Received: 13 December 2025 | Revised: 19 March 2026 | Accepted: 23 March 2026

DOI: 10.1002/asi.70079

RESEARCH ARTICLE

JASIST WILEY

# Does language bias GenAI academic evaluation in humanities and social sciences? A mixed-methods study based on Chinese-language HSS papers

**Yu Zhu** | **Yujie Jia** | **Yumeng Zhu** | **Jiyuan Ye**

School of Information Management, Nanjing University, Nanjing, China

**Correspondence**
Yu Zhu, School of Information Management at Nanjing University, No. 163, Xianlin Avenue, Nanjing, China.
Email: zhu.yu@smail.nju.edu.cn

**Funding information**
Graduate Research and Innovation Projects of Jiangsu Province, Grant/Award Number: KYCX25_0130; National Social Science Fund of China, Grant/Award Number: 24&ZD323

**Abstract**

As generative AI (GenAI) systems are increasingly deployed in cross-language research evaluation, whether GenAI evaluates multilingual scholarship without language-induced bias remains unclear. This study examines language bias patterns in GenAI evaluation of humanities and social sciences (HSS) research across models and disciplines. Using a within-subjects design, 1150 expert-selected papers from 23 disciplines were evaluated by GPT-4o and DeepSeek-V3 in Chinese and English. Results reveal opposite language biases depending on model type: GPT-4o favors English (Cohen's $d = 1.10$), while DeepSeek-V3 favors Chinese (Cohen's $d = -0.87$), persisting across all disciplines. Thematic analysis reveals a systematic decoupling between scores and evaluative reasoning: both models generate more critical comments for English papers, yet arrive at opposite scores through different rhetorical strategies—GPT-4o tends to moderate its positive assessments of Chinese papers while DeepSeek-V3 amplifies them. This decoupling suggests that bias is embedded in the multi-layered pathways through which models generate and aggregate evaluations. This study provides controlled evidence that language bias in GenAI evaluation is bidirectional and model-dependent, with scores not directly reflecting evaluative justifications. The findings have implications for designing fairer multilingual academic evaluation systems and for understanding the limitations of GenAI as scholarly evaluation infrastructure.

## 1 | INTRODUCTION

Generative AI (GenAI) is rapidly entering academic evaluation contexts, from manuscript screening and peer review to quality assessment and funding decision support (Thelwall, 2025a, 2025e; Zhu, Lu, et al., 2026). As these systems become increasingly embedded in academic decision-making infrastructure, their fairness, transparency, and reliability have emerged as unavoidable concerns (Hicks et al., 2015). Research has revealed that GenAI exhibits systematic biases across gender (Hall & Ellis, 2023), race (Noseworthy et al., 2020), institutional prestige, career stage, and discipline (Thelwall & Kousha, 2023; Zhu, Haunschild, et al., 2026). These biases not only affect individual researchers but also structurally impact knowledge production systems. Yet a crucial and long-overlooked attribute is writing language. As cross-language submissions, international collaborations, and multilingual review processes grow more common, departments increasingly rely on GenAI to handle

cross-language evaluation tasks, making language fairness concerns all the more pressing.

Language serves not merely as a medium for expressing academic content but as a carrier of cultural frameworks, epistemological traditions, and research paradigms. The long-standing dominance of English in the global academic publishing ecosystem creates structural asymmetries. Non-English academic outputs consistently face disadvantages in acceptance rates, citation performance, and international visibility (Hamel, 2007), disparities that cannot be fully attributed to research quality alone (Ammon, 2001). For scholars from non-English-speaking regions, language functions not only as a communication barrier but as a structural factor affecting their academic status and discursive power (Z. Xu, 2025), particularly in the humanities and social sciences (HSS) (Fourcade, 2009), and to some extent across all fields (Salö, 2015, 2018).

In the GenAI era, this issue may be further amplified (V. Brown et al., 2025). Current mainstream large language models are primarily trained on English corpora; their syntactic preferences, semantic embeddings, disciplinary knowledge distributions, and evaluative frameworks are all shaped by English-dominated academic culture (Y. Xu et al., 2025). Consequently, GenAI may mistake English-contextualized academic norms for universal quality standards, thereby reshaping and consolidating language inequality at the algorithmic level, especially in the HSS, which are highly sensitive to language (Ben-David, 1971) and cultural context (Moed et al., 2004). For HSS research, concepts are deeply embedded in linguistic and cultural contexts. Cross-language expression often entails semantic loss, framework misalignment, or conceptual dilution (Buden et al., 2009), making such research more vulnerable to GenAI language bias.

Despite the critical importance of language, systematic empirical research on language bias in GenAI academic evaluation remains scarce (Thelwall & Kousha, 2023). Existing studies on GenAI cross-language performance focus primarily on general tasks (Ahuja et al., 2023) and cannot answer the core question facing academic evaluation: under conditions of equivalent research quality, does GenAI provide different evaluations solely due to language differences? This study aims to fill this gap. Using expert-selected high-quality HSS samples, we employ a within-subjects design with different language presentations of identical research content, providing the most rigorous test of language effects to date. Through large-scale experiments and mixed-methods analysis, this study provides the first large-scale evidence of cross-language evaluation bias in the HSS under controlled quality conditions, reveals cross-disciplinary heterogeneity in language bias, and explains its semantic origins through mechanism analysis.

## 2 | THEORETICAL BACKGROUND AND HYPOTHESES

### 2.1 | Linguistic bias and algorithmic fairness in academic evaluation

The core of algorithmic fairness lies in ensuring that system decisions are independent of task-irrelevant protected attributes such as gender, race, and nationality (Mitchell et al., 2021). In academic evaluation, this requirement closely aligns with the universalism principle in Mertonian scientific norms (Merton, 1942): scientific evaluation should be based on impersonal criteria rather than the attributional characteristics of individuals or groups (Langfeldt et al., 2020). The *San Francisco Declaration on Research Assessment* (DORA, 2012) and the *Leiden Manifesto* (Hicks et al., 2015) reaffirm this principle, emphasizing that academic evaluation should transcend disciplinary, national, and linguistic boundaries, applying unified standards of academic quality.

Empirical research, however, demonstrates that language bias is systematically prevalent in human peer review. Manuscripts written in non-native or non-standard English receive systematically lower review scores even when content quality is equivalent (Politzer-Ahles et al., 2020). In academic publishing, acceptance rates for native English-speaking authors are significantly higher than for non-native authors (Yen & Hung, 2019), with some reviewers explicitly acknowledging bias against manuscripts that fail to meet native-like English standards (Strauss, 2019). This evidence reveals a core problem that academic content of equivalent quality receives differential evaluation due to language differences, directly violating fairness principles in academic assessment.

With the widespread application of GenAI in academic evaluation, language bias may persist or even amplify in new forms. Research has demonstrated GenAI's considerable potential across multiple evaluation tasks, including assisting peer review (Checco et al., 2021; Heaven, 2018), assessing academic quality (Thelwall, 2024; Thelwall & Cox, 2025; Zhu, Lu, et al., 2026), originality (Huang et al., 2025), and generating alternative citation metrics (Thelwall, 2025d, 2025e). However, researchers also stress the need for responsible use of GenAI in evaluation (Zhu, Lu, et al., 2026), warning against potential biases in its assessments (Thelwall & Kurt, 2025).

Yet current research on the algorithmic fairness of GenAI in academic evaluation pays insufficient attention to language bias as a core dimension. To our knowledge, while existing studies have revealed evaluation biases potentially triggered by attributes such as gender, nationality (Thelwall & Kurt, 2025), and field (Zhu, Lu, et al., 2026), few systematically examine whether language, a task-irrelevant attribute, triggers evaluation unfairness, and fewer still provide in-depth explanations of its underlying mechanisms. This gap is particularly critical in the HSS, and constitutes the core motivation for this study.

## 2.2 | Mechanisms of linguistic bias in GenAI

The formation of language bias in GenAI can be understood at two levels: imbalance in training data and the consolidation of spurious correlations during statistical learning processes (Y. Xu et al., 2025).

Training corpora of mainstream GenAI models are highly imbalanced, with English possessing overwhelming advantages in scale, quality, domain coverage, and accessibility (Dong et al., 2025; Shen et al., 2024). The English-dominated international publishing system, databases, and open-access resources constitute the primary semantic foundation of these models (Meneghini & Packer, 2007), while non-English corpora remain limited in scale and diversity (Liu, 2017).

The consolidation of statistical associations further transforms data imbalance into systematic bias. During pre-training, GenAI models learn co-occurrence structures through maximum likelihood estimation. This mechanism not only captures linguistic patterns but also absorbs social biases present in training corpora. Since contemporary academic evaluation systems commonly use English publication (especially in high-impact journals) as a proxy for quality or internationalization (Nakatumba-Nabende et al., 2025), training data exhibit stable statistical patterns in which English texts co-occur frequently with high-quality evaluative vocabulary. Models do not distinguish genuine causal relationships from noise; they directly map such co-occurrence patterns into parameter space, forming an implicit English = high quality association (Hofmann et al., 2024). Cultural norms, evaluative vocabulary, and implicit biases embedded in cross-language corpora are likely absorbed and encoded in model representation spaces (Caliskan et al., 2017), leading to inconsistent reasoning paths and judgment standards when processing different languages.

These two mechanisms reveal a critical issue: in academic evaluation, GenAI may not judge based on objective standards of content quality but may be systematically influenced by language cues. Accordingly, we propose the first hypothesis:

> **H1.** GenAI assigns significantly higher scores to English-presented content than to Chinese-presented content.

