*Acta Informatica Pragensia*
2027, Volume 16, Issue 1, Forthcoming article
https://doi.org/10.18267/j.aip.336

PRAGUE UNIVERSITY OF ECONOMICS AND BUSINESS

**Article** Open Access

# Self-Preference or Position Bias? A Controlled Multilingual Study of Evaluation Bias in LLM-as-Judge Pairwise Comparison

**Tomáš Bruckner, Filip Vencovský, Oleg Svatoš**

Department of Information Technology, Faculty of Informatics and Statistics, Prague University of Economics and Business, Prague, Czech Republic

Corresponding author: Tomáš Bruckner (bruckner@vse.cz)

## Editorial Record

**First submission received:**
June 4, 2026

**Revisions received:**
September 10, 2026
October 22, 2026

**Accepted for publication:**
October 23, 2026

**Academic Editor:**
Jost Bartol
University of Ljubljana, Slovenia

This article was accepted for publication by the Academic Editor upon evaluation of the reviewers' comments.

**How to cite this article:**
Bruckner, T., Vencovský, F., & Svatoš, O. (2027). Self-Preference or Position Bias? A Controlled Multilingual Study of Evaluation Bias in LLM-as-Judge Pairwise Comparison. *Acta Informatica Pragensia*, 16(1), Forthcoming article. https://doi.org/10.18267/j.aip.336

**Copyright:**
© 2027 by the author(s). Licensee Prague University of Economics and Business, Czech Republic. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution License (CC BY 4.0).

## Abstract

**Background:** Large language models (LLMs) are increasingly used as automated evaluators, but concerns remain about systematic biases such as self-preference. Prior findings may be confounded by stylistic similarity, shared quality preferences, or presentation order rather than genuine authorship-based bias.

**Objective:** We test whether LLM judges favor their own generated texts in blind pairwise comparison beyond what shared quality preferences would predict, and we quantify position bias as a competing explanation.

**Methods:** We conduct a controlled, blind, multilingual experiment. Five near-frontier LLMs from five developers generate short texts in five European languages across three text types, with five generations per cell (375 texts). Each model then blindly compares its own text against each other model's text, in both presentation orders, yielding 3,000 evaluations. Self-selection is benchmarked against cross-model selection of the same text as a quality-controlled baseline, and we apply equivalence testing to bound the size of small effects.

**Results:** Across 3,000 evaluations, a correctly specified conditional-logit model (Bradley–Terry candidate strengths plus a self-bias term, controlling for both candidate quality and presentation order) revealed a modest but robust overall own-text preference (odds ratio or OR=1.26; $p<.001$). This was concentrated in two of the five judges (Gemini OR=2.09, Mistral OR=3.20; both $p<.001$); one judge showed the reverse (GPT OR=0.43; $p<.001$), and two showed no significant effect (Claude, Qwen). Model-specific position biases were strong (up to $\phi=0.62$) and constitute the single largest distortion.

**Conclusion:** Evaluation reveals a genuine own-text preference beyond shared quality, alongside strong position bias; both must be actively mitigated in LLM-as-judge pipelines. We do not attribute the own-text preference specifically to authorship recognition rather than familiarity, which our design does not separate.

## Index Terms

Large language model; Self-preference bias; Position bias; Evaluation bias; Multilingual evaluation.

## 1 INTRODUCTION

Large language models have rapidly expanded beyond their original role as text generators to serve as automated evaluators of text quality (Liu et al., 2023; Zheng et al., 2023). The so-called LLM-as-judge paradigm, in which one or more language models assess, compare, or rank outputs produced by other systems, has become a cornerstone of modern NLP evaluation methodology.

This approach offers substantial advantages over human evaluation: it is scalable, reproducible, relatively inexpensive, and capable of providing evaluations across multiple dimensions simultaneously (Chiang & Lee, 2023; Zheng et al., 2023). Prominent benchmarks such as MT-Bench (Zheng et al., 2023), AlpacaEval (Dubois et al., 2024), and FLASK (Ye et al., 2024) now rely on LLM-based evaluation as a primary or complementary assessment mechanism.

However, the validity of LLM-based evaluation rests on a critical assumption: that models can provide unbiased assessments of text quality, irrespective of the source of the texts being evaluated. Several forms of systematic bias have been identified in LLM evaluators, including sensitivity to factual accuracy (Y.-S. Chen et al., 2025), presentation order (Li et al., 2024; Wang et al., 2024), verbosity preference (Dubois et al., 2024), and stylistic favoritism (Koo et al., 2024; Stureborg et al., 2024). Among these concerns, one particularly insidious possibility has attracted growing attention: self-preference bias — the tendency of a model to prefer its own outputs when serving as an evaluator.

The present study originated from a practical observation during our ongoing research on multilingual text generation. When evaluating model outputs across several non-English languages, LLM evaluators appeared to systematically favor texts produced by themselves, particularly in linguistically challenging settings where none of the candidate texts were unambiguously superior. This raised the question of whether models harbor an inherent tendency to prefer their own linguistic patterns — a kind of evaluative narcissism that could undermine the validity of automated evaluation pipelines.

Such a bias, if it exists, has serious implications beyond evaluation methodology. If models consistently favor their own outputs, feedback loops may emerge in which models trained or fine-tuned on self-evaluated data progressively amplify their own stylistic tendencies, creating *echo chambers* of model-specific linguistic preferences (Alemohammad et al., 2024; Shumailov et al., 2024). We expand on this connection to model collapse in Section 2.5.

