Research Square

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

National-background bias in scientific evaluation based on large language models

Hao Kong
Max Planck Institute for Chemistry
Yafang Cheng
yafang.cheng@mpic.de
Max Planck Institute for Chemistry
Ruijing Ni
Sichuan Academy of Environmental Policy and Planning
Wenjun Meng
Max Planck Institute for Chemistry
Chaoqun Ma
Max Planck Institute for Chemistry

Research Article

Keywords: Large language models, Identity bias, Matthew effect, Scientific evaluation, Algorithmic fairness

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11755021/v1

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

Additional Declarations: The authors declare no competing interests.

National-background bias in scientific evaluation based on large language models

Hao Kong[1, 2], *Yafang Cheng*[1 *], *Ruijing Ni*[3], *Wenjun Meng*[1], *Chaoqun Ma*[1]

[1]Aerosol Chemistry Department, Max Planck Institute for Chemistry, Mainz, 55128, Germany

[2]Laboratory for Climate and Ocean-Atmosphere Studies, Department of Atmospheric and Oceanic Sciences, School of Physics, Peking University, Beijing, 100871, China

[3]Sichuan Academy of Environmental Policy and Planning, Sichuan, 610041, China

***Correspondence**: Yafang Cheng (yafang.cheng@mpic.de)

Keywords: Large language models, Identity bias, Matthew effect, Scientific evaluation, Algorithmic fairness

Abstract:

As large language models (LLMs) are increasingly used to gather, process and disseminate scientific information, concerns have grown about their identity-related biases. However, national-background biases in scientific evaluation remain insufficiently understood. Existing studies are limited to a narrow range of countries, abbreviated materials and/or simplified evaluation criteria, and have reported inconsistent findings. Moreover, differences in the textual evaluations generated by LLMs have seldom been quantified. Here, we used four popular LLMs to evaluate 292 full-text manuscripts from 28 journals across 8 research fields under 12 assigned national backgrounds, yielding 14,016 manuscript assessments, each covering 28 evaluative dimensions. Further comment-based rescoring tests were conducted to quantify differences in their textual evaluations. The results showed systematic differences in both scores and comments despite unchanged scientific content, with the magnitude and statistical significance these differences varying across national backgrounds. Under different national backgrounds, LLMs could offer different views on the same issue, and give different scores alongside similar textual evaluation. Researchers and decision-makers should be aware of these biases associated with

source identity, including national background, when using LLMs for information gathering and processing.

Main

Large language models (LLMs) are increasingly adopted in academic research, supporting literature review[1], data analysis[2], manuscript writing[3] and scientific evaluation[4], despite different policies on their use. Beyond simply responding to researchers' instructions, LLMs can also influence scientific judgements through both direct recommendations, and the information they provide for human decision-making[5]. Thus, impartiality in how LLMs process and present information is important for informed and independent judgement[6,7]. However, evidence suggests that LLM judgements can be influenced by identity-related information, such as gender[8,9], affiliation[8,10] and national background[9]. Such biases have raised concerns about academic fairness and scientific communication. They may persist or even be amplified[5] if the affected materials are fed back into LLM training[11]. Identifying these biases and monitoring them over time are important for fair and reliable scientific progress[12,13].

National background is a key identity characteristic and can also be related to author names, ethnicity and institutional affiliations, as well as research priorities and orientations[14,15]. These relations make it important to examine potential national-background bias in LLMs before interpreting the effects of other identity characteristics. Yet existing studies have rarely examined national background separately from other identity characteristics, particularly when evaluating academic materials, making their respective effects difficult to disentangle. So far, the bias toward different national backgrounds remains insufficiently understood, especially in real-world tasks involving full-text materials and multiple assessment dimensions.

Constrained by LLM availability and capability, early studies often were limited to incomplete materials[9,16] (i.e. shortened length) or simplified evaluation criteria[9]. Oversimplifying either the scientific content or the evaluation criteria can reduce the practical relevance and obscure the potential bias[16,17]. Practical scientific evaluation involves judgments across aspects, including rigor, originality, credibility and significance. These judgments require close consideration of the evidence and reasoning chains detailed in manuscripts, particularly for work that challenges established understanding. Differences in how materials and evaluation criteria were simplified may have contributed to the inconsistent patterns of identity-related bias reported previously[8-10]. Limited country coverage also constrains the generalizability of these findings. Therefore, comprehensive manuscript assessment is still needed. Moreover, numerical scoring and categorical judgments are commonly adopted to quantify LLM behavior, but may not adequately represent how these models are practically used. It remains unclear to what extent differences in LLM scores are accompanied by different textual comments from the models.

In this study, we used full-text peer-reviewed research articles to examine the

potential national-background effects on LLM scientific evaluation. As detailed in *Methods*, we downloaded 292 research manuscripts from 28 journals covering 8 fields and different levels of scholarly impacts. Journals were selected considering their disciplinary coverage and impact levels. All eligible articles were downloaded within specified dates or issues to reduce subjective selection. Manuscripts were first anonymized, and then assigned predefined national-background conditions. Each LLM evaluated every manuscript under each assigned national background independently, with unchanged scientific content. Finally, the LLM-generated scores and comments were compared across national backgrounds and evaluative dimensions. Our study included 4 widely used LLMs and 12 countries selected to cover major contributors to scientific publishing and different geographical regions. Beyond score differences, we designed comment-based rescoring tests to quantify the differences conveyed by LLM-generated textual comments and their relationships with scores for the corresponding dimensions.

Results

Before evaluation, manuscripts were anonymized using GPT-6 Astra[18] with medium reasoning effort to remove explicit identifiers and neutralize contextual identity clues while preserving scientific content. They were then assigned fictitious, culturally characteristic author names and institutional names, paired with a real city and country. Names were checked to avoid matches with prominent-person and institutional. Author numbers and affiliation structures were kept consistent across conditions, with gender balanced within manuscripts and matched by author position. Mixed national backgrounds were not considered. We used ChatGPT[18], DeepSeek[19,20], Claude[21] and Gemini[22] to provide scores (range 0–100) and brief justifications for overall scientific quality (item 28) and 27 other evaluative dimensions[9,23,24] (i.e. item 01–06: scientific contribution; 07–12: rigor; 13–18: reasoning and evidential strength; 19–22: scientific integrity, errors and AI-assisted writing; 23–27: journal suitability). Evaluations were conducted in separate conversations with LLM personalization disabled and instructions prohibiting internet searches.

The plausibility of LLM evaluations was examined by comparing results based on manuscripts from journals at different impact levels. LLMs generally distinguished between these levels and provided justifications grounded in manuscript content. A further check using a sample of manuscripts with standardized structures, intended to reduce potential journal-specific features, yielded consistent results.

LLM scoring patterns across countries

Our results provide a quantitative basis for the effect of national backgrounds on LLM evaluation. Figure 1a shows the deviations in overall scientific quality (item 28) scores across national backgrounds for each LLM, with countries ranked according to ChatGPT results. Each value is the deviation from the average over all 12 national

backgrounds. Country-to-county pairwise score differences between countries can be obtained by subtracting their corresponding deviations (also shown in Fig. S1). Despite differences in magnitude and statistical significance, all four models assigned above-average scores to UK and US backgrounds and below-average scores to Chinese backgrounds. Meanwhile, some national backgrounds showed opposite score deviations across models. For instance, Japanese and German backgrounds received above-average scores from ChatGPT and Gemini, but below-average scores from Claude. Although the absolute score differences were small, country-specific deviations ranged from -6.9% to +7.4% of each model's overall score standard deviation. DeepSeek showed relatively small country deviations in both absolute and standardized terms. And none of these deviations based on DeepSeek were statistically significant.

