AMCIS 2026 Reno

Structural Biases in LLM-as-a-Judge Systems: Implications for reliable IS Adoption

Full Paper

Jonathan Piaget
University of Applied Sciences and Arts Western Switzerland (HES-SO)
jonathan.piaget@he-arc.ch

Ulysse Rosselet
University of Applied Sciences and Arts Western Switzerland (HES-SO)
ulysse.rosselet@he-arc.ch

Cédric Gaspoz
University of Applied Sciences and Arts Western Switzerland (HES-SO)
cedric.gaspoz@he-arc.ch

Abstract

This study investigates the LLM-as-a-Judge paradigm as a critical Information System (IS) whose reliability and neutrality determine organizational adoption. Using dictionary definition evaluation as a methodologically neutral testbed, we conduct 8,000 blind pairwise comparisons between definitions from five established English dictionaries and four large language models, judged by the same four LLMs. Our results reveal a pronounced position bias (Definition A preferred 65.7% of the time), a massive self-preference bias (up to +73.3 percentage points when a model judges its own output), and moderate inter-judge agreement (Fleiss' kappa = 0.357). Lexical diversity analysis shows that LLM-generated definitions are shorter yet lexically comparable to human dictionaries. These structural biases constitute barriers to trust and adoption. Explicit bias measurement, diversified blind evaluation pipelines, and human-calibrated assessment are essential for reliable deployment of LLM-as-a-Judge systems in organizational contexts.

Keywords

LLM-as-a-Judge; Algorithmic bias; AI adoption; Automated decision-making; Self-preference; Lexicography

Introduction

The deployment and adoption of information systems (IS) based on artificial intelligence (AI) and large language models (LLMs) have become critical strategic concerns for organizations. These systems, particularly in decision-support or automated evaluation roles, constitute a new generation of tools whose reliability and neutrality are determining factors for successful adoption or rejection by human users. The LLM-as-a-Judge paradigm lies at the heart of this challenge, offering a scalable, fast, and cost-effective automated evaluation solution intended to replace or assist human annotation and judgment in high-stakes domains.

The LLM-as-a-Judge framework, where an LLM is used to evaluate candidate options, is increasingly relevant. Its high human-agreement rates support its adoption in complex IS including specialized applications such as applicant screening, content evaluation, and expert jury assistance in legal or medical fields (Li et al., 2025; Wang et al., 2025; Zheng et al., 2023). However, the adoption of these intelligent IS is contingent upon user-perceived trust, and this trust is directly threatened by the presence of systematic algorithmic biases.

Among the barriers to adoption and reliability, self-preference bias (favoring one's own outputs) and AI-AI bias (preferring AI-generated content over human-curated content) are particularly concerning. These biases are not mere technical artifacts; they introduce structural inequity and organizational risk that undermine the perceived legitimacy of the tool and, consequently, hinder its adoption. Simulated recruitment pipelines, for instance, have shown that applicants using the same LLM as the recruiter to generate application materials can be 23 to 60% more likely to be shortlisted for a position despite equal qualifications (Xu, J. et al., 2025).

Our work uses the LLM-as-a-Judge paradigm applied to lexicographic quality evaluation as a methodologically neutral testbed. Unlike more subjective evaluation tasks such as summarization or dialogue generation, where different humans produce varied outputs, the evaluation of dictionary definitions—which are objective and stable reference sources—allows us to isolate and measure the impact of self-preference and AI-AI bias. This objective task, which results in a single common definition regardless of the human generator, decouples bias measurement from the inherent subjectivity of other typical evaluation methods.

This article is positioned at the intersection of ethical AI research and IS adoption models. The core question we address is:

To what extent do intrinsic biases - such as AI-AI bias, self-preference and position - pose risks to reliable acceptance of the LLM-as-a-Judge paradigm in organizational context ?

