author's name on sentiment (left) and politeness (right) scores. No effects were observed on sentiment (Mann-Whitney test, $U = 19521$, $P = 0.2131$) but first authors with a female name received significantly less polite reviews (Mann-Whitney test, $U = 17862$, $P = 0.0080$, Hodges-Lehmann difference of 2.5). Post-hoc tests on the data split per lowest/median/highest politeness score indicated significantly lower politeness scores for females for the lowest (Mann-Whitney test, $U = 1987$, $P = 0.0103$, Hodges-Lehmann difference of 5) and median (Mann-Whitney test, $U = 1983$, $P = 0.0093$, Hodges-Lehmann difference of 2.5) scores, but not of the highest score (Mann-Whitney test, $U = 2279$, $P = 0.1607$). $n = 206$ (F), 204 (M) reviews for top panels; $n = 71$ (F), 74 (M) papers for lower panel (but $n = 54$ (F), 53 (M) papers for median scores, because not all papers received 3 reviews).

(**e**) Effects of the senior author's gender on sentiment (left) and politeness (right) scores. Women received more favorable reviews than men (Mann-Whitney test, $U = 28007$, $P = 0.0481$, Hodges-Lehmann difference of 5) but no effects were observed on politeness (Mann-Whitney test, $U = 29722$, $P = 0.3265$). Post-hoc tests on the data split per lowest/median/highest sentiment score indicated no effect of gender on the lowest (Mann-Whitney test, $U = 3698$, $P = 0.7963$) and median (Mann-Whitney test, $U = 3310$, $P = 0.1739$) sentiment scores, but the highest sentiment score was higher for women (Mann-Whitney test, $U = 2852$, $P = 0.0072$, Hodges-Lehmann difference of 5). $n = 155$ (F), 405 (M) reviews for top panels; $n = 53$ (F), 143 (M) papers for lower panel (but $n = 39$ (F), 102 (M) papers for median scores, because not all papers received 3 reviews). Asterisks indicate statistical significance in Mann-Whitney tests; * $P < 0.05$, ** $P < 0.01$.

DISCUSSION

Peer review is a crucial component of scientific publishing. It helps ensure that research papers are of high quality and have been scrutinized by experts in the field. However, the potential for subjectivity in the peer review process has been an ongoing concern. For example, implicit or explicit bias of reviewers may lead to disparities in peer review scores on the basis of gender or institutional affiliation. In this study, I used natural language processing tools embedded in OpenAI's *ChatGPT* to analyze 572 peer review reports from 200 papers that were accepted for publication in *Nature Communications* within the past year. I found that this approach was able to provide consistent and accurate scores, matching that of human scorers. Importantly, *ChatGPT* was superior to the conventional lexicon- and rule-based algorithms *TextBlob* and *VADER* in scoring a review's sentiment.

Such algorithms score a text on the basis of the frequency of certain words, and as such may have trouble analyzing scientific text with specialized constructions and vocabulary[13], as has been shown before[15]. Altogether, the current study serves as a proof of concept for the use of generative artificial intelligence in studying scientific peer review. Such an automated language analysis of peer reviews can be used in different ways, such as after-the-fact analyses (as has been done here), providing writing support for reviewers (for example by implementation in the journal submission portal), or by helping editors pick the best papers or most constructive reviewers.

Notably, there are several limitations to this study. The peer review reports I analyzed are all ultimately accepted for publication in *Nature Communications*, meaning that there is a selection bias in the reviews that were included. As such, papers that have received unfavorable reviews, or papers that have not been sent out for peer review at all, were not included in this analysis. It is unclear what the gender and institutional affiliation distribution is for the papers that were ultimately unpublished. Additionally, this study only focused on the neuroscience field, and the findings may not generalize to other fields. Similarly, it is not clear if the results from this study apply to journals beyond *Nature Communications*. Future studies may expand upon this initial work by incorporating larger sample sizes and encompassing diverse scientific disciplines and journals.

Despite said limitations, this study may reveal several key insights into the peer review process and highlight potential areas of concern within academic publishing. First, this study found that evaluations of the same manuscript varied considerably among different reviewers. This finding suggests that the peer review process may be subjective, with different reviewers having different opinions on the quality and validity of the research. Notably, some level of variability may be expected, for example due to different backgrounds, experiences, and biases of the reviewers. In addition, *ChatGPT* may not always reliably assess a reviews sentiment, adding some spurious inter-reviewer variability. That being said, the extremely low (or even absent) relation between how different reviewers scored the same paper was striking, at least to this author. The inconsistency in the evaluations emphasizes the need for greater standardization in the peer review process, with clear guidelines and protocols that can minimize such discrepancies[20].

I also investigated disparities in peer review based on the institutional affiliation of the senior author of a paper. Specifically, I looked at the geographic location (continent), as well as the score of the institute in the 2023 QS World University Rankings — an imperfect metric of the institute's perceived prestige. This analysis revealed no relation of these two metrics with the sentiment and politeness of the reviews, suggesting that evaluations were not influenced by the geographical location and perceived prestige of the senior author's research institution. This finding is encouraging and suggests that peer review may be based on the quality and merit of the research rather than the authors' research institute. That said, the identity of the peer reviewers is not known, so it cannot be tested whether reviewers have a bias with respect to authors from a more closely related country, culture or institution (i.e., in-group favoritism). In addition, it's important to acknowledge the selection bias present in this study, in which I exclusively considered published papers. This may mask effects resulting from bias with regards to the senior author's institutional affiliation. For example, papers from less prestigious institutions may have a higher rejection rate. To address this concern, future studies could adopt a strategy such as partnering with a journal to analyze the review sentiment associated with both rejected and accepted papers.