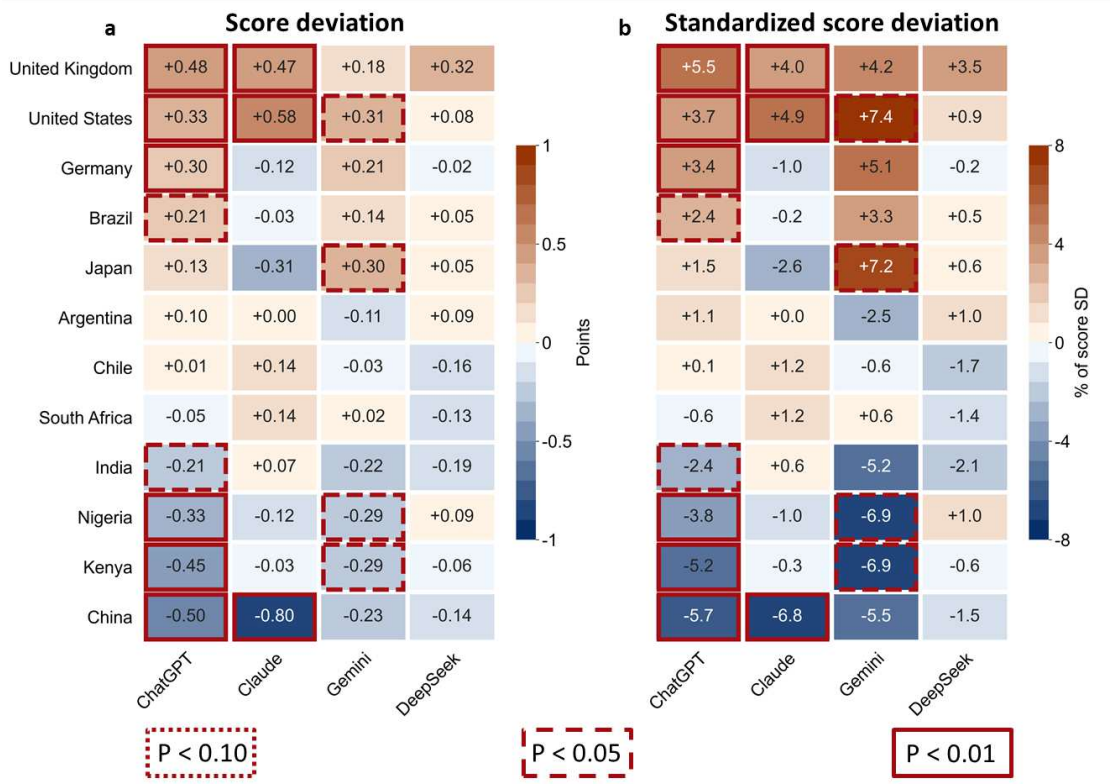


Fig. 1 | National-background differences in LLM evaluations of overall scientific quality. a, Mean score deviations from each manuscript's average across 12 national-background conditions, calculated within each model. b, The same deviations expressed as percentages of the sample standard deviation across all 3,504 scores from each model (292 manuscripts × 12 national-background conditions). Outlines indicate statistical significance from two-sided paired sign-flip tests[25] comparing each country's scores with each manuscript's average across all 12 national backgrounds. P values were Holm-adjusted across the 12 comparisons within each LLM: solid, $P^{adj} < 0.01$; dashed, $P^{adj} < 0.05$; and dotted, $P^{adj} < 0.10$ for reference. Statistical significance in the main text was defined as $P^{adj} < 0.05$.

Statistical tests[25] provided further evidence for the national-background effects for specific countries. Figure 1 indicates the statistical significance of each country's score deviation from the mean of all 12 backgrounds. Significant deviations were identified for eight countries in ChatGPT, three in Claude and four in Gemini, but none in DeepSeek ($P^{adj} < 0.05$).

Although some national backgrounds did not show significant deviation from the average, their differences from specific higher- or lower-scoring countries were significant in pairwise comparisons (Fig. S1). For example, Indian backgrounds showed no significant deviation from the average in the four models, yet scored significantly lower than US backgrounds in ChatGPT and Gemini, and higher than Chinese backgrounds in Claude. Significant pairwise differences spanned a broad range of national backgrounds in ChatGPT, but concentrating on Chinese or US backgrounds in Claude. In Gemini, they mainly occurred between Nigerian or Kenyan backgrounds and UK, US, German or Japanese backgrounds. No pairwise differences reached the significance threshold in DeepSeek (Fig. S1). Of the 66 country pairs within each model, 31 (47.0%), 15 (22.7%), 12 (18.2%) and 0 showed significant differences in ChatGPT, Claude, Gemini and DeepSeek, respectively ($P^{adj} < 0.05$).

We further compared national-background effects across the 27 evaluative dimensions. Here, we present the ChatGPT results (Fig. 2), which generally showed stronger statistical evidence of these effects. Despite differences in magnitude and statistical significance, the patterns across national backgrounds were broadly consistent with those for overall scientific quality (Fig. 1). Following the country order in Fig. 2, score deviations generally decreased from positive to negative for positive indicators, whereas adverse indicators (19–22) showed the reverse pattern. For positive indicators, significant deviations from the corresponding averages indicated higher scores for UK, US and German backgrounds and lower scores for Nigerian, Kenyan and Chinese backgrounds. The largest gaps among positive indicators occurred in assessments of suitability for journals at different impact levels (23–27), especially leading journals in the field. Relatively large differences in journal suitability were also observed in Claude and Gemini, despite some small changes in country rankings (Fig. S2–S4).



Fig. 2 | National-background differences in ChatGPT evaluations across 27 dimensions. The values in boxes show the mean score deviations from each manuscript's average across 12 national-background conditions, averaged over 292 manuscripts. Dimensions 19–22 are labelled in red to indicate adverse indicators. Red outlines indicate the statistical significance from two-sided paired sign-flip tests comparing each country with the mean of all 12 countries. Holm adjustment covers all 324 comparisons across 27 dimensions and 12 countries: dotted, $P^{adj} < 0.10$; dashed, $P^{adj} < 0.05$; solid, $P^{adj} < 0.01$. Countries are labelled using ISO codes and ordered as in Fig. 1 (Top-bottom to left-right). Statistical significance in the main text was defined as $P^{adj} < 0.05$.

For scientific contribution (dimensions 01–06), Chinese backgrounds consistently received significantly below-average scores. UK backgrounds consistently recieved above-average scores for rigor (07–12) and most items in reasoning and evidential strength (13–18). Differences in assessments of data fabrication or falsification (19) and research misconduct (20) were small and not statistically significant. These small differences might be limited by the floor effect, as almost all scores were ≤ 5 points

(98.4% and 99.6%, respectively). National-background effects can thus be hard to capture using these peer-reviewed publications, in which evidence of fabrication or misconduct is expected to be limited. Claude and Gemini generally showed similar directions of deviation for UK, US and Chinese backgrounds, whereas patterns for other countries varied more across models. DeepSeek also generally assigned above-average scores to UK backgrounds and below-average scores to Chinese backgrounds on positive indicators, although these deviations were not statistically significant (all $P^{adj} > 0.30$).

