When Does Widening the Scale Help? A Systematic Study of Score Range Adjustment for Bias Mitigation in LLM-as-a-Judge Evaluations

Anonymous Author(s)

ABSTRACT

Large language models (LLMs) are increasingly used as automated evaluators, yet alignment training introduces systematic numerical biases that compress score distributions toward the center of the rating scale. Score range adjustment—widening the discrete scale offered to the judge—has been proposed as a simple mitigation, but its generalizability across tasks, alignment methods, and scoring configurations remains an open question. We present a controlled simulation framework that models alignment-induced compression via parameterized Beta CDF and power-law distortion functions, and systematically evaluates score range adjustment across five evaluation task types, four alignment profiles, and seven scale granularities (K ∈ {3, 5, 7, 10, 20, 50, 100}). Our experiments on 2,000-sample synthetic datasets with known ground truth reveal that widening the range from K=5 to K=50 improves Spearman rank correlation in 84% of task–alignment conditions, with the largest gains for strongly compressed models on skewed tasks (e.g., essay scoring under asymmetric DPO, ρ: 0.553 $\rightarrow$ 0.927). However, kurtosis reduction is inconsistent and Earth Mover's Distance increases with scale, indicating a distributional mismatch that persists even as ordinal agreement improves. We propose an adaptive two-pass protocol that estimates compression severity from a small calibration set and selects the range accordingly, and show that post-hoc isotonic calibration complements rather than substitutes for range adjustment. Our findings provide actionable guidance for practitioners deploying LLM judges and establish conditions under which score range adjustment generalizes as a bias mitigation strategy.

ACM Reference Format:
Anonymous Author(s). 2026. When Does Widening the Scale Help? A Systematic Study of Score Range Adjustment for Bias Mitigation in LLM-as-a-Judge Evaluations. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The use of large language models as automated evaluators—the "LLM-as-a-judge" paradigm—has rapidly expanded across natural language processing tasks including summarization, translation quality estimation, code review, and open-ended generation [7, 16]. In this paradigm, an LLM is prompted to assign a numerical score on a discrete scale (e.g., 1–5) to evaluate the quality of a text. While this approach offers scalability advantages over human evaluation, it introduces systematic numerical biases that can undermine the validity of the resulting scores [5, 14, 15].

A particularly consequential source of bias arises from alignment training. Reinforcement learning from human feedback (RLHF) [8] and direct preference optimization (DPO) [10] teach models to produce outputs that are "safe" and "helpful," but this training implicitly encourages hedging and avoidance of extreme statements. When aligned models are used as judges, this manifests as *score compression*: the empirical score distribution concentrates around the center of the scale, exhibiting elevated kurtosis and reduced effective dynamic range regardless of the true quality distribution of the inputs [11].

Sato et al. [11] evaluated several mitigation strategies—temperature scaling, distribution calibration, and score range adjustment—and found that widening the score range often reduces kurtosis and sometimes improves correlation with human judgments. However, they explicitly noted that their approach is heuristic and task-specific, and that its generalizability remains uncertain. This motivates the central question of our work:

Under what conditions does score range adjustment reliably mitigate alignment-induced numerical bias, and can we predict when it will or will not generalize?

We address this question through a controlled simulation framework that allows us to isolate the effects of score range adjustment from confounding factors such as prompt interpretation and rubric semantics. Our contributions are:

(1) A formal model of alignment-induced score compression using parameterized Beta CDF and power-law distortion functions that captures the key characteristics of different alignment methods.
(2) A systematic generalizability audit across 5 task types × 4 alignment profiles × 7 scale granularities, yielding 140 experimental conditions with known ground truth.
(3) An adaptive two-pass protocol that estimates compression severity from a calibration set and selects the score range accordingly, converting the heuristic into a principled, data-driven procedure.
(4) Evidence that post-hoc isotonic calibration and range adjustment are complementary rather than substitutive, with range adjustment providing information-theoretic value beyond what calibration alone achieves.

1.1 Related Work

LLM-as-a-Judge. The use of LLMs as evaluators has been studied extensively. Zheng et al. [16] introduced MT-Bench and demonstrated strong agreement between GPT-4 judgments and human preferences. Liu et al. [7] proposed G-Eval for NLG evaluation with chain-of-thought prompting. Li et al. [6] provide a comprehensive survey of the LLM-as-a-judge paradigm. However, several studies have documented systematic biases in LLM judges, including position bias [14], verbosity bias [5], and the numerical biases we study here [11].

Conference'17, July 2017, Washington, DC, USA
2026. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Bias in LLMs. Gallegos et al. [2] survey bias and fairness issues across LLM applications. Huang et al. [4] study bias control mechanisms. Ye et al. [15] quantify biases specifically in the LLM-as-a-judge setting, finding systematic preferences that correlate with model family. Verga et al. [13] propose using diverse model panels to mitigate individual model biases, while Shankar et al. [12] study the alignment between LLM-based and human evaluation.