Literature Review

The LLM-as-a-Judge paradigm is an automated evaluation information system (IS) that employs a LLM to assess, rank, or critique the output of human work, other systems, or other LLMs, aiming to emulate human judgment in a scalable, fast, and cost-effective manner (Murugadoss et al., 2025; Wang et al., 2025; Zheng et al., 2023). Human-agreement rates often exceeding 80% attest to its technical performance (Bavaresco et al., 2025; Li et al., 2025). Technology acceptance models such as TAM (Davis, 1989) and UTAUT (Venkatesh et al., 2016) have been widely mobilized to analyze the adoption of information systems. Their recent extension to the context of artificial intelligence shows that the adoption of AI-based systems and generative models depends not only on their utility and ease of use, but also on critical factors such as trust, transparency, data quality, and perceived reliability (Choung et al., 2022; Wanner et al., 2022). Regarding LLM-as-a-Judge, adoption is underway in high-stakes organizational domains—recruitment screening, scholarship attribution, legal and medical jury assistance—where LLMs score, rank, or summarize applicants dossiers (Gu et al., 2026; Head et al., 2023; Li et al., 2025; Sandan et al., 2025). Despite these capabilities, widespread adoption faces fundamental trust barriers: systematic biases, prompt sensitivity, and output variability generate resistance and ethical misalignment with organizational values (Dorner et al., 2025; Shi et al., 2025; Thakur et al., 2025).

Self-preference bias refers to an LLM judge systematically assigning higher scores to its own outputs or those of closely related models, beyond what genuine quality differences can explain (Koo et al., 2024; Panickssery et al., 2024; Wataoka et al., 2024). AI-AI bias generalizes this to a preference for LLM-generated content over human-generated content, risking "anti-human" discrimination in automated pipelines (Laurito et al., 2025; Xu, J. et al., 2025).

Four interacting mechanisms explain self-preference: (1) *self-recognition*—LLM evaluators recognize their own outputs with non-trivial accuracy, and this capability correlates linearly with self-preference strength (Panickssery et al., 2024); mere perceived-author labels can shift preferences regardless of content quality (Saraf et al., 2025); (2) *perplexity-based familiarity*—models assign higher ratings to texts with low perplexity (a measure of how surprising a text is to the model) relative to their own distribution, suggesting the fundamental bias is toward distributional familiarity rather than self-generation per se (Wataoka et al., 2024); (3) *cognitive shortcuts*—provenance and recency cues injected into prompts trigger systematic biases without being mentioned in justifications (Marioriyad et al., 2025); and (4) *model strength*—more powerful models show more self-preference, partly reflecting genuinely superior outputs, but they are also particularly poor at recognizing their own errors (Chen et al., 2025).

Meanwhile, AI-generated and human-written texts are increasingly hard to distinguish (Khera et al., 2025), raising the risk of hidden AI-AI bias. In hiring contexts, self-preference induces non-demographic but

consequential inequity (Laurito et al., 2025; Xu, J. et al., 2025), and if LLM judges prefer AI-generated content that is itself demographically biased, downstream systems may amplify both source-based and social biases (Fang et al., 2024).

While no single mitigation solution is fully satisfactory, proposed strategies include inference-time scaling (Chen et al., 2025), activation-based steering vectors (Roytburg et al., 2025), self-recognition reduction (Panickssery et al., 2024; Xu, J. et al., 2025), increasing benchmark data diversity (Xu, W. et al., 2026), blind evaluation protocols (Saraf et al., 2025; Spiliopoulou et al., 2025), diverse judge panels (Verga et al., 2024), and human-calibrated score normalization (Spiliopoulou et al., 2025; Zheng et al., 2023). Studies emphasize the need to continuously calibrate and validate LLM judges, integrate hybrid pipelines (LLM + human), and develop more robust metrics (Dietz et al., 2025; Li et al., 2025; Wang et al., 2025; Wei et al., 2025). The use of LLM judges in high-stakes domains must remain cautious given the potential consequences of errors or biases (Croxford et al., 2025; Hu et al., 2025).

Despite extensive research confirming the presence of systematic biases in the LLM-as-a-Judge paradigm, a critical gap remains in the ability to isolate and objectively quantify these structural threats to organizational trust and IS adoption. Current studies primarily rely on subjective evaluation tasks (e.g., summarization or dialogue), where inherent output variability makes it challenging to decouple genuine quality differences from algorithmic preference biases. Consequently, there is a lack of clear, methodologically neutral evidence to precisely measure how intrinsic factors like self-recognition or distributional familiarity translate into systematic, consequential inequity. This paper addresses this void by employing lexicographic quality evaluation as a rigorously objective testbed, allowing for the precise measurement and systematic isolation of position, self-preference, and AI-AI biases. By achieving this methodological neutrality, we aim to provide the necessary empirical grounding to inform robust deployment and mitigation strategies for LLM-as-a-Judge systems in high-stakes organizational settings.

