International Journal of Computer Applications (0975 - 8887)
Volume 187 - No.121, June 2026

Quantifying Label-Induced Bias in Large Language Model Self and Cross Evaluations

Muskan Saraf
Data Science Department
Actual Reality Technologies
Toledo, USA

Sajjad Rezvani Boroujeni
Data Science Department
Actual Reality Technologies
Toledo, USA

Justin Beaudry
Software Department
Actual Reality Technologies
Toledo, USA

Hossein Abedi
Data Science Department
Actual Reality Technologies
Toledo, USA

Tom Bush
CEO
Actual Reality Technologies
Maumee, USA

ABSTRACT

Large language models (LLMs) are increasingly relied upon to evaluate text quality in research, industry, and automated content workflows. However, their judgments may not be as objective as assumed. This study systematically examined whether LLMs exhibit bias when assessing text attributed to different model "authors." Blog posts were generated by three leading LLMs, ChatGPT, Gemini, and Claude, and each model evaluated every post under three conditions: with no author label, with a correct author label, and with deliberately incorrect author labels. The results reveal substantial bias driven by perceived authorship rather than actual content quality. Posts labeled as "Claude," regardless of who produced them, consistently received elevated scores, while posts labeled as "Gemini" were systematically downgraded. In many cases, false author labels not only shifted absolute scores but reversed preference rankings entirely, with swings as large as 50 percentage points. Additional behavioral patterns emerged: Gemini tended to be unusually harsh when evaluating its own work, whereas Claude tended to rate its own writing more favorably. These effects appeared not only in overall preferences but also across detailed quality dimensions such as coherence, informativeness, and conciseness. Taken together, the findings indicate that LLM evaluation is highly sensitive to author attribution cues and may be influenced by implicit reputational priors associated with model identities. The results suggest that evaluators do not consistently separate content quality from perceived authorship, leading to systematic score inflation for some labels and penalties for others. These observations call into question the reliability of LLM-based assessment methods commonly used for benchmarking, content moderation, and automated review pipelines. To mitigate these risks, future evaluation frameworks should incorporate blind assessment protocols, multi-model consensus scoring, and statistical safeguards designed to detect label-induced bias.

General Terms

Artificial Intelligence, Evaluation, Bias, Human Factors

Keywords

Large Language Models, AI Evaluation Bias, Label Effects, Cross-Model Evaluation, Benchmarking Fairness

1. INTRODUCTION

Large language models (LLMs) such as ChatGPT, Gemini, and Claude are increasingly deployed not only for content generation but also for content evaluation. This dual role raises a critical question: can LLMs evaluate outputs impartially, or are their judgments influenced by perceived authorship? Previous studies have shown that both humans and models exhibit systematic bias, often favoring certain sources or stylistic patterns [1, 2]. When evaluators are aware of the source, their ratings may be shaped by prior expectations—a phenomenon known as source bias [3]. In LLMs, this bias may manifest as self-preference bias, where a model rates its own outputs higher, or label-induced bias, where a model's name affects evaluation regardless of quality [4].

The present study investigates these biases by analyzing how three leading LLMs—ChatGPT-4o, Gemini 2.5 Flash, and Claude Sonnet 4—evaluate blog posts authored by themselves and each other under four controlled conditions: no labels, true labels, and two false-label scenarios. Two complementary scoring approaches are employed: percentage-based preference scoring and point-based quality scoring for Coherence, Informativeness, and Conciseness, with the latter converted to percentages for direct comparison.

The findings reveal striking asymmetries. The "Claude" label consistently boosts scores regardless of content, while the "Gemini" label consistently depresses them. False labels produce swings of up