## 2.3 | Moderating factors

This study further focuses on two moderating factors that provide mechanistic clues for understanding the boundary conditions of language bias.

### 2.3.1 | Disciplinary indigeneity

Language is deeply embedded in cultural contexts, with its semantic structures, conceptual systems, and discursive modes shaped by specific sociocultural frameworks (Salö, 2018). Drawing on anthropological research (Merlan, 2009), we define this embeddedness as disciplinary indigeneity—the degree to which a discipline's core concepts and knowledge representations depend on specific cultural contexts. Strong-indigeneity disciplines rely on particular social narratives, cultural codes, and institutional backgrounds. Their knowledge expressions are more susceptible to semantic attenuation, conceptual drift, or incomplete mapping of cultural connotations during cross-language conversion (Canagarajah, 2002; Venuti, 2017). In contrast, weak-indigeneity disciplines possess conceptual systems with higher cross-cultural consistency, where cross-language expressions typically approach semantic equivalence (Montgomery, 2013). Accordingly, we propose the following hypothesis:

> **H2.** Language bias intensifies as disciplinary indigeneity increases.

### 2.3.2 | GenAI model differences

GenAI models are sociotechnical products with cognitive boundaries constrained by their developmental ecosystems (Lewandowski et al., 2024). Training corpora, as the foundation for GenAI model learning, directly influence a model's capacity to understand and process language through their content, distribution, and quality (Caliskan et al., 2017). Because different models exhibit significant variations in corpus distributions and development ecosystems, this heterogeneity may moderate the direction and

intensity of language bias (Muennighoff et al., 2023). Accordingly, we propose the following hypothesis:

> **H3.** Compared to English-dominant GenAI models, Chinese-dominant GenAI models exhibit weaker bias or bias in the opposite direction.

## 3 | METHODS

This study employs a within-subject design (Shadish et al., 2001), with each paper evaluated by GenAI under both Chinese and English language conditions. The core advantage is that each paper serves as its own control: language effects are identified through scoring differences across conditions, thereby controlling for confounds such as academic quality, disciplinary characteristics, author identity, and publication year.

### 3.1 | Data and sample

#### 3.1.1 | Data source

Metadata were collected from the Classic Literature Database (CLD) maintained by the Book and Newspaper Information Center of Renmin University of China (2025a). The CLD covers academic literature in the HSS from 2013 to 2022. Its core objective is to identify mainstream, classic, and essential high-value publications, with evaluation criteria focusing on theoretical innovation, knowledge contribution, disciplinary impact, and practical guidance. The CLD is curated by over 1300 senior scholars with an acceptance rate of approximately 0.5%, constituting a high-quality subset of Chinese academic literature.

The CLD was selected because this study aims to examine language bias in GenAI evaluation rather than its quality identification capability. Using expert-certified high-quality samples ensures quality homogeneity, allowing observed evaluation differences to be attributed to language rather than quality variation. This provides the most rigorous testing conditions for examining language neutrality in academic evaluation—examining whether GenAI still exhibits systematic differences due to language presentation mode, even for publications widely recognized as exemplary by the Chinese academic community.

#### 3.1.2 | Disciplinary scope

This study covers 23 first-level disciplines in the HSS from the CLD, comprehensively representing the disciplinary ecology of Chinese HSS research. The 23 disciplines are classified into three categories: strong-indigeneity disciplines ($n = 7$) are those whose core concepts are deeply rooted in Chinese cultural contexts and difficult to translate with complete correspondence, such as Chinese Language & Literature, Chinese History, and Art. Medium-indigeneity disciplines ($n = 8$) combine indigenous practical characteristics with international theoretical frameworks, such as Philosophy, Law, and Information Resources Management. Weak-indigeneity disciplines ($n = 8$) have highly internationalized conceptual systems and strong Chinese-English terminological correspondence, such as Applied Economics, Business Administration, and Psychology.

#### 3.1.3 | Sampling strategy

We employed stratified random sampling, selecting 50 papers per discipline to form a fully balanced experimental design. The sampling time window follows these principles:

- Primary time window (2020–2022): Applied to disciplines with sufficient CLD sample sizes. This period was selected for two considerations: first, academic writing standards in recent papers are more mature, ensuring higher quality of bilingual abstracts; second, a three-year pool sufficiently supports representative random sampling while avoiding potential changes in academic conventions from an excessively broad time span.
- Extended time window (2013–2022): For disciplines with limited samples after applying screening criteria within 2020–2022, the sampling frame was extended to 2013–2022 to ensure sufficient sample pools.

All candidate papers were required to have complete bilingual titles, keywords, and abstracts that were author-original rather than later supplements. Papers failing to meet these criteria were replaced through random supplementation from alternative sample pools. We did not intervene in translation quality to maintain ecological validity and reflect the authentic bilingual presentation of Chinese academic journals. The final sample includes 1150 papers.

### 3.2 | Model setup

We selected GPT-4o (OpenAI, 2024) and DeepSeek-V3 (DeepSeek-AI et al., 2025) for comparison, representing state-of-the-art GenAI capabilities in the English and

Chinese ecosystems (Chiang et al., 2024). DeepSeek-V3 is a large language model developed by the Chinese company DeepSeek, demonstrating strong performance across multiple benchmarks with relatively balanced Chinese-English training data (Guo et al., 2025). GPT-4o is OpenAI's flagship model, with superior performance relative to GPT-4o-mini commonly used in large-scale academic evaluation research (Thelwall, 2025c), and a parameter scale closer to DeepSeek-V3, making it a well-matched choice for this study. We obtained evaluation data through API interfaces, ensuring each call was an independent single-turn event, thereby precluding memory effects and data contamination that might occur with web interfaces.

We established two input conditions—Chinese and English—each presenting titles and abstracts. Prompt language matches input content, aiming to test the influence of complete language environments on GenAI evaluation.

We designed prompts based on LangGPT (Wang et al., 2024), with core elements including: (1) clear role positioning as an academic review expert; (2) explicit definition of four evaluation dimensions and their connotations; (3) specification of a structured JSON output format; (4) detailed scoring criteria; (5) explicit emphasis on universalism principles, requiring that GenAI evaluation to be independent of irrelevant attributes such as language, author identity, and institutional background.

Models were asked to follow the evaluation criteria of the CLD (Book and Newspaper Information Center of Renmin University of China, 2025b), encompassing (1) academic contribution, assessing theoretical innovation and knowledge contribution of papers; (2) theoretical innovation, evaluating novelty levels in concepts, methods, or perspectives; (3) research rigor, assessing research design, data quality, and argumentative adequacy; (4) academic normativity, evaluating the normative degree of literature review, logical structure, and academic expression. Scoring standards are 1–3 (poor), 4–5 (fair), 6–7 (good), 8–9 (excellent), and 10 (outstanding). After scoring, the GenAI was required to provide a brief overall judgment text explaining its rationale, supplying the foundation for subsequent qualitative analysis.

To enhance scoring stability, each paper underwent five independent evaluations under each condition, with the arithmetic mean serving as the final score. Prior research indicates that repeated GenAI evaluations with averaging significantly reduce random fluctuations in complex evaluation tasks, and that five repetitions balance reliability gains against computational cost (Thelwall, 2024). We use the mean rather than the mode or median, although scores typically possess ordinal properties, because previous research found that this approach, followed by scaling, is an effective predictive mechanism (Thelwall, 2025c).

## 3.3 | Analysis

### 3.3.1 | Hypothesis testing

To test H1, we used paired-samples *t*-tests to compare scoring differences between the English and Chinese conditions. Effect size was calculated using Cohen's d (Cohen, 2013) to quantify the practical significance magnitude of language bias. Tests were conducted at the full sample level, across three disciplinary categories, and within all 23 disciplines to reveal overall patterns and disciplinary differences in bias.

To test H2 and H3, we constructed a mixed-effects model:

$$Score \sim Condition \times Indigeneity \times Model + (1|Discipline) + (1|Paper)$$

where *Condition* represents input condition, *Indigeneity* represents disciplinary indigeneity category, and *Model* represents the GenAI model. Random intercepts for $(1|Discipline)$ and $(1|Paper)$ control for nested structures at the respective levels. Key tests include main effect of *Condition* (H1), the $Condition \times Indigeneity$ interaction effect (H2), and the $Condition \times Model$ interaction effect (H3). The model was estimated with Restricted Maximum Likelihood.

For multiple comparisons, False Discovery Rate (FDR) correction was applied to control Type I error inflation (Benjamini & Hochberg, 1995).

### 3.3.2 | Thematic analysis

To understand the mechanisms underlying language bias more deeply, we conducted thematic analysis of GenAI-generated evaluation texts, examining the cognitive processes and decision-making logic behind the statistical patterns.

1. **Analytical framework**: We employed Appraisal Theory (Martin & White, 2005) to systematically code GenAI evaluation texts. Originating from systemic functional linguistics, Appraisal Theory reveals how evaluative meaning is constructed in discourse and has been widely applied in academic discourse analysis to depict how evaluators express stance (Hood, 2010; Lam & Crosthwaite, 2018), negotiate

certainty, and adjust evaluative intensity. Given the specificity of the peer review genre and the coding workload, we contextually adapted the original framework: the attitude system retains the tripartite structure of appreciation, judgment, and affect; the engagement system retains the basic distinction between monogloss and heterogloss, capturing how GenAI negotiates evaluative certainty and dialogic space under different language conditions; and the graduation system selects force to quantify intensity differences.