Recent work has begun to address self-preference directly. Panickssery et al. (2024) reported evidence that LLM evaluators can recognize and favor their own generations. However, Wataoka et al. (2024) argue that apparent self-preference operates through *perplexity* and *familiarity*: judges favor text that is more probable under their own distribution, whether or not they authored it. This is a crucial alternative hypothesis, because it predicts elevated self-selection even in the complete absence of authorship-based bias. A critical methodological challenge therefore lies in separating self-preference from *text quality* and from *shared distributional preferences*: a model that produces objectively better — or merely more fluent and distributionally central — text will naturally choose its own output more often, even without any authorship bias.

To address this confound, we designed a controlled experiment with the following properties: (a) five near-frontier LLMs generated multiple texts per condition in five languages across three text types; (b) each model blindly evaluated pairwise combinations of its own text and each other model's text, in both presentation orders; and (c) each model's self-selection rate was compared against how often *other* models chose that same text, providing a quality-controlled baseline. This symmetric design enables a decomposition of self-selection into a component shared across all evaluators (quality and shared preferences) and a component specific to the author–evaluator (candidate authorship bias).

Our study addresses the following research questions:

- **RQ1:** Do LLMs exhibit authorship-based self-preference in blind pairwise text evaluation, i.e., do they select their own texts more often than warranted by how other models rate the same texts?
- **RQ2:** Are there other systematic biases — particularly position bias — that affect LLM evaluation in this setting, and how do they relate to the quality gap between candidates?
- **RQ3:** Do these effects generalize across languages and text types (examined descriptively, not as confirmatory claims)?

We do not attempt to isolate the familiarity mechanism of Wataoka et al. (2024), which would require access to token-level log-probabilities unavailable for most proprietary judges; we return to this limitation in the Discussion.

Our contributions are: (1) a **quality-controlled, counterbalanced experimental design** that separates candidate authorship bias from quality and position; (2) an adequately powered, quality- and position-controlled analysis (a conditional-logit Bradley–Terry model with a self-bias term) showing a modest but robust own-text preference

concentrated in two of five judges, where a naive proportion test over-attributes it to two further judges; (3) evidence on **model-specific position biases** and their relationship to the candidate quality gap; and (4) a multilingual perspective across five European languages.

## 2 RELATED WORK

### 2.1 The LLM-as-Judge Paradigm

The use of language models as automated text evaluators has evolved rapidly. Liu et al. (2023) proposed G-Eval, demonstrating that GPT-4-based evaluation achieved higher correlation with human judgments than traditional automatic metrics. Zheng et al. (2023) formalized the LLM-as-judge paradigm through MT-Bench and Chatbot Arena. Subsequent work has explored reference-free evaluation (Kim et al., 2024), multi-dimensional assessment (Ye et al., 2024), evaluation awareness (Needham et al., 2025), and the use of LLM judges in reinforcement learning from human feedback pipelines (Rafailov et al., 2023).

### 2.2 Self-Preference Bias

Self-preference bias refers to the hypothesis that models systematically favor outputs they themselves generated. Panickssery et al. (2024) provided initial evidence that several LLMs could distinguish their own outputs and preferentially selected them. Xu et al. (2024) reported that self-bias can amplify during iterative self-refinement. However, these studies have been critiqued for potential confounds: if a model's outputs possess distinctive stylistic features that the model itself values — not because it recognizes them as its own, but because those features align with its implicit quality criteria — this would produce artifacts even without genuine authorship-based bias. Z.-Y. Chen et al. (2025) introduce a more nuanced measurement framework, arguing that apparent self-preference often emerges from latent stylistic similarity rather than explicit self-recognition. Similarly, Chhun et al. (2024) find that models tend to assign higher scores to texts that align with their own generative style, though this effect is highly sensitive to prompting.

Most directly, Wataoka et al. (2024) propose that self-preference is mediated by perplexity: judges prefer text they find more probable, which tends to be their own. Under this account, our quality-controlled design should show *no* residual authorship effect, because a text that is low-perplexity to its author is often low-perplexity to other capable judges as well — self-preference and cross-model consensus become, in part, the same signal measured twice. We take this seriously as the leading alternative explanation for any null we observe, and we discuss in Section 5.1 why our design can bound authorship bias but cannot exclude the familiarity mechanism.

### 2.3 Position Bias in LLM Evaluation

A well-documented bias in LLM pairwise evaluation is sensitivity to presentation order. Wang et al. (2024) demonstrated significant position bias across several LLMs. Li et al. (2024) proposed split-and-merge mitigation strategies. Koo et al. (2024) situated position bias within a broader framework of cognitive biases, drawing parallels with primacy and recency effects. Wu & Aji (2025) investigate the preference for style over content. Shi et al. (2025), in a large-scale study of roughly 150,000 evaluation instances across fifteen judges, show that position bias is strongly modulated by the *quality gap* between candidates: the larger the gap, the more (or less) an evaluator's decision can be swung by order. This is directly relevant to our design, because a steep quality hierarchy among candidates can itself inflate apparent position effects. We address this by selecting near-frontier models of comparable capability (Section 3.1) and by explicitly relating position effects to the measured quality gap (Section 4.3). Our design incorporates both presentation orders, enabling detection and quantification of position bias as a within-pair quantity.