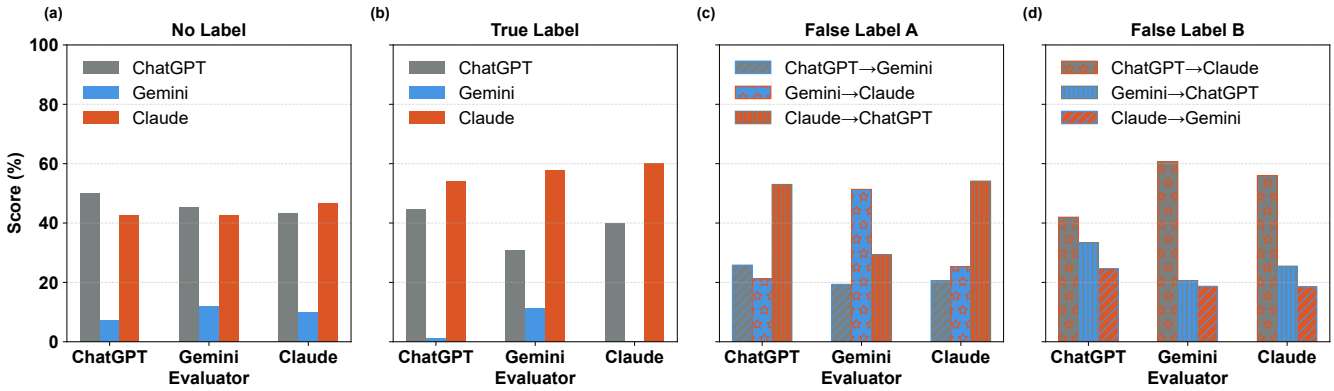

Fig. 1. Percentage-based overall scores for ChatGPT, Gemini, and Claude models evaluated under four conditions: (a) No Label, (b) True Labels, (c) False Labels – Scenario 1 (ChatGPT labeled as Gemini, Gemini as Claude, Claude as ChatGPT), and (d) False Labels – Scenario 2 (ChatGPT labeled as Claude, Gemini as ChatGPT, Claude as Gemini).

to 50 percentage points in preference scores and up to 12 percentage points in quality ratings. This work provides: (i) a controlled, multi-condition analysis of self- and cross-model evaluation bias, (ii) quantitative evidence comparing label effects across preference and quality dimensions, and (iii) recommendations for mitigating bias through blind or multi-model evaluation protocols.
The paper is organized as follows: Section II reviews related work; Section III describes the methodology; Section IV presents the results; Section V discusses implications for LLM benchmarking; and Section VI concludes with recommendations for future research.

2. RELATED WORK

Bias in automated language model evaluation has garnered growing attention in recent years. Research on self-preference bias shows that LLMs favor their own outputs, with models demonstrating measurable self-recognition capabilities that correlate with stronger self-favoritism [5, 6]. Complementary work on label-induced evaluation bias reveals how LLMs may be swayed by perceived authorship regardless of content quality. Wang et al. demonstrate that systematic bias based on response position can manipulate rankings, even making weaker models outperform stronger ones under certain prompt orderings [7]. Chen et al. further investigate whether self-preference reflects genuine superiority or signaling bias, finding that harmful bias persists even in stronger models [8]. Researchers have also examined implicit versus explicit evaluation dynamics, revealing inconsistencies in how models consciously versus unconsciously express bias [9]. Similar concerns emerge across deep learning domains [10], where pre-trained [12] and architecture-specific models [13] despite high accuracy, often inherit systematic biases from training regimes and structural choices [14, 15]. These examples underscore a broader challenge: deep learning systems across language and vision domains are susceptible to implicit and structural biases that influence evaluation. Broader surveys categorize bias into intrinsic and extrinsic types and emphasize mitigation strategies across data, model, and output layers [11]. These reviews underscore the relevance of our dual-method approach, which examines both overall preference and fine-grained quality criteria across controlled labeling experiments.

3. METHODOLOGY

This study investigates label-induced and self-preference bias in LLM evaluations using a controlled, multi-model, multi-condition design involving three stages: blog generation, evaluation under manipulated label conditions, and dual-method scoring analysis. Three LLMs-ChatGPT-4o, Gemini 2.5 Flash, and Claude Sonnet 4-generated blog posts using a fixed prompt template: "You are a professional blog writer. Write a concise blog post (around 200 words) for the title 'insert your title here'. The style should be engaging and suitable for an online audience. Return only the blog content, no extra text." Ten distinct titles covering diverse topics with similar complexity were used, with each model generating one blog per title, yielding 30 blog posts total.
Each model then evaluated all blogs, including its own, under four labeling conditions: no labels (no author attribution), true labels (correct attribution), False Label Scenario 1 (ChatGPT labeled as Gemini, Gemini as Claude, Claude as ChatGPT), and False Label Scenario 2 (ChatGPT labeled as Claude, Gemini as ChatGPT, Claude as Gemini). Two scoring systems were used: a percentage-based preference score and a point-based quality score for Coherence, Informativeness, and Conciseness, converted to percentages for direct comparison. Analyses were performed at three levels: intra-condition, cross-condition, and metric-specific evaluations. To capture how label manipulations altered judgments relative to their original scores, condition-to-condition differences (Δ-values) were computed using simple score subtraction,

$$\Delta = S_{\text{target}} - S_{\text{baseline}}.$$

This allows the study to quantify label-induced shifts with precision and compare the magnitude of bias across evaluators and label mappings.