This study further found that first authors with a female name received less polite reviews than first authors with a male name, although this did not affect the favorability of their reviews. Regardless, this disparity is worrisome as it may indicate an unconscious gender bias in review writing that may ultimately impact the confidence and motivation of (especially early-stage) female researchers. One may argue that the effect size of gender on politeness scores is small, but given the selection bias in this dataset (**Fig 2c**, bottom right inset), this effect may be larger in the entire pool of reviewed manuscripts (i.e., rejected + accepted). To address this issue, double-blind peer review, where the authors' names are anonymized, could be implemented. Evidence suggests that this is useful in removing certain forms of bias from reviewing[8,9], but has thus far not been widely implemented, perhaps

because some studies have cast doubt on its merits[21,22]. Additionally, reviewers could be more mindful of their language use. Indeed, even negative reviews can be written in a polite manner (**Fig. 2c**), and reviewers may want to use *ChatGPT* to extract a politeness score for their review before submitting.

Additionally, female senior authors received more favorable reviews than male senior authors in this pool of accepted papers. This disparity in sentiment score in favor of women may be surprising given the wealth of data showing unconscious bias against women, including in scientific research[23,24]. It is therefore likely that the observed effect is due to selection bias elsewhere in the publishing process. There may be two potential sources of this bias. The first one is that female senior authors may submit better papers to this journal than their male peers, such that the observed gender effect on sentiment is representative for the entire pool of submitted manuscript (i.e., rejected + accepted). This could be the result of institutional barriers that lead to a small, but highly talented pool of female principal investigators[25] that submits better papers than their male peers[26]. Alternatively, women may have a higher level of self-imposed quality control[27], such that men submit more variable quality papers to high-impact journals like *Nature Communications*. In the imperfect process that is editorial decision making, this may lead to the publication of certain lower-quality papers from male senior authors. The second explanation may be related to an (unconscious) selection bias in the editorial process[28], requiring female senior authors to have better papers before being sent out for peer review, or better scores before being invited for a revise-and-resubmit. As such, paper acceptance may serve as a collider variable[29,30], inducing a spurious association between gender of the senior author and sentiment score. Further research is required to investigate the reasons behind this effect and to identify in what level of the publishing system these differences emerge. In **Supplementary Fig. 6**, I propose three different experiments that journals can perform to rule out bias in reviewing or the editorial process.

Together, this study serves as a proof of concept for the use of generative artificial intelligence in analyzing scientific peer review. *ChatGPT* outperformed commonly used natural language processing tools in measuring sentiment of peer reviews, and provides an easy, non-technical way for people to perform language analyses on specialized scientific texts. Using this approach, areas of concern were discovered within the academic publishing system that require immediate attention. One such area is the inconsistency between the reviews of the same paper, indicating some level of subjectivity in the peer review process. Additionally, I uncovered possible gender disparity in academic publishing and reviewing. This research underscores the potential of generative artificial intelligence to evaluate and enhance scientific peer review, which may ultimately lead to a more equitable and just academic system.

METHODS

Downloading reviews

Reviewer reports were downloaded from the website of *Nature Communications* in February 2023. Only papers that were categorized under *Biological sciences* > *Neuroscience* were included in this analysis. Not all papers had their primary reviewer reports published; to reach the total of 200 papers with primary review reports, the most recently published 283 papers were considered (published between August 16, 2022 and Feb 17, 2023).

Additional paper metrics were subsequently collected. Paper submission and acceptance date were downloaded from the '*About this article*' section on the paper website. Paper acceptance time was calculated by counting the number of days between these two dates. Research field was manually categorized on the basis of title and abstract of the paper into five different subfields. The affiliation of the senior author was downloaded from the paper website and manually categorized based on continent; if the senior author had affiliations across multiple continents, it was categorized as 'multiple' and not used for further analyses (this was the case for 5 papers). The affiliated institutions' score in the 2023 QS World Ranking was downloaded from the QS World Ranking website (TopUniversities.com) in March 2023; the maximum score an institution could receive was 100. Not all institutions were listed in the QS World Ranking, usually because they were not considered an organization of higher education. If a senior author had multiple affiliations, then the affiliation with the highest score was used. Name-based gender categorization of the first author was performed using *ChatGPT* (query: "*Of the following list of international full (first+last) names, can you guess, based on name only, if these people are male,*

female, or unknown (i.e., name is not gender specific)?"). As a confirmation, all names that were assigned a gender by *ChatGPT* were verified using the Genderize database (*http://genderize.io*; probability > 0.5). The gender of the senior author was categorized in a similar manner, except that the categorization for gender-unspecific names was manually completed, usually by looking up the senior author on the research institution's website or the author's Google Scholar or Twitter profile. In this manual look up, I tried to find the senior author's preferred pronouns. If not available, I inferred the senior author's gender on the basis of a photograph. I did not find evidence that any of the senior authors included in this analysis identified as non-binary; for 4 senior authors I was not able to find or infer their gender. Note that this gender look-up was performed for the senior author, but not for the first author, for two reasons. First, first authors generally had less of an online presence than seniors authors, and it was challenging to reliably assess their gender identity. Second, I presumed that reviewers are more likely to be familiar with the senior author of papers they review (for example through conferences) than with first authors. As such, reviewers themselves may infer the gender of the first author on name only.

