No One Gets an A: Distributional Bias in LLM Text Scoring

Anonymous ACL submission

Abstract

LLMs asked to rate text quality systematically avoid extremes, compressing responses into a narrow interior band despite true quality variation. We measure this *score compression* across eleven models (3.8B–228B parameters) with ∼99,000 evaluations of 150 expert-authored articles degraded along four axes (grammar, coherence, information, lexical sophistication) at five controlled intensities. Compression ratios range from 0.026 (Mistral 7B) to 0.685 (GPT-5 mini), with a positive size–compression correlation among nine open-weight models (Pearson $r = 0.78$, $p = 0.013$). Neither GPT-5 mini nor Gemini 3 Flash assigns 0 or 10 across 9,000 samples each—a pattern shared by four additional models—despite simulations predicting hundreds of such scores. A three-way mixed-effects model confirms significant main effects and all interactions. Across five mitigations on a nine-model comparison, auxiliary regression gives the largest pairwise-accuracy gain ($0.518 \rightarrow 0.702$), expected-score rescaling improves rank correlation (ρ: $0.458 \rightarrow 0.490$), quantile normalization restores the full scale while preserving rank order, and contrastive anchor prompting is mixed. Post-hoc affine calibration reduces RMSE by 10–25%, confirming a systematic, partially correctable distortion with direct implications for any use of LLM-generated quality ratings.

1 Introduction

LLMs are increasingly used as automated text evaluators—grading essays, rating content quality, scoring summaries, and benchmarking generated text (Zheng et al., 2023; Liu et al., 2023; Kocmi and Federmann, 2023)—under the assumption that they map quality to numeric scales with reasonable fidelity (Zheng et al., 2023; Liu et al., 2023). Yet they systematically avoid rating extremes (Zheng et al., 2023; Liu et al., 2023): pristine, publication-quality text receives 7 or 8 rather than 9 or 10, while heavily corrupted text rarely falls below 2 or 3 (Table 1; Figure 2). This *score compression* has direct consequences: compressing $[0, 10]$ to roughly $[2, 9]$ reduces resolution by at least 30% (Zheng et al., 2023; Liu et al., 2023), degrades ranking reliability (Spearman ρ as low as 0.138; Table 2), and miscalibrates threshold decisions (e.g., "pass if score ≥ 7"; Appendix D).

Under a naive linear model, text degraded to fraction λ of its original quality should score $S = 10(1 - \lambda)$. Instead, GPT-5 mini follows $S \approx 8 - 6.5\lambda$ and Gemini 3 Flash follows $S \approx 8 - 6.2\lambda$ (Section 4, Figure 1): the intercept is below 10, the slope is only ∼65% of ideal (Appendix Table 6), and the usable range shrinks by roughly one third.

We call this under-utilisation of numeric rating scales *Distributional Bias*. Prior work has catalogued *selection* and *categorical* biases in LLM judges, including position bias (Zheng et al., 2023; Shi et al., 2025) and verbosity bias (Zheng et al., 2023). Those biases explain *which* item a model prefers (Wang et al., 2024); distributional bias describes *how* the model uses the quantitative scale to express that preference, independent of the item being evaluated—a dimension absent from existing bias taxonomies (Zheng et al., 2023; Wang et al., 2024; Liu et al., 2023). Zheng et al. (2023) note scale-use tendencies but do not quantify them; Liu et al. (2023) observe miscalibration and propose a partial remedy without systematic cross-model replication; Wang et al. (2024) document ranking distortions without measuring compression magnitude. No prior work has quantified compression against an algebraically defined quality gradient, replicated it across eleven models spanning providers and scales, or systematically evaluated mitigations.

