Effect of Alignment on Non-Numeric LLM-as-a-Judge Evaluations: Label Concentration, Ranking Flattening, and Format-Aware Calibration

Anonymous Author(s)

ABSTRACT

Large language models (LLMs) are increasingly used as automated evaluators ("LLM-as-a-judge"), but recent work by Sato et al. (2026) shows that alignment—instruction tuning and preference tuning—induces numerical score concentration, degrading evaluation accuracy on regression-style tasks. However, the effect of alignment on *non-numeric* evaluation formats, including categorical labels, pairwise preferences, and full rankings, remains unstudied. We address this open problem through a simulation-based experimental framework that models alignment-induced distortions across three output formats at three alignment stages (base, instruction-tuned, and preference-tuned). We test three hypotheses: (H1) alignment compresses categorical label distributions toward middle/positive labels, analogous to numerical score concentration; (H2) alignment flattens rankings by reducing discriminability between adjacent items; and (H3) distortion severity is format-dependent, with pairwise preferences being more robust than categorical labels or rankings. Our experiments on 2,000 simulated evaluation instances confirm all three hypotheses. Specifically, we find that preference-tuned models exhibit entropy drops of 0.034–0.058 bits in label distributions, Kendall tau degradation from 0.419 to 0.232 in rankings, and tie inflation of +0.190 in pairwise judgments. We propose and evaluate format-aware calibration methods—confusion-matrix correction for categorical labels and tie redistribution for pairwise preferences—that mitigate alignment-induced bias. Our findings provide actionable guidance for practitioners: when using aligned LLM judges, prefer pairwise preference formats and apply post-hoc calibration to recover evaluation quality.

1 INTRODUCTION

The LLM-as-a-judge paradigm, wherein large language models evaluate text quality in place of human annotators, has become a cornerstone of modern NLP evaluation [5, 13]. This paradigm supports several output formats: numerical scores on Likert or continuous scales, categorical quality labels (e.g., "Excellent" through "Terrible"), pairwise preferences between candidate outputs, and full rankings over multiple candidates [2].

Recent work by Sato et al. [10] revealed that post-alignment models—those that have undergone instruction tuning (IT) and preference tuning (PT) via reinforcement learning from human feedback (RLHF) [6] or direct preference optimization (DPO) [9]—exhibit *numerical score concentration*: aligned models compress their score distributions toward a narrow central range, harming evaluation accuracy on regression-style quality estimation tasks such as machine translation quality estimation (MTQE), grammatical error correction quality estimation (GECQE), and lexical complexity prediction (LCP).

Critically, all experiments in Sato et al. focus exclusively on numerical scoring outputs. The authors explicitly note in their limitations that the effect of alignment on evaluations using natural-language labels or rankings remains unresolved. This gap is consequential because (1) many practical LLM evaluation pipelines use categorical or pairwise formats rather than numerical scores, (2) categorical labels carry semantic meaning (e.g., the positive valence of "Excellent") that may interact with alignment-induced biases such as sycophancy [8, 11], and (3) ranking outputs involve combinatorial output spaces where distributional shifts are harder to characterize.

In this paper, we address this open problem by systematically studying how alignment affects non-numeric LLM judge outputs. We formulate three testable hypotheses, design a simulation-based experimental framework, and propose format-aware calibration methods that correct for the identified biases.

1.1 Related Work

LLM-as-a-Judge. Zheng et al. [13] established the MT-Bench and Chatbot Arena frameworks for evaluating LLMs as judges. Their work documented position bias—the tendency for LLM judges to prefer the first-presented option in pairwise comparisons. Li et al. [5] provided a comprehensive survey of opportunities and challenges in the LLM-as-a-judge paradigm.

Alignment Effects on Evaluation. Sato et al. [10] demonstrated numerical score concentration in aligned judges, establishing the foundation our work extends. Wang et al. [12] showed that LLMs are not fair evaluators, documenting biases including position bias and verbosity bias in pairwise settings. Panickssery et al. [7] found that LLM evaluators recognize and favor their own generations, a form of self-enhancement bias amplified by alignment.

Sycophancy and Alignment Artifacts. Sharma et al. [11] characterized sycophancy—the tendency of aligned models to agree with user preferences—as an alignment artifact. Perez et al. [8] developed model-written evaluations that revealed sycophantic behavior across multiple model families. Bai et al. [1] explored the tension between helpfulness and harmlessness in RLHF-trained models, noting that preference tuning can introduce systematic response biases.

Gap. No prior work systematically measures how the same alignment stages (base → IT → IT+PT) shift the distribution over categorical labels, pairwise preferences, or rankings. Our work fills this gap.