Methodology

Lexicographic Reference Corpus

To establish a lexicographic reference corpus, we selected the 100 most frequent English words from the Corpus of Contemporary American English (COCA) frequency list (Davies, 2008). Definitions for each word were collected from the following five sources: dico1: Cambridge Dictionary; dico2: Collins English Dictionary (Complete and Unabridged); dico3: Merriam-Webster; dico4: Oxford Advanced Learner's Dictionary; dico5: Wiktionary (English edition). These sources were chosen to represent a spectrum of lexicographic traditions—prescriptive and descriptive, commercial and collaborative—ensuring that the evaluation captures diverse editorial approaches rather than a single standard. The complete set of definitions collected from the five dictionaries constitutes the corpus of 500 human-curated definitions.

LLM Definition Generation

Four LLMs were selected based on diversity and market representativeness. Two commercial models—llmb: GPT-5 (OpenAI) and llmd: Mistral Large (Mistral AI)—and two open-weight models—llma: DeepSeek R1 (DeepSeek) and llmc: Llama 3 8B (Meta)—were retained based on popularity rankings for text generation on Hugging Face (HuggingFace, 2025). For each word, each model was prompted via LiteLLM, a provider-agnostic API library. The system prompt was: *"You are a dictionary editor. For each word received, produce a dictionary definition entry. Include: part(s) of speech, word senses with definitions, and usage examples where helpful."* The word to be defined was passed directly as the user message. This process resulted in 400 LLM-generated definitions (100 words from each of the 4 models), expanding the total corpus to 900 definitions (comprising 500 human-generated and 400 LLM-generated entries).

Corpus Analysis

To assess the quality of LLM-generated definitions relative to human-curated definitions , we measured lexical diversity using Rényi entropy (a generalization of Shannon entropy that captures vocabulary richness) and the Giraud coefficient (a ratio of unique words to the square root of text length), and discourse coherence using DiscoScore, a metric built on the BERT language model that evaluates how well sentences

connect within a text (Zhao et al., 2023). These judge-independent indicators complement the pairwise judgments.

Definition length varies substantially across sources (Figure 1), yet lexical diversity remains within a narrow band (Figure 2). LLM sources are interleaved with dictionaries on the diversity axis, confirming that LLM-generated definitions are shorter on average but lexically comparable to their human-curated counterparts. DiscoScore lexical chain analysis (Figure 3) confirms that LLM definitions achieve discourse coherence comparable to or exceeding human-curated dictionaries, with llmb and llmd obtaining the highest scores (0.54 and 0.51 respectively). The corpus thus offers a representative sample of varied lexicographic styles, enabling a fair confrontation with human-curated sources.

