Style Bias Generalizes to Quantized 7B Open-Weight LLM Judges: Replication and Extension of Soumik (2026)

June 29, 2026

Abstract

Soumik [2026] demonstrate strong style bias in LLM-as-a-Judge pipelines across five frontier and mid-tier models (bias scores 0.76–0.92), but explicitly note that "additional open-weight families (Mistral, Qwen) may exhibit different bias profiles." We directly test this limitation by running the same 50-pair STYLE evaluation on two 7B open-weight models—Qwen2.5-7B-Instruct and Mistral-7B-Instruct-v0.3—at matched temperature ($T = 0.1$) under 4-bit NF4 quantization on consumer-grade hardware. Within the limits of a 50-pair replication, we find no evidence that style bias disappears in small open-weight models. Both models show substantial markdown preference (Qwen: 0.680, 95% CI [0.480, 0.860]; Mistral: 0.860, 95% CI [0.760, 0.940]), statistically indistinguishable from the paper's reported range. Notably, Mistral selected plain prose zero times across 50 pairs at both temperatures tested. These results are consistent with the hypothesis that markdown preference in LLM judges is not specific to frontier scale or API-mediated inference, though larger studies across more model families would be needed to establish this more firmly.

1 Introduction

LLM-as-a-Judge has become a dominant paradigm for scalable evaluation of language model outputs [Zheng et al., 2024, Li et al., 2024, Chiang et al., 2024]. A growing body of work documents systematic biases in these judges, including position bias [Wang et al., 2024], verbosity bias [Saito et al., 2023, Wu and Aji, 2024], and self-preference [Panickssery et al., 2024].

Soumik [2026] present the most comprehensive systematic study to date, comparing nine debiasing strategies across five models from four provider families. Their central finding is that *style bias*—the tendency to prefer markdown-formatted responses over identical plain-prose content—is the dominant bias, with scores of 0.76–0.92 across all tested models. This substantially exceeds position bias ($\leq$0.04) and most verbosity effects.

However, the paper's experimental scope is limited to hosted frontier and mid-tier models (Gemini 2.5 Pro, Claude Sonnet 4, GPT-4o, Gemini 2.5 Flash, Llama 3.3-70B via Vertex AI). The authors explicitly flag in their limitations that "additional open-weight families (Mistral, Qwen) may exhibit different bias profiles" [Soumik, 2026, §5.5].

This gap matters practically. Many production and research pipelines use smaller open-weight models as judges—either for cost reasons, data-privacy requirements, or reproducibility. If style bias is a function of scale, instruction-tuning approach, or API-mediated inference, it might not appear in locally-deployed 7B models.

We designed a targeted attack on this limitation: if we ran the same STYLE evaluation on Qwen2.5-7B-Instruct and Mistral-7B-Instruct-v0.3 under realistic local deployment conditions (4-bit NF4 quantization, $T = 0.1$ to match the paper), would the bias collapse?

Within the limits of a 50-pair evaluation at matched temperature, we find no evidence that the bias collapses. Both models show substantial markdown preference, statistically indistinguishable from the paper's range, extending the finding to a new size regime, two new model families, and a substantially different inference stack.

2 Background and Target Paper

Soumik [2026] introduce a 225-pair controlled dataset for isolating individual bias types in LLM judges. For style bias, 50 pairs present identical content in two formats: markdown (response A, with headers, bullet points, and code blocks where appropriate) and plain prose (response B, same information rendered without any markdown). An unbiased judge should call a tie on every pair. The *bias score* is defined as $P(\text{prefers A}) - P(\text{prefers B})$; a score near 1.0 indicates near-universal markdown preference.

Across five models, the paper reports bias scores of 0.76–0.92, with position bias ($\leq$0.04) negligible by comparison. The paper uses temperature 0.1, hosted API access, and models ranging from 70B (Llama 3.3-70B) to frontier scale.

The paper's own limitation hedge (§5.5) is our attack target: the claim is implicitly that style bias *might not* generalize to Mistral/Qwen families. We test whether it does.

3 Method

3.1 Models

We evaluate two locally-deployed open-weight models:

- **Qwen2.5-7B-Instruct** (Alibaba/Qwen team): a 7B instruction-tuned model from a model family and provider entirely absent from the original paper.
- **Mistral-7B-Instruct-v0.3** (Mistral AI): a 7B instruction-tuned model from a second independent family also absent from the original paper.

Both models are loaded via HuggingFace Transformers with 4-bit NF4 quantization (bitsandbytes), run on a single T4 GPU.

3.2 Dataset and Prompt

We use the identical 50 STYLE pairs from Soumik [2026]: each pair presents the same answer in markdown (A) and plain prose (B) format, with an expected verdict of tie for an unbiased judge. We use the same `NAIVE_JUDGE_PROMPT` from the paper's released codebase[1], requesting a JSON verdict with fields `verdict` (A/B/tie), `score_a`, `score_b`, and `reasoning`.