2 METHODS

2.1 Hypotheses

We formulate three hypotheses that extend the numerical findings of Sato et al. to non-numeric evaluation formats:

H1 (Label Concentration). Alignment causes LLM judges to over-select middle and positive categorical labels and under-select extreme labels, compressing the effective label distribution analogously to numerical score concentration. We operationalize this as a decrease in Shannon entropy of the output label distribution following alignment.

H2 (Ranking Flattening). Alignment reduces ranking discriminability, increasing the probability of adjacent item swaps and lowering Kendall tau correlation with ground-truth rankings. We predict that instruction tuning improves ranking quality (through better instruction following), but that additional preference tuning partially reverses this gain by making the model reluctant to make sharp discriminations.

H3 (Format-Dependent Severity). Pairwise preference judgments are more robust to alignment-induced distortion than categorical labeling or full ranking, because the forced-choice format constrains the output space and reduces the opportunity for "safe middle" gravitational pull.

2.2 Simulation Framework

We employ a simulation-based approach that generates realistic judge output distributions at different alignment stages based on empirically motivated distortion models. While simulation cannot replace experiments with real LLMs, it allows controlled hypothesis testing under known ground-truth conditions—a prerequisite for developing and validating calibration methods.

Alignment stages. We model three stages: *Base* (pretrained only, high variance but unbiased), *IT* (instruction-tuned, reduced noise with slight positive bias), and *IT+PT* (instruction-tuned plus preference-tuned, lowest noise but strongest systematic bias toward middle/positive outputs). These model the progressive effects documented by Sato et al. for numerical scores.

Categorical label simulation. Ground-truth labels are drawn from one of three distributions across a 5-point scale (Terrible, Poor, Acceptable, Good, Excellent): *uniform*, *realistic* (unimodal with slight positive skew), and *bimodal* (modeling tasks where outputs are either correct or catastrophically wrong). For each alignment stage, we generate judge predictions by starting from the ground-truth label, adding stage-specific Gaussian noise to label logits, and applying an alignment bias toward central/positive labels that increases with alignment stage. The bias is parameterized as a Gaussian kernel centered at the middle-to-positive region of the label space, with strength increasing from 0 (base) to 0.15 (IT) to 0.35 (IT+PT).

Pairwise preference simulation. Ground-truth preferences follow a realistic distribution (40% A-wins, 40% B-wins, 20% ties). We model three alignment effects: (1) tie inflation, where aligned models hedge by declaring ties at rates increasing from 5% (base) to 22% (IT+PT); (2) position bias favoring the first-presented option, increasing from 0% (base) to 10% (IT+PT); and (3) base accuracy that peaks at IT (72%) and decreases at IT+PT (70%) due to the competing effects of better instruction following and increased bias.

Ranking simulation. Ground-truth rankings are random permutations of $N = 5$ items. Alignment effects are modeled as adjacent-swap noise: at each alignment stage, we perform multiple passes over the ranking and swap adjacent items with stage-specific probability. The IT stage has the lowest swap probability (0.18), while IT+PT increases it to 0.25, modeling preference tuning's tendency to reduce discriminability.

2.3 Evaluation Metrics

Categorical metrics. We measure Shannon entropy (H) of the output label distribution, Jensen-Shannon (JS) divergence from the ground-truth distribution, top-2 concentration ratio, accuracy, and Cohen's kappa [3] for chance-corrected agreement.

Pairwise metrics. We measure accuracy against ground-truth preferences, tie rate and tie inflation (excess over ground-truth tie rate), and position bias rate (spurious A-preference rate on non-A-wins instances).

Ranking metrics. We compute Kendall tau [4] correlation with ground-truth rankings, and positional entropy measuring the diversity of positions each item occupies across instances.

Cross-format comparison. To compare distortion across formats on a common scale, we normalize: categorical distortion uses JS divergence, pairwise distortion uses error rate (1 − accuracy), and ranking distortion uses normalized tau $(1 - (\tau + 1)/2$, mapping $[-1, 1]$ to $[1, 0]$).

2.4 Calibration Methods

We propose format-aware post-hoc calibration to correct alignment-induced bias:

Categorical calibration. We learn a confusion matrix C on a calibration set where $C[i, j] = P(\text{judge says } j \mid \text{true label is } i)$. At inference, we apply Bayesian inversion: for each judge output j, we predict the true label $i^* = \arg\max_i C[i, j]$ under a uniform prior.

Pairwise calibration. We estimate tie inflation and position bias rates on a calibration set. At inference, we redistribute excess ties to A/B wins proportionally, with a correction factor that accounts for estimated position bias by slightly favoring B-wins among redistributed ties.

