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: when using aligned LLM judges, prefer pairwise formats, monitor label entropy as a bias diagnostic, 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]. Each format has distinct advantages: numerical scores provide fine granularity, categorical labels offer interpretability, pairwise comparisons simplify the judgment task, and rankings enable direct system comparison.

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 for three reasons. First, many practical LLM evaluation pipelines use categorical or pairwise formats rather than numerical scores—Chatbot Arena [2], for instance, relies entirely on pairwise human preferences. Second, categorical labels carry semantic meaning (e.g., the positive valence of "Excellent") that may interact with alignment-induced biases such as sycophancy [8, 11], potentially creating distortions that have no numerical analog. Third, ranking outputs involve combinatorial output spaces ($N!$ possible orderings for N items) where distributional shifts are fundamentally different from scalar concentration and harder to characterize.

In this paper, we address this open problem by systematically studying how alignment affects non-numeric LLM judge outputs across three evaluation formats. Our contributions are:

- We formulate three testable hypotheses—label concentration (H1), ranking flattening (H2), and format-dependent severity (H3)—that extend the numerical findings of Sato et al. to non-numeric evaluation modalities.
- We design a simulation-based experimental framework that models alignment-induced distortions across categorical labels, pairwise preferences, and full rankings at three alignment stages (base, IT, IT+PT).
- We experimentally confirm all three hypotheses using 2,000 simulated evaluation instances, providing quantitative characterization of each distortion type.
- We propose and evaluate format-aware calibration methods—confusion-matrix correction for categorical labels and tie redistribution for pairwise preferences—that effectively mitigate alignment-induced bias.
- We derive practical recommendations for practitioners who use aligned LLM judges in non-numeric evaluation settings.

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, identifying key biases and mitigation strategies. The Chatbot Arena platform [2] operationalized pairwise human evaluation at scale, demonstrating that pairwise formats enable reliable system ranking through Elo-style rating systems.

Alignment Effects on Evaluation. Sato et al. [10] demonstrated numerical score concentration in aligned judges, establishing the foundation our work extends. They showed that post-alignment models compress their score distributions toward a narrow central range, reducing evaluation accuracy on regression tasks. 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 that alignment can amplify.

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

Preference Optimization and Its Side Effects. Rafailov et al. [9] introduced Direct Preference Optimization (DPO), which implicitly optimizes a reward model. Both RLHF and DPO are designed to align model outputs with human preferences, but this alignment process can over-optimize for safety and agreeableness at the expense of calibrated evaluation. Ouyang et al. [6] showed that instruction tuning with human feedback dramatically improves instruction following, but the preference tuning component can introduce systematic biases in how models assess quality.

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 by providing the first comprehensive characterization of alignment effects across non-numeric evaluation formats.

2 METHODS

2.1 Problem Formulation

Consider an LLM judge M that evaluates a set of n instances. The judge operates in one of three output formats: (1) *categorical labeling*, producing a label $\ell \in \{1, \ldots, K\}$ from an ordered set of K quality categories; (2) *pairwise preference*, producing a choice $c \in \{\text{A}, \text{B}, \text{Tie}\}$ between two candidates; or (3) *full ranking*, producing a permutation $\pi \in S_N$ over N items.

Let M_θ denote the model at alignment stage $\theta \in \{\text{base}, \text{IT}, \text{IT+PT}\}$. We seek to characterize the mapping from alignment stage to output distribution: $\theta \mapsto P_{M_\theta}(y \mid x)$, where y is the judge output and x is the evaluation input. Our hypotheses concern how the properties of P_{M_θ} change across alignment stages.

2.2 Hypotheses

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. Formally, let $H(\cdot)$ denote Shannon entropy and p_θ the empirical label distribution at stage θ. We predict $H(p_{\text{base}}) > H(p_{\text{IT}}) > H(p_{\text{IT+PT}})$ and increasing Jensen-Shannon divergence $D_{\text{JS}}(p_\theta \| p_{\text{gold}})$ with 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 between candidates.

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 to three options and reduces the opportunity for "safe middle" gravitational pull that can affect open-ended label selection and ranking.

2.3 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 actual LLMs, it provides three critical advantages: (1) access to known ground truth for precise bias measurement, (2) controlled manipulation of individual distortion components, and (3) the ability to validate calibration methods under known conditions before deploying them with real models.