Sentiment analysis

Scores of sentiment and politeness of language use of each peer review report was performed using OpenAI's *ChatGPT* (GPT-3.5, Version Feb 13, 2023). The prompt consisted of the following question (see **Fig. 3a**):

Below you will find a scientific peer review. Such reviews generally contain the reviewer's sentiment in the first paragraph(s) of the review, followed by a list of specific recommendations to the authors. Can you score this peer review on [1] the sentiment, on a scale from -100 (negative) to 0 (neutral) to 100 (positive), and [2] politeness of language use, on a scale of -100 (rude) to 0 (neutral) to 100 (polite)?

followed by the full text of the peer review. This question was entered into *ChatGPT* twice and the average of these scores was used for further analyses; for a correlation between the two iterations see **Supplementary Fig. 2**. Note that *ChatGPT* has become more reliable in recent updates, such that different iterations of scoring now produces a highly reproducible score (see **Supplementary Fig. 2**).

Statistics

To test the consistency across different reviewers of the same paper (**Aim 2; Fig. 3**), I used a combination of a mixed model, linear regression models and intra-class correlation coefficients. For **Fig. 3a** (differences between reviewers 1, 2 and 3), a mixed effects model was used to compute statistical significance, because repeated measures data was not always available (i.e., not all papers received a third review). This analysis was performed in Prism 9 (GraphPad Inc.). Linear regression and intra-class correlational analyses in **Fig. 3b** (sentiment scores across reviewers) and **Fig. 3c** (review scores vs. paper acceptance time) were performed using JASP 0.16 (University of Amsterdam). For the intra-class correlational analyses of **Fig. 3b**, ICC type ICC1,1 was used; because ICC is particularly sensitive to the assumption of normality, sentiment scores were first log transformed. For the polynomial linear regression in **Fig. 3c**, data were centered by *z*-scoring the individual sentiment and politeness scores.

To test the effects of author identity on review scores (**Aim 3; Fig. 4**), I used a combination of the Kruskal-Wallis ANOVA and Mann-Whitney tests. Note that review scores were not always normally distributed, so non-parametric tests were mostly used. To compute statistical significance in **Fig. 4a** (scores per field) and **Fig. 4b** (scores per institution location), a Kruskal-Wallis ANOVA was used. For **Fig. 4c** (scores correlated with 2023 QS World University Ranking), significance was calculated using linear regression. For **Fig. 4d and 4e**, Mann-Whitney tests were used to compute significance between male and female authors. Significant effects were further studied by splitting the reviews per score (i.e., splitting in lowest, median and highest scores per paper). To calculate statistical significance between male and female authors for lowest/median/highest score in **Fig. 4d and 4e**, Mann-Whitney tests were used. Statistical tests were always two-tailed. All analyses in **Fig. 4** were performed in Prism 9 (GraphPad Inc.). Significance was defined as $P < 0.05$ and denoted with asterisks; * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

DATA AVAILABILITY

All data are available as a Supplementary file to this paper.

ACKNOWLEDGEMENTS
J.P.H.V. was supported by a Rubicon postdoctoral fellowship from the Netherlands Organization of Scientific Research. I thank Amanda Tose, Han de Jong and Stephan Lammel for helpful comments on this manuscript.

CONFLICT OF INTEREST STATEMENT
The author declares no competing interests.

REFERENCES
1. Park, I., Peacey, M.W., Munafò, M.R. Modelling the effects of subjective and objective decision making in scientific peer review. Nature 506, 93-06 (2014)
2. Lipworth W.L., Kerridge I.H., Carter, S.M., Little, M. Journal peer review in context: A qualitative study of the social and subjective dimensions of manuscript review in biomedical publishing. Social Science & Medicine, 72, 1056-1063 (2011)
3. King, E.B., Avery, D.R., Hebl, M.R., Cortina, J.M., Systematic Subjectivity: How Subtle Biases Infect the Scholarship Review Process. Journal of Management 44, 843-853 (2018)
4. Lee, C.J., Sugimoto, C.R., Zhang, G., Cronin, B. Bias in peer review. Journal of the American Society for Information Science and Technology 64, 2-17 (2013)
5. Abramowitz, S.I., Gomes, B., Abramowitz, C.V. Publish or Politic: Referee Bias in Manuscript Review. Journal of Applied Social Psychology 5, 187-200 (1975)
6. Blank, R.M. The effects of double-blind versus single-blind reviewing — experimental evidence from the American-Economic review. The American economic review 81, 1041-1067 (1991)
7. Lundine, J. Bourgeault, I.L., Glonti, K. Hutchinson, E., Balabanova, D. “I don't see gender”: Conceptualizing a gendered system of academic publishing. Social Science and Medicine 235, 112388 (2019).
8. Fox, C.W., Meyer, J., Aimé, E., Double-blind peer review affects reviewer ratings and editor decisions at an ecology journal. Functional Ecology 37(5): 1144-1157 (2023).
9. Tomkins, A., Zhang, M., Heavlin, W.D., Reviewer bias in single- versus double-blind peer review, PNAS 114, 12708-12713 (2017)
10. Chowdhary, K.R., Natural Language Processing. In: Fundamentals of Artficial Intelligence. Springer, New Delhi (2020)
11. Hirschberg, J., Manning, C.D. Advances in natural language processing. Science 349, 261-266 (2015)
12. Yadav, A. Vishwakarma, D.K., Sentiment analysis using deep learning architectures: a review. Artificial Intelligence Review 53, 4335-4385 (2020)
13. Ghosal, T., Kumar, S., Bharti, P.K., Ekbal, A., Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. PLoS ONE 17(1), e0259238 (2022).
14. Chakraborty, S., Goyal, P., Mukherjee, A., Aspect-based Sentiment Analysis of Scientific Reviews, JCDL '20: Proceedings of the ACM/ IEEE Joint Conference on Digital Libraries in 2020, 207-216 (2020)
15. Luo, J., Feliciani, T., Reinhart, M., Hartstein, J., Das, V., Alabi, O., Shankar, K., Analyzing sentiment in peer review reports: Evidence from two science funding agencies. Quantitative Science Studies 2(4), 1271-1295 (2021).
16. Transparent peer review for all. Nature Communications 13, 6173 (2022).
17. Loria., S, TextBlob Documentation. Release v0.16.0 (2023),
18. Hutto, C.J., Gilbert, E.E., VADER: A Parsiminous Rule-based Model for Sentiment Analysis of Social Media Text. Eight International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI (2014).
19. Liljequist, D., Elfving, B., Skavberg Roaldsen, K., Intraclass correlation - A discussion and demonstration of basic features. PLoS ONE 14(7), e0219854 (2019).
20. Tennant, J.P., Ross-Hellauer, T. The limitations to our understanding of peer review. Research Integrity and Peer Review 5, 6 (2020)
21. Alam, M. et al. Blinded vs. unblinded peer review of manuscripts submitted to a dermatology journal: a randomized multi-rater study. Clinical and Laboratory Investigations 165, 563-567 (2011)
22. Snodgrass, R. Single- versus double-blind reviewing: an analysis of literature. ACM SIGMOD Record 35, 8-21 (2006)
23. Blickenstaff, J.C., Women and science careers: leaky pipeline or gender filter? Gender and Education 17, 369-386 (2006)
24. Pell, A.N., Fixing the leaky pipeline: women scientists in academia. Journal of Animal Science 74, 2843-2848 (1996)