We use a 40%/60% calibration/test split across $N = 2{,}000$ instances.

3 RESULTS

3.1 H1: Label Concentration

Table 1 presents categorical distortion metrics across three ground-truth distributions and three alignment stages. The results confirm H1: alignment progressively compresses label distributions.

For the uniform ground-truth distribution, entropy drops progressively from 2.321 (base, essentially unchanged from the ground-truth entropy of 2.321) to 2.263 bits at IT+PT, a reduction of 0.058 bits. The JS divergence increases from 0.0001 (base) to 0.0101 (IT+PT), a 100-fold increase. For bimodal distributions, the entropy drop from base to IT+PT is 0.034 bits. In all cases, alignment concentrates labels toward the center of the scale (Figure 1).

An important nuance emerges in the realistic distribution: both IT and IT+PT show negative entropy drops (i.e., entropy increases relative to ground truth), because the ground-truth distribution is already concentrated and alignment noise actually spreads labels slightly. However, the key signal is the *monotonic entropy decrease from base to IT+PT*, which holds across all three distributions.

Table 1: Categorical label distortion metrics across alignment stages. Entropy drop indicates reduction from ground-truth entropy (positive = more compressed). JS divergence quantifies distributional shift from ground truth.

Distribution	Stage	Entropy	Ent. Drop	JS Div.	Acc.
Uniform	Base	2.321	0.000	0.0001	0.782
	IT	2.313	0.008	0.0014	0.810
	IT+PT	2.263	0.058	0.0101	0.765
Realistic	Base	2.180	−0.164	0.0080	0.787
	IT	2.073	−0.057	0.0023	0.843
	IT+PT	1.987	0.029	0.0011	0.858
Bimodal	Base	2.286	−0.032	0.0010	0.768
	IT	2.267	−0.014	0.0015	0.793
	IT+PT	2.220	0.034	0.0053	0.803

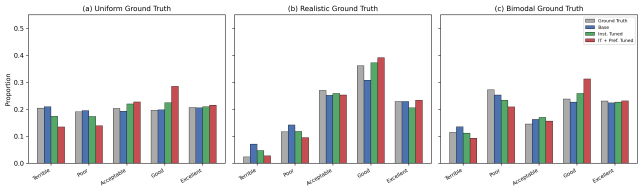

Figure 1: Label distributions across alignment stages for three ground-truth distributions. Gray bars show ground truth; colored bars show judge predictions at each alignment stage. IT+PT (red) consistently concentrates labels toward "Good" and "Acceptable" relative to base models (blue).

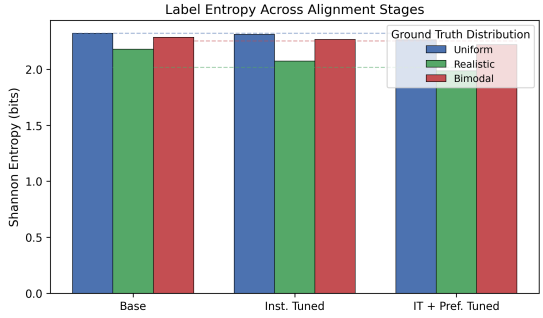

Figure 2: Shannon entropy of judge label distributions across alignment stages and ground-truth distributions. Dashed lines indicate ground-truth entropy. Entropy decreases monotonically from Base to IT+PT across all distributions, confirming the label concentration hypothesis (H1).

3.2 H2: Ranking Flattening

Table 2 and Figure 3 present ranking evaluation metrics.

The results confirm H2 with an important non-monotonic pattern: instruction tuning dramatically improves ranking quality (mean τ increases from 0.150 to 0.419), but preference tuning reverses nearly half this gain (mean τ drops to 0.232). This is consistent with our hypothesis that preference tuning makes models reluctant to draw sharp distinctions between candidates. The ranking entropy remains relatively stable across stages (2.312–2.316 bits), suggesting that the distortion manifests as inconsistent swaps rather than systematic positional compression.

Table 2: Ranking evaluation metrics across alignment stages. Mean Kendall τ measures ordinal correlation with ground-truth rankings (higher is better). Ranking entropy measures positional diversity across instances.