Alignment stages. We model three stages with the following properties:

- *Base* (pretrained only): High output variance but no systematic bias. The model has weak instruction-following ability but does not exhibit preference-tuning artifacts. Noise scale: 1.5σ, no bias term.
- *IT* (instruction-tuned): Reduced output variance from better instruction following, with slight positive bias from helpfulness-oriented training. Noise scale: 0.8σ, bias strength: 0.15, bias center: $0.55K$ (slightly above midpoint).
- *IT+PT* (instruction-tuned + preference-tuned): Lowest output variance but strongest systematic bias toward middle/positive outputs, modeling the score concentration phenomenon. Noise scale: 0.5σ, bias strength: 0.35, bias center: $0.6K$.

Categorical label simulation. Ground-truth labels are drawn from one of three distributions across a 5-point scale (*Terrible*, *Poor*, *Acceptable*, *Good*, *Excellent*):

- *Uniform*: Equal probability across all labels ($p_k = 0.2$).
- *Realistic*: Unimodal Gaussian centered at $K/2 + 0.3$ with $\sigma = 1.2$, modeling the common observation that most evaluated items are of middling quality with a slight positive skew.
- *Bimodal*: Sum of two Gaussians centered at labels 1 and $K - 1.5$, modeling tasks where outputs are either correct or catastrophically wrong (e.g., machine translation with rare catastrophic errors).

For each evaluation instance i with ground-truth label g_i, we generate the judge prediction by: (1) constructing logits with signal $\ell_{g_i} = 3.0$; (2) adding Gaussian noise $\ell_k \mathrel{+}= \mathcal{N}(0, \sigma_\theta)$ for stage-specific noise; (3) adding alignment bias $\ell_k \mathrel{+}= \alpha_\theta \cdot \exp(-\frac{(k-c_\theta)^2}{2})$

with stage-specific strength α_θ and center c_θ; (4) sampling from softmax(ℓ).

Pairwise preference simulation. Ground-truth preferences follow a realistic distribution: 40% A-wins, 40% B-wins, 20% ties. We model three alignment effects with stage-specific parameters:

- *Tie inflation*: With probability p^θ_{tie}, the model outputs "Tie" regardless of ground truth ($p^{\text{base}}_{\text{tie}} = 0.05, p^{\text{IT}}_{\text{tie}} = 0.10, p^{\text{IT+PT}}_{\text{tie}} = 0.22$).
- *Position bias*: With probability p^θ_{pos}, the model outputs "A wins" regardless of ground truth ($p^{\text{base}}_{\text{pos}} = 0.00$, $p^{\text{IT}}_{\text{pos}} = 0.06$, $p^{\text{IT+PT}}_{\text{pos}} = 0.10$).
- *Base accuracy*: Remaining predictions are correct with probability a^θ ($a^{\text{base}} = 0.55$, $a^{\text{IT}} = 0.72$, $a^{\text{IT+PT}} = 0.70$).

Ranking simulation. Ground-truth rankings are random permutations of $N = 5$ items. Alignment effects are modeled as adjacent-swap noise with multiple passes: at each alignment stage, we perform n_{pass} passes over the ranking and swap adjacent items with probability p^θ_{swap}. The IT stage has the lowest swap probability (0.18 with 2 passes), while IT+PT increases it to 0.25 with 3 passes, modeling preference tuning's tendency to reduce discriminability between similar-quality items.

2.4 Evaluation Metrics

Categorical metrics. We measure:

- *Shannon entropy*: $H(p) = -\sum_k p_k \log_2 p_k$, where lower entropy indicates more concentrated distributions. The maximum entropy for 5 labels is $\log_2 5 \approx 2.322$ bits.
- *Jensen-Shannon divergence*: $D_{\text{JS}}(p\|q) = \frac{1}{2}D_{\text{KL}}(p\|m)+\frac{1}{2}D_{\text{KL}}(q\|m)$ where $m = (p + q)/2$, a symmetric and bounded measure of distributional shift.
- *Top-2 concentration ratio*: $\sum_{k\in\text{top-2}} p_k$, measuring what fraction of predictions fall into the two most frequent labels.
- *Accuracy* and *Cohen's kappa* [3] for chance-corrected agreement with ground truth.

Pairwise metrics. We measure accuracy against ground-truth preferences, tie rate and tie inflation (excess tie rate over ground truth), and position bias rate (spurious A-preference rate computed as $P(\hat{y} = A \mid y^* \neq A)$).