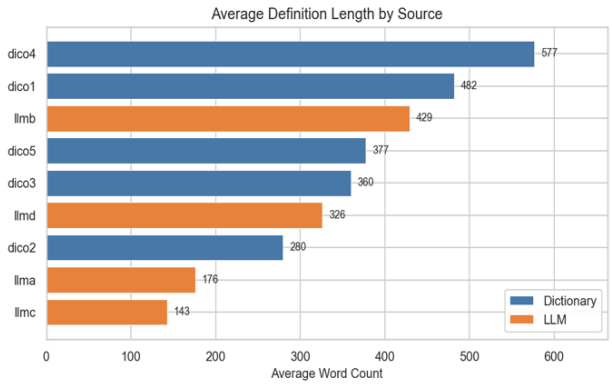

Figure 1: Average definition length by source

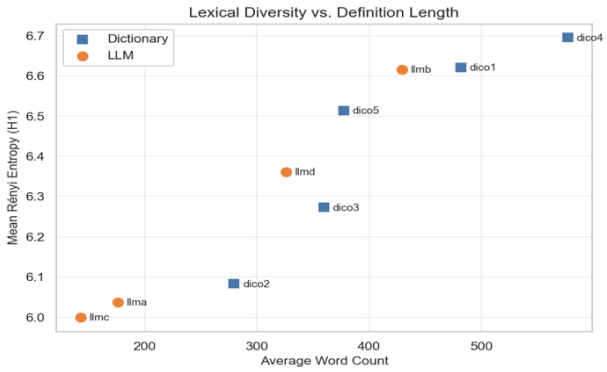

Figure 2: Lexical diversity against average definition length

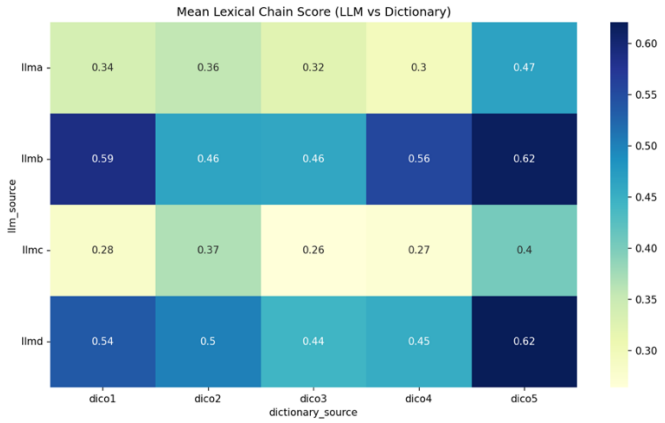

Figure 3: DiscoScore lexical chain coherence between LLM-generated definitions and human dictionary references. Higher scores indicate stronger discourse coherence.

Evaluation Protocol

We adopt a blind pairwise comparison design. Each LLM-generated definition is compared against each human-curated definition for the same word, producing a Cartesian product of 20 pairs per word (4 LLMs × 5 dictionaries), yielding 2,000 comparisons per judge model. The same four LLMs serve as judges, enabling analysis of self-preference bias. In total, 8,000 judgments were collected.

To ensure impartiality, definitions are presented anonymously as "Definition A" and "Definition B," with presentation order randomized per pair (fixed seed = 42). The judge receives a system prompt assigning it the role of a dictionary editor choosing which definition to include; no provenance information is disclosed. Two response modes are used: *simple mode* (choice only) for the full corpus, and *detailed mode* (choice + justification) for the first 10 comparisons per session. All responses are structured in JSON with schema validation.

Results and Analysis

Overall Win Rates

The overall win rate for human-curated definitions is 44.8%, meaning that in more than half of the 8,000 judgments, LLM judges preferred a synthetic definition over a dictionary entry (Table 1). The win rate represents the percentage of pairwise confrontations won by each source across all judges.

Rank	Source	Type	Win Rate
1	llmb	LLM	85.5%
2	llmd	LLM	62.5%
3	dico4	Dictionary	56.0%
4	llma	LLM	53.0%
5	dico3	Dictionary	50.0%
6	dico1	Dictionary	46.1%
7	dico2	Dictionary	40.3%
8	dico5	Dictionary	31.8%
9	llmc	LLM	19.6%

Table 1: Win rates by source.

llmb is the top-performing source, surpassing all dictionaries, underscoring its ability to produce definitions that align with the normative expectations of other LLMs (clarity, structure, examples). dico4 is the strongest source (56.0%) and the only dictionary exceeding 50%, while dico5—a collaborative, volunteer-edited resource—ranks second to last (31.8%). llmc, a smaller open-weight model, achieves only 19.6%, confirming that model capacity strongly influences definition quality as perceived by LLM judges. Notably, judge behavior also varies: as a judge, llmb is the most favorable to dictionaries (60.2% human-curated definitions win rate), while llmd is the most lenient toward synthetic productions (only 33.4% human-curated definitions wins).

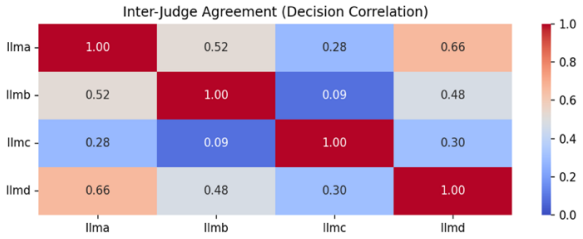

Figure 4: Inter-judge correlation matrix.