### 2.4 Bias and Task Dependence

Recent work shows that LLM evaluation biases are highly task-dependent. Y.-S. Chen et al. (2025) demonstrate that while LLMs act as biased evaluators in general settings, this bias largely disappears in fact-centric retrieval-augmented generation, where grounding constrains subjective variation. By restricting evaluation to purely linguistic quality and removing factual and interpretative dimensions, we create a setting analogous to a low-

ambiguity evaluation regime, allowing us to test whether self-preference persists when subjective degrees of freedom are minimized.

### 2.5 Model Collapse and Echo Chambers

The concern about self-preference connects to *model collapse*. Model collapse refers to the progressive degradation of model quality that occurs when generative models are trained on data produced by earlier generations of models rather than on human-authored text. Shumailov et al. (2024) show that such *recursively generated data* causes distributions to lose their tails: rare events and minority patterns disappear first, and successive generations converge toward a narrow, self-reinforcing subset of the original distribution. For example, a translation model repeatedly trained on its own outputs may drift toward a small set of "safe" phrasings and lose lexical variety. Alemohammad et al. (2024) document an analogous failure they term *model autophagy disorder*, in which self-consuming generative loops amplify artifacts and reduce diversity over time. If evaluators preferentially select their own outputs, this could accelerate such degradation by creating closed feedback loops in which model-specific idiosyncrasies are rewarded rather than genuine quality. Our study provides evidence on whether this mechanism operates at the evaluation stage.

### 2.6 Multilingual Evaluation

Most existing studies of LLM evaluation bias have focused on English. Ahuja et al. (2023) highlighted systematic performance gaps between English and other languages. Our study examines self-preference and position bias across five European languages, providing evidence on whether evaluation biases are language-dependent within this scope.

## 3 METHODS

### 3.1 Model Selection

We selected five instruction-tuned LLMs from five different developers and three geographies, chosen to be of comparable near-frontier capability so as to minimize the quality gap between candidates (Table 1). Minimizing the quality gap directly addresses the finding of Shi et al. (2025) that large quality gaps inflate position effects, and avoids a confound in which a single lightweight model would drive an artificially steep hierarchy.

***Table 1.*** *Models used in the study, accessed via the OpenRouter.ai API. Exact version identifiers were pinned against the live API on the access date for reproducibility; all five span five developers and three geographies at comparable near-frontier capability.*

| Short name | API identifier | Developer (geography) | Type |
|---|---|---|---|
| Claude | *anthropic/claude-sonnet-5* | Anthropic (US) | Proprietary |
| GPT | *openai/gpt-5.1* | OpenAI (US) | Proprietary |
| Gemini | *google/gemini-3.8-flash* | Google (US) | Proprietary |
| Qwen | *qwen/qwen3.8-flash* | Alibaba (CN) | Open-weight |
| Mistral | *mistralai/mistral-large-2512* | Mistral AI (EU) | Open-weight |

All models are instruction-tuned, which is essential for reliable instruction-following across languages and for consistent structured output. Two of the five (Qwen and Mistral) are open-weight.

### 3.2 Language Selection

We study five European languages: **English** (en), **German** (de), **Czech** (cs), **Spanish** (es), and **Finnish** (fi). These span four language families — Germanic (en, de), Slavic (cs), Romance (es), and Uralic (fi) — and a range of morphological complexity (from analytic English to agglutinative Finnish) and digital-resource availability. We deliberately do *not* claim broad "typological diversity": all five are European and all five use the Latin script. Our sample contains no isolating or tonal language, no non-Latin writing system, and no genuinely low-resource

language. We treat the extension of this design to such languages as important future work (Section 6), because it would require native-speaker validation of both generation and evaluation quality that is outside the scope of the present, fully automated study.

## 3.3 Text Types

Three text types were selected to elicit stylistically diverse short texts (≈120 words each): (1) a **formal apology** to a customer for a shipment delay; (2) an **advertising copy** for ecological coffee; and (3) an **encyclopedic definition** of "freedom" for a children's encyclopedia. These span functional registers requiring pragmatic appropriateness, creativity, and clarity respectively. We use these three types to elicit stylistic variety, and we report results by text type only descriptively (Section 4.5); three types cannot support confirmatory claims about text-type effects, and we make none.

## 3.4 Generation Protocol

Each model generated $k = 5$ independent texts per (language × text type) cell at generation temperature 0.7, giving 5 models × 5 languages × 3 types × 5 replicas = 375 texts. Replication lets us estimate consistency of decisions within cells — impossible with a single generation per cell — and separates stable model capability from sampling variation across draws. Prompts were translated into each target language; the full set of generation prompts is provided in the released materials (see Data availability).

## 3.5 Blind Pairwise Evaluation Protocol

For each anchor model (the evaluator, which is also the author of one candidate), we compared the anchor's own text against each other model's text within each language–text-type condition, using matched replicas: replica $i$ of the anchor was compared with replica $i$ of the other model. Each such directed pair was evaluated in *both* presentation orders — once with the anchor's text first, once second — rather than at a single random position. This makes position a controlled, within-pair variable that can be subtracted directly. The design yields 5 × 4 directed author pairs × 15 conditions × 5 replicas × 2 orders = 3,000 evaluations.