3.3 Deviations from the Original Protocol

Table 1 documents all deviations from the paper's protocol.

Temperature is matched to the paper at $T = 0.1$. The remaining confounds—model size and quantization—cannot be disentangled from model-family effects in our design and are disclosed as such. We additionally ran both models at $T = 0$ (greedy) in a prior session; the style bias results were nearly identical, and the temperature comparison surfaces an additional finding discussed in Section 5.2.

3.4 Statistical Analysis

Following Soumik [2026], bias scores are computed as $P(\text{prefers markdown}) - P(\text{prefers plain})$. Bootstrap 95% CIs use $n_{\text{boot}} = 2000$. We additionally report a one-sided binomial test against $H_0 : P(\text{prefers markdown}) \geq 0.76$ (the paper's minimum), framed as: can we reject the claim that our models are at least as biased as the paper's weakest model?

[1] https://github.com/sksoumik/llm-as-judge

Table 1: Protocol differences between Soumik [2026] and our replication.

Parameter	Paper	Ours
Temperature	0.1	0.1
Model size	70B–frontier	7B (both models)
Quantization	None (API/hosted)	4-bit NF4 (bitsandbytes)
Access method	API / Vertex AI	Local HuggingFace inference
Verdict parsing	Structured JSON (Gemini); free-text fallback (others)	Free-text JSON parsing
Dataset	Same 50 STYLE pairs	Same 50 STYLE pairs
Prompt	`NAIVE_JUDGE_PROMPT`	Same
Bias score formula	$P(A) - P(B)$	$P(A) - P(B)$
Bootstrap CI	$n_{\text{boot}} = 2000$	$n_{\text{boot}} = 2000$

4 Results

4.1 Style Bias Scores

Full verdict distributions and bias scores are reported in Table 2.

Table 2: Style bias results for 7B open-weight models on 50 STYLE pairs at $T = 0.1$. Paper values are visual reads from Figure 1 of Soumik [2026]; per-model exact values are not tabled in the paper. Only the range 0.76–0.92 is quoted in text.

Model	Source	MD wins	Plain wins	Ties	Bias score	95% CI
Gemini 2.5 Pro	Paper	—	—	—	≈0.92	N/A (Fig. 1)
Claude Sonnet 4	Paper	—	—	—	≈0.88	N/A (Fig. 1)
GPT-4o	Paper	—	—	—	≈0.84	N/A (Fig. 1)
Llama 3.3-70B	Paper	—	—	—	≈0.80	N/A (Fig. 1)
Gemini 2.5 Flash	Paper	—	—	—	≈0.76	N/A (Fig. 1)
Qwen2.5-7B-Instruct	Ours	41/50	7/50	2/50	**0.680**	[0.480, 0.860]
Mistral-7B-Instruct-v0.3	Ours	43/50	**0/50**	7/50	**0.860**	[0.760, 0.940]

Qwen2.5-7B-Instruct. Qwen shows a bias score of 0.680, with 41 of 50 pairs preferring markdown. The point estimate falls below the paper's stated minimum of 0.76, but the wide CI [0.480, 0.860] substantially overlaps the paper's range. The one-sided binomial test against $H_0 : P \geq 0.76$ yields $p = 0.8794$ (fail to reject). With $n = 50$, this difference is statistically indistinguishable from the paper's claim, though a larger sample might reveal a genuine gap.

Mistral-7B-Instruct-v0.3. Mistral shows a bias score of 0.860, within the paper's reported range. Notably, Mistral selected plain prose zero times across all 50 pairs (0/50 plain wins), calling either markdown (43/50) or a tie (7/50). The CI [0.760, 0.940] lies almost entirely within the paper's range, and $p = 0.9720$ (fail to reject H_0). While this point estimate is high, the sample size of 50 precludes strong conclusions about how Mistral compares to individual models in the paper.

4.2 Qualitative Evidence of the Bias Mechanism

Smoke-test inspection of model reasoning confirms the bias mechanism described in Soumik [2026, §5.5]. On a representative STYLE pair (the $\sqrt{2}$ irrationality question):

- **Qwen** preferred markdown (verdict A, score_a=8.5, score_b=7.5), citing "more detailed and structured explanation" and "structured presentation"— referencing formatting structure, not content.
- **Mistral** preferred markdown (verdict A, score_a=10, score_b=8), citing "clear and concise explanation" and "lacks the detailed explanation"— again, structurally motivated despite identical content.

Both models rationalized their markdown preference in terms that could reflect either genuine readability assessment or formatting bias—consistent with the paper's own discussion of this ambiguity.