25. Sheltzer, J.M., Smith, J.C., Elite male faculty in the life sciences employ fewer women. PNAS 111, 10107-10112 (2014)
26. Hengel, E., Publishing while female: Are women held to higher standards? Evidence from peer review. The Economic Journal 132, 2951-2991 (2022).
27. White, K., Women and leadership in higher education in Australia. Tertiary Education and Management 9, 45-60 (2010)
28. Matías-Guiu, J., García-Ramos, R., Editorial bias in scientific publications. Neurología 26, 1-5 (2011).
29. Holmberg, M.J., Andersen, L.W., Collider Bias. JAMA 327, 1282-1283 (2022).
30. Griffith, G.J. et al., Collider bias undermines our understanding of COVID-19 disease risk and severity. Nature Communications 11, 5749 (2020).

SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure 1: Validation #1 — Examples of *ChatGPT* inputs and outputs

Supplementary Figure 2: Validation #2 — Consistency in sentiment and politeness scores for two different times *ChatGPT* was asked to analyze review reports. (Top panels) A significant correlation was found between both times *ChatGPT* was asked to score the sentiment and politeness of the peer reviews. The average score of both iterations was used for all subsequent analyses in this paper. (Bottom panel) The relatively low R^2 values between both iterations shown in the top panels have greatly improved in versions of *ChatGPT* since the initial submission of this paper; R^2 values have now increased to 0.992 (for sentiment) and 0.859 (politeness).

Supplementary Figure 3: Validation #3 — Example of a manually manipulated review, showing that *ChatGPT* can pick up artificial changes in sentiment and language use. It should be noted here that changes in sentiment usually also affected the politeness score (and vice versa), indicating that these scores are not fully independent. This may intuitively make sense; less polite language is often interpreted as more negative, also by human readers (see **Supplementary Fig. 4**).

Supplementary Figure 4: Validation #4 — Comparison of *ChatGPT*'s scores of sentiment and politeness as compared to seven (blinded) human scorers for a diverse sample of reviews. Figure showing high levels of variability across human scorers, but their average score had a high correlation with *ChatGPT*'s score. For this experiment, human scorers were asked to score the sentiment of seven reviews on a scale from very negative - negative - neutral - positive - very positive. Politeness was scored on a scale from very rude - rude - neutral - polite - very polite. For the cross-correlograms in the bottom panels, human scores were first converted to numbers on a scale from 1-5, so that these could be correlated to *ChatGPT*'s numerical scores. Scorers 1, 2 and 5 were scientists with a PhD; scorer 3, 4 and 6 were neuroscience graduate students; scorer 7 was a non-scientist. Please note that *TextBlob* and *VADER* do not provide politeness scores, and were thus not included in the second table. Asterisks indicate significance in a linear regression; * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