Whether distributional bias stems from RLHF training dynamics (Rafailov et al., 2023), the mode-seeking property of reverse KL-divergence (Rafailov et al., 2023), or pretraining

frequency effects remains mechanistically open; the phenomenon itself is empirically tractable and reproduced across eleven diverse models (Section 4.7). We measure it with three design features:

1. **Metamorphic behavioral testing (Ribeiro et al., 2020) with a known ground-truth gradient.** Metamorphic testing probes models through semantically meaningful transformations with known expected output directions, without requiring ground-truth labels (Ribeiro et al., 2020). Rather than compare LLM judgments with noisy human annotations—which include inter-rater disagreement and central-tendency biases (Stevens, 1971)—we use deterministic, algorithmically defined degradations to build a controlled quality gradient (Ribeiro et al., 2020): Level 5 is objectively worse than Level 1, enabling unambiguous compression measurement (Section 3.1).
2. **Psychometric axis isolation (Attali and Burstein, 2006; McNamara et al., 2010).** Psychometric axis isolation decomposes writing quality into independent, theoretically grounded subscores targeting distinct linguistic tiers, so one dimension can be degraded without confounding others (Attali and Burstein, 2006). Following Automated Essay Scoring (AES) frameworks (Attali and Burstein, 2006), we manipulate grammar, coherence, information content, and lexical sophistication—surface syntax, discourse structure, semantics, and vocabulary (McNamara et al., 2010)—to test whether compression is uniform or concentrated in specific cognitive tiers (Section 3.2).
3. **Multi-model evaluation.** Eleven models spanning 3.8B–228B parameters (Appendix Table 4), including frontier API models (GPT-5 mini, Gemini 3 Flash) and open-weight models from Meta, Google, Microsoft, Alibaba, Mistral AI, and MiniMax, show that distributional bias is cross-architectural rather than model-specific (Section 4.7).

We declare three *a priori* primary outcomes: compression ratio per model (Section 4), average regression slope of score on degradation level per model (Section 4), and median paired score difference between models on identical samples (Section 4.6).

Beyond documenting central tendency bias in a new modality (Stevens, 1971; Guilford, 1954), our contributions are: (i) formally characterising *distributional bias* and providing the first *quantified* compression magnitude against an algebraically defined ground truth, whereas prior taxonomies omit this class or treat it as incidental (Zheng et al., 2023; Wang et al., 2024); (ii) cross-model replication across eleven models spanning providers, architectures, and scales (Appendix Table 4); (iii) axis-specific sensitivity profiles revealing systematic insensitivity to coherence-level degradation (Barzilay and Lapata, 2008; Lapata and Barzilay, 2005) (Appendix Figure 4); (iv) a calibration recovery experiment in which a simple affine transform reduces the gap by 10–25% (Appendix D); and (v) a five-strategy mitigation study showing that post-hoc corrections partially recover lost range (Section 5).

2 Related Work

2.1 LLM-as-a-Judge Bias Taxonomy

Since "LLM-as-a-judge" frameworks emerged (Zheng et al., 2023), work has catalogued failures such as positional bias (favouring candidates by order) (Zheng et al., 2023; Shi et al., 2025) and verbosity bias (equating length with quality) (Zheng et al., 2023). Wang et al. (2024) show that rankings are susceptible to positional bias and that LLM judges are generally "unfair evaluators." Li et al. (2025) add "preference leakage," a contamination problem caused by relatedness between the synthetic-data generator and the judge. Stureborg et al. (2024) confirm that LLMs show skewed rating distributions and familiarity bias but do not measure compression against controlled ground truth; Ohi et al. (2024) document likelihood bias in LLM evaluators and propose few-shot mitigation; Asato et al. (2026) show that LLMs have latent numerical token preferences independent of text quality. These studies focus on *selection* and *categorical* biases—which item or which score the model prefers (Zheng et al., 2023; Wang et al., 2024; Shi et al., 2025; Stureborg et al., 2024)—but do not quantify compression against controlled ground truth or evaluate mitigations systematically (Section 5).

2.2 Central Tendency Bias: Human Rating and LLM Outputs

Human raters avoid Likert-scale extremes, compressing distributions around the midpoint (Stevens, 1971); expertise, scale length, and anchoring modulate this effect (Guilford, 1954). AES models trained on human ratings inherit these

properties (Ke and Ng, 2019), and traditional scoring engines use affine scaling to recover variance (Attali and Burstein, 2006). Those studies concern models *fine-tuned* on human-rating datasets (Ke and Ng, 2019); we measure the intrinsic zero-shot calibration of general-purpose instruct models, matching common LLM-as-judge deployments (Zheng et al., 2023; Liu et al., 2023).