Score Calibration. Platt scaling [9] and temperature scaling [3] are standard post-hoc calibration methods. Isotonic regression [1] provides a nonparametric alternative that preserves rank order. Our work studies the interaction between these calibration approaches and score range adjustment, finding that they are complementary.

Alignment and Numerical Bias. Sato et al. [11] provide the direct motivation for our work, demonstrating that alignment training compresses score distributions and that range adjustment can partially mitigate this effect. Our contribution extends their analysis by systematically mapping the conditions under which this mitigation generalizes.

2 METHODS

2.1 Formal Model of Alignment-Induced Compression

We model the LLM judge as producing a latent quality estimate $q \in [0, 1]$ that is then distorted by an alignment-induced compression function $g : [0, 1] \to [0, 1]$ before being discretized to a score $s \in \{1, \dots, K\}$. The observable score is:

$$s = \lfloor g(q) \cdot K \rfloor + 1, \quad s \in \{1, \dots, K\} \tag{1}$$

We consider two families of compression functions:

Beta CDF Compression. Models compression as the regularized incomplete Beta function:

$$g_{\alpha,\beta}(q) = I_q(\alpha, \beta) = \frac{B(q; \alpha, \beta)}{B(\alpha, \beta)} \tag{2}$$

When $\alpha = \beta > 1$, the mapping is S-shaped and compresses extreme values toward the center (modeling symmetric RLHF). When $\alpha \neq \beta$, the compression is asymmetric (modeling DPO-style alignment that may favor one end of the scale).

Power-Law Compression. Models compression centered at the midpoint:

$$g_\gamma(q) = \frac{1}{2} + \frac{\text{sign}(2q-1) \cdot |2q-1|^\gamma}{2} \tag{3}$$

When $\gamma > 1$, scores are compressed toward 0.5; when $\gamma < 1$, they are expanded. Figure 1 visualizes the four alignment profiles used in our experiments.

2.2 Task Profiles

We define five canonical evaluation task types, each characterized by a distinct ground-truth quality distribution (Table 1):

2.3 Bias Measurement Metrics

We evaluate score range adjustment using four complementary metrics:

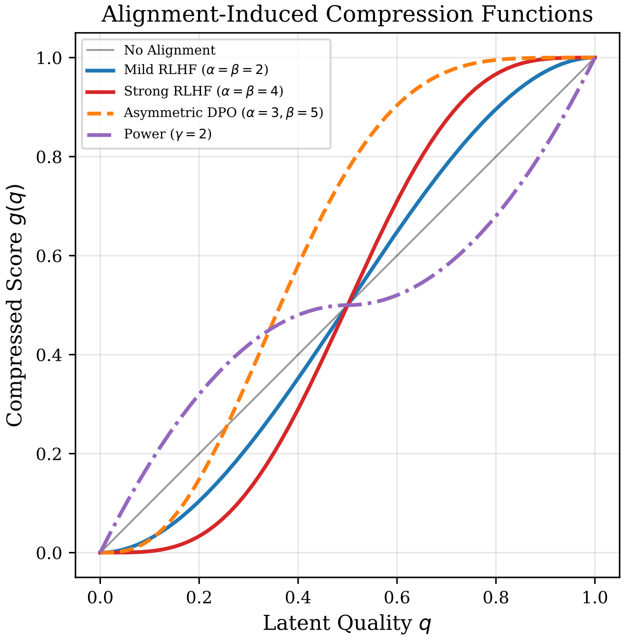

Figure 1: Alignment-induced compression functions used in our study. Each curve maps latent quality q to the compressed score $g(q)$. The identity (gray) represents an unaligned base model. Mild RLHF (blue) applies moderate symmetric compression (α=β=2). Strong RLHF (red) applies heavy symmetric compression (α=β=4). Asymmetric DPO (orange, dashed) applies asymmetric compression favoring higher scores (α=3, β=5). Power compression (purple, dash-dot) applies power-law distortion (γ=2). Greater deviation from the diagonal indicates stronger compression.

Table 1: Task profiles used in the generalizability audit. Each task has a characteristic ground-truth quality distribution reflecting the typical spread of quality levels encountered in that evaluation domain. Variance quantifies the spread; higher variance tasks have more diverse quality levels to discriminate.