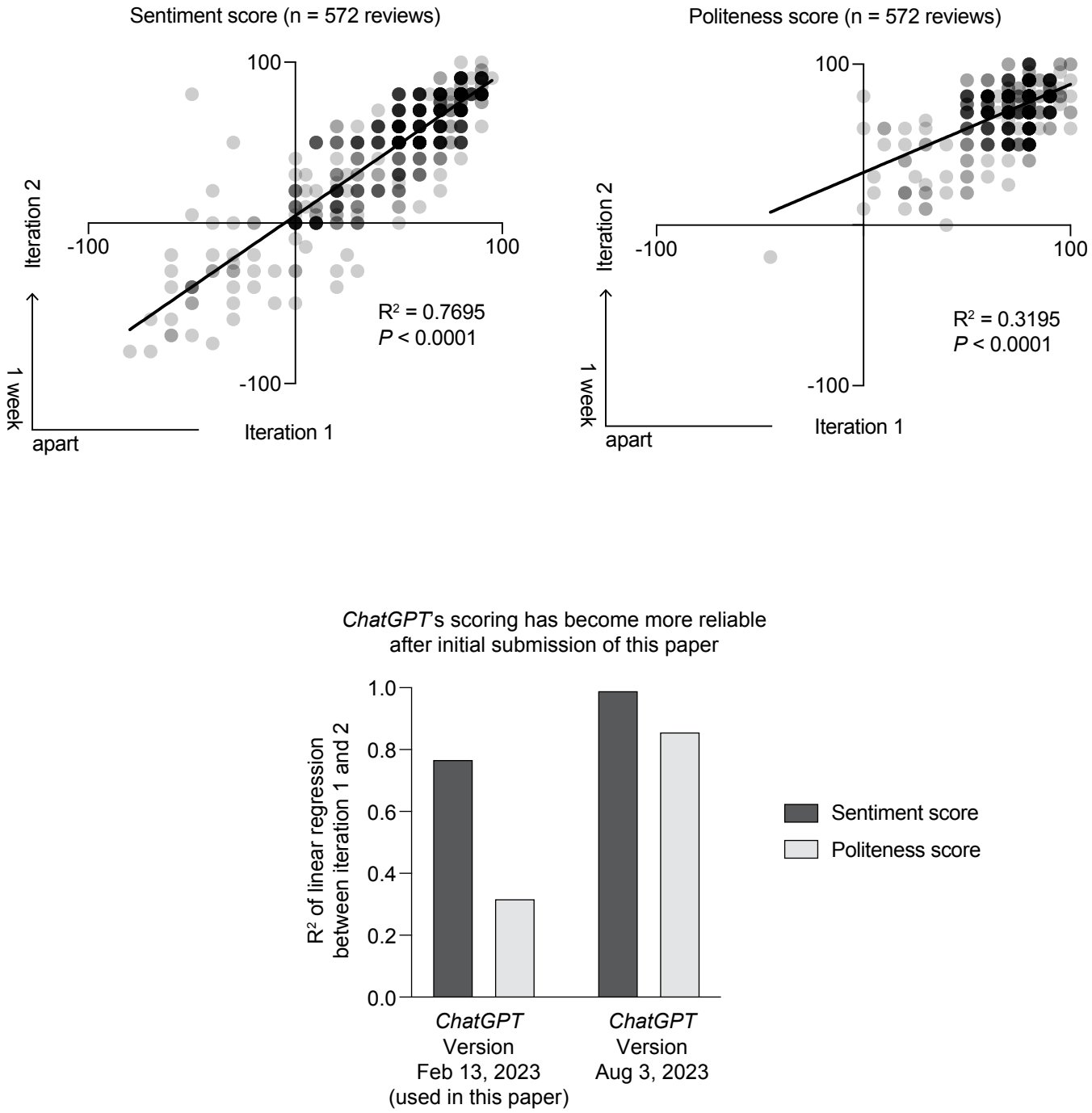
Supplementary Figure 5: Sentiment and politeness scores for papers in different gender groups. As an example, 'F/M' indicates that the first author had a female name and that the senior author was male. Two-way ANOVAs on these data indicated no significant first × last author gender interaction effects for both sentiment ($F(1,397) = 0.05291$, $P = 0.8182$) and politeness ($F(1,397) = 0.02808$, $P = 0.8670$). $n = 73$ reviews for F/F, 126 for F/M, 44 for M/F, 158 for M/M.

Supplementary Figure 6: Experiments that journals can perform to rule out gender bias in reviewing and editorial decision making. If these experiments show no evidence of bias, gender differences likely emerge higher up in the academic pipeline.