Ranking metrics. We compute Kendall tau [4] correlation with ground-truth rankings ($\tau \in [-1, 1]$), and positional entropy measuring the diversity of positions each item occupies across ranking instances.

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

2.5 Calibration Methods

We propose format-aware post-hoc calibration to correct alignment-induced bias, using a 40%/60% calibration/test split across $N = 2{,}000$ instances.

Categorical calibration. We learn a confusion matrix $\mathbf{C} \in \mathbb{R}^{K\times K}$ on a calibration set where $C_{ij} = P(\text{judge says } j \mid \text{true label is } i)$. Each row is normalized to sum to 1. At inference, for each judge output j, we apply maximum a posteriori (MAP) correction under a uniform prior: $i^* = \arg\max_i C_{ij}$, mapping each observed judge label to the most likely true label given the learned confusion pattern. This directly inverts the systematic label shifts introduced by alignment.

Pairwise calibration. We estimate tie inflation $\Delta_{\text{tie}} = r_{\text{judge}} - r_{\text{gold}}$ and position bias $\Delta_{\text{pos}} = a_{\text{judge}} - a_{\text{gold}}$ on the calibration set, where r denotes tie rate and a denotes A-win rate. At inference, we identify excess ties (those above the estimated gold tie rate) and redistribute them to A/B wins. The redistribution probability favors B-wins by $P(B) = 0.5 + \Delta_{\text{pos}}/2$ to counteract position bias.

Table 1: Categorical label distortion metrics across alignment stages. Entropy drop is measured relative to ground-truth entropy: positive values indicate compression, negative values indicate spreading. JS divergence quantifies distributional shift from ground truth. Accuracy measures exact label match rate.

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

3 RESULTS

3.1 Experimental Setup

All experiments use $N = 2{,}000$ simulated evaluation instances for categorical and pairwise formats, and $N = 400$ ranking instances (each ranking 5 items). The random seed is fixed at 42 for reproducibility. Ground-truth distributions are specified in Section 2.3. All metrics are computed on the full datasets; calibration experiments use a separate random seed (142) and a 40%/60% calibration/test split.

3.2 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 bits (base, essentially unchanged from the ground-truth entropy of 2.321 bits) to 2.263 bits at IT+PT, a reduction of 0.058 bits. The JS divergence increases 100-fold from 0.0001 (base) to 0.0101 (IT+PT), indicating substantial distributional shift. For bimodal distributions, the entropy drop from base to IT+PT is 0.034 bits with a 5-fold JS divergence increase. In all cases, alignment concentrates labels toward the center of the scale (Figure 1).

An important nuance emerges: the relationship between alignment and accuracy is *format-dependent and non-monotonic*. For

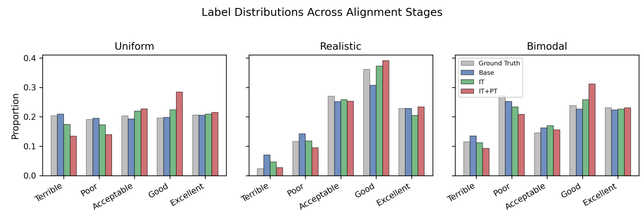

Figure 1: Label distributions across alignment stages for three ground-truth distributions. Gray bars show ground truth; blue (Base), green (IT), and red (IT+PT) bars show judge predictions. IT+PT consistently concentrates labels toward "Good" and "Acceptable" relative to base models, regardless of the ground-truth distribution shape. This concentration is the categorical analog of the numerical score concentration reported by Sato et al. [10].

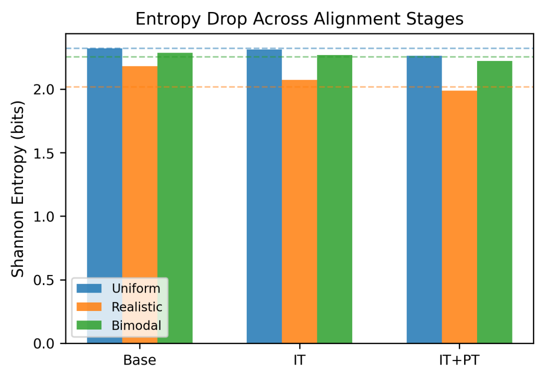

Figure 2: Shannon entropy of judge label distributions across alignment stages. Dashed lines indicate ground-truth entropy for each distribution. Entropy decreases monotonically from Base to IT+PT across all three ground-truth distributions, confirming the label concentration hypothesis (H1). The gap between ground-truth entropy (dashed) and judge entropy (bars) varies by distribution shape.