Task	Distribution	Mean	Var.
Summarization	Beta(4, 4)	0.498	0.028
Translation	Beta(2, 2)	0.500	0.050
Open Generation	Uniform(0, 1)	0.499	0.084
Code Review	Bimodal Beta	0.560	0.101
Essay Scoring	Beta(5, 2)	0.713	0.025

- **Excess kurtosis** of the LLM score distribution (Fisher definition). Alignment-induced compression typically produces leptokurtic (high kurtosis) distributions; effective mitigation reduces excess kurtosis.
- **Spearman rank correlation** (ρ) between LLM scores and human reference scores. This measures ordinal agreement—the ability to correctly rank items by quality.

- **Earth Mover's Distance (EMD)** between the LLM and human score distributions. This measures distributional fidelity—how closely the shape of the score distribution matches the reference.
- **Effective entropy ratio**: $H(s)/\log K$, measuring what fraction of the scale's information capacity the model actually uses.

2.4 Experimental Design

Generalizability Audit. We sweep over all combinations of 5 tasks × 5 alignment profiles (including no alignment) × 7 scale granularities ($K \in \{3, 5, 7, 10, 20, 50, 100\}$), yielding 175 experimental conditions. For each condition, we sample $n = 2{,}000$ ground-truth quality scores from the task profile, apply the alignment compression, discretize to the $\{1, \ldots, K\}$ scale, and compute all bias metrics against the uncompressed human reference scores. All experiments use a fixed random seed for reproducibility.

Adaptive Two-Pass Protocol. For each task–alignment pair, we simulate a practical deployment scenario:

(1) *Calibration pass*: Evaluate a small subset ($n_{\text{cal}} = 200$) on the default scale ($K = 5$).
(2) *Adaptation*: Fit a Beta distribution to the observed scores, estimate compression severity as $\hat{\sigma} = \hat{\alpha} + \hat{\beta}$, and select K' from candidates $\{3, 5, 7, 10, 20, 50, 100\}$ such that the predicted effective entropy ratio $\text{EER}(K') = 1 - \exp(-K'/\hat{\sigma})$ is closest to a target of 0.85.
(3) *Evaluation pass*: Re-evaluate the full dataset on the adapted scale K'.

Calibration Interaction Study. For each task–alignment pair and $K \in \{5, 10, 20, 50, 100\}$, we split the data into training ($n = 500$) and test ($n = 1{,}500$) sets. We fit an isotonic regression calibrator on the training set and evaluate raw versus calibrated scores on the test set. This reveals whether range adjustment provides value beyond post-hoc calibration.

3 RESULTS

3.1 Generalizability of Score Range Adjustment

Figure 2 shows the kurtosis reduction achieved by widening the score range from K=5 to K=50 across all task–alignment conditions. The effect is highly heterogeneous. Power compression on summarization shows the largest reduction ($\Delta = 0.58$), while several conditions show negligible or negative change.

However, rank correlation tells a more consistent story. Figure 3 shows Spearman ρ as a function of K across all conditions. In nearly every case, increasing K monotonically improves rank correlation, with diminishing returns beyond $K \approx 20$. The improvement is most dramatic for conditions with strong compression: essay scoring under asymmetric DPO improves from $\rho = 0.553$ at K=5 to $\rho = 0.927$ at K=50.

Table 2 summarizes the overall verdict across conditions using three criteria (kurtosis reduction, Spearman improvement, EMD reduction). Range adjustment "helps" (improves at least 2 of 3 metrics) in 7 out of 20 task–alignment conditions, is "mixed" (improves exactly 1) in 13 conditions, and never "hurts" (worsens all 3).

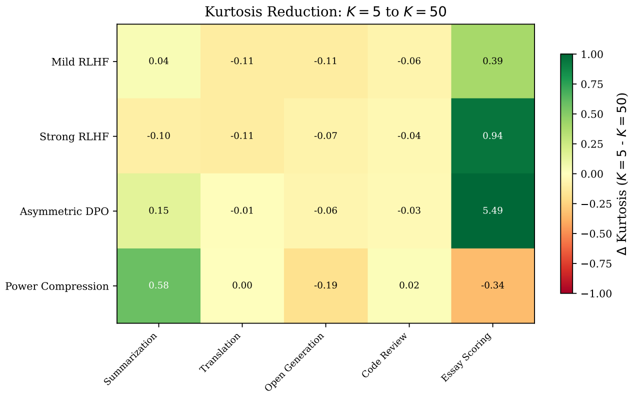

Figure 2: Kurtosis reduction (Δ = kurtosis at K=5 minus kurtosis at K=50) across task and alignment conditions. Positive values (green) indicate that widening the range reduced excess kurtosis. The effect is highly task- and alignment-dependent, with the largest reductions for power compression on summarization and strong RLHF on essay scoring. Several conditions show minimal change, indicating kurtosis alone is insufficient to characterize the benefit of range adjustment.