2. **Sample selection**: We used stratified purposeful sampling to select 45 papers from each indigeneity category from the 1150 papers, totaling 135 papers for qualitative analysis. Sampling criteria were: (a) bias intensity: within each indigeneity category, we selected 15 papers exhibiting strong bias (effect size in the top 20%), moderate bias (near the average effect size), and weak bias (bottom 20%) to capture bias intensity gradients; (b) model differences: we included cases where GPT-4o bias exceeds DeepSeek-V3's, cases where the two are similar, and cases where DeepSeek-V3 bias is stronger. Since each paper received five repeated evaluations under each language condition, we selected the judgment text whose scores was closest to the median as the representative sample. The median is insensitive to extreme values and more robustly represents typical evaluation patterns. When multiple median evaluations existed, the most detailed text was selected to yield richer analytical material.
3. **Disciplinary coverage**: All 23 disciplines are represented to enhance the disciplinary robustness of findings. The final sample yields 540 evaluation texts, ensuring theoretical saturation while maintaining feasibility of in-depth coding analysis.
4. **Coding procedure and reliability**: Two independent coders first jointly studied the theoretical literature and coding manuals from Martin and White (2005), then jointly coded 10 samples to align their understanding of the framework and revise coding guidelines. After this trial coding, 10% (54 texts) were randomly selected for double coding to assess intercoder reliability. Cohen's Kappa coefficients were 0.86 (attitude system), 0.84 (engagement system), and 0.91 (graduation system), all indicating good consistency. Disagreements were resolved through discussion; the remaining texts were completed independently.
5. **Analytical strategy**: After coding was completed, we conducted three levels of analysis. First, we operationalized the three Appraisal systems into measurable indicators for cross-condition comparison. The attitude system was captured by negative attitude proportion (percentage of negative markers among all attitude markers), with negative academic quality appreciation and negative capability judgment further distinguished. The engagement system was captured by monoglossic proportion (percentage of bare assertions without dialogic alternatives) and heteroglossic proportion (percentage of hedged or dialogically expanded statements). The graduation system was captured by graduation intensity (mean force score on a 1–5 scale). Second, we calculated distribution frequencies of these indicators under each condition to identify systematic difference patterns and tracked evaluative resource changes for the same paper across language conditions through case comparison. Third, we integrated coding findings with quantitative results to examine how attitude polarity, engagement strategies, and graduation intensity jointly shape scoring outcomes.

# 4 | RESULTS

## 4.1 | Descriptive statistics

Figure 1 presents the score distributions of the four evaluation dimensions under Chinese and English conditions. Overall, GenAI scores are concentrated at the high end, consistent with the curated nature of the CLD samples. Under the Chinese condition, the overall mean across four dimensions is 6.99 (SD = 0.34), with academic contribution receiving the highest scores (M = 7.35, SD = 0.52) and research rigor the lowest (M = 6.65, SD = 0.80). The overall mean for the English condition is 6.85 (SD = 0.38), with scores across all four dimensions slightly lower than those under the Chinese condition.

Scores across all dimensions are primarily concentrated in the 7–10 range (51.3%), with medians around 7.0. The Chinese condition shows slightly higher scores and smaller standard deviations across all four dimensions, indicating higher scoring consistency. This preliminary observation suggests that language conditions may exert systematic effects on GenAI evaluation.

## 4.2 | Language bias patterns

### 4.2.1 | H1: Overall pattern of language bias

Paired-samples *t*-test results for the full sample are shown in Figure 2. The two GenAI models exhibit diametrically opposite language bias patterns.

For GPT-4o, English scores (M = 7.161, SD = 0.413) are significantly higher than Chinese scores (M = 6.842, SD = 0.456), with a mean difference of 0.319 points,

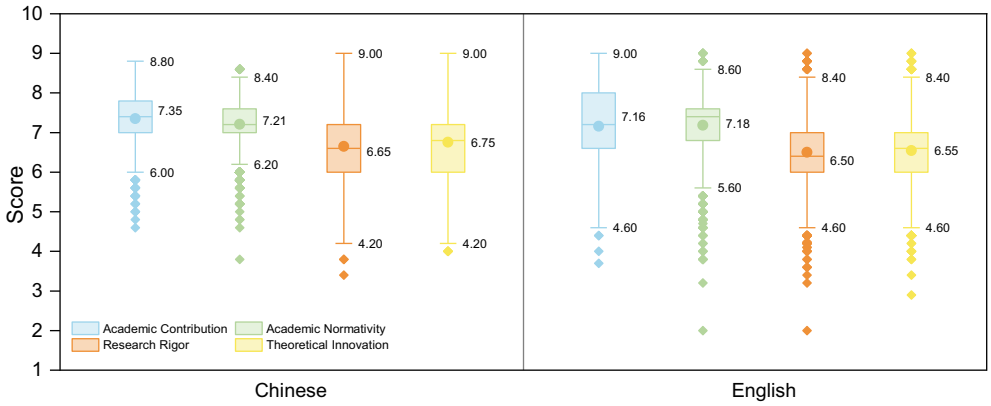



**FIGURE 1** Score distributions of four evaluation dimensions. Box represents interquartile range (IQR, Q1–Q3), horizontal line shows median, whiskers extend to 1.5 × IQR, dots represent outliers.

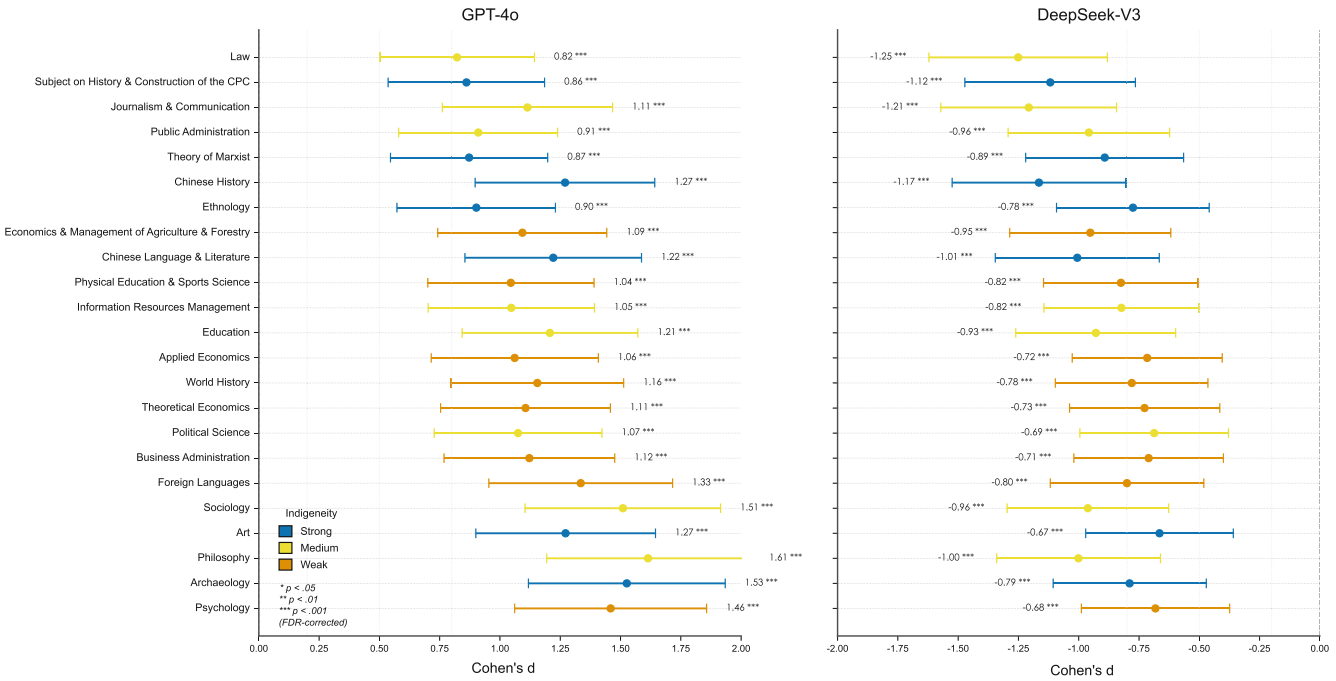



**FIGURE 2** Paired-samples *t*-test results for full sample (by discipline).

t(1149) = −37.27, $p < 0.001$, Cohen's d = 1.099, 95% CI [1.026, 1.172], a large effect size. 86.3% of papers ($n = 992$) received higher scores under the English condition, with only 7.5% of papers ($n = 86$) receiving higher scores under the Chinese condition. A Wilcoxon signed-rank test confirms this finding (Z = −27.52, $p < 0.001$).

Conversely, DeepSeek-V3 demonstrates a significant Chinese preference: Chinese scores (M = 7.140, SD = 0.556) are significantly higher than English scores (M = 6.535, SD = 0.715), with a mean difference of 0.605 points, t(1149) = 29.42, $p < 0.001$, Cohen's d = −0.867, 95% CI [−0.935, −0.800]. 82.0% of papers (943/1150) received higher scores under Chinese conditions, with only 16.5% scoring higher scores under the English condition. A Wilcoxon signed-rank test likewise confirms this pattern (Z = −24.47, $p < 0.001$).

Language bias is consistent patterns across all four evaluation dimensions. For GPT-4o, bias is strongest for theoretical innovation (Cohen's d = 0.902, mean difference = 0.393), followed by academic contribution (d = 0.805, mean difference = 0.296) and academic normativity (d = 0.770, mean difference = 0.378), with research rigor showing the weakest bias (d = 0.438, mean difference = 0.208). For DeepSeek-V3, bias is strongest for academic contribution (d = −1.085, mean difference = −0.681) and theoretical innovation (d = −1.071, mean difference = −0.812) dimensions, with academic normativity the weakest (d = −0.459, mean difference = −0.424). All dimensional effects reach statistical significance for both models ($p < 0.001$).