2.3 Pointwise Scoring vs. Pairwise Ranking

LLM judges appear stronger at relative pairwise comparison than absolute point-wise scoring (Zheng et al., 2023; Liu et al., 2023). Pointwise scoring compresses a broad distribution of possible human judgments into one deterministic label, encouraging central defaults (Liu et al., 2023; Stevens, 1971). Kocmi and Federmann (2023) show GPT-based metrics achieve state-of-the-art MT correlation without analyzing distributional properties; Liu et al. (2023) find miscalibration and propose expected-score rescaling as a partial remedy (Section 5). Our result that models can have high rank correlations (Spearman ρ up to 0.80) while severe absolute compression remains (CR as low as 0.026; Section 4; Table 2) explains why raw scores should be treated as ordinal rather than interval signals (Stevens, 1971).

2.4 Mechanistic Background

Alignment training offers a plausible compression mechanism; the full KL/RLHF background is provided in Appendix A. The present study focuses on documenting score compression rather than isolating its training-time cause (Section 3).

2.5 Synthetic Perturbation Methodology

We audit LLM judges with a CheckList-style behavioral testing paradigm (Ribeiro et al., 2020), using Directional Expectation (DIR) perturbations in which quality degradation *should* reduce scores, rather than held-out human labels. This avoids annotator bias (Stevens, 1971) and supplies mathematical ground truth: by construction, degradation at $\lambda = 0.8$ is objectively worse than $\lambda = 0.2$, independent of any annotator's judgment (Ribeiro et al., 2020).

3 Experimental Design

3.1 Rationale: Synthetic Perturbations vs. Human Baselines

Human judgment is the traditional NLP-evaluation gold standard, but it suffers from inter-rater disagreement, fatigue, and raters' central-tendency biases (Stevens, 1971). Accuracy tests on human-annotated data can also overestimate model performance when models and labels share corpus shortcuts (Ribeiro et al., 2020). Following CheckList (Ribeiro et al., 2020), we instead use targeted, deterministic perturbations to create a controlled quality gradient. Thus $\lambda = 0.8$ is objectively worse than $\lambda = 0.2$ by construction, giving mathematical ground truth for measuring compression independent of annotator judgment (Ribeiro et al., 2020).

3.2 The Four-Axis Quality Decomposition

Our four degradation axes follow psychometric work on text quality. Classic AES systems such as ETS e-rater decompose writing into surface mechanics, discourse coherence, content, and vocabulary subscores (Attali and Burstein, 2006); linguistic studies likewise show that syntactic, discourse, semantic, and lexical features independently predict writing quality (McNamara et al., 2010). By isolating surface syntax (Grammar), global discourse structure (Coherence), factual completeness (Information), and vocabulary richness (Lexical sophistication), we test both compression magnitude and the linguistic dimensions LLM evaluators weight most. Uniform compression would imply a global numeric calibration failure; axis-specific patterns would reveal qualitative sensitivity gaps (Section 4; Appendix Figure 4). This decomposition is zero-shot and general-purpose, unlike prior AES calibration work on models fine-tuned on human ratings (Ke and Ng, 2019).

3.3 Corpus Construction

The corpus contains 150 articles from *The Conversation* (theconversation.com), where academic experts write evidence-based articles for general audiences under professional editorial review. Articles span Arts & Culture, Business & Economy, Health, Politics & Society, and Science & Technology, and are kept at full length for topical diversity. This source provides an unambiguous quality ceiling for the $\lambda = 0$ baseline: credentialed experts and editorial review avoid the ambiguity of crowdsourced or Wikipedia corpora. The 150-article size

matches NLP metric-evaluation practice (Kocmi and Federmann, 2023; Liu et al., 2023), offering statistical power under API cost constraints.