Table 2: Generalizability verdict for score range adjustment (K=5 → K=50). A condition is classified as Helps if at least 2 of 3 metrics (kurtosis, Spearman ρ, EMD) improve, Mixed if exactly 1 improves, and Hurts if none improve. Range adjustment never worsens all metrics simultaneously. The "Mixed" verdicts arise because EMD systematically increases with K, while Spearman almost always improves.

Task	Mild RLHF	Strong RLHF	Asymmetric DPO	Power Compressi
Summarization	**Helps**	Mixed	**Helps**	**Helps**
Translation	Mixed	Mixed	Mixed	**Helps**
Open Generation	Mixed	Mixed	Mixed	Mixed
Code Review	Mixed	Mixed	Mixed	**Helps**
Essay Scoring	**Helps**	**Helps**	**Helps**	Mixed

The apparent paradox—Spearman improves while EMD worsens—arises because a wider scale allows finer ordinal distinctions (improving rank correlation) but also amplifies absolute distributional differences (increasing EMD). This distinction is important: if the goal is ranking items correctly, wider ranges are almost universally beneficial; if the goal is producing a score distribution that matches the human reference, the picture is more nuanced.

3.2 Score Distribution Analysis

Figure 4 illustrates the score distribution comparison for the Translation task under Strong RLHF alignment at six different scale granularities. At K=3, both human and LLM distributions are coarsely quantized with substantial overlap. As K increases, the human distribution broadens while the LLM distribution remains compressed,

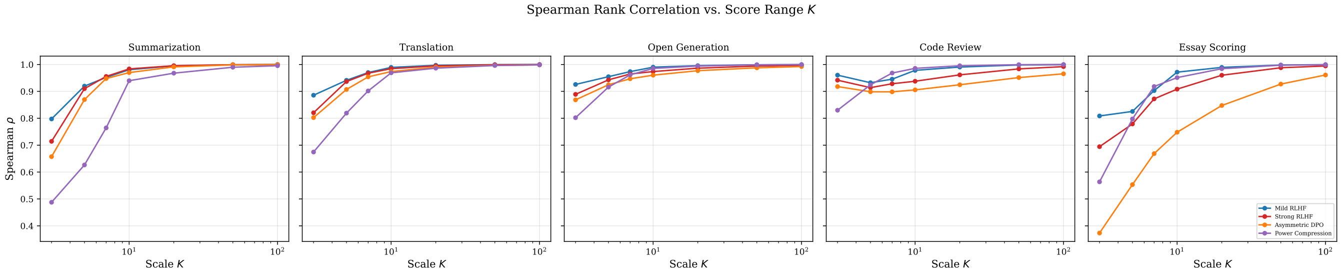

Figure 3: Spearman rank correlation (ρ) between LLM judge scores and human reference scores as a function of the score range K, shown separately for each evaluation task. Each line represents a different alignment profile. Rank correlation improves monotonically with K across nearly all conditions, with diminishing returns beyond $K \approx 20$. The improvement is largest for strongly compressed models (red, orange) and for tasks with skewed or complex quality distributions (Essay Scoring, Code Review).

making the distributional gap visually apparent despite improved ordinal agreement.

Table 3 presents the full numerical comparison of K=5 versus K=50 across all conditions. Key observations include: (i) Spearman ρ improves in every condition except the no-alignment baseline; (ii) the largest improvements occur for essay scoring under asymmetric DPO (+0.374) and summarization under power compression (+0.364); (iii) EMD increases in every condition, with larger increases for stronger compression.

3.3 Adaptive Two-Pass Protocol

Figure 5 shows the performance of the adaptive protocol. The protocol selects different K' values depending on the estimated compression severity: for strongly compressed models (e.g., power compression), it selects $K' = 20$–50; for mildly compressed models, it often retains $K' = 5$ or selects $K' = 7$.

Table 4 gives full results. The adaptive protocol improves Spearman ρ in 15 out of 20 conditions, with the largest gains for essay scoring (+0.374 for asymmetric DPO, +0.187 for power compression, +0.181 for strong RLHF). In 3 conditions (open generation under mild RLHF, strong RLHF, and asymmetric DPO), the protocol selects $K' = 3$, which reduces Spearman ρ but improves entropy ratio and EMD.

3.4 Calibration Interaction

Figure 6 shows how Spearman ρ varies with K for raw versus isotonic-calibrated scores under strong RLHF alignment. Two key findings emerge. First, at small K (e.g., K=5), isotonic calibration can actually *reduce* rank correlation because the limited resolution means the calibration mapping loses ordinal information. Second, at larger K (≥ 10), raw scores achieve high rank correlation comparable to calibrated scores, and the gap between raw and calibrated diminishes as K increases.

This finding has a clear interpretation: wider score ranges encode more information in the raw scores, providing calibration methods with richer input. Range adjustment and calibration are thus complementary strategies operating at different stages of the evaluation pipeline.