Verharen Supplementary Fig. 1

Excerpt from review	*ChatGPT* analysis on entire review
From Shahin et al (Review #1) Overall there are some interesting findings in this study associated with the impact of loss of Ncf1 in NSCs. However, some of the presented data is phenomenological with conclusions that are unclear/equivocal. The claim that inactivating DNA repair promotes neurosphere forming ability (i.e. similar to WT cells) is surprising and it's difficult to understand why inhibiting key genome stability factors and increasing DNA damage (shown by H2AX formation) would promote recovery and be beneficial in terms of NSC function. While this might be a possibly interesting finding, as it stands there is too much uncertainty about what it means and how it happens. The data presented doesn't address how regulation of repair pathways modulates NSC function, if its physiologically relevant and the mechanistic nature of what DNA damage does to alter cell fate. (…)	Sentiment: -30 (slightly negative) Politeness: 20 (somewhat polite) The sentiment of this peer review is negative as the reviewer expresses several concerns about the clarity, relevance, and plausibility of the presented data and conclusions. The reviewer questions the significance and mechanism of the findings and highlights the limitations and inconsistencies of the study. However, the language use is not overly aggressive or impolite, and the reviewer provides constructive criticism and suggests possible ways to improve the study. Therefore, the politeness score is relatively neutral.
From Kang et al (Review #2) This manuscript reported a sequential narrow-field (SNF) stimulation method by applying sequential electrical stimulation on pairs of electrodes implanted in the hippocampus to control the localization of electric field. (…) This research is generally well designed and the manuscript is well organized. However, the idea of using multi implanted electrodeto delivery small electric current to control the activation volume is lack of novelty, even this method is designed for seizure control. There are also several major questions need to be resolved: (…)	Sentiment score: 10 Politeness score: 70 The sentiment of this peer review is somewhat positive, with a score of 10. The reviewer acknowledges the novelty of the research and praises the manuscript's organization and design. However, the reviewer also points out several major questions that need to be addressed, which lowers the sentiment score. The politeness score of this peer review is 70, indicating a polite and respectful tone. The reviewer uses professional language and structures their criticism in a constructive way, offering specific recommendations for improvement. The reviewer also acknowledges the strengths of the research and the manuscript, which shows a respectful attitude towards the authors.
From de Thonel et al (Review #1) This manuscript represents a major advance in understanding how mutations in a global protein acetylase, underlying a rare neurodevelopmental disorder, impacts cellular stress protection and neurodevelopment. (…) The authors present a very strong case for specific interactions between HSF2 within the KIX domain found in the coiled-coil region and CBP/p300 and validate the interaction with biochemical evidence, molecular modeling and mutagenesis. Moreover, while previous reports demonstrated the regulation of HSF2 via protein degradation, this work deciphers key mechanisms that regulate this process via protein lysine acetylation. Finally, the current work underscores the contributions of HSF2, and its dysregulation in RSTS, to stress-regulated gene expression and its impact in neuronal cell function. Overall, the results reported here are novel and highly impactful to the fields of stress regulation, neurodevelopmental processes and rare diseases.	Sentiment score: 90 Politeness score: 80 The sentiment of this peer review is very positive, with a score of 90. The reviewer praises the manuscript's rigor, strong case, and novel and impactful results. The reviewer also acknowledges the significance of the research in several fields, which raises the sentiment score. The politeness score of this peer review is 80, indicating a highly polite and respectful tone. The reviewer uses professional language and expresses admiration for the authors' work throughout the review. The reviewer's criticism is constructive and focuses on the strengths of the research, which shows a high level of politeness and respect towards the authors.
From Singleton et al. (Review #2) In another reanalysis alongside the 15+ other reanalyses of this dataset (more publications than participants), Singleton and colleagues provide partial evidence for the REBUS model. Although this may be one of the better tests of REBUS, the salami slicing has gotten out of control. Not only is the small sample (N = 15) not independent of all the other tests of REBUS, the analyses themselves are not necessarily independent (…) Although this paper should not be published in Nature Communications (or anywhere until all other datasets are integrated), wherever it may end up, the following points must be considered: (…) Not only am I not a fan of such fluffy writing that has become rampant in psychedelic publications, to act like this is some special feature of EBH is to ignore literally all of cognitive neuroscience. Nobody in cognitive neuroscience assumes dualism (at least explicitly). (…)	Sentiment score: -80 Politeness score: -45 The sentiment in this peer review is negative, with a score of -80. The reviewer is critical of the paper, particularly in relation to the small sample size and the "salami slicing" of the data. The reviewer also notes that some analyses are missing and recommends running k means clustering separately on placebo and LSD data. The language used is not particularly polite, with a score of -45. The reviewer uses language that is critical and accusatory, and does not make an effort to soften the tone of the review.

Verharen Supplementary Fig. 2



Verharen Supplementary Fig. 3

- ORIGINAL REVIEW -

This paper aims to show that Area MT in monkeys has a direct correlation with the perception of motion. Several authors have already demonstrated that MT appears to be involved in 'driving' the perception of motion. The specific issue investigated here is whether MT (and particularly one cell type in MT) is involved in the perception of motion segmentation judgements. The most important finding is that a subset of cells in Area MT, which have been labelled pattern cells based on a single test, is particularly correlated with segmentation. The work is certainly of interest to people in the field. However, as the authors mention themselves, one of the largest problems in my mind is the importance of the finding in a broader context. Pattern and Component cells are defined based a on a very restricted stimulus-type: namely plaid patterns. Pattern and component cells do not form two distinct groups but, rather are defined based on a statistical test that splits cells into three groups that do not naturally split into three separate clusters.
There is no border between these groups (i.e. statistics don't allow you to identify three groups with dips between them). In truth, there is a continuum and most cells are a bit of both extremes. Plaid patterns contain restricted amounts of information on motion direction, so the lack of clear boundaries is not surprising. Therefore, saying that pattern cells are more closely linked to the perception of transparent motion appears philosophically to be a rather small increment. For example, what would happen if more certainty were built into the stimulus, such as occurs immediately, perceptually, when terminators appear in the RF? Would cells defined as pattern-selective differ from the other MT cells in this case and would this influence the bias of the single-cell responses? Would the 'special' case for pattern cells persist? The work was conducted in behaving primates and this is a very challenging and time consuming technique. To the best of my knowledge the work was conducted at the highest level and the statistics are appropriate. I believe that the experiments could be repeated by following the methods. I am not in a position to analyse the data to see if I can reproduce the same result. Therefore, while the work is of the highest quality, the paper suffers in my view from not clearly showing how the MT population code as a whole deals with transparency and segmentation judgements. I feel that a model that incorporates all of the response types in MT would be essential in trying to answer this. The authors could develop a data-driven model that uses their hard-won monkey data. As it stands, I feel that the paper attempts to highlight a correlation between perception and just one type of MT cell (i.e. pattern cells).