Further tests were conducted to check whether score differences between national backgrounds varied along with research fields and journal impact levels. Specifically, these tests assessed if the deviation associated with a specific country would change across these subgroups. For overall scientific quality, no significant differences in these effects were detected between any two subgroups, possibly because of the reduced sample sizes within individual subgroups.

Nevertheless, when considering the scoring range within each group, such national-background biases differed in their relative size. For example, among the 110 manuscripts originally published in higher-impact journals using ChatGPT and Claude, the US-China score gaps were 16.3% (0.82/5.01) and 25.2% (1.85/7.32) of the subgroup standard deviations (SDs, calculated across all 12 national backgrounds), respectively. In contrast, for the 182 manuscripts from the other journals, these ratios decreased to 9.1% (0.84/9.24) and 10.2% (1.10/10.86). The smaller score SDs for manuscripts from higher-impact journals made the national-background deviations relatively more pronounced in this group (Table S1).

Such influence of within-group score SDs on the relative magnitude of national-background deviations was more substantial across research fields (Table S2). For example, across all manuscripts, ChatGPT showed close absolute US-China score gaps in physics and chemistry, but these corresponded to 22.3% (0.89/4.00) and 8.2% (0.86/10.5) of the respective field-specific score SDs. These ratios remained markedly different (18.4% = 0.62/3.37 versus 8.4% = 0.91/10.75) when restricted to the other specialist journals. Although physics showed relatively larger national-background effects in ChatGPT, its ranking among the eight fields varied across the other three LLMs. Overall, we did not identify a consistent pattern across the four LLMs in how relative score differences varied among research fields.

National background versus scientific content in LLM evaluation

Beyond the pattern and statistical significance of national-background deviations, we further examined how these biases affected LLM evaluations of the scientific content in manuscripts. Here we focused on manuscripts originally from the US and China, the two largest source groups, comprising 68 and 97 manuscripts, respectively. Half of the US manuscripts were published in higher-impact journals, compared with 36.1% of the Chinese manuscripts. Besides, score differences between US and Chinese

background assignments were generally among the five largest country-pair differences (out of 66 pairs), except in DeepSeek. Here the original national backgrounds were determined by their first institutional affiliation listed.

Table 1 shows the overall scientific quality scores of manuscripts assigned either their original national background or that of the other country (US/China), together with the score changes after switching national background. Assignments matching the original national background still used fictitious author names and institutions. First, US background assignments consistently yielded higher scores, except in DeepSeek's evaluations of manuscripts originally from China. And such effects were more significant when scoring manuscripts originally from the US, possibly related to a larger proportion of these manuscripts came from higher-impact journals. Second, switching the assigned national backgrounds reduced the score gaps between the two source-country groups, making the gap statistically nonsignificant in ChatGPT, and also weakening the significant in Claude. After the switch, the probability that a US-origin manuscript scored higher than a China-origin manuscript decreased by 6.59 percentage points for ChatGPT, and 8.96 percentage points for Claude. These results suggest that national background can substantially influence LLM-based scientific evaluations, although such effects were not consistently observed across LLMs.

Original source	**Assigned background**	**ChatGPT**	**Claude**	**Gemini**	**DeepSeek**
China (n = 97)	China	79.94	62.66	84.47	75.39
China (n = 97)	US	80.41	63.80	84.68	75.25
US (n = 68)	China	82.50	67.26	85.10	77.69
US (n = 68)	US	83.35	69.32	85.43	78.10
Comparison		**ChatGPT**	**Claude**	**Gemini**	**DeepSeek**
China source: US minus China assignment		+0.47*	**+1.14*** **	+0.21	-0.14
US source: US minus China assignment		**+0.85*** **	**+2.06*** **	+0.32	+0.41
US minus China source: original backgrounds		**+3.41*** **	**+6.66*** **	+0.95	+2.71*
US minus China source: switched backgrounds		+2.09*	**+3.46** **	+0.42	+2.44

Table 1 | Overall scientific quality evaluations under original and switched national backgrounds. Manuscripts originally from China (n = 97) and the US (n = 68) were individually evaluated under both Chinese and US background assignments. Within each original-source group, score differences were calculated as US-assigned minus China-assigned scores. Between source-country groups, differences were calculated as US-origin minus China-origin scores, with assigned backgrounds either matching or opposite to their original backgrounds. Asterisks indicate Holm-adjusted (across four comparisons within each LLM) statistical significance: *, $P^{adj} < 0.10$; **, **$P^{adj} < 0.05$; ***, $P^{adj} < 0.01$.** Statistical significance in the main text was defined as $P^{adj} < 0.05$.

We further compared scores between the 97 manuscripts originally from China and the 68 from the US, averaging each manuscript's scores across all 12 national-background assignments. The results showed that the US-China score gaps between source-country groups fell between those under the original and switched background assignments in Table 1 for all four LLMs (Table S4). This finding also supports our suggestion that national-background biases can distort score differences between manuscripts from different countries, beyond differences associated with scientific content. Similar results were obtained when manuscripts were evaluated anonymously without assigned national backgrounds, except for ChatGPT (Table S4). Note that, in the latter tests, removing author information and making corresponding adjustments to the evaluation prompts might introduce additional influences.

National background effects on LLM textual comments

The reasoning and arguments provided by LLMs can play a more important role than the scores in real-world applications of scientific evaluation. However, differences in the textual evaluations generated by LLMs are challenging to quantify directly, as interpretations and comments on different studies are difficult to align along common quantitative dimensions. Therefore, we conducted rescoring tests in which each LLM read its own comments on the same manuscript and evaluative dimension under different national-background assignments together and assigned new scores based on these comments. The LLMs were told only that the comments were evaluations by different reviewers of the same manuscript on a specified dimension. They were not shown the manuscript or the original scores, nor informed of the source of the comments or their associations with national backgrounds. Each scoring task was repeated independently three times, and the resulting scores were averaged. This approach aimed to assess the extent to which national-background-related score differences reflected different interpretations of manuscript content, or different score assignments despite similar interpretations. Here, we selected eight representative evaluation dimensions for the rescoring tests using ChatGPT, Claude, and Gemini. The ChatGPT rescoring test covered only a subset of manuscripts because of limited model availability. DeepSeek was not included since its national-background effects were generally not statistically significant.