4. RESULTS

The results are presented in percentage terms for both evaluation formats: (1) overall preference votes and (2) point-based ratings converted to percentages, enabling direct comparison of label effects across evaluation methods.
Fig. 1 summarizes the evaluation patterns across the four label conditions: no label, true label, False Label Scenario 1 (ChatGPT labeled as Gemini, Gemini as Claude, Claude as ChatGPT), and False

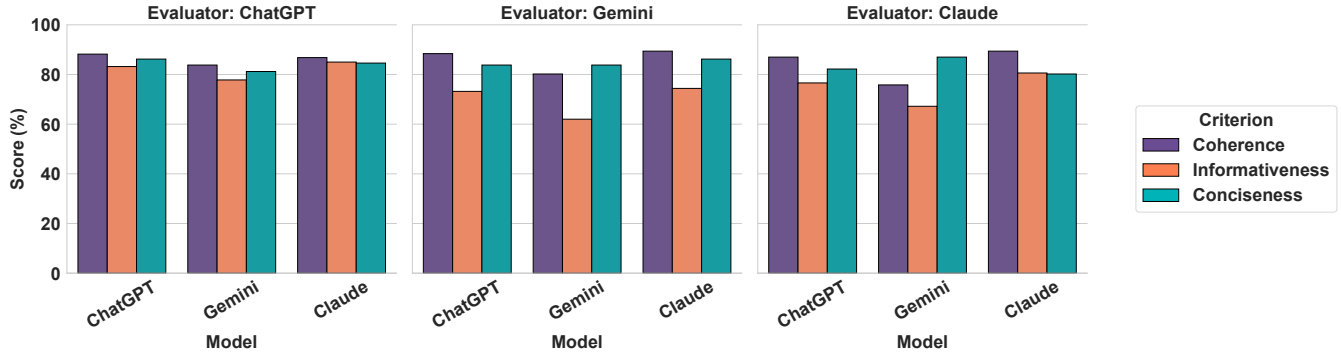

Fig. 2. Point-based scores under the *No Label* condition.

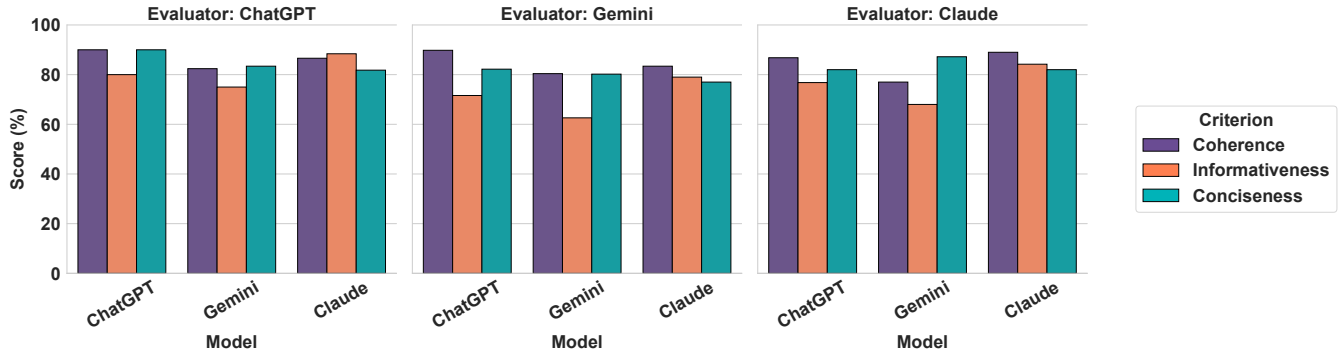

Fig. 3. Point-based scores under the *True Label* condition.