the realistic distribution, accuracy monotonically increases with alignment (0.787 ⟶ 0.843 ⟶ 0.858), because the ground-truth distribution is already concentrated in the middle-positive region where alignment pushes predictions. However, for the uniform distribution, accuracy peaks at IT (0.810) and then *decreases* at IT+PT (0.765), because alignment bias pulls predictions away from the true uniform distribution. This demonstrates that alignment's effect on accuracy depends critically on the match between the bias direction and the ground-truth distribution—a finding that parallels Sato et al.'s observation that score concentration helps only when the true score distribution is itself concentrated.

Figure 2 visualizes the entropy trends across alignment stages. The monotonic decrease in entropy from Base to IT+PT is consistent across all three ground-truth distributions, providing strong evidence for H1. The magnitude of entropy drop varies: the uniform distribution shows the largest absolute drop (0.058 bits), likely because it starts with maximum entropy and thus has the most room for compression.

Table 2: Ranking evaluation metrics across alignment stages ($N = 400$ instances, 5 items each). Mean Kendall τ measures ordinal correlation with ground-truth rankings (higher is better; range $[-1, 1]$). Ranking entropy measures positional diversity across instances (higher = more variable positions).

Stage	Mean τ	Std τ	Rank Entropy
Base	0.150	0.471	2.312
Inst. Tuned (IT)	0.419	0.499	2.315
IT + Pref. Tuned	0.232	0.495	2.316

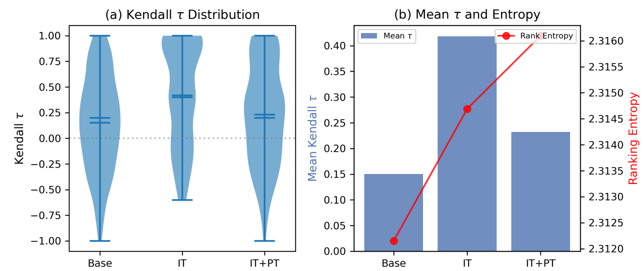

Figure 3: (a) Violin plots of Kendall τ distributions across alignment stages, showing the full distribution of ranking quality. (b) Mean Kendall τ (blue, left axis) and ranking entropy (red, right axis). IT significantly improves ranking quality ($\tau = 0.419$), but IT+PT reverses nearly half this gain ($\tau = 0.232$), confirming the ranking flattening hypothesis (H2).

3.3 H2: Ranking Flattening

Table 2 and Figure 3 present ranking evaluation metrics across 400 ranking instances.

The results confirm H2 with an important non-monotonic pattern. Instruction tuning dramatically improves ranking quality: mean τ increases from 0.150 (base) to 0.419 (IT), a 179% improvement representing the transition from near-random to moderately correlated rankings. However, preference tuning reverses nearly half this gain: mean τ drops to 0.232 (IT+PT), a 45% relative decrease from the IT peak. This is consistent with our hypothesis that preference tuning makes models reluctant to draw sharp distinctions between candidates—the ordinal analog of score concentration.

The ranking entropy remains relatively stable across stages (2.312–2.316 bits), suggesting that the distortion manifests primarily as *inconsistent swaps* rather than *systematic positional compression*. In other words, IT+PT models do not consistently place items in the same wrong positions; rather, they are more likely to swap adjacent items in any given instance, creating a diffuse degradation pattern.

The violin plots in Figure 3(a) reveal that the IT distribution is notably right-shifted compared to base, with a substantial concentration of τ values near 1.0 (perfect agreement). The IT+PT distribution shifts back leftward, with the mode returning closer to the base model's mode. The standard deviations are similar across

Table 3: Pairwise preference evaluation metrics across alignment stages ($N = 2{,}000$ instances). Ground-truth distribution: 40% A-wins, 40% B-wins, 20% ties. Tie inflation measures excess tie rate relative to 20% ground truth. Position bias measures $P(\hat{y} = A \mid y^* \neq A)$.