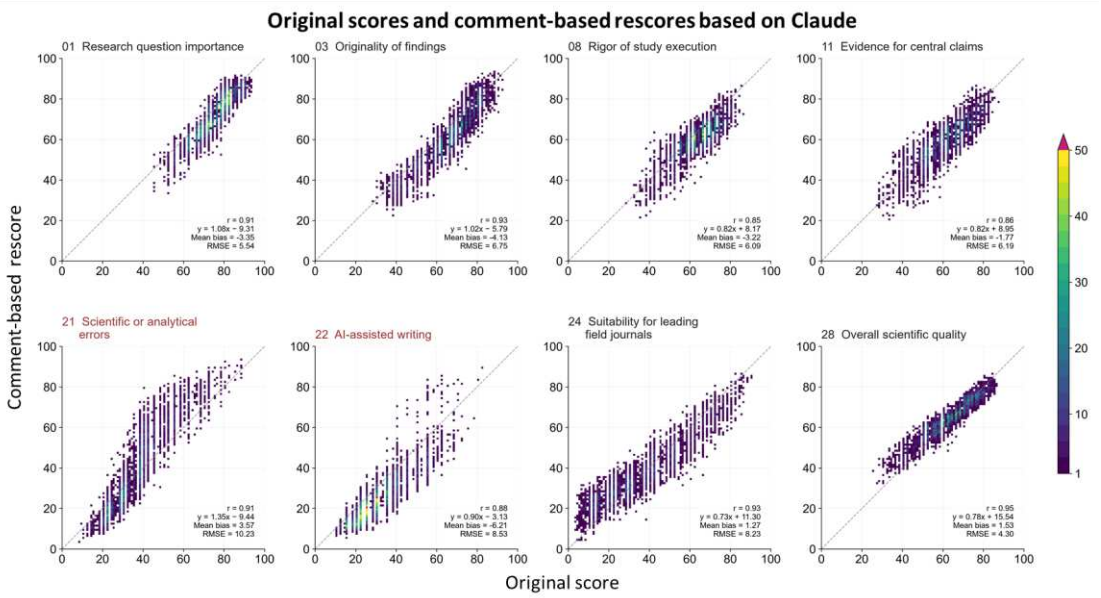


Figure 3 | Scatter plots of original versus comment-based scores assigned by Claude. Each subplot compares the original scores with the mean of three independent rescores for one evaluative dimension. Each sample represents a manuscript assigned a specific national background (292 manuscripts × 12 backgrounds). Colors indicate the sample counts. Dashed lines indicate a 1:1 relationship. Results based on ChatGPT and Gemini are shown in Fig. S5-S6.

As shown in Fig. 3, comment-based rescoring closely reproduced the original evaluations, even without access to manuscript content. Across all eight selected dimensions, rescores were highly correlated with (0.85–0.95) with the original, with relatively small mean differences. Each rescoring task was conducted independently in a separate temporary conversation without access to previous chat history, consistent with our brief preliminary checks. Explicit instructions also prohibited the use of chat history or saved memories. The close agreement may partly be related to the use of descriptions for each 10-point score range in both the original evaluations and rescoring, as some of the original comments may have used similar expressions from these descriptions. Vertical stripes are shown in Fig. 3, reflecting the tendency of original scores to cluster at certain values and the greater variability in rescores, likely due to less evidence and fewer constraints for scoring without the manuscripts.

As shown in Fig. 4a, c significant national-background effects in Claude were more concentrated in pairwise comparisons involving the US and China than in ChatGPT (Figs. 1 and S1). Similarly, for positive indicators, Claude generally assigned above-average scores to UK and US backgrounds and below-average scores to Chinese backgrounds. A notable difference was that German and Japanese backgrounds generally received above-average scores in ChatGPT assessments but below-average scores in Claude assessments.

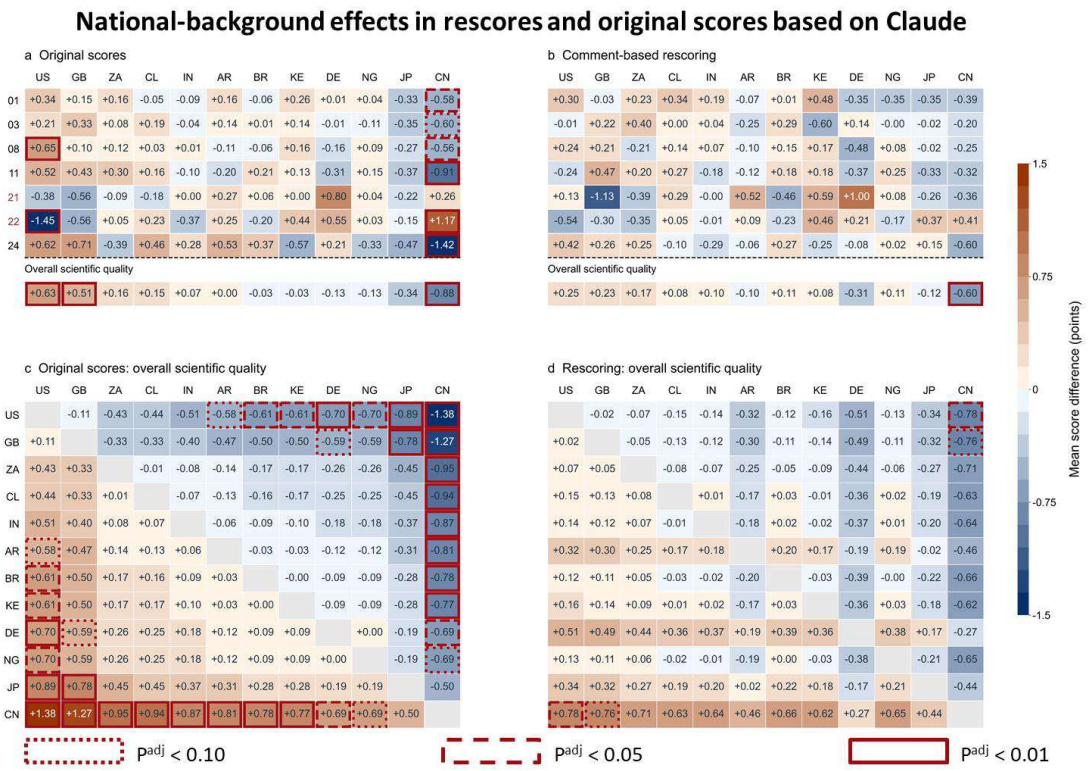


Figure 4 | National-background effects in comment-based rescores and original scores using Claude. (a, b) Mean score differences between each assigned national background and all 12 backgrounds within each manuscript, averaged across manuscripts, for the eight evaluative dimensions. (c, d) Pairwise national-background differences in overall scientific quality scores (column minus row). Panels a and c show the original scores; and panels b and d show the mean of three independent comment-based rescores. Countries are ordered by decreasing original mean overall scientific quality score. Panel a shows the same values and significance annotations as the corresponding results in Fig. 1 and Fig. S2. Panel c shows the same values and significance annotations as Fig. S1b with a different country order. Red outlines indicate statistical significance from two-sided paired sign-flip tests, with scores paired within each manuscript. Holm adjustment covered 324 comparisons (27 dimensions × 12 countries) for dimensions 01–24 shown in a, and 84 comparisons (7 dimensions × 12 countries) for those shown in b. For overall scientific quality, P values were adjusted separately across 12 countries in each figure. In c and d, Holm adjustment covered 66 unique country pairs separately. Dotted, dashed and solid outlines indicate $P^{adj} < 0.10$, $P^{adj} < 0.05$ and $P^{adj} < 0.01$, respectively. Statistical significance in the main text was defined as $P^{adj} < 0.05$. Results based on ChatGPT and Gemini are shown in Fig. S7-S8.

The rescores showed national-background effects similar to those in the original scores (Fig. 4), with similar country rankings for overall scientific quality but smaller score gaps and weaker statistical significance. For overall scientific quality, only the US-China score difference remained statistically significant. And The UK-China difference had a P^{adj} value of 0.057. The rescoring gaps were 56.4% and 59.9% of the

original gaps for the US-China and UK-China comparisons, respectively. Linear regression showed that the comment-based rescoring gaps explained 27.9% and 28.6% of the variance in the corresponding original score differences, respectively. These results suggest that differences in evaluative reasoning, and in the scores assigned to comparable reasoning both contributed substantially to the national-background effects observed in our study. However, across manuscripts and evaluative dimensions, their relative contributions could differ substantially.