Each evaluation was a fresh API call with no conversation history, ensuring the comparison was entirely blind. The system prompt instructed the model to act as an expert linguist and evaluate strictly linguistic quality along four dimensions: grammatical accuracy, lexical appropriateness, stylistic fluency, and coherence (full prompt in Appendix A). Output was restricted to a single character (*1* or *2*). All evaluation calls used temperature 0.2.

## 3.6 Data Structure

Each observation was encoded as:

$$(\text{anchor},\ \text{other},\ \text{forSelf},\ \text{wasFirst},\ \text{lang},\ \text{type},\ \text{rep})$$

where forSelf ∈ {0,1} indicates whether the anchor selected its own text, wasFirst ∈ {0,1} its (controlled) position, and rep the replica index.

## 3.7 Statistical Analysis Framework

The central challenge is distinguishing authorship bias from text quality. For each model $M$ we compute

$$\begin{aligned} \hat{p}_{\text{self}}(M) &= P(\text{forSelf} = 1 \mid \text{anchor} = M), \\ \hat{p}_{\text{other}}(M) &= P(\text{forSelf} = 0 \mid \text{other} = M), \end{aligned}$$

where the other-selection rate captures how often *other* models choose $M$'s text, serving as a quality-controlled baseline. If $\hat{p}_{\text{self}}(M) \approx \hat{p}_{\text{other}}(M)$, self-selection is explained by quality; if $\hat{p}_{\text{self}}(M) > \hat{p}_{\text{other}}(M)$, the model exhibits residual self-preference.

Two-proportion $z$-tests of $\hat{p}_{\text{self}}$ versus $\hat{p}_{\text{other}}$ provide an interpretable model-level first pass, but they ignore presentation order and treat the 3,000 evaluations as independent even though each generated text is reused across many comparisons. Our *primary* confirmatory test is therefore a conditional-logit model of the paired choice (Bradley & Terry, 1952): for each comparison we model the log-odds of choosing the first-shown text as a function of

(i) Bradley–Terry candidate strengths (one per model, capturing quality), (ii) a position term, and (iii) a self-bias term, and its interaction with the judge, which is the own-text preference *net of* candidate quality and position. Modelling the difference between the two candidates is essential: a model that scores each text without reference to its opponent mis-attributes a strong model's easier opponents to authorship (in our data it inflated one judge's odds ratio roughly tenfold). We report the self-bias odds ratio per judge with 95% confidence intervals (Bates et al., 2015).

**Equivalence testing.** A non-significant self-preference test is, by itself, only an absence of evidence. To make a positive claim, we pre-specify an equivalence margin of $\pm 5$ percentage points on the self-minus-other difference and apply two one-sided tests (TOST) (Lakens, 2017). Rejecting the TOST null lets us assert that any authorship bias is *at most* of trivial size. This also guards against the opposite failure mode: with $N = 3{,}000$, a negligible difference can become "statistically significant", so every headline claim is paired with an effect size and the equivalence bound.

**Position bias.** With both orders present, the position effect for model $M$ is the within-pair difference $P(\text{self} \mid \text{first}) - P(\text{self} \mid \text{second})$, estimated with cluster-robust intervals. Following Shi et al. (2025), we further relate the magnitude of each model's position effect to the measured quality gap between the paired candidates.

Position-bias contingency analyses use chi-square and Fisher's exact tests; the quality hierarchy uses a Bradley–Terry model (Bradley & Terry, 1952) with bootstrap confidence intervals. All tests are two-sided with $\alpha = .05$; Bonferroni correction is applied within families of per-model comparisons. Effect sizes are reported as odds ratios with 95% CIs and Cohen's φ (Agresti, 2002; Cohen, 1988).

# 4 RESULTS

## 4.1 Overall Self-Preference Test

Of 3,000 evaluations, models selected their own text in 1,649 cases (55.0%). A one-sample binomial test against $\pi = 0.50$ was significant ($z = 5.44$, $p < .001$; 95% CI $[.532, .567]$), and a two one-sided test against a $\pm 5$ pp equivalence margin did *not* establish equivalence to chance. There is thus evidence of aggregate self-preference, in contrast to our earlier low-powered pilot ($N = 300$) which used a slightly different setup, model set, and analysis and found none.

## 4.2 Quality-Controlled Self-Preference Analysis

As a descriptive first pass, Table 2 compares each model's self-selection rate with its text's selection rate when evaluated by other models. At this level, four of five models appear to select their own text more often than others select the same text (Mistral +24.2 pp, Gemini +16.2 pp, Claude +8.3 pp, Qwen +7.7 pp; GPT shows the reverse, −6.7 pp). However, this pooled comparison ignores presentation order and the reuse of texts across comparisons; the correctly specified confirmatory model below shows that two of these four effects do not survive those controls.

***Table 2.*** *Self-selection rate vs. other-selection rate (quality-controlled self-preference test), with two-proportion z-test and odds ratio. Δ in percentage points; positive = own-text preference. None of the effects are statistically equivalent to zero at a ±5 pp margin.*

| Model | $\hat{p}_{\text{self}}$ | $\hat{p}_{\text{other}}$ | $\Delta$ | OR [95% CI] | $p$ |
|---|---|---|---|---|---|
| Mistral | .502 | .260 | +24.2 | 2.86 [2.24, 3.64] | < .001 |
| Gemini | .818 | .657 | +16.2 | 2.35 [1.80, 3.07] | < .001 |
| Claude | .490 | .407 | +8.3 | 1.40 [1.12, 1.76] | .004 |
| Qwen | .493 | .417 | +7.7 | 1.36 [1.08, 1.71] | .008 |
| GPT | .445 | .512 | −6.7 | 0.77 [0.61, 0.96] | .021 |

The Pearson correlation between model-level self- and other-selection rates was $r = .705$ ($t(3) = 1.72$, $p = .184$; $R^2 = .497$) — far from perfect, consistent with a genuine authorship component that quality alone does not explain.