Inter-Judge Agreement

Fleiss' kappa is a statistical measure used to assess the degree of agreement among multiple raters when assigning categorical ratings, beyond what would be expected by chance. Fleiss' kappa across the four judge models is 0.357, indicating moderate agreement (Landis & Koch, 1977). The pairwise correlation matrix (Figure 4) reveals that llma and llmd share the highest correlation (r = 0.663), while llmb and llmc show nearly independent judgment patterns (r = 0.086).

Consensus Analysis

To evaluate agreement among the four judges, we aggregated their votes for each pair of definitions. A consensus was defined as at least three out of four judges selecting the same option (i.e., a 3–1 or 4–0 split). Using this criterion, the panel reaches consensus in 84.7% of the 2,000 evaluated cases, indicating

that most comparisons lead to a clear majority decision. . At the committee level, LLM definitions win by majority in 50.5% of cases versus 34.2% for dictionaries. However, this overall result hides substantial variation across sources. Table 2 shows the most significant results in terms of consensus analysis. Among human-curated sources, dico4 performs comparatively well, reaching near parity with LLMs (48.2% human consensus). In contrast, dico5 is less competitive, with consensus favoring LLM definitions in 66.8% of cases. Among the LLM-based sources, results are highly polarized. llmb shows a very strong advantage, with consensus in its favor in 90.2% of cases, suggesting consistently preferred outputs. By contrast, llmc performs poorly, with consensus favoring human definitions in 82.6% of cases.

Definition source	Consensus for Dictionary	Consensus for LLMs (%)	Split (2-2)
dico4	48.2%	37.0%	14.8%
dico5	19.8%	66.8%	13.5%
llmb (Gen)	2.8%	90.2%	7.0%
llmc (Gen)	82.6%	7.2%	10.2%

Table 2: Per-source consensus across 2,000 evaluated pairs.

Systematic Biases

Position Bias

Of the 8,000 total judgments, Definition A, which was presented first, was selected by judges 65.7% of the time, irrespective of its actual quality or source. To statistically confirm the influence of the presentation order, a chi-square test of independence was performed for each judge model. As shown in Table 3, this test was used to determine if the judges' choice was significantly correlated with the definition's position or if the choice was independent of the order.

Judge Model	Human win (if 1st)	Human win (if 2nd)	χ^2 statistic	p-value
llmc	87.2%	4.7%	1373.05	< 0.001 ***
llmd	45.6%	21.4%	131.65	< 0.001 ***
llma	50.8%	29.4%	95.20	< 0.001 ***
llmb	58.8%	61.6%	1.58	0.21 (NS)

Table 3: Chi-square independence tests for position bias per judge model.

For three of the four models, presentation order significantly influences the judge's decision ($p < 0.001$). llmc shows extreme position bias ($\chi^2 = 1373.05$): when the human-curated definition appears first, it wins 87.2% of the time; when second, only 4.7%. By contrast, llmb shows no significant position bias ($p = 0.21$), confirming it as the most impartial judge with respect to presentation order. The randomization of presentation order in our protocol ensures that this bias does not systematically favor one source type in aggregate, but it contributes substantially to inter-judge disagreement.

Self-Preference Bias

Model used for Judgement	Rejection of other LLMs in favor of own definition	Rejection of own definition in favor of dictionary	Difference (percentage points)	p-value
llmb	78.5%	5.2%	+73.3	< 0.001 ***
llmd	40.1%	13.2%	+26.9	< 0.001 ***
llma	39.5%	41.8%	-2.3	0.36 (NS)
llmc	44.4%	49.6%	-5.2	0.04 *

Table 4: Self-preference bias per judge model.

The llmb model displays a strong bias towards its own output: it favors dictionary definitions only 5.2% of the time when evaluating its own definitions, compared to a much higher 78.5% when evaluating definitions from other models. This represents a significant self-preference swing of +73.3 percentage points. llmd shows a similar but less extreme pattern (+26.9 pp). llma is the most neutral judge, with no statistically significant self-preference. Conversely, llmc shows a significant reverse bias ($p = 0.04$): it is harsher toward its own definitions than toward definitions from other models, possibly because its short outputs fail its own quality criteria when judging. These findings are consistent with the self-recognition and perplexity-based familiarity mechanisms identified in the literature (Panickssery et al., 2024; Wataoka et al., 2024). They also confirm that model sophistication does not guarantee impartiality; it may actually increase a model's ability to recognize and favor its own probabilistic signature.