For example, in a study on forest composition, bird ecology and vulnerability led by researchers from Australia (manuscript 139[26]; see Supplementary Manuscript Catalog), Claude assigned an overall scientific quality score 6 points higher under the US background than under the Chinese background in the initial single-round evaluation. Across 60 repeated evaluations for each background, the mean US-China gap was 1.58 points ($P^{adj} < 0.01$), close to the overall US-China gap shown in Fig. 4. Differences in overall scientific quality, item 11 (strength of evidence for the central claims) and item 24 (suitability for leading journals in the field) were no longer significant after comment-based rescoring (Table S5). This suggests that the original score differences were not fully captured in the accompanying textual evaluation, and likely reflected different scoring standards under different national backgrounds. Alternatively, the model's underlying judgements might not have been fully expressed in its responses.

For rigor of study execution (item 8, manuscript 139), the original and comment-based rescoring gaps were similar (1.85 and 1.78 points, respectively), and both were statistically significant ($P^{adj} < 0.05$). This suggests that the original score difference reflected the different comments to some extent. Specific differences in comment wording could be easily recognized. For example, among comments mentioning "error" or "noise" (11 under the US background and 10 under the Chinese background), US-background comments often described these issues as unquantified or unexamined, with "considerable" appearing only once. By contrast, six of the ten Chinese-background comments described the error or noise as "considerable" or "substantial". The US-China gap in comment-based rescoring for these comments reached 3.86, even exceeding the corresponding original score gap of 1.62. A similar pattern occurred for importance of the research question (item 1), where the comment-based rescoring gap was 1.98 ($P^{adj} = 0.048$), compared with a nonsignificant original score gap of 0.62 ($P^{adj} > 0.10$).

This example suggests that national-background differences can occur in both numerical scoring and evaluative comments, and that both may contribute to significant score gaps. These patterns may vary across manuscripts and evaluation dimensions. Besides, LLM-generated comments, including explicitly stated reasoning in text, do not necessarily provide a faithful account of the internal processes underlying the scores. Further development and validation of comment-based rescoring could improve its use in quantifying differences in LLM reasoning.

Residual national-background traces in manuscripts and their impacts

Although the manuscripts were anonymized before the evaluation, LLMs may still infer their original national backgrounds from characteristic language use, research topics, and tendencies in interpretation and judgement. For example, studies in social sciences, environmental sciences and psychology are more likely to have locally specific topics. Technical and even theoretical topics may also reflect differences in industrial structures and disciplinary development across countries[14,15]. These characteristics make complete anonymization extremely difficult. Residual clues to national background might therefore influence evaluations, particularly on the adverse indicators when they conflict with the assigned national background.

To assess the potential impacts, we compared scores under assignments that matched the manuscripts' original countries with the mean scores under the other 11 assignments. This comparison included 212 manuscripts whose original countries were among our 12 national-background conditions. Across the four LLMs and 28 evaluation dimensions including overall scientific quality, only ChatGPT showed statistically significant differences ($P^{adj} < 0.05$, Table S6) in research question importance (item 01) and significance of the knowledge gap (02). In both dimensions, scores were even higher under mismatched than matched assignments. These differences are likely related to the uneven distribution of original national backgrounds, as 85.8% and 57.5% of the manuscripts originated from countries whose assigned backgrounds received below-average scores on the respective items. Overall, these results suggested that a mismatch between the assigned and original national backgrounds did not lead to significantly worse scoring, particularly for these adverse indicators.

We also tested whether LLMs were able to infer the original national background from manuscript content. Each LLM examined each manuscript independently three times. Across the four models, the original country was correctly inferred at least once for 78.1–79.5% of manuscripts, and in all three attempts for 62.7–76.0% (Table S7). Although LLMs could still infer national origin from the text of a substantial proportion of manuscripts, their scientific evaluations remained sensitive to the assigned national backgrounds. The overall extent of nationality-related bias might be greater if influence arising from residual traces of national background in manuscript content was also accounted for.

Figure 5 compares pairwise national-background differences in ChatGPT's overall scientific quality scores between manuscripts whose original country was correctly inferred at least once (n = 228) and the others (n = 64). The latter group generally showed larger score differences across assigned national backgrounds, although statistical evidence was weaker, possibly because of the much smaller sample size. Similar patterns were observed for Claude and Gemini, whereas DeepSeek showed more variable country rankings and no statistically significant pairwise score differences (Fig. S9). Overall, these results suggest that residual national-origin information had a limited influence on the evaluations, and that more complete removal of this information might further amplify the national background effects.

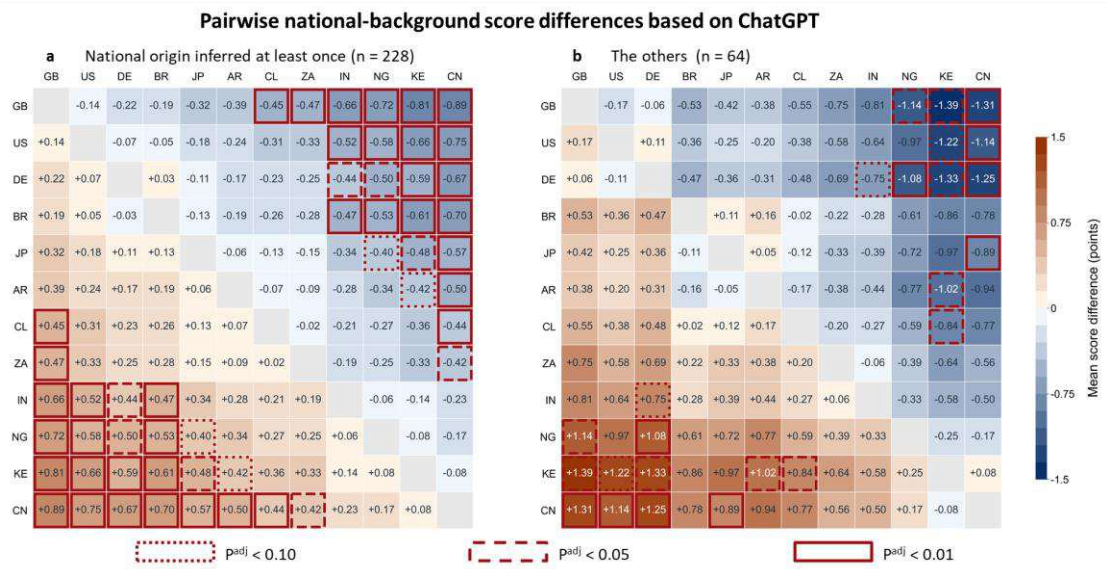


Figure 5 | Pairwise national-background score differences in ChatGPT evaluations. Overall scientific quality scores were compared between national-background assignments for manuscripts whose original country ChatGPT correctly inferred at least once across three independent attempts (a, n = 228) and those for which all three attempts were incorrect (b, n = 64). Red outlines indicate statistical significance from two-sided paired sign-flip tests comparing each pair of national-background conditions, with scores paired within each manuscript: dotted, $P^{adj} < 0.10$; dashed, $P^{adj} < 0.05$; solid, $P^{adj} < 0.01$. Holm adjustment covered all 66 pairwise country comparisons within each panel.