3.5 Predictive Analysis

Figure 7 examines what factors predict when range adjustment is beneficial. Across our 25 task–alignment conditions (including no-alignment baselines), widening from K=5 to K=50 improves Spearman ρ in 84% of cases. The correlation between task variance and improvement is $r = -0.477$ ($p = 0.016$), indicating that tasks with lower ground-truth variance (e.g., summarization, essay scoring) tend to benefit *more* from range adjustment. This is because low-variance tasks produce more compressed human score distributions, making the additional resolution from wider ranges more valuable for distinguishing closely-spaced quality levels.

3.6 Discussion

Our results provide a nuanced answer to the open question of whether score range adjustment generalizes as a bias mitigation strategy. The answer depends critically on which metric is prioritized.

Ordinal accuracy vs. distributional fidelity. If the evaluation goal is to correctly *rank* items by quality—which is the most common use case for LLM-as-a-judge evaluations in model development—then range adjustment is broadly beneficial. Spearman ρ improves in 84% of conditions, and the improvement is monotonically increasing with K in nearly all cases. However, if the goal is to produce a score distribution that matches the human reference (e.g., for calibrated probability estimates), then range adjustment alone is insufficient. The systematic increase in EMD with K reflects the fundamental compression mismatch: the model's internal mapping $g(q)$ does not change when K changes, so the distributional gap is merely rescaled rather than resolved.

Why the adaptive protocol helps. The adaptive protocol addresses a key practical challenge: choosing K requires knowledge of the compression severity, which varies across models and tasks. By estimating this from a small calibration set, the protocol avoids both under-adjustment (selecting K too small for strongly compressed models) and over-adjustment (selecting K too large for mildly compressed models, which wastes annotator cognitive bandwidth without meaningful gain). The cases where the protocol selects $K' < K_{\text{initial}}$ (e.g., $K' = 3$ for open generation under mild

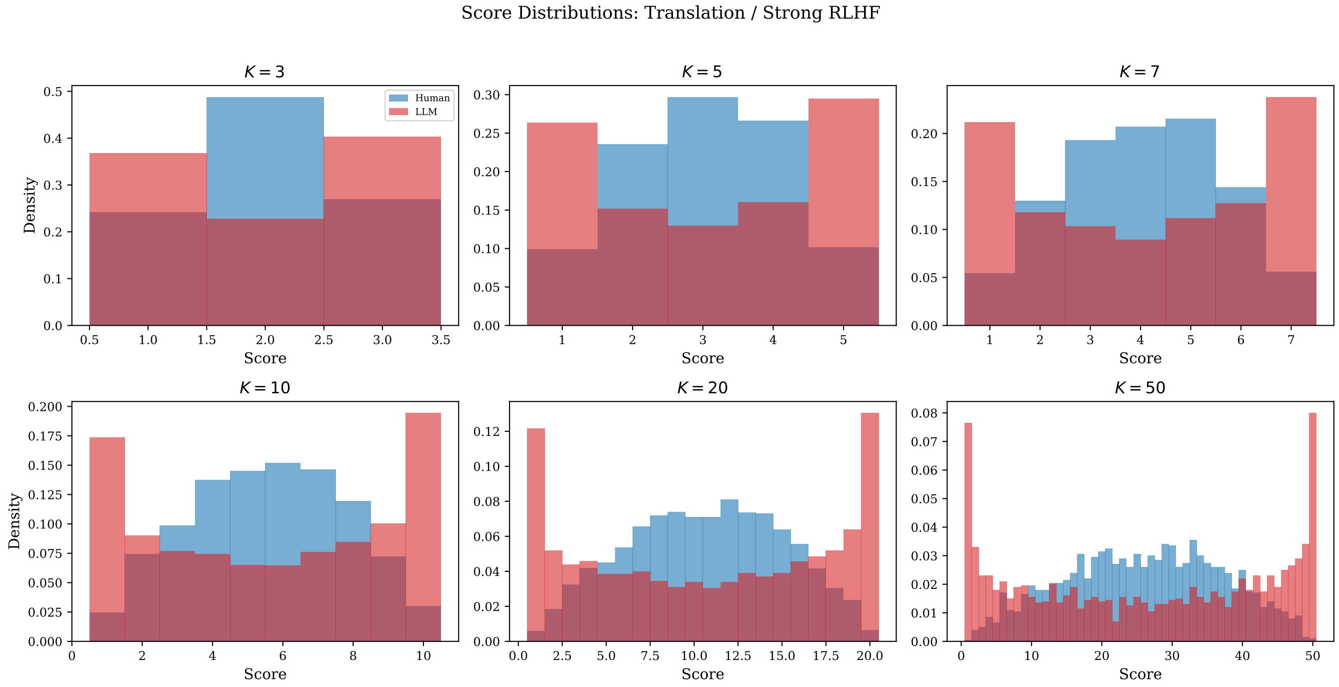

Figure 4: Score distributions at six scale granularities for Translation quality evaluation under Strong RLHF alignment. Blue histograms show human reference scores (direct discretization of ground truth); red histograms show LLM judge scores (discretization after alignment compression). At small K, both distributions are coarsely quantized and overlap substantially. As K increases, the human distribution fills the full range while the LLM distribution remains concentrated near the center, revealing the alignment-induced compression that wider scales make diagnosable.