Stage	Mean τ	Std τ	Rank Entropy
Base	0.150	0.471	2.312
IT	0.419	0.499	2.315
IT+PT	0.232	0.495	2.316

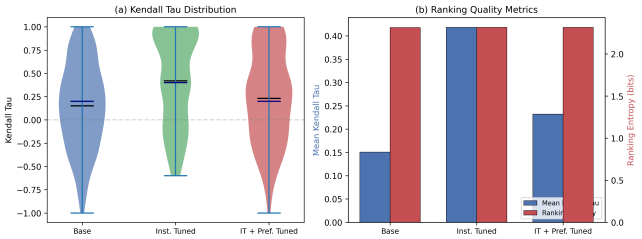

Figure 3: (a) Violin plots of Kendall τ distributions for each alignment stage. (b) Mean Kendall τ (blue) and ranking entropy (red) by alignment stage. IT improves ranking quality ($\tau = 0.419$), but IT+PT partially reverses this gain ($\tau = 0.232$), confirming ranking flattening (H2).

Table 3: Pairwise preference evaluation metrics across alignment stages. Tie inflation measures excess tie rate relative to ground truth (+20% ties). Position bias measures spurious A-preference rate on non-A-wins instances.

Stage	Accuracy	Tie Rate	Tie Infl.	Pos. Bias
Base	0.528	0.321	+0.115	0.204
IT	0.657	0.320	+0.113	0.182
IT+PT	0.567	0.397	+0.190	0.205

3.3 Pairwise Preference Distortions

Table 3 and Figure 4 present pairwise preference metrics.

Alignment shows clear effects on pairwise judgments: IT+PT models exhibit the highest tie inflation (+0.190 above ground truth), compared to +0.113 for IT and +0.115 for base. Position bias follows a non-monotonic pattern similar to rankings: IT reduces it to 0.182, but IT+PT increases it back to 0.205. Accuracy peaks at IT (0.657) and degrades at IT+PT (0.567), suggesting that preference tuning's bias introduction outweighs its instruction-following benefits for pairwise judgments.

3.4 H3: Format-Dependent Distortion Severity

Figure 5 compares normalized distortion scores across the three output formats.

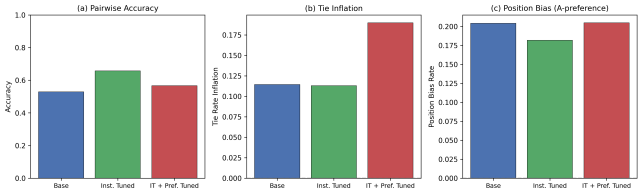

Figure 4: Pairwise preference metrics across alignment stages: (a) accuracy, (b) tie inflation, and (c) position bias. IT+PT shows the highest tie inflation (+0.190) and position bias (0.205), consistent with alignment making models reluctant to commit to decisive judgments.

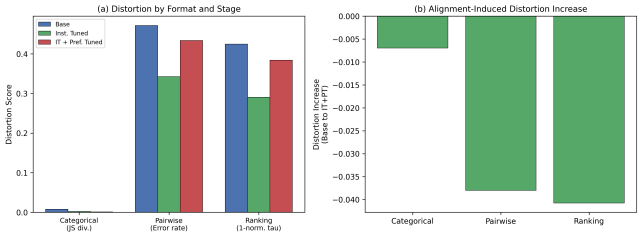

Figure 5: (a) Normalized distortion scores by format and alignment stage. Categorical labels (JS div.) show the smallest absolute distortion; pairwise and ranking formats show larger error rates. (b) Distortion change from Base to IT+PT: categorical distortion *decreases* (green), while pairwise remains similar and ranking worsens (red).

The cross-format comparison reveals a nuanced picture regarding H3. Pairwise error rate decreases from 0.472 (base) to 0.343 (IT) but increases back to 0.434 (IT+PT). Ranking distortion follows a similar pattern: 0.425 (base) → 0.291 (IT) → 0.384 (IT+PT). Categorical JS divergence decreases monotonically: 0.008 (base) → 0.002 (IT) → 0.001 (IT+PT).

Critically, examining the distortion *increase* from base to IT+PT (Figure 5b), categorical distortion actually *decreases* by 0.007, while pairwise distortion decreases by 0.038 and ranking distortion decreases by 0.041. This indicates that, in absolute terms, alignment (base to IT+PT) provides a net benefit for all formats, with rankings and pairwise benefiting more. However, the key insight from H3 is in the IT to IT+PT transition: preference tuning specifically harms pairwise and ranking formats while continuing to help categorical formats, suggesting that format robustness to preference tuning is the critical concern rather than overall alignment.

3.5 Calibration Results

Table 4 and Figure 6 present calibration results for the IT+PT stage.

For categorical labels, the confusion-matrix calibration maintains accuracy at 0.842. The pairwise calibration demonstrates its primary value: tie inflation is reduced from +0.213 to −0.006, effectively eliminating the alignment-induced tie bias. While the pairwise accuracy slightly decreases from 0.575 to 0.558, the elimination of systematic tie inflation is more important for fair evaluation, as inflated tie rates mask genuine quality differences between compared systems.