Label Scenario 2 (ChatGPT labeled as Claude, Gemini as ChatGPT, Claude as Gemini). In the no-label condition, each model showed mild self-preference, with ChatGPT selecting its own outputs 50 percent of the time, Gemini 45.3 percent, and Claude 46.7 percent. Cross-model evaluations also showed consistent undervaluation of Gemini, which received only 7–12 percent of preference votes from ChatGPT and Claude. Introducing true labels amplified these tendencies. Claude's self-preference increased to 60 percent, and its outputs were preferred by all evaluators, receiving 54–60 percent of votes. Gemini's scores collapsed under true labels, receiving 0 percent from Claude, 1.34 percent from ChatGPT, and only 11.32 percent from itself. ChatGPT's self-preference under true labels was moderate at 44.66 percent, but it sharply penalized Gemini's labeled outputs. The false-label conditions produced the strongest label-driven distortions. Under False Label Scenario 1 (ChatGPT labeled as Gemini, Gemini as Claude, Claude as ChatGPT), evaluators favored content they believed to be their own: Gemini's preference for content mislabeled as Claude rose to 51.35 percent, and Claude's preference for content mislabeled as ChatGPT reached 54.15 percent. Under False Label Scenario 2 (ChatGPT labeled as Claude, Gemini as ChatGPT, Claude as Gemini), the "Claude" label consistently drew the highest ratings, while the "Gemini" label produced severe penalties for example, Claude-authored content mislabeled as Gemini fell from 60 percent (true label) to 18.48 percent. Across all four conditions, three trends were stable: the Claude label attracted strong positive bias, the Gemini label consistently depressed scores, and label identity strongly influenced evaluations regardless of actual content quality.

In Fig. 2, under the no-label condition, all three evaluators rated Claude highest on coherence (86–89%) while ChatGPT and Gemini received moderately lower but comparable scores. Informativeness displayed the strongest variation: Gemini consistently ranked lowest (62–78%), while Claude and ChatGPT scored higher and more uniformly across evaluators. Conciseness remained the most stable metric, showing only small differences across models (generally 81–86%). When true labels were introduced, as shown in Fig. 3, these patterns intensified. Claude's outputs received the highest scores from all evaluators, with informativeness rising to 79–88%. In contrast, Gemini experienced a marked drop in ratings. Claude scored Gemini near 0% in preference comparisons and gave it substantially lower informativeness scores, while ChatGPT's informativeness for its own content dropped significantly relative to the no-label condition. Overall, true labels magnified existing preferences: Claude gained the strongest boost, Gemini suffered the largest penalty, and informativeness remained the metric most sensitive to label visibility.

Fig. 4 and 5 present the false-label conditions, where outputs were intentionally mislabeled to test the strength of label-driven bias. These conditions produced the most dramatic distortions. Under False Label Scenario 1 (ChatGPT labeled as Gemini, Gemini as Claude, Claude as ChatGPT), evaluators tended to favor outputs they *believed* to be their own, regardless of true authorship. For example, Gemini's preference for text mislabeled as "Claude" increased sharply from 11.32% in the true-label condition to 51.35% and Claude's ratings for content mislabeled as "ChatGPT" rose to 54.15%. Similarly, informativeness scores increased by 8–10 points whenever a model encountered content labeled as its own. Under