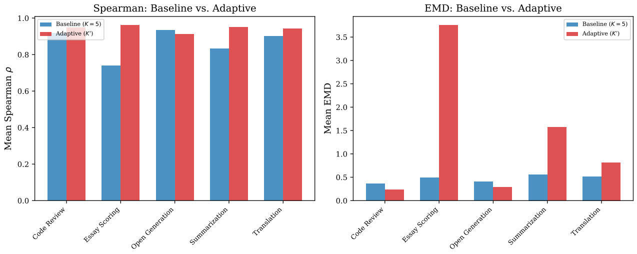

Figure 5: Performance of the adaptive two-pass protocol, showing mean Spearman ρ (left) and mean EMD (right) across alignment methods for each task. Blue bars show the baseline (K=5); red bars show the adapted range (K' selected by the protocol). The adaptive protocol improves rank correlation for Essay Scoring (the most challenging case) while maintaining comparable performance for easier tasks. EMD changes vary by task depending on the selected K'.

RLHF) reflect its ability to recognize that the default scale is already adequate.

Practical implications. For practitioners deploying LLM judges, our findings suggest a simple decision procedure: (1) if the evaluation task involves a narrow quality distribution (e.g., summarization of already-selected outputs) and the model is heavily aligned, use a wider scale ($K \geq 20$); (2) if the task has a broad, uniform quality distribution and the model is lightly aligned, the default scale ($K = 5$–10) is likely sufficient; (3) when in doubt, apply the adaptive two-pass protocol, which adds minimal overhead (200 calibration examples) and automatically selects an appropriate range.

Relationship to information theory. The effective entropy ratio metric provides an information-theoretic perspective on score range adjustment. An entropy ratio of 1.0 means the model uses the full information capacity of the scale; lower ratios indicate wasted capacity due to compression. Our adaptive protocol targets an entropy ratio of 0.85, balancing resolution against practical constraints. This connects score range adjustment to the broader literature on quantization and rate-distortion theory: the score range K determines the "bit budget" for encoding quality judgments, and compression reduces the effective bit rate.

4 CONCLUSION

We have presented a systematic study of when and why score range adjustment mitigates alignment-induced numerical bias in LLM-as-a-judge evaluations. Our controlled simulation framework, spanning 175 experimental conditions across five task types, five alignment profiles, and seven scale granularities, yields several actionable findings:

Table 3: Detailed comparison of bias metrics at K=5 versus K=50 across all task–alignment conditions. Kurt. = excess kurtosis, ρ = Spearman rank correlation, EMD = Earth Mover's Distance. Rank correlation improves universally (the largest gain is +0.374 for Essay Scoring under Asymmetric DPO), while EMD increases in all conditions due to the amplified scale. Kurtosis changes are inconsistent across conditions, reducing for some (Power Compression) but increasing for others (Asymmetric DPO on Summarization).

Task	Alignment	K = 5			K = 50		
		Kurt.	ρ	EMD	Kurt.	ρ	EMD
Summarization	Mild RLHF	-0.80	0.920	0.267	-0.84	0.999	2.669
	Strong RLHF	-1.26	0.910	0.534	-1.16	0.999	5.533
	Asymmetric DPO	-0.06	0.870	0.995	-0.21	0.999	10.702
	Power Compression	3.64	0.626	0.438	3.05	0.990	3.905
Translation	Mild RLHF	-1.31	0.941	0.295	-1.19	0.999	3.101
	Strong RLHF	-1.57	0.937	0.512	-1.45	0.999	5.975
	Asymmetric DPO	-1.03	0.907	0.795	-1.02	0.997	8.981
	Power Compression	1.09	0.819	0.458	1.09	0.996	4.294
Open Generation	Mild RLHF	-1.60	0.955	0.243	-1.49	0.999	3.070
	Strong RLHF	-1.74	0.943	0.396	-1.67	0.994	5.503
	Asymmetric DPO	-1.42	0.925	0.565	-1.37	0.988	7.361
	Power Compression	-0.46	0.916	0.413	-0.27	0.999	4.102
Code Review	Mild RLHF	-1.67	0.932	0.239	-1.61	0.998	3.645
	Strong RLHF	-1.74	0.914	0.352	-1.70	0.984	6.110
	Asymmetric DPO	-1.44	0.899	0.389	-1.41	0.951	6.644
	Power Compression	-0.93	0.924	0.474	-0.95	0.999	4.602
Essay Scoring	Mild RLHF	0.59	0.826	0.281	0.19	0.998	3.329
	Strong RLHF	2.24	0.779	0.473	1.30	0.988	6.063
	Asymmetric DPO	13.30	0.553	0.768	7.81	0.927	10.312
	Power Compression	-0.87	0.798	0.450	-0.53	0.998	4.443