At the disciplinary level, language bias for both models reaches statistical significance across all

**TABLE 1** Key mixed-effects model analysis results.

| Effect | β | SE | *t* | *p* |
|---|---|---|---|---|
| Condition × Indigeneity | 0.005 | 0.007 | 0.748 | 0.455 |
| Condition × Model | −0.232 | 0.006 | −39.537 | 0.000*** |
| Three-way interaction | 0.023 | 0.007 | 3.122 | 0.002** |

***$p < 0.001$,**$p < 0.01$.

23 disciplines (after FDR correction, $p < 0.001$) (Figure 2). For GPT-4o, the largest effect sizes appear in Philosophy (d = 1.614), Archaeology (d = 1.526), and Sociology (d = 1.510), and the smallest in Law (d = 0.822), Subject on History & Construction of the CPC (d = 0.861), and Theory of Marxist (d = 0.872). For DeepSeek-V3, the largest absolute effect sizes appear in Law (d = −1.251), Journalism & Communication (d = −1.207), and Chinese History (d = −1.165), and the smallest in Art (d = −0.665), Psychology (d = −0.682), and Political Science (d = −0.687).

Overall, H1 receives partial support. Systematic language bias exists in GenAI evaluation, but its direction of bias varies by GenAI model. GPT-4o exhibits significant English bias, while DeepSeek-V3 demonstrates significant Chinese bias. Disciplinary-level variations suggest that indigeneity may moderate bias intensity, providing preliminary evidence for testing H2.

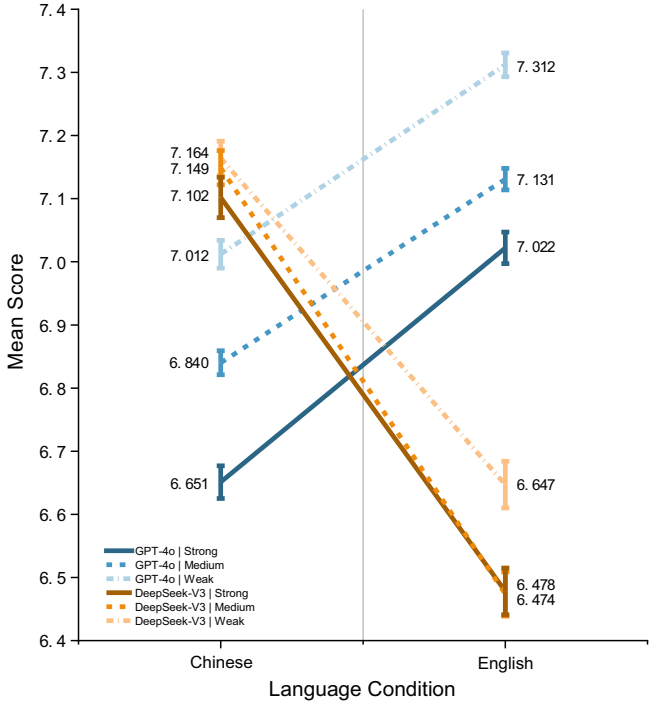



**FIGURE 3** Interaction effect plot.

### 4.2.2 | H2 and H3: Moderating effects of disciplinary indigeneity and GenAI model differences

The mixed-effects model successfully converged, with a random intercept variance of 0.132 for papers and a residual variance of 0.158, yielding an intraclass correlation coefficient of 0.455, indicating that 45.5% of score variance is attributable to between-paper differences. Key interaction effects are shown in Table 1; complete model output including all fixed-effect coefficients and model fit statistics is provided in Appendix B (Tables B1 and B2).

All three main effects are significant. The main effect of language condition (β = −0.072, SE = 0.006, z = −12.23, $p < 0.001$) indicates that after controlling for other factors, English scores are on average 0.14 points lower than Chinese scores. The main effect of disciplinary indigeneity (β = 0.111, SE = 0.015, z = 7.33, $p < 0.001$) indicates that weak-indigeneity disciplines receive higher overall scores than strong-indigeneity disciplines. The main effect of GenAI model (β = −0.080, SE = 0.006, z = −13.57, $p < 0.001$) shows that DeepSeek-V3 assigns higher overall scores than GPT-4o.

The *Condition* × *Indigeneity* interaction is not significant (β = 0.005, $p = 0.455$), failing to support H2. Simple slope analysis reveals that score differences between English and Chinese conditions are for strong-, medium-, and weak-indigeneity disciplines respectively—a spread of only 0.02 points that does not reach statistical significance.

The *Condition* × *Model* interaction is highly significant (β = −0.232, SE = 0.006, z = −39.54, $p < 0.001$), strongly supporting H3. As shown in Figure 3, the two models exhibit diametrically opposite patterns: GPT-4o's score are 0.32 points higher than its Chinese scores (β = 0.160, $p < 0.001$), whereas DeepSeek-V3's English scores are 0.61 points lower than its Chinese scores (β = −0.304, $p < 0.001$). This interaction effect size (β = −0.232) is the largest fixed effects, indicating that model identity is the most important moderator of language bias.

The three-way interaction effect reaches statistical significance (β = 0.023, SE = 0.007, z = 3.12, $p = 0.002$),

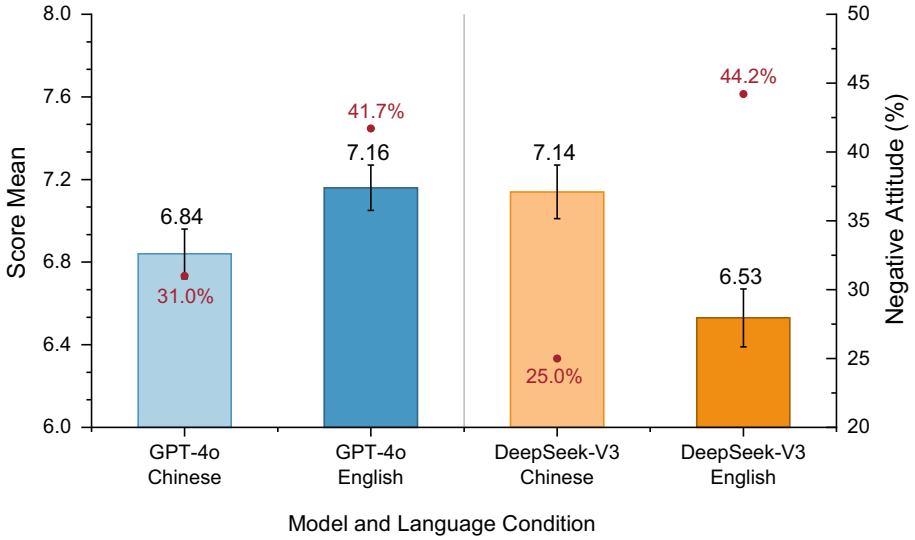



**FIGURE 4** Comparison of attitude resource allocation. Bar charts (left Y-axis) represent average evaluation scores with standard error bars; red dots with percentage labels (right Y-axis) show the percentage of negative attitudes in evaluation texts.

but as evident from Figure 3, this effect is relatively weak. Although the moderating role of indigeneity differs slightly between the two models, both patterns are dominated by the strong *Condition* × *Model* main effect.

## 4.3 | Cognitive mechanisms of language bias

Qualitative coding covered 540 evaluation texts, averaging 11.7 attitude markers, 3.9 engagement markers, and 8.3 graduation markers per text.

### 4.3.1 | Model-specific evaluative resource allocation patterns

The scoring results show that the two models exhibit opposite biases, yet qualitative coding reveals that both generate higher proportions of negative attitudes toward English papers in their evaluative justifications. As shown in Figure 4, this contradiction manifests clearly across the four condition combinations.

Both models show significantly higher negative attitudes (the percentage of negative attitude markers among all attitude markers) under English conditions: GPT-4o increases from 31.0% under the Chinese condition to 41.7% under the English condition ($d = -0.74$, $p < 0.001$), while DeepSeek-V3 increases from 25.0% to 44.2% ($d = -1.06$, $p < 0.001$).

The key mechanism behind this paradox lies in differentiated deployment of engagement strategies. GPT-4o's heteroglossic proportion (the percentage of evaluative statements that hedge certainty or acknowledge alternatives) under the English condition reaches 51.9%, far exceeding the 27.2% under the Chinese condition; by contrast, DeepSeek-V3's heteroglossic proportion rises only from 12.4% to 32.5%. Meanwhile, graduation intensity increases significantly under English conditions for both models (GPT-4o: 3.36 vs. 2.94, $d = -1.07$, $p < 0.001$; DeepSeek-V3: 3.50 vs. 3.28, $d = -0.49$, $p < 0.001$).

Taking typical expressions of negative academic quality appreciation as examples, Table 2 presents cases systematically selected for scoring patterns consistent with overall model-level bias directions and substantially divergent negative attitude patterns between models. These cases demonstrate how resource allocation operates at the micro-level.

Paper No. 5 illustrates how appraisal resources jointly shape scores at the micro level. GPT-4o increases negative appraisals from 3 to 5 under the English condition, yet the score rises to 6.5—because these criticisms are delivered heteroglossically (e.g., hedged within broader acknowledgment), attenuating their scoring impact. Conversely, DeepSeek-V3 generates 7 negative appraisals under the Chinese condition (66.7% delivered monoglossically, i.e., as bare assertions without hedging) with elevated graduation intensity, translating directly into a scoring penalty; under the English condition negative appraisals decrease to 5, yet the score drops further to 5.0. Paper No.1068 displays a similar pattern: GPT-4o shows 5 negative appraisals in Chinese versus 3 in English, yet the English score is higher (7.0 vs. 6.0). These cases indicate that scores are not determined by attitude polarity alone but by the coordinated deployment of engagement and graduation resources—a mechanism further elaborated in the Discussion.