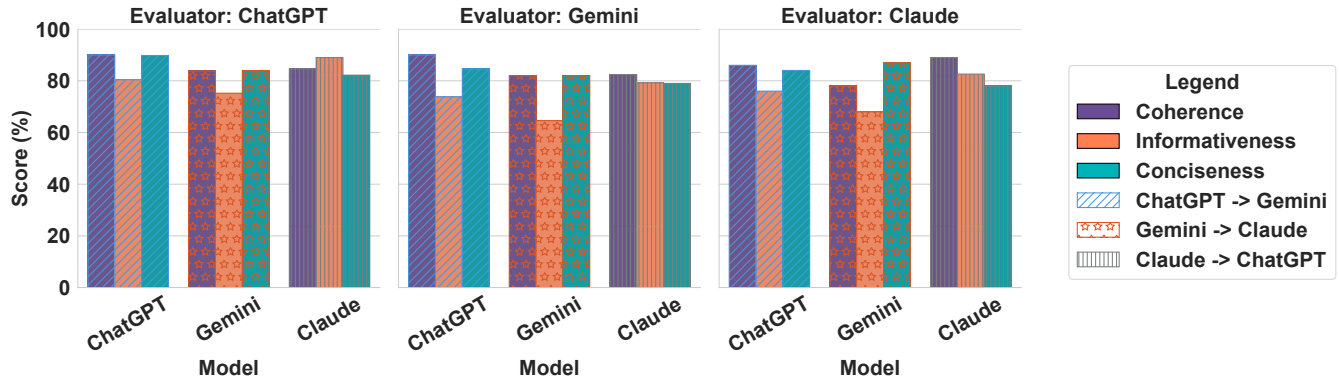

Fig. 4. Point-based scores under *False Label Scenario 1*.

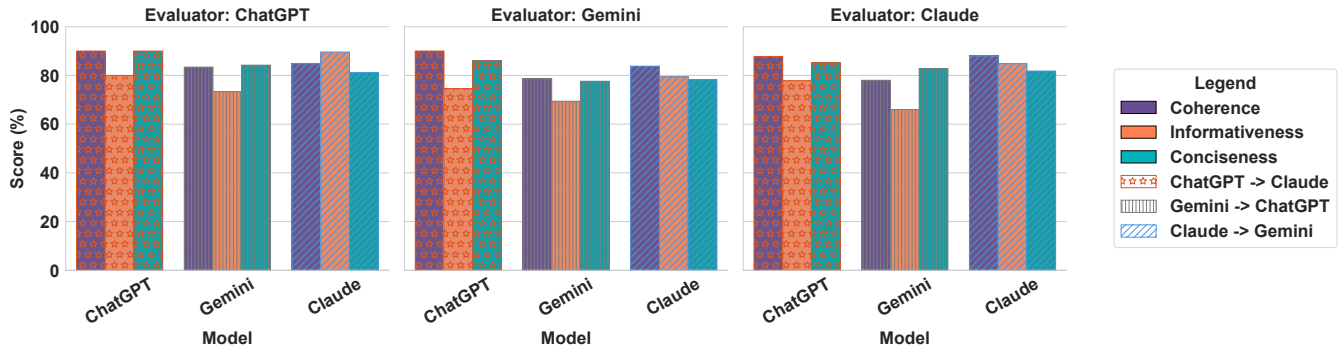

Fig. 5. Point-based scores under *False Label Scenario 2*.

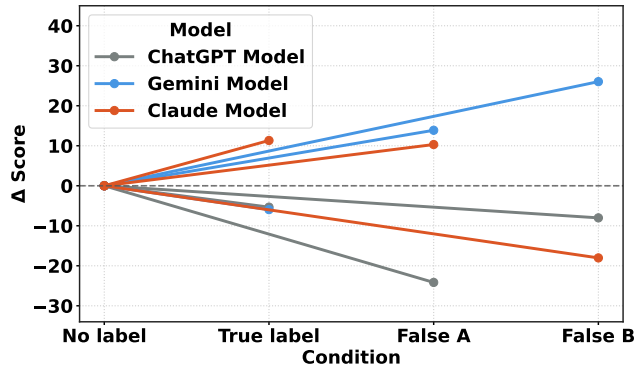

Fig. 6. Change of judgment relative to No Label baseline for evaluator *ChatGPT*.

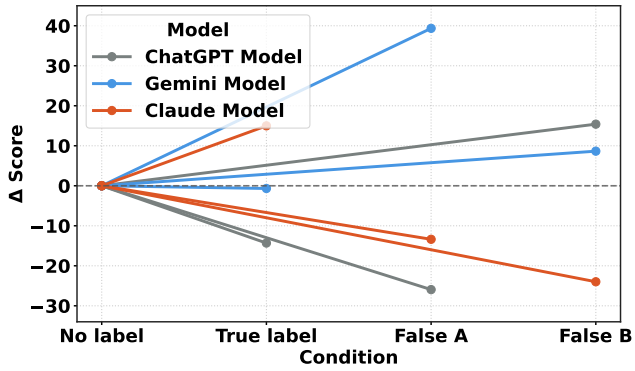

Fig. 7. Change of judgment relative to No Label baseline for evaluator *Gemini*.