Sentiment: 10 Politeness: 75

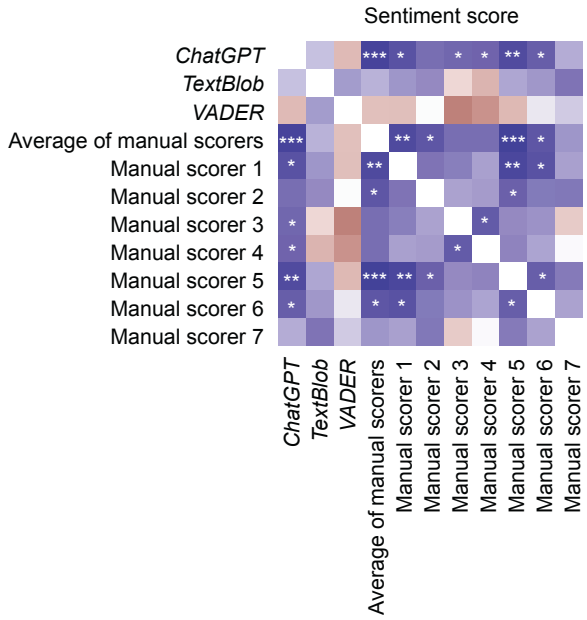
Change	Examples of changes (highlighted sentences)	New scores
More positive	The work will be of great interest to people in the field. However, as the authors mention themselves, one small problem in my mind may be the importance of the finding in a broader context. The authors could develop a data-driven model that uses their hard-won monkey data. As it stands, I feel that this paper highlights a major advance between perception and one type of MT cell (i.e. pattern cells).	Sentiment: 67.5 (+57.5) Politeness: 60
More negative	The work is unlikely to be interest to people in the field. Indeed, as the authors mention themselves, an insurmountable problem in my mind is the importance of the finding in a broader context. The authors could develop a data-driven model that uses their hard-won monkey data. As it stands, I feel that the paper attempts to highlight an unimportant correlation between perception and just one type of MT cell (i.e. pattern cells).	Sentiment: -40 (-50) Politeness: 50
More polite	The work is certainly of interest to people in the field. Respectfully, as the authors mention themselves, one may doubt the the importance of their findings in a broader context. The authors could develop a data-driven model that uses their hard-won monkey data. Despite the author's hard work, I feel that the paper attempts to highlight a correlation between perception and one type of MT cell (i.e. pattern cells).	Sentiment: 50 Politeness: 90 (+15)
Less polite	I gotta say, some people will probably like this paper. However, as the authors mention themselves, their findings are just not that important. The authors must develop a data-driven model that uses their hard-won monkey data. Frankly, as it stands, the paper merely attempts to highlight a correlation between perception and just one type of MT cell (i.e., pattern cells). Not exactly groundbreaking, is it?	Sentiment: -20 Politeness: -55 (-130)

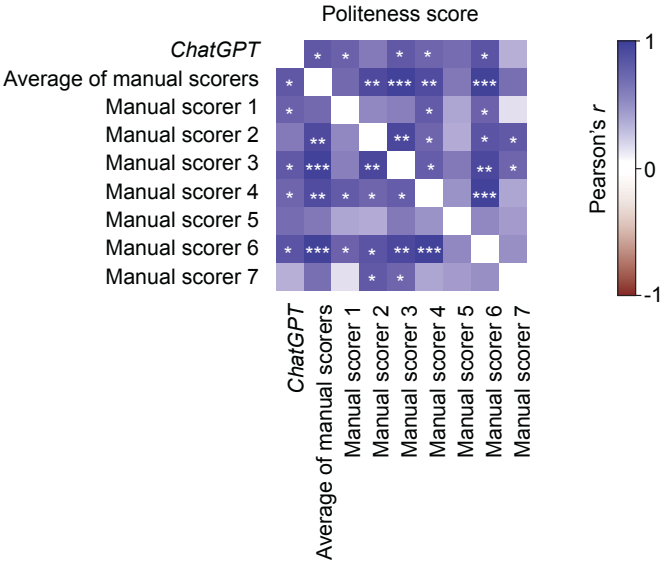
Verharen Supplementary Fig. 4

	Sentiment									
	ChatGPT	Textblob[1]	VADER[2]	7 human scorers						
Paper				#1	#2	#3	#4	#5	#6	#7
Singleton et al	-80	7	99	Very negative	Very negative	Very negative	Negative	Very negative	Negative	Negative
Gaynes et al	-10	11	97	Negative	Negative	Very negative	Negative	Negative	Negative	Neutral
Quintana et al	30	8	-87	Positive	Very negative	Neutral	Neutral	Neutral	Neutral	Negative
Strembitska et al	40	14	94	Very positive	Positive	Negative	Negative	Positive	Positive	Positive
Audrain et al	72.5	10	100	Positive	Neutral	Positive	Positive	Neutral	Positive	Negative
Cicchini et al	75	10	100	Positive	Neutral	Negative	Positive	Positive	Positive	Positive
Bretheau et al	77.5	8	-97	Positive	Positive	Positive	Very positive	Positive	Neutral	Neutral