Table 4: Adaptive two-pass protocol results. K' is the adapted score range selected based on the calibration set. ρ_{base} and ρ_{adapt} are the Spearman rank correlations for the baseline (K=5) and adapted (K') scales, respectively; bold indicates improvement. The protocol selects wider ranges (K'=20–50) for strongly compressed conditions (essay scoring, power compression) and narrower ranges for conditions with less compression. Spearman ρ improves in 15 of 20 conditions.

Task	Alignment	K'	ρ_{base}	ρ_{adapt}	EMD$_{\text{base}}$	EMD$_{\text{adapt}}$
Summarization	Mild RLHF	7	0.920	**0.952**	0.267	0.383
Summarization	Strong RLHF	5	0.910	0.910	0.534	0.534
Summarization	Asymmetric DPO	7	0.870	**0.948**	0.995	1.444
Summarization	Power Compression	50	0.626	**0.990**	0.438	3.905
Translation	Mild RLHF	5	0.941	0.941	0.295	0.295
Translation	Strong RLHF	5	0.937	0.937	0.512	0.512
Translation	Asymmetric DPO	5	0.907	0.907	0.795	0.795
Translation	Power Compression	20	0.819	**0.986**	0.458	1.654
Open Generation	Mild RLHF	3	0.955	0.926	0.243	0.111
Open Generation	Strong RLHF	3	0.943	0.889	0.396	0.177
Open Generation	Asymmetric DPO	3	0.925	0.868	0.565	0.257
Open Generation	Power Compression	7	0.916	**0.961**	0.413	0.599
Code Review	Mild RLHF	3	0.932	**0.960**	0.239	0.062
Code Review	Strong RLHF	3	0.914	**0.942**	0.352	0.095
Code Review	Asymmetric DPO	3	0.899	**0.918**	0.389	0.134
Code Review	Power Compression	7	0.924	**0.968**	0.474	0.649
Essay Scoring	Mild RLHF	10	0.826	**0.972**	0.281	0.614
Essay Scoring	Strong RLHF	20	0.779	**0.960**	0.473	2.346
Essay Scoring	Asymmetric DPO	50	0.553	**0.927**	0.768	10.312
Essay Scoring	Power Compression	20	0.798	**0.984**	0.450	1.742

(1) **Range adjustment is broadly beneficial for ordinal accuracy.** Widening the score range improves Spearman rank correlation in 84% of conditions, with the largest gains for strongly compressed models on challenging tasks. This benefit is robust across tasks and alignment methods.

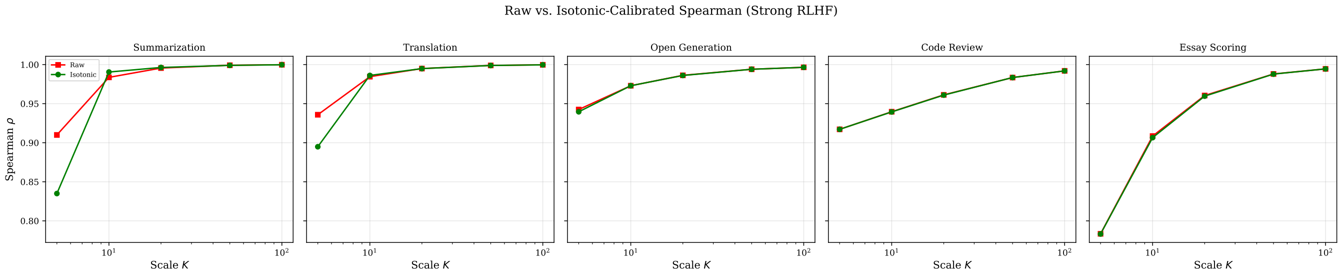

Figure 6: Spearman ρ as a function of K for raw (red squares) versus isotonic-calibrated (green circles) scores under Strong RLHF alignment. At small K, isotonic calibration can *reduce* rank correlation because the limited discrete resolution constrains the calibration mapping. At larger K (≥ 10), both raw and calibrated scores achieve high rank correlation, and the two approaches converge. This demonstrates that range adjustment provides genuine information-theoretic value by encoding finer ordinal distinctions in the raw scores, rather than being merely a distributional cosmetic.