False Label Scenario 2 (ChatGPT labeled as Claude, Gemini as ChatGPT, Claude as Gemini) as displayed in Fig. 5, the label effects became even stronger: the "Claude" label consistently received the highest coherence and informativeness scores (often 85–90%), while anything carrying a "Gemini" label was heavily penalized, with Claude-as-Gemini dropping from 60% to only 18.48%. These shifts confirm that label identity, not underlying content, dominates quality judgments when labels are misleading, and that the "Claude" label functions as a strong positive cue while the "Gemini" label reliably depresses evaluations across all criteria.

Fig. 6–8 illustrate the label-induced score shifts (Δ-scores) across conditions, beginning at the no-label baseline and showing how evaluations diverge once labels are introduced. Gemini exhibits the strongest positive swings: its Δ-scores rise by roughly 10–15 points under true labels and increase further under False Label A, reaching peaks of 15–20 points, especially when Gemini-authored content is falsely labeled as Claude. These trends correspond closely to several of the largest positive shifts in Table I, including the +40.03 change from true labels to False Label A and the +39.35 shift from no-label to False Label A for the Gemini→Claude map-

Table 1. Top 10 Largest Label-Induced Evaluation Shifts Across Conditions

Transition	Evaluator	Label Mapping	Δ Value
True → False-B	Claude	Claude → Gemini	−41.52
False-A → False-B	Gemini	ChatGPT → Gemini ChatGPT → Claude	+41.35
True → False-A	Gemini	Gemini → Claude	+40.03
No → False-A	Gemini	Gemini → Claude	+39.35
True → False-B	Gemini	Claude → Gemini	−38.97
False-A → False-B	Claude	Claude → ChatGPT Claude → Gemini	−35.67
False-A → False-B	Claude	ChatGPT → Gemini ChatGPT → Claude	+35.50
True → False-B	ChatGPT	Gemini → ChatGPT	+31.99
False-A → False-B	Gemini	Gemini → ChatGPT Gemini → Claude	−30.70
True → False-B	Gemini	ChatGPT → Claude	+29.72

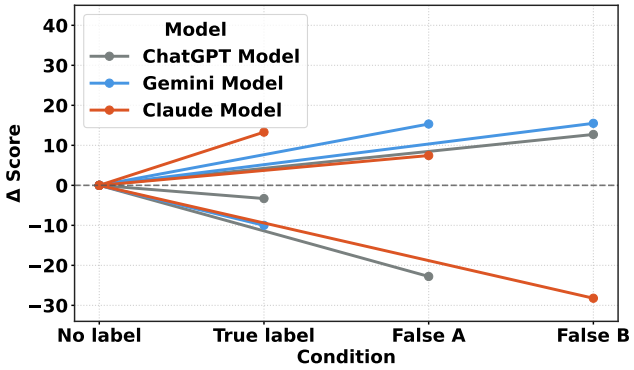

Fig. 8. Change of judgment relative to No Label baseline for evaluator *Claude*.

ping. In contrast, ChatGPT's evaluations display the steepest declines. As shown in Fig. 6–8, ChatGPT's Δ-score drops by about -15 under true labels and falls to around -25 under False Label A, remaining negative through False Label B except when artificially boosted by mislabeling. Claude shows a different pattern: it rises by approximately 12 points under true labels and remains positive under False Label A, but undergoes the most severe penalty under False Label B. As shown in Figures 6–8, Claude's Δ-score falls by more than -30 when its own outputs are mislabeled as Gemini, matching the largest negative entry in Table 1, the -41.52 shift for the Claude→Gemini mapping in the true→False-B transition. Together, Figures 6–8 and Table 1 show that mislabeling produces the most extreme distortions, with the Gemini label generating the strongest penalties, the Claude label producing the largest boosts, and ChatGPT's evaluations varying most sharply across conditions.

5. DISCUSSION

The findings show that label identity strongly shaped evaluations across all models, often overshadowing the underlying content. The *Claude* label consistently acted as a positive signal, while the *Gemini* label reliably depressed scores, even when the text was identical. These effects appeared not only in the manipulated-label conditions but also in the no-label baseline, suggesting that evaluators may carry pre-existing reputational priors learned during training or alignment.