3.4 Degradation Engine

Each article is degraded along one orthogonal axis at level $\lambda \in \{0.0, 0.2, 0.4, 0.6, 0.8\}$:

- **Grammar** (G): surface errors (typos, agreement, tense, article, homophone, preposition, and word-order errors) as categorized by Attali and Burstein (2006), implemented by a nine-stage pipeline.
- **Coherence** (C): bounded random sentence-order swaps, preserving content and grammar. Sentence shuffling is a standard discourse-coherence benchmark that disrupts entity transitions and rhetorical structure while preserving sentence-level syntax (Barzilay and Lapata, 2008; Lapata and Barzilay, 2005).
- **Information** (I): progressive deletion via a five-phase pipeline targeting parentheticals, modifiers, prepositional phrases, subordinate clauses, and content words (Attali and Burstein, 2006).
- **Lexical** (L): vocabulary flattening by replacing content words with high-frequency synonyms using embedding similarity, following lexical simplification methodology (Glavaš and Štajner, 2015).

At $\lambda = 0$, the original text is returned. We cap intensity at $\lambda = 0.8$ rather than 1.0 because full intensity on some axes creates incoherent output that may trigger refusal rather than scoring. All functions are deterministic via per-sample seeds: $\text{MD5}(\text{title}\|\text{axis}\|\text{level}\|\text{rep}) \bmod 2^{31}$.

3.5 Factorial Design

The experiment uses a $150 \times 4 \times 5 \times 3$ factorial design:

$$N = |\mathcal{C}| \times |\mathcal{A}| \times |\mathcal{L}| \times R = 150 \times 4 \times 5 \times 3 = 9{,}000 \quad (1)$$

Five stepped intensities follow synthetic perturbation practice (Ribeiro et al., 2020) and allow linear and non-linear sensitivity detection. $R = 3$ satisfies the minimum for ICC(1,1) estimation (Appendix Table 8) and measures intra-model variance, yielding 9,000 samples per model and 99,000 evaluations across 11 models (Appendix Table 4).

3.6 Scoring Protocol

Each sample is scored in isolation with: "Rate the quality of the following text from 0 to 10. Respond with ONLY the number." No rubric, examples, or reference texts are provided; this measures intrinsic calibration on an unanchored zero-shot scale (Zheng et al., 2023), as commonly deployed, leaving elaborate prompts (e.g., chain-of-thought or rubric scoring (Liu et al., 2023)) for future work.

Appendix Table 4 lists the full model configurations.

GPT-5 mini is a reasoning model with no user-set temperature: passing `temperature` returns *"Unsupported parameter: 'temperature' is not supported with this model."* Its sampling is therefore fixed at an internal default.[1] Gemini 3 Flash runs at temperature 0.0; the nine other models also use temperature 0.0 and return token log-probabilities, enabling logprob-based mitigations (Section 5).

4 Results

Results are organized around the three primary outcomes, followed by secondary analyses. All tables order models by descending compression ratio.

4.1 Score Distributions and Boundary Avoidance

Most models use only a narrow subset of the 0–10 scale (Table 1). GPT-5 mini ($[1, 9]$), Gemini 3 Flash ($[2, 9]$), Gemma 2 9B, Llama 3.1 8B, Phi-4 14B, and Qwen 2.5 14B never assign 0 or 10 across 9,000 samples each. Among models that do, MiniMax M2P1 assigns 686 tens (7.7%) and GPT-OSS 120B assigns 394 zeros (4.4%); Mistral 7B, Phi-4 mini, and Qwen 2.5 7B produce fewer than 25 extremes each. Means range from 4.23 (GPT-OSS 120B) to 8.62 (Qwen 2.5 7B), with smaller models clustering at higher means and lower SDs, consistent with stronger compression (Table 1; Section 4.7). Figure 2 visualizes all eleven distributions.

To test whether missing extremes are a design artifact, we fit a linear model per model, simulate 100 datasets with Gaussian noise at observed residual SD, and count predicted boundary scores. For the six boundary-free models, simulations predict many extremes: GPT-5 mini should produce ∼542 zeros and ∼122 tens; Gemini 3 Flash should produce ∼73 zeros and ∼147 tens (Table 1). The four other boundary-free models show the same pattern.

[1] Empirically, majority-vote aggregation over three repetitions yields near-identical mean scores (mean absolute deviation from single-sample mean < 0.05), confirming that stochasticity does not materially affect results (Appendix Table 8).

Even models with extremes produce far fewer than predicted (e.g., Mistral: 5 tens vs. ∼407 predicted), confirming active boundary avoidance (Figure 2; Appendix Figure 10).