**Confirmatory conditional-logit analysis.** Our primary test models the paired choice directly, with Bradley–Terry candidate strengths, a position term, and a per-judge self-bias term (§3.7; Table 3). Pooled across judges, there is a

modest but robust own-text preference net of quality and position ($\mathrm{OR} = 1.26$, 95% CI [1.16,1.36], $p < .001$), alongside a strong tendency to choose the second-shown text (OR = 0.68 for the first slot, $p < .001$). Per judge, the own-text preference is robust only for **Gemini** (OR = 2.09, $p < .001$) and **Mistral** (OR = 3.20, $p < .001$); **GPT** shows a significant *anti*-self preference (OR = 0.43, $p < .001$); and **Claude** (OR = 1.17, $p = .26$) and **Qwen** (OR = 1.11, $p = .44$) show no significant own-text preference once quality, position, and the opponent are controlled. The Claude and Qwen effects visible in Table 2 are thus artefacts of the uncontrolled comparison. The estimated Bradley–Terry strengths (Gemini $+0.95$, GPT 0, Claude $-0.11$, Qwen $-0.08$, Mistral $-0.35$ in log-odds) reproduce the quality hierarchy of §4 independently.

***Table 3.*** *Confirmatory per-judge own-text preference from the conditional-logit model (Bradley–Terry strengths + position + self-bias), i.e. the odds of choosing one's own text net of candidate quality and presentation order. OR > 1 indicates own-text preference.*

| Judge | Self-bias OR [95% CI] | $p$ |
|---|---|---|
| Mistral | 3.20 [2.37, 4.32] | < .001 |
| Gemini | 2.09 [1.55, 2.83] | < .001 |
| Claude | 1.17 [0.89, 1.54] | .258 (ns) |
| Qwen | 1.11 [0.85, 1.46] | .443 (ns) |
| GPT | 0.43 [0.32, 0.56] | < .001 (anti-self) |
| **Pooled** | **1.26 [1.16, 1.36]** | **< .001** |

## 4.3 Position Bias Analysis

Pooled across models, self-selection was lower when the anchor's text appeared first (46.5%) than second (63.4%; $\chi^2(1) = 86.2$, $p < .001$, $\varphi = .17$): an aggregate recency pull toward the second slot. This pooled figure, however, masks strong and opposing model-specific effects (Table 4).

***Table 4.*** *Position bias per anchor model, as the within-pair difference between self-selection when the anchor's text appears first vs. second. Bonferroni-corrected $\alpha = .01$.*

| Model | $P_{1st}$ | $P_{2nd}$ | OR | $\varphi$ | $p$ |
|---|---|---|---|---|---|
| Mistral | .190 | .813 | 0.05 | .623 | < .001 |
| GPT | .303 | .587 | 0.31 | .285 | < .001 |
| Gemini | .863 | .773 | 1.84 | .117 | .004 |
| Claude | .440 | .540 | 0.67 | .100 | .014 |
| Qwen | .530 | .457 | 1.34 | .073 | .072 |

**Mistral** exhibits an extreme recency bias ($\varphi = .62$): it selects its own text in only 19.0% of cases when that text is shown first, but 81.3% when shown second. **GPT** shows a strong recency bias ($\varphi = .29$) and **Gemini** a mild primacy bias ($\varphi = .12$); **Claude** a mild recency effect, and **Qwen** no significant position bias after correction. Position bias is therefore the single largest distortion in the data, exceeding the authorship effects above in magnitude for several models.

**Position bias and the quality gap.** Following Shi et al. (2025), we related the magnitude of each model's position effect to the measured quality gap between paired candidates. The association is inconsistent across judges (Pearson $r$ ranging from $+.74$ for Claude to $-.57$ for Gemini; with only four opponents per judge, these values are indicative only), offering no uniform support for a simple gap-drives-position rule in this near-frontier, low-gap regime; if anything, position bias here appears largely intrinsic to the judge rather than a function of the candidate gap.

## 4.4 Text Quality Hierarchy

Table 5 reports the pairwise win matrix and normalized Bradley–Terry strengths.

***Table 5.*** *Pairwise win matrix (cell = times the row model's text was preferred over the column model's; 300 comparisons per unordered pair), overall win rate, and normalized Bradley–Terry strength.*

| | Cl | GP | Ge | Qw | Mi | Win | BT $\hat{\lambda}$ |
|---|---|---|---|---|---|---|---|
| Claude | – | 156 | 76 | 142 | 164 | .448 | .150 |
| GPT | 144 | – | 85 | 163 | 182 | .478 | .166 |
| Gemini | 224 | 215 | – | 230 | 216 | **.738** | **.412** |
| Qwen | 158 | 137 | 70 | – | 181 | .455 | .153 |
| Mistral | 136 | 118 | 84 | 119 | – | .381 | .119 |

Gemini's text is clearly preferred (win rate .738; BT $\hat{\lambda} = .41$), while the remaining four models are tightly clustered (win rates .38–.48). The hierarchy is flat — no single model dominates — confirming that the near-frontier model set keeps the quality gap moderate, a precondition for interpreting the position and authorship effects without a dominant-candidate confound.