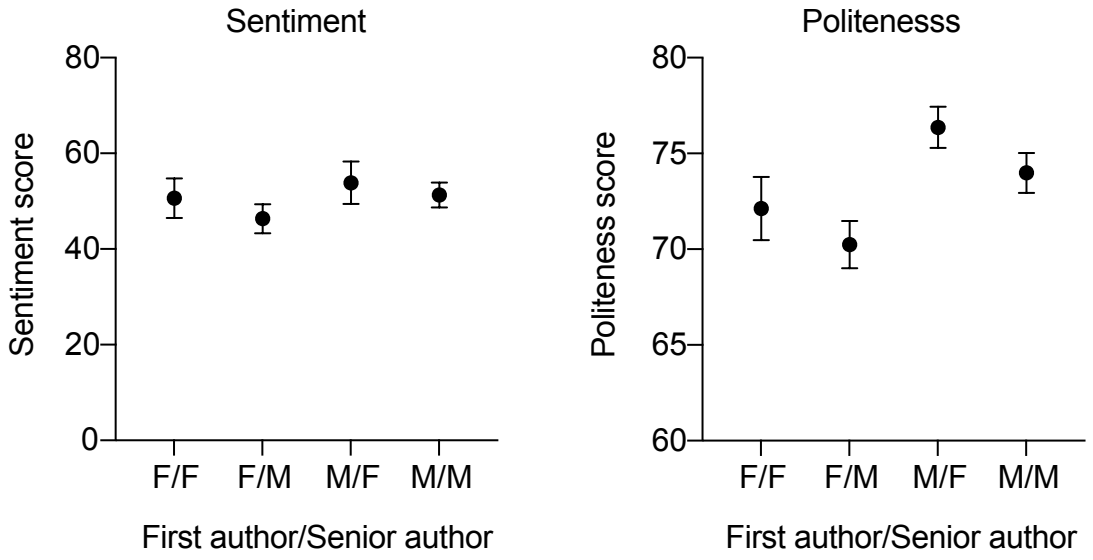
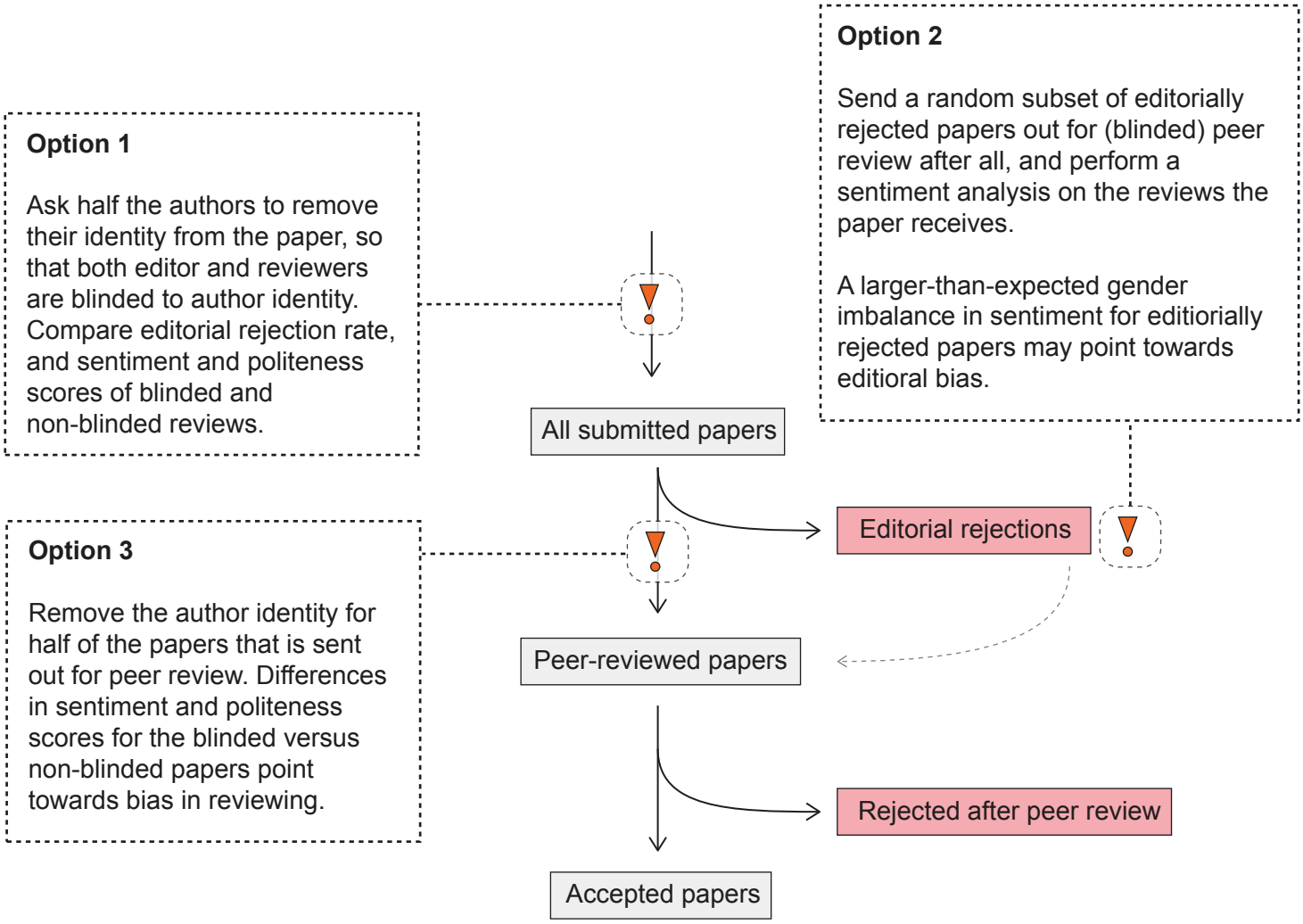
[1] 'Polarity' in Textblob ranges between [-1, 1]. This value was multiplied by 100 to match *ChatGPT*'s score range [-100, 100].
[2] 'Compound polarity score' in VADER ranges between [-1, 1]. This value was multiplied by 100 to match *ChatGPT*'s score range [-100, 100].

	Politeness									
	ChatGPT	Textblob	VADER	7 human scorers						
Paper				#1	#2	#3	#4	#5	#6	#7
Singleton et al	-32.5	n/a		Very rude	Very rude	Very rude	Very rude	Rude	Very rude	Neutral
Gaynes et al	50			Neutral	Very rude	Rude	Rude	Polite	Rude	Neutral
Bretheau et al	50			Neutral	Polite	Polite	Polite	Neutral	Polite	Very polite
Strembitska et al	62.5			Very polite	Polite	Polite	Polite	Neutral	Polite	Polite
Quintana et al	70			Polite	Rude	Neutral	Polite	Neutral	Polite	Rude
Audrain et al	77.5			Polite	Polite	Very polite	Very polite	Very polite	Very polite	Very polite
Cicchini et al	85			Neutral	Polite	Very polite	Neutral	Polite	Polite	Very polite





Verharen Supplementary Fig. 5



Verharen Supplementary Fig. 6