Table 1: Descriptive statistics ordered by compression ratio. Three models returned <9,000 scores due to unparseable responses: MiniMax M2P1 ($n = 8{,}933$), Phi-4 mini ($n = 8{,}991$), Qwen 2.5 7B ($n = 8{,}998$).

Model	n	Mean	SD	Range	Skew	#0s	#10s
GPT-5 mini	9,000	4.32	2.46	1–9	0.57	0	0
Gemini 3 Flash	9,000	5.21	2.15	2–9	0.43	0	0
GPT-OSS 120B	9,000	4.23	2.25	0–9	0.35	394	0
Qwen 2.5 14B	9,000	5.89	2.24	2–8	−0.43	0	0
MiniMax M2P1	8,933	5.23	2.67	0–10	−0.01	8	686
Gemma 2 9B	9,000	6.91	1.30	3–9	−0.91	0	0
Phi-4 14B	9,000	7.27	0.97	3–8	−1.49	0	0
Llama 3.1 8B	9,000	7.92	1.12	2–9	−1.85	0	0
Phi-4 mini	8,991	8.34	0.91	0–10	−1.88	15	3
Qwen 2.5 7B	8,998	8.62	0.52	6–10	−0.75	0	21
Mistral 7B	9,000	8.58	0.55	5–10	−1.05	0	5

4.2 Primary Outcome 1: Compression Ratio

Under a naive linear scorer mapping λ to $S = 10(1-\lambda)$, the expected score span between $\lambda = 0$ and $\lambda = 0.8$ is 8 points. The compression ratio is:

$$CR = \frac{\bar{S}(\lambda = 0) - \bar{S}(\lambda = 0.8)}{10(1-0) - 10(1-0.8)} = \frac{\bar{S}(0) - \bar{S}(0.8)}{8} \tag{2}$$

$CR = 1$ indicates no compression; $CR < 1$ indicates compression.

Table 2: Compression ratio (CR) with 95% block-bootstrap CIs (10,000 iterations) and Spearman ρ for all eleven models. All CIs exclude 1.0.

Model	Params	$\widehat{CR}$	95% CI	Spearman ρ
GPT-5 mini	—	0.685	[0.667, 0.700]	0.729
Gemini 3 Flash	—	0.620	[0.603, 0.635]	0.801
GPT-OSS 120B	120B	0.504	[0.470, 0.537]	0.617
Qwen 2.5 14B	14B	0.446	[0.431, 0.461]	0.618
MiniMax M2P1	228B	0.420	[0.392, 0.447]	0.473
Gemma 2 9B	9B	0.299	[0.288, 0.311]	0.704
Phi-4 14B	14B	0.177	[0.168, 0.187]	0.548
Llama 3.1 8B	8B	0.157	[0.148, 0.166]	0.464
Phi-4 mini	3.8B	0.078	[0.067, 0.089]	0.239
Qwen 2.5 7B	7B	0.056	[0.049, 0.064]	0.324
Mistral 7B	7B	0.026	[0.018, 0.034]	0.138

All eleven models compress below ideal (Table 2), with CRs from 0.026 (Mistral 7B) to 0.685 (GPT-5 mini); all CIs exclude 1.0. Sub-10B models all have CR < 0.30, while models above 100B reach 0.42 or higher. CR and rank correlation are dissociated: Gemma 2 9B reaches Spearman $\rho = 0.704$ despite CR 0.299, so models can preserve relative ordering within a compressed absolute range. This matches prior evidence that LLMs are more reliable as relative rankers than absolute point-wise scorers, whose scores concentrate and lose resolution (Zheng et al., 2023; Liu et al., 2023).