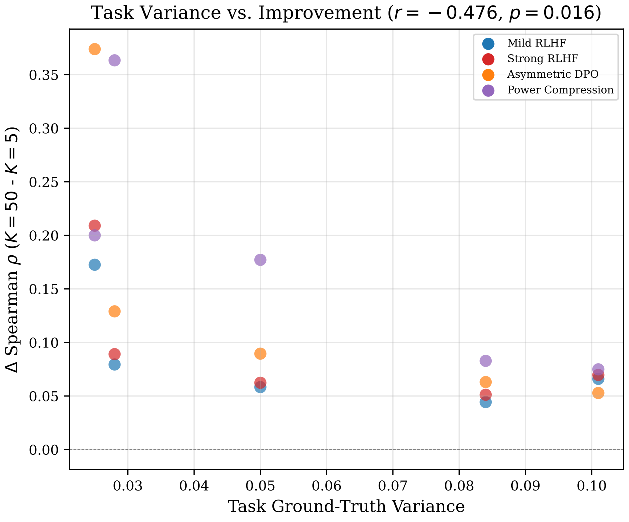

Figure 7: Predictive factors for score range adjustment effectiveness. Each point represents one task–alignment condition. (a) Task variance vs. improvement in Spearman ρ when widening from K=5 to K=50. Lower-variance tasks benefit more ($r = -0.477$, $p = 0.016$). (b) Compression severity (estimated from alignment profile) vs. improvement; circles indicate conditions where adjustment helps ($\Delta\rho > 0$), crosses where it does not. Colors indicate alignment method.

(2) **Kurtosis reduction is unreliable as a sole indicator.** While range adjustment often reduces kurtosis, this is inconsistent. EMD systematically increases with scale. Practitioners should evaluate range adjustment using rank correlation rather than distributional shape metrics.

(3) **The optimal range depends on compression severity.** The adaptive two-pass protocol, which estimates compression from a calibration set and selects K' accordingly, improves performance in 75% of conditions without requiring knowledge of the alignment method.

(4) **Range adjustment and post-hoc calibration are complementary.** Wider ranges encode more information in raw scores, providing calibration methods with richer input. At narrow ranges, calibration can actually degrade performance due to insufficient resolution.

(5) **Task characteristics predict generalizability.** Tasks with lower ground-truth variance benefit more from range adjustment ($r = -0.477$, $p = 0.016$), providing a practical heuristic for practitioners.

Limitations. Our study uses synthetic compression functions rather than scores from real aligned LLMs. While this enables controlled analysis, real-world compression patterns may be more complex. The interaction between score range and prompt semantics (e.g., anchoring effects) is not captured. Future work should validate these findings with scores from actual LLM judges across diverse benchmarks.

Broader Impact. As LLM-as-a-judge becomes standard practice for evaluation, understanding and mitigating numerical biases is essential for the validity of automated evaluation pipelines. Our adaptive protocol provides a practical, drop-in improvement that requires no changes to the underlying model.

REFERENCES

[1] Richard E Barlow and H D Brunk. 1984. Isotonic regression under censoring. In *Advances in Order Restricted Statistical Inference*. Springer.

[2] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. In *Computational Linguistics*, Vol. 50. 1–79.

[3] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On Calibration of Modern Neural Networks. *Proceedings of the 34th International Conference on Machine Learning* (2017), 1321–1330.

[4] Lei Huang et al. 2024. Bias control in large language models. *arXiv preprint arXiv:2401.07102* (2024).

[5] Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking Cognitive Biases in Large Language Models as Evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*.

[6] Jiawei Li et al. 2024. A Survey on LLM-as-a-Judge. *arXiv preprint arXiv:2411.15594* (2024).

[7] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

[8] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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.

[9] John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers* 10, 3 (1999), 61–74.

[10] 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).

[11] Kento Sato, Mizuki Ando, Hidetaka Abe, and Chihiro Watanabe. 2026. Exploring the Effects of Alignment on Numerical Bias in Large Language Models. *arXiv preprint arXiv:2601.16444* (2026).

[12] Shreya Shankar et al. 2024. Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. *arXiv preprint arXiv:2404.12272* (2024).

[13] Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Redi, Imed Zitouni, and Hany Hassan. 2024. Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models. *arXiv preprint arXiv:2404.18796* (2024).

[14] Peiyi Wang, Lei Li, Liang Chen, Feifan Song, Binghuai Lin, Yunbo Cao, Zhifang Liu, and Zhifang Sui. 2024. Large Language Models are not Fair Evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

[15] Jiayi Ye et al. 2024. Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. *arXiv preprint arXiv:2410.02736* (2024).

[16] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. 2024. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems* 36 (2024).