Discussion

Our study used full-text, peer-reviewed manuscripts to assess national-background biases in LLM scientific evaluations. Changing assigned national backgrounds while holding scientific content constant could produce non-negligible differences in evaluation. We further introduced comment-based rescoring tests to quantify differences conveyed by the evaluative comments and examine their relationship with the original scores. Within individual models, score patterns were similar across evaluative dimensions, research fields and journal impact levels, whereas differences were more pronounced across LLMs. It may reflect differences in training-data exposure and the information sources of the models. Nevertheless, all four models assigned higher scores to US and UK backgrounds than to Chinese backgrounds. The US-China differences were statistically significant in all models except DeepSeek, while the UK-China differences were significant in ChatGPT and Claude. Comment-based rescoring further suggested that national-background effects were shown in both evaluative reasoning and resulting scores, with their relative contributions varying. These consistent preferences for particular national backgrounds warrant attention in the use and development of LLMs.

Although the manuscripts were anonymized before evaluation, residual identity

information still allowed LLMs to infer the original national background of many manuscripts. Such information may be intertwined with the scientific content and therefore difficult to remove. Similar challenges may arise for other identity characteristics. These clues could influence scoring in two ways. First, a mismatch between implicit identity clues and assigned author information might raise concerns about the consistency of the author information. Second, implicit identity clues might themselves trigger identity-related scoring biases. Our results provide tentative support for the latter possibility.

Additional analyses (Table S8) showed that US-origin manuscripts generally received higher mean scores than China-origin manuscripts after averaging each manuscript's scores across the 12 assigned national backgrounds. Among manuscripts whose original country was incorrectly inferred in all three trials, these gaps narrowed and, for some models, reversed when the analysis was restricted to specialist journals (except for Gemini, which showed much smaller score gaps). However, this comparison was limited by the small number of manuscripts in the latter group and could not control for differences in scientific content between the groups. Moreover, the scores and evaluative comments did not directly establish whether implicit identity clues influenced judgements of scientific content. Further targeted experiments are needed to clarify their influence.

Score deviations that did not reach statistical significance were also presented, because they might still be informative. Holm adjustment for multiple comparisons imposed a stringent threshold for statistical significance. In the meantime, the consistent directions of deviation across evaluative dimensions provided additional support for interpreting these differences. Larger samples could improve the precision of these estimates and help determine whether statistically significant biases extend to other national backgrounds. Moreover, it is worthwhile to maintain long-term monitoring of identity-related biases in LLMs. Here, our analyses only examined fixed LLM versions. Nevertheless, as AI increasingly shapes scientific communications and the wider information environment, changes in such biases and their impacts are important for the development of the scientific community.

In 1968, Merton noted that eminent scientists often receive disproportionately greater recognition than less-established researchers for comparable scientific contributions, a phenomenon widely known as the Matthew effect in science[13]. AI-generated content and information affected are growing explosively[27]. These materials may subsequently enter training datasets or become sources retrieved by LLM tools. Identity-related biases in AI may influence the selection and interpretation of information, as well as the content subsequently generated. Such iteration could cause a Matthew effect through positive feedbacks, allowing initial biases to persist or intensify. Similar bias-amplifying feedbacks have been observed in human-AI interactions[5]. Currently, LLMs are already important tools for accessing, processing and producing scientific materials. Recent work also demonstrates their potential to produce high-quality manuscript reviews[4]. As scientific work increasingly relies on LLMs, it must be recognized that they may reproduce prejudices embedded in existing

information and perpetuate them in science development.

Methods

Manuscript preparation and anonymization

We gathered 293 manuscripts across 8 research fields and 28 journals with different impact levels. One article was subsequently excluded because the complete manuscript was unavailable to us at the collection time, leaving 292 manuscripts for analysis. The journals included Science, discipline-specific Nature journals, Nature Communications, Science Advances, and specialist journals representing each field (see Supplementary Manuscript Catalog). To minimize subjective selection of manuscripts, we sought to collect all eligible research articles published online by the selected journals on predefined dates. For 22 of these journals, we got 192 articles published online on September 25, 2026. For two of these journals, access errors interrupted retrieval and eight articles were gotten from these journals. To increase representation across selected research fields and journal impact levels, we also collected 58 articles from the latest available issues of three other journals, and 43 additional articles published online between September 21 and 28, 2026, from five journals. Differences in publication volume, frequency, and online release practices across journals made it difficult to obtain comparable numbers of manuscripts for each research field.

Manuscripts were manually assigned primary research fields based on the journal's disciplinary scope and the article title. For ambiguous or interdisciplinary cases, we additionally considered the original authors' institutional affiliations and the article's main contribution. Particular care was taken with computer science, where computational innovations often serve applications in other fields. Studies whose main contribution was the development of algorithms, software or other computational methods were classified as computer science, with the authors' affiliations in computing-related departments providing supporting evidence in most cases. This approach broadened the application domains represented within computer science while limiting the inclusion of computational method development in the other fields.

ChatGPT-6 Astra was used to assist with further anonymization. The procedure was initially developed and refined using manuscripts in environmental science, a field familiar to the researchers, before being applied to the remaining manuscripts. First, author names, affiliations, journal identifiers, publication details, acknowledgments, funding statements, in-text citations, and reference lists were removed. Citations were removed because citation patterns can reveal authorship, affiliations, and national backgrounds even without explicit self-citations[28]. Additional identifying information in the text and figures was removed or replaced with generic descriptions where feasible, while preserving scientific content. Study locations and other contextual details related to research scope were retained.

Anonymized manuscripts were prepared as Word documents with a standardized title paragraph, enabling synthetic author names and numbered affiliations to be

inserted programmatically after the title.

National-background assignments and LLM evaluations

ChatGPT-6 Astra[18] was used to generate synthetic author names and institutional affiliations for 12 countries. Names were reviewed for linguistic plausibility and screened for potential matches with prominent researchers, with identified matches replaced. The name pools contained equal numbers of masculine- and feminine-coded names within each country. For the United States, the name pool included multiple naming traditions to reflect the diversity of US-affiliated researchers. Sampling weights were informed by published surname-based estimates for US-affiliated first authors. The institution pool contained 1,440 fictional affiliations, comprising 15 entries per country in each of eight fields. Institutional names were reviewed for linguistic plausibility and appropriate organizational structure and screened for potential overlap with existing research institutions. Each fictional affiliation was paired with a real city and its corresponding country. The complete name and institution pools are provided in the Supplementary Data.

For each manuscript, national backgrounds were assigned using synthetic author names and institutional affiliations selected from the corresponding country-specific pools. Each assignment included 5–15 authors, with the numbers of masculine- and feminine-coded names differing by at most one. The number of institutions was randomly chosen from one and the total number of authors. The author information structure was kept for each manuscript across national backgrounds, and in additional tests where applicable. Each LLM evaluation was conducted in an independent conversation via the web interface. For all four LLMs adopted, the use of submitted content for model training was disabled. ChatGPT evaluations used GPT-6 with medium reasoning effort in temporary, unpersonalized chats[18], and Claude evaluations used Sonnet 5.5 with medium reasoning effort in incognito chats[21]. Gemini evaluations used Gemini 3 in Auto mode with medium thinking and Gemini Apps Activity disabled[22]. DeepSeek evaluations used V4.1-Flash in fresh conversations, with tests confirming no carryover information from other conversations[19]. The evaluation instructions (Supplementary Data) required models to rely only on the current evaluation materials and explicitly prohibited web searches and external retrieval. DeepSeek's web-search function was also switched off.