Statistical testing reveals that, except for negative capability judgment, all coding dimensions exhibit more negative evaluative justifications under English conditions (Table 3).

**TABLE 2** Evaluative resource allocation and scoring comparison of typical cases.

| Paper No. | Discipline | Model | Condition | Number of negative attitudes | Excerpt from the judgments | Score |
|---|---|---|---|---|---|---|
| 5 | World History | GPT-4o | Chinese | 3 | 但缺乏新的理论框架的构建 [But lacks the construction of new theoretical frameworks] | 6 |
| | | | English | 5 | Does not provide groundbreaking theoretical frameworks | 6.5 |
| | | DeepSeek-V3 | Chinese | 7 | 但未明确突破现有学界对保守主义实用性的普遍认知 [But does not clearly break through the prevailing scholarly consensus on the practicality of conservatism] | 6 |
| | | | English | 5 | Lacks significant theoretical innovation | 5 |
| 1068 | Foreign Language | GPT-4o | Chinese | 5 | 在理论创新性方面表现一般 [Performs ordinarily in theoretical innovation] | 6 |
| | | | English | 3 | There are areas for improvement | 7 |
| | | DeepSeek-V3 | Chinese | 3 | 在研究严谨性上缺乏具体案例分析或实证支撑 [Lacks specific case analysis or empirical support in terms of research rigor] | 7.75 |
| | | | English | 3 | Lacks specific methodological details and empirical evidence | 6.75 |

*Note*: Chinese and English excerpts within each row are independently generated evaluation texts under the respective language conditions, not translation pairs. Bracketed translations are provided for accessibility.

**TABLE 3** Comparison of language bias in evaluation dimensions between two models.

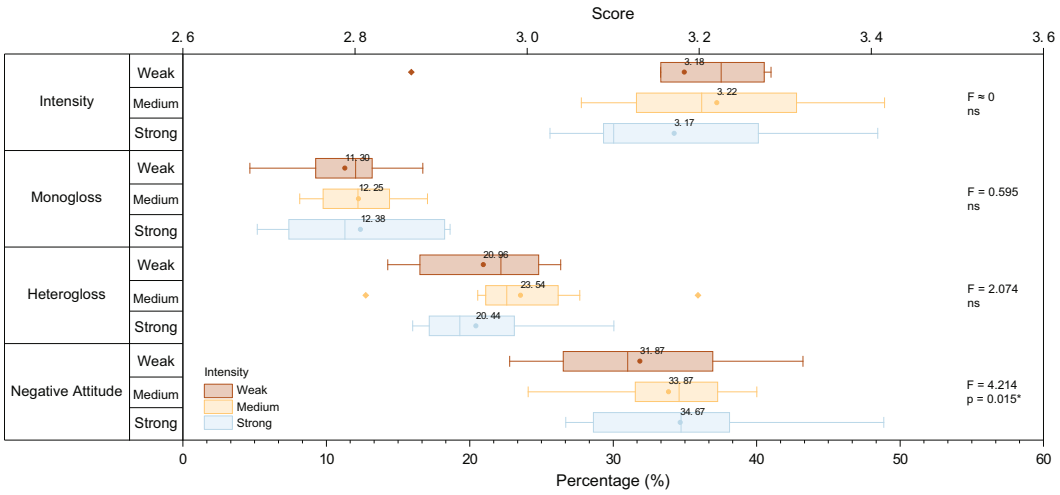
| Coding dimension | GPT-4o bias | DeepSeek-V3 bias | GPT-4o'd | DeepSeek-V3'd |
|---|---|---|---|---|
| Negative Attitude Proportion | −10.66% | −19.20% | −0.74 (***) | −1.06 (***) |
| Negative Academic Quality | −1.38 | −1.39 | −0.94 (***) | −0.73 (***) |
| Negative Capability Judgment | +0.01 | −0.10 | 0.05 (ns) | −0.36 (**) |
| Monoglossic Proportion | −4.26% | −7.40% | −0.30 (*) | −0.28 (*) |
| Graduation Intensity | −0.42 | −0.22 | −1.07 (***) | −0.49 (***) |

*Note*: Bias = Chinese condition mean − English condition mean; negative values indicate higher or stronger levels under English conditions. Negative attitude proportion and monoglossic proportion are reported as percentage point differences; negative academic quality, negative capability judgment, and graduation intensity are reported as differences in per-text means. Coding dimensions correspond to the Appraisal Theory framework: negative attitude proportion, negative academic quality, and negative capability judgment belong to the attitude system; monoglossic proportion belongs to the engagement system; graduation intensity belongs to the graduation system. See Section 3.3.2 for operational definitions of each indicator. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, ns = not significant.

Only negative capability judgment shows an effect in the opposite directions, but for GPT-4o it is not significant (d = 0.046, $p = 0.703$). Across all other dimensions, evaluative justifications under English conditions are more negative, more certain, and more intense for both models, yet run counter to the scoring directions reported in the previous section.

### 4.3.2 | Engagement strategies behind Indigeneity's null effect

Disciplinary indigeneity significantly affects GenAI's attitude judgments but does not influence engagement strategy selection. Figure 5 presents the distributions of the three indigeneity categories across key evaluation dimensions. Negative attitude proportions differ significantly across categories (F = 4.214, $p = 0.015$), with strong-indigeneity disciplines exhibiting significantly higher negative proportions (M = 37.6%, SD = 18.2%) than weak-indigeneity disciplines (M = 32.4%, SD = 17.7%). However, the engagement dimension box plots highly overlap: monoglossic proportions (Strong: M = 12.8%, SD = 23.3%; Medium: M = 14.0%, SD = 22.6%; Weak: M = 11.4%, SD = 20.8%) show no significant between-category differences (F = 0.595, $p = 0.552$), and heteroglossic proportions are likewise unaffected by indigeneity (F = 2.074, $p = 0.127$).




Lorem ipsum

**FIGURE 5** Distribution and comparison of evaluation strategies. Horizontal box plot bottom axis displays proportions of negative attitude, monogloss, and heterogloss; top axis displays graduation intensity scores.

Graduation intensity boxes are almost completely overlapping (Strong: 3.26; Medium: 3.28; Weak: 3.26), further confirming that GenAI's judgments of evaluative intensity are not moderated by disciplinary indigeneity.

As shown in Table 4, discipline-level analysis further reveals this attitude-strategy separation phenomenon. Negative attitude proportion spans 27.7 percentage points (Theory of Marxist 50.7% vs. Business Administration 23.0%), exhibiting a clear indigeneity gradient: strong-indigeneity disciplines are generally positioned at the high end of the distribution, while weak-indigeneity disciplines concentrate at the low end. However, monoglossic proportion varies unsystematically, spanning 18.2 percentage points within strong-indigeneity disciplines (Theory of Marxist 22.5% vs. Art 4.3%) and 15.1 percentage points within weak-indigeneity disciplines (Applied Economics 20.9% vs. Economics & Management of Agriculture & Forestry 5.8%), with no statistically significant between-group differences (all $p > 0.10$ after post-hoc comparisons). This indicates that GenAI is disciplinarily sensitivity in evaluation content (attitude selection) but applies a uniform paradigm in evaluation approach (engagement strategy). The complete distribution across all 23 disciplines is provided in Appendix A (Table A1).

## 5 | DISCUSSION

### 5.1 | Relationship with prior research

This study provides the first systematic confirmation that, in authentic academic evaluation tasks, scores generated by large language models exhibit structural language bias. This finding is consistent with recent research on cross-language capability differences in GenAI: multiple benchmarks demonstrate that mainstream GenAI models perform significantly better on English tasks than on other languages (Qin et al., 2025; Y. Xu et al., 2025), a disparity attributed to deep imbalances in language distribution within training corpora (T. Brown et al., 2020; Caliskan et al., 2017). Notably, this imbalance is bidirectional. Beyond the dominance of English in global academic publishing, the limited adoption of open access in Chinese academic publishing restricts the availability of Chinese-language scholarship as training material. Despite policy efforts, China lacks a national OA mandate, and most Chinese-language journals remain behind access barriers (CAST & STM, 2022), resulting in structural underrepresentation in model training corpora. When scholarly output is structurally inaccessible, training data imbalances become self-reinforcing, and models develop evaluative preferences aligned with the more accessible linguistic tradition. Against this backdrop, our validation extends the study of language bias from general task scenarios to the core information processing of academic quality judgment, demonstrating that language systematically influences model assessments of paper quality.

This study advances existing understanding in three aspects:

First, bias direction exhibits clear model dependence. Prior research assumed GenAI systems favor English content (Chua et al., 2024; Privitera et al., 2024), but we find DeepSeek-V3 demonstrates a significant Chinese preference ($d = -0.867$), in stark contrast with GPT-4o's English preference ($d = 1.099$). This finding challenges the simplified assumption that language bias necessarily favors English and confirms H3.