### 4.5 Exploratory: Language and Text Type

For completeness, self-selection rates differ only modestly across languages (en .530, de .538, cs .557, es .610, fi .513) and text types (apology .536, advertisement .539, definition .574); the largest gap is between Spanish and Finnish (≈10 pp). These pooled rates are not adjusted for candidate quality or presentation order and mix judges with very different baseline selection rates (Table 2), so they cannot be read as language- or type-specific self-preference. With three text types and this design these are not confirmatory analyses, and we draw no conclusions about text-type or language-level effects.

To describe how consistent individual judges are within the design, we also summarize the repeated decisions in each of the 75 judge×language×text-type cells (40 binary decisions per cell: 4 opponents × 5 replicas × 2 orders). For each cell we computed the Bernoulli variance of the self-selection indicator, $p(1-p)$, where $p$ is the cell's self-selection rate. This quantity is a deterministic function of $p$: it equals 0 when a judge always or never selects its own text and reaches its maximum of $.25$ when the judge is split exactly 50:50. For a single cell it therefore carries no information beyond $p$ itself and is reported only descriptively. The mean across cells was $.213$. Because the mean of $p(1-p)$ equals $\bar{p}(1-\bar{p})$ minus the variance of $p$ across cells, this value can be compared with the $.2475$ expected if every cell had the overall rate $\bar{p} = .55$: the difference implies a cross-cell variance of $p$ of about $.035$ (SD ≈ .19), several times the binomial sampling variance of ≈ .005. Self-selection rates thus differ systematically between cells, partly reflecting the judge-level differences documented in Table 2, and individual decisions within a cell are rarely unanimous.

Under independence, the standard error of a single cell's self-selection rate is $\sqrt{.213/40} \approx .07$. Since each generated text is reused across several comparisons, this figure likely understates the true uncertainty. We therefore treat cell-, language- and type-level rates as exploratory, whereas the confirmatory results rest on 600 evaluations per judge and on the conditional-logit model of Section 3.7.

## 5 DISCUSSION

### 5.1 Self-Preference Bias

Our central question is whether LLM judges favor their own text beyond what shared quality perception predicts. The correctly specified confirmatory model answers this with nuance. There is a modest but robust own-text preference overall (OR = 1.26), but it is not universal: it is clear and moderate for Gemini and Mistral (OR = 2.1 and 3.2), reversed for GPT (which disfavours its own text, OR = 0.43), and absent for Claude and Qwen once candidate quality and presentation order are controlled. The larger self-versus-other gaps that a naive proportion test reports for Claude and Qwen do not survive control for position and text reuse. The own-text preference is thus genuine but modest and judge-specific — present for some judges, reversed or absent for others — and smaller than a naive proportion test would suggest.

We are deliberately cautious about the mechanism. As Wataoka et al. (2024) argue, an own-text preference may operate through familiarity rather than authorship recognition: a model's own output is by construction low-perplexity to itself, and if capable judges share preferences for fluent, distributionally central text, then $p_{\text{self}}$ and $p_{\text{other}}$ are not fully independent. Our design bounds the own-text preference *net of the cross-model quality baseline*, but it cannot separate "the model prefers its text because it recognizes it as its own" from "the model prefers its text because that text is most probable under its own distribution." We therefore report a robust own-text preference beyond shared quality, without attributing it specifically to authorship recognition. This nuance is consistent with Z.-Y. Chen et al. (2025), who trace apparent self-preference to latent stylistic alignment, and refines rather than contradicts Panickssery et al. (2024).

### 5.2 Position Bias as the Dominant Confound

Consistent with Wang et al. (2024), Li et al. (2024), and Shi et al. (2025), position bias — not authorship — is the largest single distortion in our data. Its magnitude dwarfs the authorship effects for several judges: Mistral's recency bias ($\varphi = .62$) swings its self-selection from 19% to 81% purely by position, and GPT's recency bias ($\varphi = .29$) is likewise larger than its (reversed) authorship effect. The direction is model-specific — recency for Mistral, GPT, and Claude; mild primacy for Gemini — and, in this low-gap near-frontier regime, does not track the candidate quality gap in any uniform way, suggesting it is largely intrinsic to the judge. Because order can be correlated with authorship in naive designs, position bias can masquerade as self-preference; counterbalancing both orders, as we do, separates the two, and shows that a genuine authorship effect survives once position is controlled.

### 5.3 Implications for LLM-as-Judge Deployment

1. **Own-text preference is real but judge-specific.** Some judges (here Gemini and Mistral) favor their own text beyond quality; others do not, and one disfavors it. A model should not be used to evaluate its own outputs without a quality- and position-controlled baseline, and the direction of any bias cannot be assumed.
2. **Position bias requires active mitigation**: counterbalanced orders are essential and, for some judges (e.g. Mistral), decisive.
3. **Evaluator selection matters**: models differ markedly in both position-invariance and own-text preference.
4. **Quality and self-selection are confounded in naive analyses**; studies reporting self-preference must control for quality and be adequately powered to detect modest effects.