AI-AI Bias

Beyond self-preference, the overall pattern of results reveals a broader AI-AI bias. Across all 8,000 judgments, LLM-generated definitions win 55.2% of pairwise confrontations against human-curated definitions (rejection of LLM-generated definition in favor of dictionary: 44.8%). Only one dictionary (dico4, 56.0%) surpasses the 50% mark. LLM judges systematically prefer the structural conventions of synthetic definitions—conciseness, standardized formatting, explicit part-of-speech labeling—over the richer but less uniform entries of human dictionaries. This preference for form over substance, where stylistic conformity to LLM norms outweighs lexicographic depth, constitutes a structural AI-AI bias that mirrors findings from recruitment pipelines (Xu et al., 2026) and binary-choice experiments (Laurito et al., 2025).

Conclusion

This study addressed a critical research gap by employing lexicographic quality evaluation as a rigorously objective testbed, allowing for the systematic isolation and quantification of structural biases that compromise automated judgment. By confronting four LLM judges with a corpus of human and synthetic definitions in 8,000 blind pairwise comparisons, we established three primary empirical and theoretical contributions.

First, we provide precise, objective quantification of fundamental systemic biases in the LLM-as-a-Judge paradigm: a pronounced position bias (65.7% preference for the first-presented option), a massive self-preference bias (up to +73.3 percentage points when a model judges its own output), and a significant AI-AI bias (55.2% overall preference for synthetic content). This empirical grounding confirms that systematic algorithmic preferences, rather than purely quality differences, are fundamental threats to reliable evaluation.

Second, theoretically, the study validates that LLM sophistication does not guarantee impartiality but may, in fact, correlate with an increased ability to recognize and favor its own probabilistic signature. This is

exemplified by the state-of-the art model (llmb), which produced the most selected definitions and which exhibited the strongest self-preference when acting as a judge. This finding is critical for understanding the "model strength" and "self-recognition" mechanisms documented in the literature.

Third, the implications for Information Systems (IS) adoption and organizational reliability are substantial. The observed biases introduce structural inequity that directly undermines user trust and the perceived legitimacy of LLM-as-a-Judge systems in high-stakes domains. The systematic preference shown for the structural conventions of synthetic definitions—conciseness and standardized formatting—over the potentially richer but less uniform entries of human dictionaries constitutes an AI-AI bias that poses a structural risk of anti-human discrimination. Consequently, the moderate inter-judge agreement (Fleiss' kappa = 0.357) reinforces the finding that no single LLM can reliably substitute for expert human judgment.

For robust and trustworthy deployment, organizations must adopt organization practices to control and mitigate the risks posed by the LLM-as-a-Judge intrinsic biases in the design of evaluation pipelines. Safe practice requires explicit bias measurement, the use of diversified blind evaluation panels (Verga et al., 2024), and continuous human-calibrated assessment. Future work should investigate whether targeted mitigation strategies can successfully reduce these structural biases (position, self-preference) without destabilizing legitimate quality judgments, and whether these findings generalize across evaluation domains beyond the objective testbed of lexicography.

Acknowledgements

This article was prepared with the assistance of generative AI tools. Claude (Anthropic) was used for code development and translation; Gemini (Google) assisted with drafting and editing. All content was reviewed and validated by the authors.

References

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., Ghaleb, E., Giulianelli, M., Hanna, M., Koller, A., Martins, A., Mondorf, P., Neplenbroek, V., Pezzelle, S., Plank, B., Schlangen, D., Suglia, A., Surikuchi, A. K., Takmaz, E., & Testoni, A. (2025). LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 238–255. https://doi.org/10.18653/v1/2025.acl-short.20

Chen, Z.-Y., Wang, H., Zhang, X., Hu, E., & Lin, Y. (2025). Beyond the Surface: Measuring Self-Preference in LLM Judgments (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2506.02592

Croxford, E., Gao, Y., First, E., Pellegrino, N., Schnier, M., Caskey, J., Oguss, M., Wills, G., Chen, G., Dligach, D., Churpek, M. M., Mayampurath, A., Liao, F., Goswami, C., Wong, K. K., Patterson, B. W., & Afshar, M. (2025). Evaluating clinical AI summaries with large language models as judges. Npj Digital Medicine, 8(1), 640. https://doi.org/10.1038/s41746-025-02005-2

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319-340.