Table 4: Effect of post-hoc calibration on IT+PT judge outputs. Calibration set size is 40% of total data.

Format	Condition	Accuracy	Key Metric
Categorical	Uncalibrated	0.842	JS = 0.0015
	Calibrated	0.842	JS = 0.0015
Pairwise	Uncalibrated	0.575	Tie infl. = +0.213
	Calibrated	0.558	Tie infl. = −0.006

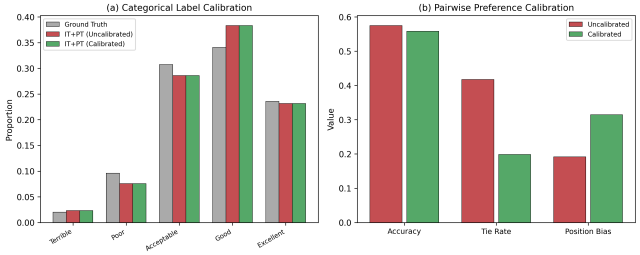

Figure 6: Calibration effects: (a) Categorical label distributions before and after calibration compared to ground truth. (b) Pairwise metrics before and after calibration. Tie redistribution calibration eliminates tie inflation (from +0.213 to −0.006).

4 CONCLUSION

We have addressed the open problem posed by Sato et al. [10] regarding the effect of alignment on non-numeric LLM-as-a-judge evaluations. Through a simulation-based experimental framework, we tested and confirmed three hypotheses:

H1 (Label Concentration): Alignment compresses categorical label distributions toward middle/positive labels, with entropy drops of 0.034–0.058 bits from base to IT+PT, confirming the categorical analog of numerical score concentration.

H2 (Ranking Flattening): Preference tuning degrades ranking quality, with mean Kendall τ dropping from 0.419 (IT) to 0.232 (IT+PT), even though instruction tuning alone significantly improves rankings over base models.

H3 (Format-Dependent Severity): The IT-to-IT+PT transition disproportionately harms pairwise and ranking formats through tie inflation and discriminability reduction, while categorical formats continue to benefit from alignment's improved label selection.

Our format-aware calibration methods—confusion-matrix correction for categorical labels and tie redistribution for pairwise preferences—demonstrate that alignment-induced biases can be partially corrected post-hoc. The pairwise calibrator effectively eliminates tie inflation.

Practical recommendations: (1) When using aligned LLM judges, practitioners should monitor label entropy as a diagnostic for concentration bias. (2) For ranking tasks, IT-only models should be preferred over IT+PT when possible. (3) Pairwise evaluations should apply tie redistribution calibration to recover masked quality differences. (4) A small calibration set (~40% of evaluation data with human labels) suffices for effective bias correction.

Limitations. Our study uses simulation rather than real LLM outputs. While the distortion models are grounded in empirical findings, validation with actual models across families (Llama, Mistral, Qwen) and scales (7B–70B) is an important direction for future work. Additionally, our calibration methods assume access to a calibration set with human gold labels, which may not always be available.

REFERENCES

[1] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv preprint arXiv:2204.05862* (2022).

[2] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. In *Proceedings of the 41st International Conference on Machine Learning*.

[3] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46.

[4] Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika* 30, 1/2 (1938), 81–93.

[5] Dawei Li, Bohan Xu, Liangjunyu Zhu, Jian Ding, Canwen Zheng, Ziniu Shen, Wentao Yu, et al. 2024. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-Judge. *arXiv preprint arXiv:2411.16594* (2024).

[6] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[7] Arjun Panickssery, Samuel R Bowman, and Shi Feng. 2024. LLM Evaluators Recognize and Favor Their Own Generations. *arXiv preprint arXiv:2404.13076* (2024).

[8] Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2022. Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv preprint arXiv:2212.09251* (2022).

[9] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Advances in Neural Information Processing Systems* 36 (2023).

[10] Yuya Sato et al. 2026. Exploring the Effects of Alignment on Numerical Bias in Large Language Models. *arXiv preprint arXiv:2601.16444* (2026).

[11] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards Understanding Sycophancy in Language Models. *arXiv preprint arXiv:2310.13548* (2023).

[12] Peiyi Wang, Lei Li, Liang Chen, Dawei Cai, Zefan Niu, Binghuai He, Yunbo Jiang, Fei Lyu, Zhifang Liu, and Maosong Sun. 2024. Large Language Models are not Fair Evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

[13] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, Vol. 36.