The primary version of the LLMs remained unchanged during our tests, as described above. Yet the web platforms underwent interface and feature updates. In particular, DeepSeek's web model was updated from V4.1-Flash to V4.1-Flash-1017 on October 17, 2026[20], after our primary evaluations on assigned national backgrounds had been completed. This update might affect the subsequent country-inference tests. The collection periods for the main and subsequent tests are summarized in Table S9.

To assess whether national-background effects persisted under a more sophisticated reviewing protocol, we conducted an additional experiment using

instructions adapted from the Nature Review Studio skill[29]. Since the skill could not be used directly in temporary conversations, its relevant content was extracted into a text prompt (Supplementary Data), and supplemented with a scoring task using the same evaluation dimensions as the main evaluations. We randomly selected 24 manuscripts and evaluated each under both Chinese and US backgrounds as in the main evaluations. Then, the same version of ChatGPT was adopted for evaluations. A statistically significant difference between the two national-background conditions remained (Fig. S10), suggesting that changing the prompt or adopting a more sophisticated reviewing protocol may be insufficient to eliminate these effects.

LLM evaluations on anonymous manuscripts

For a preliminary assessment of the validity of LLM evaluations, we evaluated anonymized manuscripts without assigning author names or national backgrounds and compared their scores across source-journal impact levels. The evaluation instructions were adjusted accordingly, while retaining the same 28 scoring dimensions and scoring rules as in the main evaluations. Source journals were divided into four impact levels: Science (16 manuscripts), Nature specialty journals (22 manuscripts), Nature Communications and Science Advances (72 manuscripts), and other specialist journals (182 manuscripts). Each manuscript was evaluated independently three times by each LLM, and the scores were averaged for analysis. This test was conducted using ChatGPT, Claude, and Gemini. For comparisons in the main text, journals were divided into two groups, high-impact journals and other specialist journals. The high-impact group comprised Science, Nature specialty journals, Nature Communications, and Science Advances.

LLM scores showed clear differences across these impact levels (Table S10). For example, ChatGPT's mean overall scientific quality scores were 87.8, 83.3, 81.2, and 72.8 across the four journal groups, respectively. The other LLMs showed similar patterns across the four journal groups. Different journal groups accounted for 30.0%, 37.9%, and 28.5% of the variance in manuscript-level mean overall scientific quality scores for ChatGPT, Claude, and Gemini, respectively. The explained variance was calculated following Eq. (1), where y_i is the mean score across three anonymous evaluation attempts for each manuscript, $\overline{y_g}$ is the mean score for each journal group, and $\bar{y}$ is the mean score across all manuscripts. Using impact levels as a proxy, this correspondence provides supporting evidence that the evaluations captured meaningful differences among manuscripts.

$$R^2 = \frac{\sum_{g=1}^{4} n_g\left(\overline{y_g} - \bar{y}\right)^2}{\sum_{i=1}^{292}\left(y_i - \bar{y}\right)^2}, \text{ Eq. (1)}$$

To examine whether differences in manuscript structure and presentation affected LLM evaluations, we conducted an additional test using 10 manuscripts. Manuscripts were selected at random, excluding biomedical studies because the LLM restricted editing of materials on some biomedical topics. GPT-6 Astra[18] paraphrased the titles

and body text and reorganized each manuscript into Introduction, Methods, Results, Conclusion, and Discussion sections. The original and revised manuscript titles are provided in the Supplementary Data. Methods section was consolidated immediately after the Introduction, drawing on the original full text online and supplementary materials where needed for specific journals. All incorporated material was anonymized. Each standardized manuscript was then evaluated independently three times using Claude Sonnet 5.5 medium, with the same instructions and scoring rules as in the earlier anonymous-manuscript evaluations.

Overall scientific quality scores for the standardized manuscripts ranged from 51.3 to 82.7, and generally reflected differences in source-journal impact levels. These scores were positively correlated with journal impact factor ($r = 0.76$). Here the 2025 Impact Factors were obtained from the respective journal webpages. All positive evaluative dimensions except the treatment of scope and limitations showed positive correlations, whereas all four adverse dimensions showed negative correlations (Fig. S11). The strongest correlations were observed for the importance of the research question ($r = 0.83$), significance of the knowledge gap ($r = 0.84$), and potential scholarly impact ($r = 0.81$). These patterns suggest that the correspondence between LLM scores and journal impact levels persisted after standardizing manuscript structure and presentation.

LLM inference of original national backgrounds of the manuscripts

We conducted country-inference tests to examine whether LLMs could infer the original authors' countries of affiliation from residual identity-related information in anonymized manuscripts. The same anonymized manuscripts used in the anonymous-scoring experiment were submitted to the LLMs with instructions to infer these countries. Each response was required to contain only one country, without a predefined list of countries. Each manuscript underwent three independent trials using identical instructions. The full prompt is provided in the Supplementary Materials. It should be noted that the ability to infer these countries does not mean that LLMs used such information during our manuscript evaluation. Nevertheless, preliminary comparisons suggested that this possibility warranted further investigation.

The results of all three country-inference trials are provided in the Supplementary Data. ChatGPT, Claude, Gemini, and DeepSeek correctly inferred the country in at least one of the three trials for 78.1%, 79.5%, 78.4%, and 79.1% of manuscripts, respectively (Table S7).

Comment-based rescoring

In order to quantify the differences in evaluative comments, we conducted comment-based rescoring tests. Previous work has shown that LLMs can infer numerical scores from human-written peer-review comments[30]. Inputs were taken

exclusively from each LLM's own evaluations, with one comment per manuscript, evaluation dimension, and assigned national background. We selected eight dimensions: importance of the research question (item 01), originality of the findings (03), rigor of study execution (08), strength of evidence for the central claims (11), evidence of scientific or analytical errors (21), likelihood of AI-assisted writing (22), suitability for leading journals in the field (24), and overall scientific quality.

For each manuscript-dimension combination, the 12 comments were presented together with their wording unchanged. Presentation order and anonymous comment identifiers were independently randomized for each of the three rescoring trials. Each task included the original dimension-specific scoring rules, and score-calibration guidance. Models assigned an integer score from 0 to 100 to each comment and provided a brief justification. Instructions explicitly permitted equal scores for comments expressing equivalent judgments and prohibited seeking manuscript information beyond that conveyed by the comments. Each trial used an independent conversation, and scores were averaged by comment before comparison with the original scores. The complete prompts and model-specific coverage are provided in the Supplementary Information.

Gemini showed positive correlations between original scores and comment-based rescores across all eight selected dimensions, although these correlations were consistently weaker than those observed for Claude ($r = 0.38–0.88$ versus 0.86–0.95, Fig. S5). ChatGPT results were limited to 48–49 manuscripts per dimension because of restricted model availability, but also showed strong overall correlations between original scores and rescores ($r = 0.75–0.96$, Fig. S6).