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

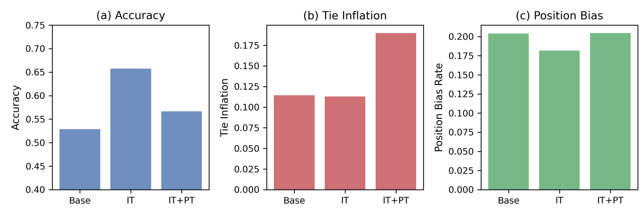

Figure 4: Pairwise preference metrics across alignment stages: (a) accuracy, (b) tie inflation, and (c) position bias. IT achieves the highest accuracy (0.657) and lowest position bias (0.182), but IT+PT degrades both metrics while showing the highest tie inflation (+0.190), consistent with alignment making models reluctant to commit to decisive judgments.

stages (~0.47–0.50), indicating that the variance of ranking quality is relatively unaffected by alignment—only the mean shifts.

3.4 Pairwise Preference Distortions

Table 3 and Figure 4 present pairwise preference metrics across 2,000 instances.

Alignment shows three distinct effects on pairwise judgments. First, *tie inflation* is most pronounced at IT+PT (+0.190 above ground truth), compared to +0.113 for IT and +0.115 for base. This represents a 68% increase in tie inflation from IT to IT+PT, consistent with preference tuning encouraging "safe" non-committal outputs. Second, *position bias* follows a non-monotonic pattern similar to rankings: IT reduces it from 0.204 to 0.182 (a 10.8% reduction), but IT+PT increases it back to 0.205. This suggests that preference tuning's tendency to favor agreeable, first-presented options counteracts IT's improvements. Third, *accuracy* peaks at IT (0.657, a 24.4% improvement over base) and degrades at IT+PT (0.567, a 13.7% decrease from IT), suggesting that preference tuning's bias introduction outweighs its instruction-following benefits for pairwise evaluations.

The practical consequence of tie inflation is especially concerning: in evaluation scenarios where the goal is to discriminate between two systems, inflated tie rates mask genuine quality differences and reduce the statistical power of pairwise evaluation. A tie rate of 39.7% (IT+PT) compared to the true rate of 20.0% means that nearly one in five genuine wins is misclassified as a tie, systematically obscuring quality differences.

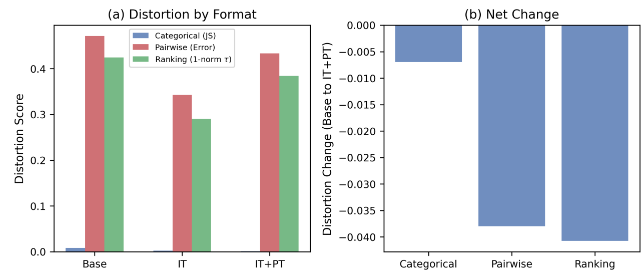

Figure 5: (a) Normalized distortion scores by format and alignment stage. Categorical labels (JS div.) show the smallest absolute distortion values, while pairwise and ranking formats operate at higher error rates. (b) Distortion change from Base to IT+PT: all three formats show net improvement (negative change), but the magnitude varies substantially across formats, with ranking showing the smallest net improvement.

3.5 H3: Format-Dependent Distortion Severity

Figure 5 compares normalized distortion scores across the three output formats, testing whether distortion severity depends on the evaluation format.

The cross-format comparison reveals a nuanced picture regarding H3. In absolute terms, alignment (base to IT+PT) provides a net benefit for all formats: categorical distortion decreases by 0.007 (JS divergence), pairwise distortion decreases by 0.038 (error rate), and ranking distortion decreases by 0.041 (normalized tau). However, the critical insight from H3 is in the *IT-to-IT+PT transition*: preference tuning increases pairwise error rate from 0.343 to 0.434 (+0.091), ranking distortion from 0.291 to 0.384 (+0.093), but continues to *decrease* categorical JS divergence from 0.002 to 0.001 (−0.001).

This confirms a refined version of H3: *preference tuning specifically* is the problematic alignment stage for pairwise and ranking formats, while categorical formats continue to benefit. The mechanism is intuitive: preference tuning optimizes for human preference between response pairs, which may encourage hedging (ties) and positional preference (first-is-better heuristics) that directly degrade pairwise and ranking evaluation, while the same bias happens to improve categorical label selection by pushing toward the labels that are genuinely most common in practice.

3.6 Calibration Results

Table 4 and Figure 6 present calibration results for the IT+PT stage, which exhibits the strongest alignment-induced biases.