**TABLE 4** Appraisal strategy distribution across disciplines.

| Discipline | Indigeneity | Negative attitude proportion | Monoglossic proportion | Graduation intensity |
|---|---|---|---|---|
| Theory of Marxist | Strong | 50.70% | 22.50% | 3.237 |
| Ethnology | Strong | 42.00% | 9.50% | 3.163 |
| Art | Strong | 38.30% | 4.30% | 3.205 |
| Applied Economics | Weak | 31.60% | 20.90% | 3.197 |
| Economics & Management of Agriculture & Forestry | Weak | 26.00% | 5.80% | 3.343 |
| Business Administration | Weak | 23.00% | 8.80% | 3.288 |

*Note*: Negative attitude proportion: percentage of negative markers among all attitude markers. Monoglossic proportion: percentage of evaluative clauses presented as bare assertions without dialogic alternatives. Graduation intensity: mean force score (1–5 scale). All three indicators are aggregated across both models and both language conditions per discipline. Arranged in descending order of negative attitude proportion; only representative disciplines are shown. See Section 3.3.2 for full operational definitions.

Second, the moderating effect of cultural context variables is far weaker than theoretical expectations. Although disciplinary indigeneity should theoretically moderate language bias intensity (Canagarajah, 2002; Salö, 2018), our results show no significant difference across indigeneity categories ($\beta = 0.005$, $p = 0.455$), failing to support H2. While the three-way interaction reaches statistical significance ($\beta = 0.023$, $p = 0.002$), the effect size is extremely small and dominated by the model main effects, suggesting that indigeneity's role in GenAI evaluation bias is nearly marginalized.

Third, this study reveals a previously unmentioned phenomenon that the inverse relationship between scoring outcomes and evaluative justifications. Both models generate higher proportions of negative justifications under English conditions (GPT-4o: 41.7% vs. 31.0%; DeepSeek-V3: 44.2% vs. 25.0%), yet their scoring directions are opposite. This mechanistic finding has important implications for understanding the transparency and explainability of GenAI evaluation decisions and is further elaborated below.

Despite the presence of language bias, our results still form complementary evidence with existing GenAI-driven academic evaluation research: even under cross-language conditions, GenAI can extract structured cues and produce meaningful predictable scores from various types of academic content (Kousha & Thelwall, 2025; Thelwall, 2024; Thelwall et al., 2025; Thelwall & Cox, 2025; Zhu, Haunschild, et al., 2026; Zhu, Lu, et al., 2026).

## 5.2 | Coordinated appraisal resources

The core paradox revealed by hypothesis testing is: GPT-4o exhibits English preference while DeepSeek-V3 demonstrates Chinese preference, yet both models generate higher proportions of negative attitudes under English conditions. The pattern of more negative attitudes coexisting with higher scores suggests that score production is not a direct mapping of attitude polarity but is systematically regulated by deeper-level evaluative resource allocation.

Differential deployment of engagement strategies provides a key explanation. GPT-4o employs 51.9% heteroglossic strategies under the English condition (27.2% under Chinese), modulating the impact of negative attitudes through expressions like "could benefit from" and "might consider." By contrast, DeepSeek-V3's heteroglossic proportions remain low under both the English (32.5%) and Chinese (12.4%) conditions, allowing negative attitudes to more directly translate into scoring penalties. This difference may stem from divergent alignment training: GPT-4o has undergone extensive human feedback optimization that softens negative expressions (Ouyang et al., 2022), while DeepSeek-V3's fine-tuning on evaluation tasks may prioritize judgment clarity.

The synergistic effect of graduation intensification further amplifies this effect. Both models show significantly higher graduation intensity under English conditions (GPT-4o: $d = -1.07$, $p < 0.001$; DeepSeek-V3: $d = -0.49$, $p < 0.001$), manifested as more frequent intensifiers like "significantly," "seriously," and "completely lacking." Yet GPT-4o softens their impact through high heteroglossic proportions ("might need further strengthening" rather than "must improve"), whereas DeepSeek-V3, operating in a low-heteroglossic environment, directly translates intensifiers into scoring penalties. Statistical testing confirms that, except for negative capability judgment, both models produce more negative, more certain, and more intense evaluations

under English conditions across all dimensions; yet differences in resource allocation lead to score reversal.

This study reveals a resource allocation paradox that models may generate more negative judgments in a given language but ultimately produce higher scores through the coordinated deployment of engagement and graduation resources. This mechanism suggests that detection of GenAI bias cannot rely solely on final scores but must also audit the logic of evaluative justifications—systematic divergence between justifications and scores may indicate hidden bias pathways.

### 5.3 | Symmetric evaluation mechanism

Disciplinary indigeneity influences GenAI evaluation through a symmetric mechanism. Indigeneity increases evaluation difficulty, but this influence occurs equally under both Chinese and English conditions and thus fails to translate into differentiated cross-language bias.

The attitude system exhibits disciplinary sensitivity. Strong-indigeneity disciplines show significantly higher negative attitude proportions (37.6%) than weak-indigeneity disciplines (32.4%). This indicates that indigeneity indeed increases evaluation difficulty: the cultural specificity of concepts makes GenAI more prone to generate negative judgments such as "insufficient theoretical elaboration" or "vague conceptual definition." Critically, this influence is not language-specific—strong-indigeneity disciplines show more negative attitudes under both conditions.

Consolidation of engagement strategies ensures symmetric score transformation. Monoglossic proportions ($F = 0.595$, $p = 0.552$) and heteroglossic proportions ($F = 2.074$, $p = 0.127$) show no significant differences across indigeneity categories. The additional negative attitudes in strong-indigeneity disciplines are therefore transformed into scoring penalties at equivalent intensity in both languages. Since engagement strategies do not vary with indigeneity, the increase in negative attitudes remains comparable across conditions, causing scores to decrease in both languages while the cross-language difference remains constant.

This symmetry can be understood as follows. Indigeneity affects baseline difficulty (scores decrease under both conditions) but does not affect language sensitivity (the Chinese-English gap remains constant). Training data distributions explain this pattern that in mainstream model training corpora, both Chinese and English academic texts in strong-indigeneity disciplines are relatively scarce. Models thus face the same challenge of cultural context identification regardless of input language. In contrast, weak-indigeneity disciplines are sufficiently represented in both academic corpora, enabling more consistent cross-language processing. This bilingual synchronous insufficiency training pattern makes indigeneity's influence uniform across language conditions, ultimately raising overall difficulty without altering relative bias.

GenAI's perception of indigeneity, thus, is holistic rather than differential. It recognizes that strong-indigeneity disciplines are harder to evaluate but cannot recognize that they are harder to evaluate after translation into English. This suggests that any moderating role indigeneity may play in human peer review is obscured in current GenAI systems by bilingual synchronous insufficiency in training data.

## 6 | LIMITATIONS

This study has several limitations.

1. Sample representativeness. Although the CLD papers, widely recognized as exemplary by the Chinese academic community, used in this study ensure quality homogeneity for rigorous testing of language bias, they also limit generalizability. The bias patterns observed may represent conservative estimates; actual application scenarios could exhibit stronger bias.
2. Language condition design. We matched matching prompt language with input content to simulate a complete language environment of authentic academic evaluation. However, this design cannot separate the independent effects of paper language and interaction language.
3. Simplification of evaluation context. We did not provide full paper contents or supplementary background information to the models. Although title-and-abstract input combination has been found to align most closely with human scoring (Thelwall, 2025b), the absence of main-text argumentation may alter how language bias manifests.

Future research cloud deepen understanding in three directions: first, extending to other language pairs and disciplinary fields; second, experimentally manipulating translation quality and separating prompt language from content language. To isolate specific sources of bias; and third, introducing human peer review as a baseline to quantify the absolute magnitude of bias rather than studying only relative differences.

## 7 | CONCLUSION

This study provides the first systematic examination of language bias in GenAI-assisted academic evaluation,

revealing how language as a form of information representation systematically affects evaluative outcomes. Using a paired experimental design with 1150 Chinese HSS papers, we find that language bias is pervasive yet highly model-dependent in direction. This associated pattern indicates that training ecosystems play an important role in bias formation. We further find that disciplinary indigeneity does not moderate language bias intensity—although indigeneity affects absolute evaluation levels, its influence is symmetrically distributed across bilingual conditions and thus does not accumulate into cross-language bias. More distinctively, this study reveals systematic divergence between scores and evaluative justifications, a phenomenon not yet identified in existing GenAI evaluation fairness research.

Theoretically, this study makes three contributions to research on GenAI system fairness and cross-cultural information behavior. First, it reveals the bidirectionality and model-specificity of language bias, demonstrating that model training ecosystems predict bias direction better than task language itself. This challenges the view of language bias as merely a technical defect and reframes it as a systematic issue modulable through model design. Second, we identify an attitude-engagement separation mechanism: models recognize the complexity of disciplinary semantics yet fail to correspondingly adjust evaluation norms, forming a semi-perceptual state that is partially sensitive to evaluated content while maintaining rigid evaluation logic. Third, we confirm the synergistic coordination of evaluative resources, whereby scores are not linear mappings of attitude polarity but are jointly shaped by engagement strategies and graduation intensity, explaining the apparent paradox of more negative attitudes alongside higher scores. These findings hold important implications for understanding GenAI's decision logic when functioning as an academic evaluation intermediary.