Davies, M. (2008). The Corpus of Contemporary American English (COCA). https://www.english-corpora.org/coca/

Dietz, L., Zendel, O., Bailey, P., Clarke, C. L. A., Cotterill, E., Dalton, J., Hasibi, F., Sanderson, M., & Craswell, N. (2025). Principles and Guidelines for the Use of LLM Judges. Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR), 218–229. https://doi.org/10.1145/3731120.3744588

Dorner, F. E., Nastl, V. Y., & Hardt, M. (2025). Limits to scalable evaluation at the frontier: LLM as judge won't beat twice the data. The Thirteenth International Conference on Learning Representations. https://openreview.net/forum?id=NO6Tv6QcDs

Fang, X., Che, S., Mao, M., Zhang, H., Zhao, M., & Zhao, X. (2024). Bias of AI-generated content: An examination of news produced by large language models. Scientific Reports, 14(1), 5224. https://doi.org/10.1038/s41598-024-55686-2

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Lin, Z., Zhang, B., Ni, L., Gao, W., Wang, Y., & Guo, J. (2026). A survey on LLM-as-a-Judge. The Innovation, 101253. https://doi.org/10.1016/j.xinn.2025.101253

Head, C. B., Jasper, P., McConnachie, M., Raftree, L., & Higdon, G. (2023). Large language model applications for evaluation: Opportunities and ethical implications. New Directions for Evaluation, 2023(178–179), 33–46. https://doi.org/10.1002/ev.20556

Hu, Y., Liu, H., Chen, Q., Zheng, N., Wang, C., Liu, Y., Clarke, C. L. A., & Shen, W. (2025). J&H: Evaluating the Robustness of Large Language Models Under Knowledge-Injection Attacks in Legal Domain. Proceedings of the AAAI Conference on Artificial Intelligence, 39(27), 28106–28115. https://doi.org/10.1609/aaai.v39i27.35029

Khera, R., Pedroso, A. F., Keloth, V. K., Xu, H., Silva, G. S., & Schwamm, L. H. (2025). Scientific Writing in the Era of Large Language Models: A Computational Analysis of AI- Versus Human-Created Content. Stroke, 56(10), 3078–3083. https://doi.org/10.1161/STROKEAHA.125.051913

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 L.-W. Ku, A. Martins, & V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024 (pp. 517–545). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.29

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310

Laurito, W., Davis, B., Grietzer, P., Gavenčiak, T., Böhm, A., & Kulveit, J. (2025). AI–AI bias: Large language models favor communications generated by large language models. Proceedings of the National Academy of Sciences, 122(31), e2415697122. https://doi.org/10.1073/pnas.2415697122

Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., Bhattacharjee, A., Jiang, Y., Chen, C., Wu, T., Shu, K., Cheng, L., & Liu, H. (2025). From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2757–2791. https://doi.org/10.18653/v1/2025.emnlp-main.138

Marioriyad, A., Rohban, M. H., & Baghshah, M. S. (2025). The Silent Judge: Unacknowledged Shortcut Bias in LLM-as-a-Judge (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2509.26072

Murugadoss, B., Poelitz, C., Drosos, I., Le, V., McKenna, N., Negreanu, C. S., Parnin, C., & Sarkar, A. (2025). Evaluating the Evaluator: Measuring LLMs' Adherence to Task Evaluation Instructions. Proceedings of the AAAI Conference on Artificial Intelligence, 39(18), 19589–19597. https://doi.org/10.1609/aaai.v39i18.34157

Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM Evaluators Recognize and Favor Their Own Generations. Advances in Neural Information Processing Systems 37, 68772–68802. https://proceedings.neurips.cc/paper_files/paper/2024/hash/7f1f0218e45f5414c79c0679633e47bc-Abstract-Conference.html

Roytburg, D., Bozoukov, M., Nguyen, M., Barzdukas, J., Fu, S., & Oozeer, N. (2025). Breaking the Mirror: Activation-Based Mitigation of Self-Preference in LLM Evaluators (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2509.03647