## 6 LIMITATIONS AND FUTURE WORK

**Shared-preference confound.** As discussed, our design bounds authorship-based bias but cannot exclude familiarity/shared-bias mechanisms (Wataoka et al., 2024); testing these would require token-level log-probabilities, available only for open-weight judges.

**LLM-based quality proxy.** Quality is proxied by other LLMs' selections, not human judgment; estimates may reflect shared inductive biases. Human validation is left to future work.

**Language and script coverage.** Our five languages are all European and Latin-script. The sample contains no isolating, tonal, non-Latin, or genuinely low-resource language. Extending the design to such languages is important but requires native-speaker validation of generation and evaluation quality, since otherwise one cannot distinguish "no bias" from "the model is simply poor in that language".

**Text length and type.** All texts were ≈120 words across three types; dynamics may differ for longer texts or reasoning/factual content.

**Model versions and API intermediation.** Models were accessed via OpenRouter.ai and are updated frequently; we pin exact versions with access dates, but longitudinal replication is needed.

**Linguistic evaluation only.** Prompts excluded content dimensions; self-preference may operate differently for helpfulness or factual accuracy.

**Debiasing instruction.** The evaluation prompt explicitly instructed judges to show no positional bias (Appendix A); position-bias estimates therefore hold under this instruction and were not compared with a prompt lacking it.

## 7 CONCLUSION

Through a controlled, counterbalanced multilingual experiment ($N = 3{,}000$) involving five near-frontier models, five languages, and three text types, we test whether LLMs exhibit authorship-based self-preference in blind pairwise evaluation. Using a conditional-logit model that controls candidate quality and presentation order, we find a modest but robust own-text preference overall (OR = 1.26) that is concentrated in two of five judges (Gemini and Mistral), reversed in one (GPT), and absent in the other two — while strong, model-specific position biases remain the single largest distortion. A naive proportion test overstates the effect.

These results suggest that, for linguistic quality evaluation, both an own-text preference and presentation order require active, judge-specific control, and that apparent self-preference should be interpreted cautiously: our design establishes its presence beyond quality for some judges but cannot attribute it specifically to authorship recognition rather than familiarity.

## ADDITIONAL INFORMATION AND DECLARATIONS

**Conflict of Interests:** The authors declare no conflict of interest.

**Author Contributions:** T.B.: Conceptualization, Methodology, Software, Data curation, Writing — original draft. F.V.: Supervision, Resources, Writing — review & editing. O.S.: Supervision, Resources, Writing — review & editing.

**Statement on the Use of Artificial Intelligence Tools:** The object of study is LLM behavior; all generated texts and evaluations reported as data were produced by the models under study, as described in the Methods. Separately, the authors used an AI coding assistant (Claude Code, based on Anthropic's Claude Opus 4.8) as a tool throughout the preparation of this work: to help design and implement the Node.js data-collection and analysis pipeline, to run the automated experiment and statistical analyses under the authors' direction, and to draft and revise portions of the manuscript. The authors conceived the study, made all methodological and interpretive decisions, verified the code, analyses, and reported results, and take full responsibility for the content of the paper.

**Data Availability:** The complete dataset — all 375 generated texts, 3,000 evaluations, prompts, and analysis outputs — is openly available on Zenodo (Bruckner et al., 2026a). The analysis pipeline (implemented in Node.js using the OpenRouter.ai API) and the analysis scripts are openly available on Zenodo (Bruckner et al., 2026b).

**Reproducibility Statement:** The pipeline is implemented in Node.js and is automated, resumable, and logged. Exact model identifiers and per-token prices are pinned against the live OpenRouter API with an access date. The model set in Table 1 was pinned on 4 September 2026, and that pinned set was used for all runs reported here. All configuration (models, languages, text types, replication, temperatures) is version-controlled. The data (N=3,000), prompts, and analysis code are openly archived on Zenodo (Bruckner et al., 2026a, 2026b).

## APPENDIX A: EVALUATION SYSTEM PROMPT

The following system prompt was used for English-language evaluations; equivalent prompts were prepared for German, Czech, Spanish, and Finnish.

> You are an expert in English linguistics and stylistics. Your task is to objectively compare two texts provided in a JSON object with the fields "text1" and "text2". Treat both texts as equal candidates and show no positional bias. Evaluate strictly linguistic quality only. Do not assess topic, factual accuracy, persuasiveness, emotional impact, ideology, politeness, verbosity, complexity, authorship, intent, or audience. Use identical standards for both texts and assess: (1) grammatical accuracy; (2) lexical appropriateness and richness; (3) stylistic fluency and naturalness; (4) coherence and structural clarity. Evaluate each text independently first, then compare. Output exactly one character: "1" if text1 is better, "2" if text2 is better. Provide no explanation or additional text.

## REFERENCES


**Agresti, A.** (2002). *Categorical data analysis* (2nd ed.). Wiley.

**Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Axmed, M., Bali, K., & Sitaram, S.** (2023). MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing,* (pp. 4775–4812). ACL.

**Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Siahkoohi, A., & Baraniuk, R. G.** (2024). Self-consuming generative models go MAD. In *Proceedings of the Twelfth International Conference on Learning Representations* (pp. 1–28). ICLR.