Practically, this study cautions against uncritical use of GenAI in academic evaluation, particularly for cross-language assessments requiring horizontal comparison. The covertness of language bias makes post-hoc score calibration difficult to implement effectively. We recommend that institutions and policymakers build multi-level governance frameworks (Zhu et al., 2023) when adopting GenAI-assisted evaluation: (1) prioritize models with balanced cross-language training at the selection stage; (2) explicitly embed language neutrality principles in prompt design; (3) establish GenAI-assisted, human-led review processes (Zhu, Lu, et al., 2026) in which experts correct potential biases and render final decisions; (4) implement evaluation-score consistency audits to monitor systematic deviations in model judgment logic. For GenAI developers, this study underscores the necessity of incorporating cross-language fairness constraints into training data curation, alignment fine-tuning, and scoring strategy design. Additionally, as GenAI systems increasingly function as academic evaluation infrastructure, the openness and representativeness of their training data become foundational concerns. Expanding open access to non-English scholarly outputs—particularly in regions where most academic literature remains behind access barriers—is not merely a publishing policy issue but a prerequisite for constructing equitable data foundations for AI-driven evaluation systems.

## ACKNOWLEDGMENTS

Authors gratefully acknowledge the grant from the major project of the National Social Science Foundation of China (No. 24&ZD323) and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. KYCX25_0130).

We employed Claude Sonnet 4.5 and Opus 4.6 for the following purposes: (1) translating parts of sections of the text into English, (2) proofreading and correcting grammatical errors. We evaluated the output by cross-referencing the translated and revised content with the original text to ensure accuracy, consistency, and alignment with the intended meaning. Additionally, we reviewed the final version to confirm that all technical terms and concepts were appropriately conveyed. The authors assume all responsibility for the content of this submission.

## DATA AVAILABILITY STATEMENT

The data that support the findings are available from the corresponding author upon reasonable request.

## ORCID

*Yu Zhu* https://orcid.org/0000-0002-2548-828X
*Yujie Jia* https://orcid.org/0000-0003-4718-0056
*Yumeng Zhu* https://orcid.org/0009-0005-2645-0653
*Jiyuan Ye* https://orcid.org/0000-0002-4232-8923

## REFERENCES

Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Ahmed, M., Bali, K., & Sitaram, S. (2023). MEGA: Multilingual evaluation of generative AI. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 4232–4267). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.258

Ammon, U. (2001). *The dominance of English as a language of science: Effects on other languages and language communities*. Walter de Gruyter.

Ben-David, J. (1971). *The scientist's role in society: a comparative study*. Prentice-Hall. With Internet Archive. http://archive.org/details/scientistsrolein00bend

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Book and Newspaper Information Center of Renmin University of China. (2025a). *Classic Literature Database*. https://zszwx.cn/

Book and Newspaper Information Center of Renmin University of China. (2025b). *Selection Criteria-Classic Literature*. https://zszwx.cn/selectlx

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Brown, V., Larasati, R., Kwarteng, J., & Farrell, T. (2025). Understanding AI and power: Situated perspectives from global north and south practitioners. *Ai & Society*. https://doi.org/10.1007/s00146-025-02731-x

Buden, B., Nowotny, S., Simon, S., Bery, A., & Cronin, M. (2009). Cultural translation: An introduction to the problem, and responses. *Translation Studies World*, *2*(2), 196–219. https://doi.org/10.1080/14781700902937730

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Canagarajah, A. S. (2002). *A geopolitics of academic writing*. University of Pittsburgh Press.

Checco, A., Bracciale, L., Loreti, P., Pinfield, S., & Bianchi, G. (2021). AI-assisted peer review. *Humanities and Social Sciences Communications*, *8*(1), 25. https://doi.org/10.1057/s41599-020-00703-8

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M. I., Gonzalez, J. E., & Stoica, I. (2024). Chatbot arena: An open platform for evaluating LLMs by human preference. Proceedings of the 41st International Conference on Machine Learning, ICML'24, 235, 8359–8388.

China Association for Science and Technology (CAST) & International Association of STM Publishers (STM). (2022). *Open Access Publishing in China* [Online resource]. figshare. https://doi.org/10.6084/m9.figshare.21708113.v2

Chua, L., Ghazi, B., Huang, Y., Kamath, P., Kumar, R., Manurangsi, P., Sinha, A., Xie, C., & Zhang, C. (2024). *Crosslingual capabilities and knowledge barriers in multilingual large language models*. NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward. https://openreview.net/forum?id=LyRpeFlxDZ

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., ... Pan, Z. (2025). *DeepSeek-V3 technical report* (arXiv:2412.19437). arXiv. https://doi.org/10.48550/arXiv.2412.19437

Dong, G., Wang, H., Sun, J., & Wang, X. (2025). Evaluating and mitigating linguistic discrimination in large language models: Perspectives on safety equity and knowledge equity. *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI), Montreal, Canada, 2025 August 16-22*, 1–12. https://doi.org/10.48550/arXiv.2404.18534

DORA. (2012). *About DORA*. https://sfdora.org/about-dora/

Fourcade, M. (2009). *Economists and societies: Discipline and profession in the United States, Britain, and France, 1890s to 1990s*. Princeton University Press. https://doi.org/10.1515/9781400833139

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., ... Zhang, Z. (2025). DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, *645*(8081), 633–638. https://doi.org/10.1038/s41586-025-09422-z

Hall, P., & Ellis, D. (2023). A systematic review of socio-technical gender bias in AI algorithms. *Online Information Review*, *47*(7), 1264–1279. https://doi.org/10.1108/OIR-08-2021-0452

Hamel, R. E. (2007). The dominance of English in the international scientific periodical literature and the future of language use in science. *AILA Review*, *20*(1), 53–71. https://doi.org/10.1075/aila.20.06ham

Heaven, D. (2018). AI peer reviewers unleashed to ease publishing grind. *Nature*, *563*(7733), 609–610. https://doi.org/10.1038/d41586-018-07245-9

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). The leiden manifesto for research metrics. *Nature*, *520*(7548), 429–431. https://doi.org/10.1038/520429a

Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). *Dialect prejudice predicts AI decisions about people's character, employability, and criminality* (arXiv:2403.00742). arXiv. https://doi.org/10.48550/arXiv.2403.00742

Hood, S. (2010). *Appraising research: Evaluation in academic writing*. Springer.

Huang, S., Huang, Y., Liu, Y., Luo, Z., & Lu, W. (2025). Are large language models qualified reviewers in originality evaluation? *Information Processing & Management*, *62*(3), 103973. https://doi.org/10.1016/j.ipm.2024.103973

Kousha, K., & Thelwall, M. (2025). Assessing the societal influence of academic research with ChatGPT: Impact case study evaluations. *Journal of the Association for Information Science and Technology*, *76*(10), 1357–1373. https://doi.org/10.1002/asi.25021

Lam, S. L., & Crosthwaite, P. (2018). APPRAISAL resources in L1 and L2 argumentative essays: A contrastive learner corpus-informed study of evaluative stance. *Journal of Corpora and Discourse Studies*, *1*, 8. https://doi.org/10.18573/jcads.1

Langfeldt, L., Nedeva, M., Sörlin, S., & Thomas, D. A. (2020). Co-existing notions of research quality: A framework to study context-specific understandings of good research. *Minerva*, *58*(1), 115–137. https://doi.org/10.1007/s11024-019-09385-2

Lewandowski, D., Haider, J., & Sundin, O. (2024). JASIST special issue editorial: Re-orienting search engine research in information science. *Journal of the Association for Information Science and Technology*, *75*(5), 503–511. https://doi.org/10.1002/asi.24845

Liu, W. (2017). The changing role of non-English papers in scholarly communication: Evidence from web of Science's three

journal citation indexes. *Learned Publishing*, *30*(2), 115–123. https://doi.org/10.1002/leap.1089

Martin, J. R., & White, P. R. R. (2005). *The language of evaluation: Appraisal in English*. Palgrave Macmillan.

Meneghini, R., & Packer, A. L. (2007). Is there science beyond English? Initiatives to increase the quality and visibility of non-English publications might help to break down language barriers in scientific communication. *EMBO Reports*, *8*(2), 112–116. https://doi.org/10.1038/sj.embor.7400906

Merlan, F. (2009). Indigeneity: Global and local. *Current Anthropology*, *50*(3), 303–333. https://doi.org/10.1086/597667

Merton, R. K. (1942). The normative structure of science. In *The sociology of science*. University of Chicago Press.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and its Application*, *8*, 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Moed, H. F., Glänzel, W., & Schmoch, U. (2004). *Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of S&T systems*. Springer.

Montgomery, S. L. (2013). *Does science need a global language?: English and the future of research*. University of Chicago Press.

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Le Scao, T., Bari, M. S., Shen, S., Yong, Z. X., Schoelkopf, H., Tang, X., Radev, D., Aji, A. F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., & Raffel, C. (2023). Crosslingual Generalization through Multitask Finetuning. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the Association for Computational Linguistics (volume 1: Long papers)* (pp. 15991–16111). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.891

Nakatumba-Nabende, J., Kagumire, S., Kantono, C., & Nabende, P. (2025). A systematic literature review on bias evaluation and mitigation in automatic speech recognition models for low-resource African languages. *ACM Computing Surveys*, *58*(4), 105:1–105:24. https://doi.org/10.1145/3769089

Noseworthy, P. A., Attia, Z. I., Brewer, L. C., Hayes, S. N., Yao, X., Kapa, S., Friedman, P. A., & Lopez-Jimenez, F. (2020). Assessing and mitigating bias in medical artificial intelligence. *Circulation. Arrhythmia and Electrophysiology*, *13*(3), e007988. https://doi.org/10.1161/CIRCEP.119.007988

OpenAI. (2024). *GPT-4o system card*. https://openai.com/index/gpt-4o-system-card/

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Politzer-Ahles, S., Girolamo, T., & Ghali, S. (2020). Preliminary evidence of linguistic bias in academic reviewing. *Journal of English for Academic Purposes*, *47*, 100895. https://doi.org/10.1016/j.jeap.2020.100895

Privitera, A. J., Ng, S. H. S., Kong, A. P.-H., & Weekes, B. S. (2024). AI and aphasia in the digital age: A critical review. *Brain Sciences*, *14*(4), 383. https://doi.org/10.3390/brainsci14040383

Qin, L., Chen, Q., Zhou, Y., Chen, Z., Li, Y., Liao, L., Li, M., Che, W., & Yu, P. S. (2025). A survey of multilingual large language models. *Patterns*, *6*(1), 101118. https://doi.org/10.1016/j.patter.2024.101118

Salö, L. (2015). The linguistic sense of placement: Habitus and the entextualization of translingual practices in Swedish academia. *Journal of SocioLinguistics*, *19*(4), 511–534. https://doi.org/10.1111/josl.12147

Salö, L. (2018). *The sociolinguistics of academic publishing: Language and the practices of homo academicus*. Palgrave Macmillan.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Cengage Learning.