Sandan, I. B., Dinh, T. A., & Niehues, J. (2025). Knockout LLM Assessment: Using Large Language Models for Evaluations through Iterative Pairwise Comparisons. In O. Arviv, M. Clinciu, K. Dhole, R. Dror, S. Gehrmann, E. Habba, I. Itzhak, S. Mille, Y. Perlitz, E. Santus, J. Sedoc, M. Shmueli Scheuer, G. Stanovsky, & O. Tafjord (Eds.), Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²) (pp. 121–128). Association for Computational Linguistics. https://aclanthology.org/2025.gem-1.10/

Saraf, M., Boroujeni, S. R., Beaudry, J., Abedi, H., & Bush, T. (2025). Quantifying Label-Induced Bias in Large Language Model Self- and Cross-Evaluations (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2508.21164

Shi, L., Ma, C., Liang, W., Diao, X., Ma, W., & Vosoughi, S. (2025). Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge. In K. Inui, S. Sakti, H. Wang, D. F. Wong, P. Bhattacharyya, B. Banerjee, A. Ekbal, T. Chakraborty, & D. P. Singh (Eds.), Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (pp. 292–314). The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. https://aclanthology.org/2025.ijcnlp-long.18/

Spiliopoulou, E., Fogliato, R., Burnsky, H., Soliman, T., Ma, J., Horwood, G., & Ballesteros, M. (2025). Play Favorites: A Statistical Method to Measure Self-Bias in LLM-as-a-Judge (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2508.06709

HuggingFace, (2025, March 27) , *Text Generation Models*. https://huggingface.co/models?pipeline_tag=text-generation&sort=likes

Thakur, A. S., Choudhary, K., Ramayapally, V. S., Vaidyanathan, S., & Hupkes, D. (2025). Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. In O. Arviv, M. Clinciu, K. Dhole, R. Dror, S. Gehrmann, E. Habba, I. Itzhak, S. Mille, Y. Perlitz, E. Santus, J. Sedoc, M. Shmueli Scheuer, G. Stanovsky, & O. Tafjord (Eds.), Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²) (pp. 404–430). Association for Computational Linguistics. https://aclanthology.org/2025.gem-1.33/

Venkatesh, V., Thong, J., & Xu, X. (2016). Unified Theory of Acceptance and Use of Technology : A Synthesis and the Road Ahead. *Journal of the Association for Information Systems*, *17*(5), 328-376. https://doi.org/10.17705/1jais.00428

Verga, P., Hofstatter, S., Althammer, S., Su, Y., Piktus, A., Arkhangorodsky, A., Xu, M., White, N., & Lewis, P. (2024). Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2404.18796

Wang, R., Guo, J., Gao, C., Fan, G., Chong, C. Y., & Xia, X. (2025). Can LLMs Replace Human Evaluators? An Empirical Study of LLM-as-a-Judge in Software Engineering. Proceedings of the ACM on Software Engineering, 2(ISSTA), 1955–1977. https://doi.org/10.1145/3728963

Wataoka, K., Takahashi, T., & Ri, R. (2024, October 12). Self-Preference Bias in LLM-as-a-Judge. Neurips Safe Generative AI Workshop 2024. https://openreview.net/forum?id=tLZZZIgPJX

Wei, H., He, S., Xia, T., Liu, F., Wong, A., Lin, J., & Han, M. (2025, March 6). Systematic Evaluation of LLM-as-a-Judge in LLM Alignment Tasks: Explainable Metrics and Diverse Prompt Templates. ICLR 2025 Workshop on Building Trust in Language Models and Applications. https://openreview.net/forum?id=CAgBCSt8gL

Xu, J., Li, G., & Jiang, J. Y. (2025). AI Self-preferencing in Algorithmic Hiring: Empirical Evidence and Insights (arXiv:2509.00462). arXiv. https://doi.org/10.48550/arXiv.2509.00462

Xu, W., Agrawal, S., Zouhar, V., Freitag, M., & Deutsch, D. (2026). Deconstructing Self-Bias in LLM-generated Translation Benchmarks (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2509.26600

Zhao, W., Strube, M., & Eger, S. (2023). DiscoScore: Evaluating Text Generation with BERT and Discourse Coherence. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 3865–3883. https://doi.org/10.18653/v1/2023.eacl-main.278

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. Advances in Neural Information Processing Systems 36, 46595–46623. https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html