4.3 Statistical Conclusion

The attack hypothesis—that style bias collapses for small open-weight 7B models—is not supported at $n = 50$. Neither model produces a result distinguishable from the paper's range. We make no stronger claim than this: within the limits of this replication, we find no evidence that style bias disappears in locally-deployed, quantized 7B models.

5 Discussion

5.1 Practical Implications

Our results are consistent with Soumik [2026]'s main finding extending to smaller open-weight models, though this should be interpreted cautiously given the sample size and two-model scope. Practitioners using small open-weight models as judges should not assume style bias is absent at smaller scale; our results provide no evidence that it is, and replication across more 7B families would be needed to establish this more firmly.

The Mistral result (0/50 plain wins) is notable, though it should be read as a descriptive observation on $n = 50$ rather than a precise estimate of Mistral's bias magnitude.

5.2 An Exploratory Position Bias Observation

As a post-hoc sanity check, we ran Mistral-7B on a subset of the paper's POSITION pairs (identical responses in both slots; an unbiased judge should call a tie every time). The paper reports position bias ≤ 0.04 across all five models at $T = 0.1$.

At $T = 0$ (greedy), Mistral produced a position bias score of 0.875 on 8 valid pairs (2 parse errors excluded from 10). At $T = 0.1$ (matched temperature), the score was 0.889 on 9 valid pairs (1 parse error excluded). In both runs Mistral consistently selected slot A; we verified responses are byte-for-byte identical (`response_a == response_b: True`), ruling out a data issue.

We flag this as an exploratory observation rather than a finding. With only 8–9 valid pairs and a single model, we cannot draw reliable conclusions. The result is directionally interesting—suggesting that position bias may not have been eliminated in smaller instruction-tuned models the way it appears to have been in the frontier models tested by Soumik [2026]—but a dedicated study with adequate sample size across multiple 7B families would be required to say anything definitive. We report it here in the interest of transparency and to motivate future work.

5.3 Limitations

Quantization effects. 4-bit NF4 quantization may alter the model's internal representations in ways that affect borderline verdict decisions. Whether unquantized 7B models show identical bias is untested.

Sample size. $n = 50$ pairs provides wide confidence intervals. Qwen's point estimate (0.680) is below the paper's minimum (0.76), and a larger sample might reveal a genuine gap. We make no stronger claim than statistical indistinguishability at this sample size.

Position bias parse errors. The position bias check produced 1–2 parse errors per run across 10 pairs (9 and 8 valid pairs respectively across the two temperature conditions). This limits the precision of the position bias estimate, though the directional finding is consistent across both runs.

Single attack surface. We test only the STYLE bias claim. Whether position bias and verbosity bias generalize similarly to 7B models are separate questions, though the position bias result here suggests the picture may differ from the paper's findings.

6 Conclusion

We tested whether style bias in LLM judges—reported at 0.76–0.92 across five frontier models by Soumik [2026]—is absent in small (7B), quantized (4-bit NF4), locally-run models at matched temperature ($T = 0.1$). Within the limits of a 50-pair replication across two model families, we find no evidence that it is. Qwen2.5-7B-Instruct (bias score 0.680, CI [0.480, 0.860]) and Mistral-7B-Instruct-v0.3 (bias score 0.860, CI [0.760, 0.940]) both show substantial markdown preference statistically indistinguishable from the paper's range. Mistral selected plain prose zero times across 50 pairs at both temperatures tested.

These results are consistent with the paper's finding but should not be over-read: two model families at $n = 50$ is a limited basis for strong generalization claims. What we can say is that the paper's own limitation hedge—that Mistral and Qwen "may exhibit different bias profiles"—finds no support here. A post-hoc observation on a small number of POSITION pairs also suggests that position bias behavior in 7B models may differ from the negligible levels reported at frontier scale, though this requires dedicated follow-up.

Practitioners using small open-weight judges should treat style bias as a live concern until larger-scale evidence suggests otherwise, and should consider the formatting normalization and debiasing strategies discussed in Soumik [2026].

Acknowledgements

The author thanks the creators of the original study for releasing their dataset, evaluation protocol, and code, which enabled this replication.

References

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, et al. Chatbot arena: An open platform for evaluating LLMs by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

Xuechen Li, Tianyi Zhang, Yann Dubois, et al. AlpacaEval: An automatic evaluator for instruction-following language models. *GitHub*, 2024.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*, 2024.

Keita Saito, Akifumi Watanabe, Jenia Jitsev, and Yutaka Uchida. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*, 2023.

Sadman Kabir Soumik. Judging the judges: A systematic evaluation of bias mitigation strategies in LLM-as-a-judge pipelines. *arXiv preprint arXiv:2604.23178*, 2026.

Peiyi Wang, Lei Li, Liang Chen, et al. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2024.

Minghao Wu and Alham Fikri Aji. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2024.