Shen, L., Tan, W., Chen, S., Chen, Y., Zhang, J., Xu, H., Zheng, B., Koehn, P., & Khashabi, D. (2024). The language barrier: Dissecting safety challenges of LLMs in multilingual contexts. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 2668–2680). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.156

Strauss, P. (2019). Shakespeare and the English poets: The influence of native speaking English reviewers on the acceptance of journal articles. *Publications*, *7*(1), 20. https://doi.org/10.3390/publications7010020

Thelwall, M. (2024). Can ChatGPT evaluate research quality? *Journal of Data and Information Science*, *9*(2), 1–21. https://doi.org/10.2478/jdis-2024-0013

Thelwall, M. (2025a). ChatGPT for complex text evaluation tasks. *Journal of the Association for Information Science and Technology*, *76*(4), 645–648. https://doi.org/10.1002/asi.24966

Thelwall, M. (2025b). ChatGPT for complex text evaluation tasks. *Journal of the Association for Information Science and Technology*, *76*(4), 645–648. https://doi.org/10.1002/asi.24966

Thelwall, M. (2025c). Evaluating research quality with large language models: An analysis of ChatGPT's effectiveness with different settings and inputs. *Journal of Data and Information Science*, *10*(1), 7–25. https://doi.org/10.2478/jdis-2025-0011

Thelwall, M. (2025d). Quantitative Methods in Research Evaluation Citation Indicators, Altmetrics, and Artificial Intelligence (arXiv:2407.00135). arXiv. https://doi.org/10.48550/arXiv.2407.00135

Thelwall, M. (2025e). Research quality evaluation by AI in the era of large language models: Advantages, disadvantages, and systemic effects – An opinion paper. *Scientometrics*, *130*(10), 5309–5321. https://doi.org/10.1007/s11192-025-05361-8

Thelwall, M., & Cox, A. (2025). Estimating the quality of academic books from their descriptions with ChatGPT. *The Journal of Academic Librarianship*, *51*(2), 103023. https://doi.org/10.1016/j.acalib.2025.103023

Thelwall, M., Jiang, X., & Bath, P. A. (2025). Estimating the quality of published medical research with ChatGPT. *Information Processing & Management*, *62*(4), 104123. https://doi.org/10.1016/j.ipm.2025.104123

Thelwall, M., & Kousha, K. (2023). Technology assisted research assessment: Algorithmic bias and transparency issues. *Aslib Journal of Information Management*, *77*(1), 175–190. https://doi.org/10.1108/AJIM-04-2023-0119

Thelwall, M., & Kurt, Z. (2025). Research evaluation with ChatGPT: Is it age, country, length, or field biased? *Scientometrics*, *130*(10), 5323–5343. https://doi.org/10.1007/s11192-025-05393-0

Venuti, L. (2017). *The translator's invisibility: A history of translation*. Routledge. https://doi.org/10.4324/9781315098746

Wang, M., Liu, Y., Liang, X., Li, S., Huang, Y., Zhang, X., Shen, S., Guan, C., Wang, D., Feng, S., Zhang, H., Zhang, Y., Zheng, M., & Zhang, C. (2024). LangGPT: Rethinking structured reusable prompt design framework for LLMs from the programming language. arXiv.org. https://arxiv.org/abs/2402.16929v2

Xu, Y., Hu, L., Zhao, J., Qiu, Z., Xu, K., Ye, Y., & Gu, H. (2025). A survey on multilingual large language models: Corpora, alignment, and bias. *Frontiers of Computer Science*, *19*(11), 1911362. https://doi.org/10.1007/s11704-024-40579-4

Xu, Z. (2025). Beyond English Hegemony: AI Academic Writing Tool Usage Among Non-Native English Speakers and International Teams. Proceedings of the ALISE Annual Conference. 10.21900/j.alise.2025.2050.

Yen, C.-P., & Hung, T.-W. (2019). New data on the linguistic diversity of authorship in philosophy journals. *Erkenntnis*, *84*(4), 953–974. https://doi.org/10.1007/s10670-018-9989-4

Zhu, Y., Chen, G., Lu, Y., & Fan, W. (2023). Generative artificial intelligence governance action framework: Content analysis based on AIGC incident report texts. *Documentation,Information & Knowledge*, *40*(4), 41–51. https://doi.org/10.13366/j.dik.2023.04.041

Zhu, Y., Haunschild, R., Ye, J., & Xie, H. (2026). Can generative AI effectively perform quality evaluation within social sciences? A case study in library and information science. *Scientometrics*. https://doi.org/10.1007/s11192-026-05570-9

Zhu, Y., Lu, Y., Xie, H., Ye, J., & Chen, M. (2026). A quasi-experimental analysis of capabilities and limitations of generative AI in academic content evaluation in social sciences. *Information Processing & Management*, *63*(1), 104365. https://doi.org/10.1016/j.ipm.2025.104365

**How to cite this article:** Zhu, Y., Jia, Y., Zhu, Y., & Ye, J. (2026). Does language bias GenAI academic evaluation in humanities and social sciences? A mixed-methods study based on Chinese-language HSS papers. *Journal of the Association for Information Science and Technology*, 1–19. https://doi.org/10.1002/asi.70079

## APPENDIX A

**TABLE A1** Complete appraisal strategy distribution across all 23 disciplines.

| Discipline | Indigeneity | Negative attitude proportion | Monoglossic proportion | Graduation intensity |
|---|---|---|---|---|
| Theory of Marxist | Strong | 50.70% | 22.50% | 3.237 |
| Ethnology | Strong | 42.00% | 9.50% | 3.163 |
| Subject on History & Construction of the CPC | Strong | 41.10% | 13.80% | 3.110 |
| Art | Strong | 38.30% | 4.30% | 3.205 |
| Chinese History | Strong | 32.90% | 15.00% | 3.362 |
| Archaeology | Strong | 31.00% | 8.70% | 3.305 |
| Chinese Language & Literature | Strong | 28.10% | 18.00% | 3.522 |
| Education | Medium | 43.60% | 14.20% | 3.304 |
| Political Science | Medium | 43.20% | 14.90% | 3.172 |
| Information Resources Management | Medium | 38.90% | 18.20% | 3.193 |
| Public Administration | Medium | 37.70% | 8.90% | 3.176 |
| Law | Medium | 34.20% | 7.50% | 3.219 |
| Philosophy | Medium | 33.70% | 18.00% | 3.450 |
| Sociology | Medium | 33.00% | 14.90% | 3.355 |
| Journalism & Communication | Medium | 24.80% | 16.40% | 3.492 |
| Psychology | Weak | 42.20% | 14.70% | 3.167 |
| Physical Education & Sports Science | Weak | 41.60% | 9.60% | 2.925 |
| World History | Weak | 36.60% | 8.50% | 3.333 |
| Theoretical Economics | Weak | 32.00% | 12.10% | 3.383 |
| Applied Economics | Weak | 31.60% | 20.90% | 3.197 |
| Foreign Languages | Weak | 30.70% | 15.30% | 3.250 |
| Economics & Management of Agriculture & Forestry | Weak | 26.00% | 5.80% | 3.343 |
| Business Administration | Weak | 23.00% | 8.80% | 3.288 |

## APPENDIX B

**TABLE B1** Fixed effects.

| Parameter | β | SE | z | *p* |
|---|---|---|---|---|
| Intercept | 6.915 | 0.012 | 565.939 | <0.001 |
| Condition | −0.072 | 0.006 | −12.229 | <0.001 |
| Indigeneity | 0.111 | 0.015 | 7.329 | <0.001 |
| Model | −0.080 | 0.006 | −13.573 | <0.001 |
| Condition × Indigeneity | 0.005 | 0.007 | 0.748 | 0.455 |
| Condition × Model | −0.232 | 0.006 | −39.537 | <0.001 |
| Indigeneity × Model | −0.052 | 0.007 | −7.213 | <0.001 |
| Condition × Indigeneity × Model | 0.023 | 0.007 | 3.122 | 0.002 |

*Note*: Condition is effect-coded (Chinese = −1, English = 1). Indigeneity is effect-coded (Strong = −1, Medium = 0, Weak = 1). Model is effect-coded (GPT-4o = −1, DeepSeek-V3 = 1). Estimated using Restricted Maximum Likelihood (REML) with random intercepts for papers.

**TABLE B2** Random effects and model fit.

| Component | Estimate |
|---|---|
| Random intercept variance (Paper) | 0.132 |
| Residual variance | 0.158 |
| ICC | 0.455 |
| Log-likelihood | −3151.825 |
| *N* (observations) | 4600 |
| *N* (groups) | 1150 |

*Note*: ICC indicates that 45.5% of total score variance is attributable to between-paper differences.