**Bates, D., Mächler, M., Bolker, B., & Walker, S.** (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

**Bradley, R. A., & Terry, M. E.** (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4), 324–345.

**Bruckner, T., Vencovský, F., & Svatoš, O.** (2026a). Self-preference or position bias? A controlled multilingual study of evaluation bias in LLM-as-judge pairwise comparison [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.22395365

**Bruckner, T., Vencovský, F., & Svatoš, O.** (2026b). Self-preference or position bias? A controlled multilingual study of evaluation bias in LLM-as-judge pairwise comparison — software [Computer software]. *Zenodo*. https://doi.org/10.5281/zenodo.22397578

**Chen, Y.-S., Jin, J., Kuo, P.-T., Huang, C.-W., & Chen, Y.-N.** (2025). LLMs are biased evaluators but not biased for fact-centric retrieval augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, (pp. 26669–26684). ACL. https://doi.org/10.18653/v1/2025.findings-acl.1369

**Chen, Z.-Y., Wang, H., Zhang, X., Hu, E., & Lin, Y.** (2025). Beyond the surface: Measuring self-preference in LLM judgments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, (pp. 1653–1672). ACL. https://doi.org/10.18653/v1/2025.emnlp-main.86

**Chhun, C., Suchanek, F. M., & Clavel, C.** (2024). Do language models enjoy their own stories? Prompting large language models for automatic story evaluation. *Transactions of the Association for Computational Linguistics*, 12, 1122–1142. https://doi.org/10.1162/tacl_a_00689

**Chiang, C.-H., & Lee, H.** (2023). Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (pp. 15607–15631). ACL.

**Cohen, J.** (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

**Dubois, Y., Galambosi, B., Liang, P., & Hashimoto, T. B.** (2024). Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *Proceedings of the Conference on Language Modeling 2024*. COLM.

**Kim, S., Suk, J., Longpre, S., Lin, B. Y., Shin, J., Welleck, S., Neubig, G., Lee, M., Lee, K., & Seo, M.** (2024). Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, (pp. 4334–4353). ACL. https://doi.org/10.18653/v1/2024.emnlp-main.248

**Koo, R., Lee, M., Raheja, V., Park, J. I., Kim, Z. M., & Kang, D.** (2024). Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024,* (pp. 517–545). ACL. https://doi.org/10.18653/v1/2024.findings-acl.29

**Lakens, D.** (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4), 355–362. https://doi.org/10.1177/1948550617697177

**Li, Z., Wang, C., Ma, P., Wu, D., Wang, S., Gao, C., & Liu, Y.** (2024). Split and merge: Aligning position biases in large language model based evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, (pp. 11084–11108). ACL. https://doi.org/10.18653/v1/2024.emnlp-main.621

**Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C.** (2023). G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, (pp. 2511–2522). ACL. https://doi.org/10.18653/v1/2023.emnlp-main.153

**Needham, J., Edkins, G., Pimpale, G., Bartsch, H., & Hobbhahn, M.** (2025). Large language models often know when they are being evaluated. arXiv:2505.23836. https://doi.org/10.48550/arXiv.2505.23836

**Panickssery, A., Bowman, S. R., & Feng, S.** (2024). LLM evaluators recognize and favor their own generations. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024).* NIPS.

**Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C.** (2023). Direct preference optimization: Your language model is secretly a reward model. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023).* NIPS.

**Shi, L., Ma, W., Wu, C., Wang, Y., Wang, Y., Xu, J., Cheng, R., Wu, Y., Wu, Y., Xu, H., Wu, Z., & Vosoughi, S.** (2025). Judging the judges: A systematic study of position bias in LLM-as-a-judge. *Proceedings of the 2025 Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL)*, (pp. 292–314). ACL. https://doi.org/10.18653/v1/2025.ijcnlp-long.18

**Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R.** (2024). AI models collapse when trained on recursively generated data. *Nature*, 631, 755–759. https://doi.org/10.1038/s41586-024-07566-y

**Stureborg, R., Alikaniotis, D., & Suhara, Y.** (2024). Large language models are inconsistent and biased evaluators. arXiv:2405.01724. https://doi.org/10.48550/arXiv.2405.01724

**Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Kong, L., Liu, Q., Liu, T., & Sui, Z.** (2024). Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* (pp. 9440–9450). ACL. https://doi.org/10.18653/v1/2024.acl-long.511

**Wataoka, K., Takahashi, T., & Ri, R.** (2024). Self-preference bias in LLM-as-a-judge. In *NeurIPS 2024 Workshop on Safe Generative AI*, (pp. 1–12). ICLR.

**Wu, M., & Aji, A. F.** (2025). Style over substance: Evaluation biases for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics,* (pp. 297–312). ACL.

**Xu, W., Zhu, G., Zhao, X., Pan, L., Li, L., & Wang, W.** (2024). Pride and prejudice: LLM amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* (pp. 15474–15492). ACL. https://doi.org/10.18653/v1/2024.acl-long.826

**Ye, S., Kim, D., Kim, S., Hwang, H., Kim, S., Jo, Y., Thorne, J., Kim, J., & Seo, M.** (2024). FLASK: Fine-grained language model evaluation based on alignment skill sets. In *Proceedings of the Twelfth International Conference on Learning Representations*, (pp. 1–54). ICLR.

**Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I.** (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems,* (pp. 46595–46623). NIPS.

*Acta Informatica Pragensia* is published by the Prague University of Economics and Business, Czech Republic | eISSN: 1805-4951