The false-label scenarios provided the clearest evidence of identity-driven bias: evaluators frequently reversed their preferences when authorship was swapped, with shifts exceeding 30–40 percentage points. Biases were most pronounced in subjective metrics, especially informativeness, while more structural qualities like conciseness remained relatively stable. This suggests that subjective judgments are particularly vulnerable to label cues, whereas objectively constrained dimensions show greater resistance to bias.

Each model exhibited a distinct bias pattern. Claude showed strong self-preference unless mislabeled as Gemini, where it faced large penalties; Gemini undervalued its own work but boosted anything labeled Claude; and ChatGPT consistently penalized Gemini-labeled content. These divergent patterns indicate that label bias is not random, but reflects model-specific internal hierarchies. In general, the results show that LLM evaluators do not consistently separate authorship signals from content quality, raising concerns about the reliability of open-label benchmarking.

The observed preference for Claude-labeled content may reflect reputational priors embedded within the evaluator models. Al-

though the evaluators were instructed to judge the content itself, the consistent score inflation associated with the Claude label suggests that model identity served as a heuristic signal during evaluation. Since modern LLMs are trained on large-scale web corpora that frequently contain comparisons, rankings, and discussions of model capabilities, labels may carry implicit associations regarding expected quality. As a result, evaluators may unconsciously incorporate these associations into their judgments, even when the underlying text remains unchanged.

The persistent penalty associated with the Gemini label provides complementary evidence of label-induced bias. Across multiple evaluators and evaluation conditions, content carrying the Gemini label was frequently assigned lower preference and quality scores regardless of its actual authorship. This finding suggests that negative attribution effects can be as influential as positive ones. The asymmetry between the Claude and Gemini labels further indicates that LLM evaluators do not merely exhibit generic label sensitivity but may maintain internal hierarchies regarding perceived model competence. Such hierarchies can distort benchmarking outcomes by rewarding or penalizing content based on reputation rather than merit.

The results also reveal distinct self-evaluation behaviors among the models. Claude frequently demonstrated self-preference, assigning higher scores to content attributed to itself, whereas Gemini often evaluated its own outputs more critically. These findings align with recent studies on self-preference bias in LLM-as-a-judge frameworks, which report that models are often capable of recognizing stylistic characteristics associated with their own generations and may systematically favor them during evaluation. Consistent with the observations of Panickssery et al. [5], Wataoka et al. [6], Wang et al. [7], and Chen et al. [8], the present results indicate that evaluator judgments can be influenced by factors unrelated to content quality. However, our findings extend prior work by demonstrating that false author labels alone can substantially alter rankings and, in some cases, completely reverse evaluation outcomes.

These observations have important implications for benchmark design and automated evaluation systems. Open-label evaluation protocols may unintentionally introduce systematic bias when model identities are disclosed. As a result, benchmark rankings derived from such evaluations may overestimate or underestimate model performance. To improve reliability, future benchmarking frameworks should incorporate blind evaluation procedures, multi-model consensus scoring, and statistical bias detection mechanisms that identify abnormal score shifts caused by author attribution. Such safeguards would help ensure that evaluations reflect genuine content quality rather than reputational effects associated with model labels.

6. CONCLUSION

This study demonstrates that LLM evaluations are strongly shaped by perceived authorship, with label identity often overriding the actual content. The Claude label consistently inflated scores, boosting preference rates by up to +40 percentage points in false-label conditions, while the Gemini label produced the steepest penalties, in some cases reducing evaluations by more than 30–40 percentage points when identical text was mislabeled. False labels were frequently sufficient to completely reverse model rankings, confirming that identity cues, not content, were driving many judgments. Subjective metrics such as informativeness showed the largest distortions, shifting by 8–12 percentage points, whereas more structural dimensions like conciseness remained comparatively stable. These findings underscore the need for blind evaluation protocols and multi-model consensus systems to counteract inherited reputational priors. Robust benchmarking must ensure that models are judged on what they generate, not on the name attached to the output. Future work should investigate whether label-induced bias persists across a broader range of tasks, including question answering, summarization, reasoning, coding, and multimodal evaluations. Additional studies involving open-source and proprietary models, such as Llama, DeepSeek, Grok, and future generations of ChatGPT, Gemini, and Claude, would help determine the generality of the observed effects. Another promising direction is the development of bias-aware evaluation frameworks that combine blind assessment, multi-model consensus scoring, and statistical bias correction techniques to improve the reliability of LLM-as-a-judge systems.