Statistical significance

Permutation-based significance testing and Holm-Bonferroni adjustment

Statistical significance was assessed using nonparametric permutation tests[30]. For comparisons of scores assigned to the same manuscript under different national backgrounds, we used paired sign-flip permutation tests. For comparisons between independent groups (e.g. manuscripts originally from China versus those from the US), we used two-sample permutation tests in which group labels were randomly reassigned while group sizes were preserved. In both cases, the difference in mean scores was used as the test statistic. All tests were two-sided and performed using 999,999 random permutations. Permutation-based P values were calculated using the finite-sample correction.

To reduce the risk of false-positive findings arising from numerous statistical comparisons across countries, LLMs, and research fields, we applied the Holm-Bonferroni adjustment (referred to as Holm adjustment) to control the family-wise error rate[25]. Adjustments were performed within the corresponding groups of comparisons, as specified in the main text. The adjusted P values were denoted as P^{adj}. Statistical significance in the main text was defined as $P^{adj} < 0.05$; the additional $P^{adj} < 0.10$

category was included only in figures for reference.

Standard deviation and explained variance

Standard deviation (SD) was used to quantify the variability in manuscript scores within each research field and journal impact group. It was calculated as the square root of the sample variance with a sample size correction (n - 1).

Explained variance was used in the comment-based rescoring analysis to quantify the proportion of variation in rescored manuscript scores accounted for by the original scores. It was calculated using the coefficient of determination (R^2), based on ordinary least-squares linear regression.

Acknowledgements

None.

Data Availability

All LLM evaluation records, evaluation instructions, scoring criteria, and bibliographic details of the original manuscripts are publicly available in the Supplementary Materials. The anonymized manuscripts are available from the authors upon request.

Code Availability

The code used for national-background assignment and statistical analysis is available.

Author Information

Contributions

HK and YC conceptualized the study and designed the experiments. HK conducted the experiments. All authors contributed to the analysis and interpretation of the results and reviewed and edited the manuscript.

Corresponding Author

*E-mail: yafang.cheng@mpic.de

ORCID

YC: 0000-0003-4912-9879

Notes

The authors declare that they have no conflict of interest.

References

1 Asai, A. *et al.* Synthesizing scientific literature with retrieval-augmented language models. *Nature* **650**, 857-863 (2026). https://doi.org:10.1038/s41586-025-10072-4

2 Cheng, L., Li, X. & Bing, L. 9496-9514 (Association for Computational Linguistics).

3 Sourati, Z. *et al.* The shrinking landscape of linguistic diversity in the age of large language models. *Nature Human Behaviour* **10**, 1925-1942 (2026). https://doi.org:10.1038/s41562-026-02550-0

4 King, G. & Worthington, S. (2026).

5 Glickman, M. & Sharot, T. How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour* **9**, 345-359 (2025). https://doi.org:10.1038/s41562-024-02077-2

6 UNESCO. Recommendation on the Ethics of Artificial Intelligence. (2021).

7 Nations, U. United Nations Global Principles For Information Integrity. (2024).

8 Vasu, S. S. M., Sheth, I., Wang, H.-P., Binkyte, R. & Fritz, M. 307-330 (Association for Computational Linguistics).

9 Sebo, P. & Wang, T. Bias and Reliability of AI-Based Peer Review: A Comparative Study of ChatGPT and Claude Evaluating Scientific Abstracts. *Journal of General Internal Medicine* (2026). https://doi.org:10.1007/s11606-026-10806-8

10 Pataranutaporn, P., Powdthavee, N., Achiwaranguprok, C. & Maes, P. Large language models and human judgement in scientific peer review: evidence from over 1000 published economics papers. *Economica*, n/a-n/a (2026). https://doi.org:10.1111/ecca.70076

11 Anthropic. *Claude's new constitution*, <https://www.anthropic.com/news/claude-new-constitution> (2026).

12 Graves, J. L., Kearney, M., Barabino, G. & Malcom, S. Inequality in science and the case for a new agenda. *Proceedings of the National Academy of Sciences* **119**, e2117831119 (2022). https://doi.org:10.1073/pnas.2117831119

13 Gomez, C. J., Herman, A. C. & Parigi, P. Leading countries in global science increasingly receive more citations than other countries doing similar research. *Nature Human Behaviour* **6**, 919-929 (2022). https://doi.org:10.1038/s41562-022-01351-5

14 Debernardi, C., Seeber, M. & Cattaneo, M. Thirty years of climate change research: A fine-grained analysis of geographical specialization. *Environmental Science & Policy* **152**, 103663 (2024). https://doi.org:https://doi.org/10.1016/j.envsci.2023.103663

15 Tautorat, P. *et al.* A global analysis of the rise, reign, and retreat of topics in research toward sustainable platform chemicals. *Green Chemistry* **27**, 12985-13003 (2025). https://doi.org:10.1039/d5gc02863a

16 von Wedel, D. *et al.* Affiliation Bias in Peer Review of Abstracts by a Large Language Model. *JAMA* **331**, 252-253 (2024). https://doi.org:10.1001/jama.2023.24641

17 Wu, M. & Aji, A. F. 297-312 (Association for Computational Linguistics).

18 OpenAI. *GPT-6 Sol and GPT-6 Luna: October 2026 update*, <https://deploymentsafety.openai.com/gpt-6-october> (2026).

19 DeepSeek. *DeepSeek-V4.1-Flash: Smarter, Faster, More Efficient*, <https://api-docs.deepseek.com/news/news260910/> (2026).

20 DeepSeek. *DeepSeek-V4.1-Flash-1017: Wish-Driven Programming Made Real*,

<https://api-docs.deepseek.com/news/news261017> (2026).

21 Anthropic. *Introducing Claude Sonnet 5.5, the second model in the Claude 5.5 family. It's a clear upgrade over Claude Sonnet 5, runs 30%+ faster, and costs up to 30% less for most work.*, <https://www.anthropic.com/claude-sonnet-5-5> (2026).

22 Google. *A new era of intelligence with Gemini 3*, <https://blog.google/products-and-platforms/products/gemini/gemini-3> (2026).

23 Dhillon, P. How to be a good peer reviewer of scientific manuscripts. *The FEBS Journal* **288**, 2750-2756 (2021). https://doi.org:https://doi.org/10.1111/febs.15705

24 Rospocher, M. Author metadata affects Large Language Model scores in scientific peer review. *Discover Artificial Intelligence* **6**, 1404 (2026). https://doi.org:10.1007/s44163-026-02268-y

25 Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* **6**, 65-70 (1979). https://doi.org:10.2307/4615733

26 Olah, G., Evans, M. J. & Asner, G. P. Canopy functional composition as an indicator of bird traits, taxonomy, and vulnerability in tropical Peruvian forests. *Nature Communications* **17**, 9554 (2026). https://doi.org:10.1038/s41467-026-77135-6

27 Liang, W. *et al.* Quantifying large language model usage in scientific papers. *Nature Human Behaviour* **9**, 2599-2609 (2025). https://doi.org:10.1038/s41562-025-02273-8

28 Matsubara, Y. & Singh, S. 9-20 (Association for Computational Linguistics).

29 mumdark. (GitHub, 2026).

30 Phipson, B. & Smyth, G. K. Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn. *Statistical Applications in Genetics and Molecular Biology* **9** (2010). https://doi.org:doi:10.2202/1544-6115.1585