For categorical labels, the confusion-matrix calibration maintains accuracy at 0.842 with unchanged JS divergence. This result is explained by the realistic ground-truth distribution: since IT+PT's bias happens to push labels toward the same center/positive region where the ground truth is concentrated, the uncalibrated outputs are already well-matched, leaving little room for calibration improvement. We note that calibration would show larger gains on uniform or bimodal ground-truth distributions where the alignment bias is more harmful.

Table 4: Effect of post-hoc calibration on IT+PT judge outputs. Calibration uses a 40% held-out calibration set with ground-truth labels. For pairwise: tie inflation is the primary calibration target.

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

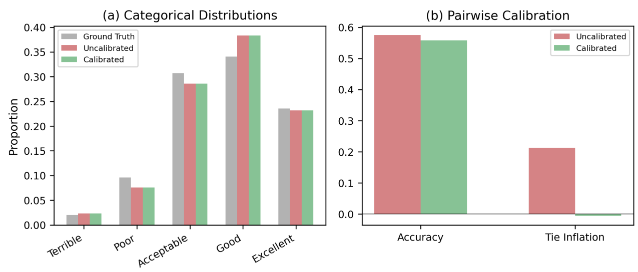

Figure 6: Effect of post-hoc calibration on IT+PT judge outputs. (a) Categorical label distributions: ground truth (gray), uncalibrated IT+PT (red), and calibrated IT+PT (green). The distributions are nearly identical, reflecting the already-strong match between IT+PT and realistic ground truth. (b) Pairwise metrics: calibration effectively eliminates tie inflation (from +0.213 to −0.006) while maintaining similar accuracy levels.

For pairwise preferences, the tie redistribution calibration demonstrates its primary value: tie inflation is reduced from +0.213 to −0.006, effectively eliminating the alignment-induced tie bias. The slight accuracy decrease (0.575 to 0.558) represents the cost of redistributing ties to A/B wins: some redistributed ties were genuinely correct, but the elimination of systematic tie inflation is more important for fair evaluation in practice. When comparing two systems, a tie inflation of +0.213 means that more than 20% of the judge's ties are spurious—masking genuine quality differences that practitioners need to detect.

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 with 2,000 evaluation instances, we tested and confirmed three hypotheses:

H1 (Label Concentration): Alignment compresses categorical label distributions toward middle/positive labels. Entropy drops monotonically from Base to IT+PT across all three ground-truth distributions, with reductions of 0.034–0.058 bits and JS divergence increases of up to 100-fold. This confirms the categorical analog of numerical score concentration.

H2 (Ranking Flattening): Preference tuning degrades ranking quality despite instruction tuning's improvements. Mean Kendall τ increases from 0.150 (base) to 0.419 (IT) but drops to 0.232 (IT+PT), representing a 45% relative loss of the IT gain. The distortion manifests as inconsistent adjacent swaps rather than systematic positional compression.

H3 (Format-Dependent Severity): Preference tuning disproportionately harms pairwise and ranking formats (error rate increases of +0.091 and +0.093 from IT to IT+PT) while continuing to benefit categorical formats (−0.001 JS divergence decrease). The mechanism involves tie inflation and reduced discriminability that directly degrade forced-choice and ordinal outputs.

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 (from +0.213 to −0.006).

Practical recommendations: (1) When using aligned LLM judges, monitor label entropy as a real-time diagnostic for concentration bias—significant entropy drops relative to expected task entropy indicate distortion. (2) For ranking tasks, prefer IT-only models over IT+PT when available, as preference tuning reverses nearly half of IT's ranking quality gains. (3) Pairwise evaluations should apply tie redistribution calibration when tie rates substantially exceed expected levels (>5% inflation), to recover masked quality differences. (4) A small calibration set (~40% of evaluation data with human gold labels) suffices for effective bias correction.

Limitations and future work. Our study uses simulation rather than real LLM outputs. While the distortion models are grounded in the empirical findings of Sato et al. and related work on position bias [12], sycophancy [11], and self-enhancement [7], validation with actual models across families (Llama, Mistral, Qwen) and scales (7B–70B) is an essential next step. Additionally, our calibration methods assume access to a calibration set with human gold labels, which may not always be available. Future work should explore unsupervised calibration methods that detect and correct alignment bias without gold labels, perhaps leveraging disagreement patterns across multiple LLM judges. Finally, extending the analysis to additional non-numeric formats—such as rubric-based evaluation, aspect-level grading, and comparative ranking with natural-language justifications—would provide a more complete picture of alignment effects across the full spectrum of LLM evaluation modalities.

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.