7. REFERENCES

[1] C. van der Lee, A. Gatt, E. van Miltenburg, S. Wubben, and E. Krahmer, "Best practices for the human evaluation of automatically generated text," in *Proc. 12th Int. Conf. Natural Language Generation*, Tokyo, Japan, Oct.–Nov. 2019, pp. 355–368. doi: 10.18653/v1/W19-8643.

[2] S. Gehrmann, H. Strobelt, and A. M. Rush, "GLTR: Statistical detection and visualization of generated text," *arXiv preprint* arXiv:1906.04043, Jun. 2019. doi: 10.48550/arXiv.1906.04043.

[3] D. Wilson and D. Sperber, "Truthfulness and relevance," *Mind*, vol. 111, no. 443, pp. 583–632, Jul. 2002. doi: 10.1093/mind/111.443.583.

[4] E. Perez *et al.*, "Discovering language model behaviors with model-written evaluations," *arXiv preprint* arXiv:2212.09251, Dec. 2022. doi: 10.48550/arXiv.2212.09251.

[5] A. Panickssery, S. R. Bowman, and S. Feng, "LLM evaluators recognize and favor their own generations," *arXiv preprint* arXiv:2404.13076, Apr. 2024. doi: 10.48550/arXiv.2404.13076.

[6] K. Wataoka, T. Takahashi, and R. Ri, "Self-preference bias in LLM-as-a-judge," *arXiv preprint* arXiv:2410.21819, Oct. 2024. doi: 10.48550/arXiv.2410.21819.

[7] P. Wang *et al.*, "Large language models are not fair evaluators," *arXiv preprint* arXiv:2305.17926, Aug. 2023. doi: 10.48550/arXiv.2305.17926.

[8] W.-L. Chen, Z. Wei, X. Zhu, S. Feng, and Y. Meng, "Do LLM evaluators prefer themselves for a reason?," *arXiv preprint* arXiv:2504.03846, Apr. 2025. doi: 10.48550/arXiv.2504.03846.

[9] Y. Zhao, B. Wang, Y. Wang, D. Zhao, X. Jin, J. Zhang, R. He, and Y. Hou, "A comparative study of explicit and implicit gender biases in large language models via self-evaluation," in *Proc. 2024 Joint Int. Conf. Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italy, May 2024, pp. 186–198.

[10] M. Saraf, A. R. Kulkarni, and M. Niamat, "Detecting Hardware Trojans: Deep Learning Solutions Combining PUF Metrics and Side-Channel Observations," in *Proc. 2025 1st Int. Conf. Secure IoT, Assured Trusted Comput. (SATC)*, Dayton, OH, USA, 2025, pp. 1–5. doi: 10.1109/SATC65530.2025.11137155.

[11] Y. Guo *et al.*, "Bias in large language models: Origin, evaluation, and mitigation," *arXiv preprint* arXiv:2411.10915, Nov. 2024. doi: 10.48550/arXiv.2411.10915.

[12] S. Rezvani Boroujeni, H. Abedi, and T. Bush, "Enhancing Glass Defect Detection with Diffusion Models: Addressing Imbalanced Datasets in Manufacturing Quality Control," *Computer and Decision Making (COMDEM)*, vol. 2, no. 1, pp. xx–xx, 2025. doi: 10.59543/comdem.v2i.14391.

[13] J. Yang, W. Cui, Y. Tao, and T. Shi, "CLNSO: A Knowledge-Aware Recommendation Algorithm Based on Comparative Learning and Negative Sample Optimization," Engineering Letters, vol. 33, no. 10, pp. 4108–4118, 2025.

[14] A. Golkarieh et al., "Breakthroughs in Brain Tumor Detection: Leveraging Deep Learning and Transfer Learning for MRI-Based Classification," Computational Demography, vol. 2, no. 1, pp. xx–xx,2024. doi: 10.59543/comdem.v2i.14243.

[15] A. Golkarieh, K. Kiashemshaki, S. R. Boroujeni, and N. A. Isakan, "Advanced U-Net Architectures with CNN Backbones for Automated Lung Cancer Detection and Segmentation in Chest CT Images," arXiv preprint arXiv:2507.09898, Jul. 2025. doi: 10.48550/arXiv.2507.09898.