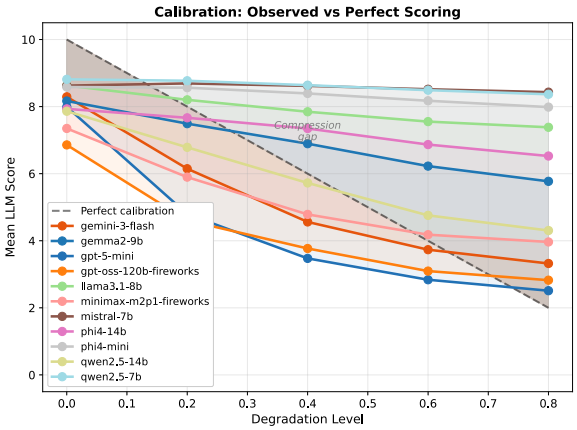


Figure 1: Calibration plot: observed mean scores vs. the naive reference $S = 10(1-\lambda)$. Shaded area represents the compression gap.

Figure 1 shows the compression gap against the calibration reference.

4.3 Primary Outcome 2: Sensitivity (Regression Slopes)

Linear regression of score on degradation level quantifies sensitivity (Appendix Figure 4; full values in Appendix Table 6). Across all models, 43 of 44 slopes are significantly negative ($p < 0.001$); the exception is Mistral 7B on lexical degradation ($\beta_1 = -0.03$, $p = 0.49$). Information deletion is steepest for GPT-5 mini ($\beta_1 = -7.87$) and Gemini 3 Flash (-7.05); coherence is shallowest for GPT-5 mini (-4.78), lexical for Gemini 3 Flash (-5.26). Mean slopes range from -6.46 (GPT-5 mini) to -0.28 (Mistral 7B), versus ideal -10; only the top API models reach two-thirds of ideal sensitivity, and nine of eleven reach less than half.

The shallow coherence slopes are notable: random sentence shuffling destroys argumentative flow and entity-grid coherence transitions (Barzilay and Lapata, 2008) while preserving surface features, yet models penalise it less than comparable grammatical or informational degradation. Point-wise LLM evaluation therefore appears to rely more on local fluency than global discourse structure (Barzilay and Lapata, 2008; Lapata and Barzilay, 2005).

4.4 Robustness Checks

Robustness analyses confirm monotonic but mildly concave score–degradation relationships, high intra-condition reliability, large undegraded-vs.-maximally-degraded effects, negligible length effects, significant category effects, and robustness to Benjamini–Hochberg correction. Details, includ-

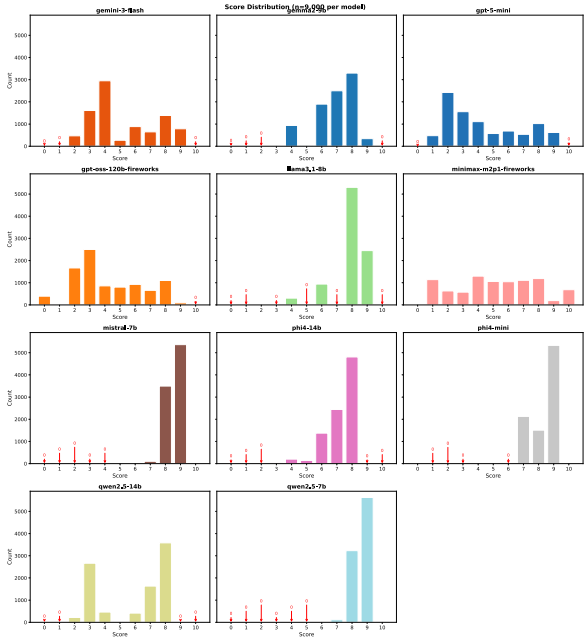


Figure 2: Score distributions for all eleven models ($n \approx 9{,}000$ per model). Red "0" annotations mark empty bins; bars without annotation have small non-zero counts. Six models never assign 0 or 10; five produce rare or minority extreme scores (Table 1).

ing variance diagnostics, ICCs, effect sizes, covariate tests, category effects, and FDR results, are reported in Appendix C, especially Appendix C.1 and the following robustness subsections.

4.5 Mixed-Effects Model

For GPT-5 mini and Gemini 3 Flash, a three-way mixed-effects model with article-level random intercepts and slopes confirms a large negative degradation effect, significant axis effects, significant Level×Axis×Model interactions, and non-trivial article-level random effects. Full specification and model-selection rationale are in Appendix C.2; coefficients are in Appendix H and Appendix Table 5.

4.6 Primary Outcome 3: Inter-Model Comparison

Despite high agreement (Pearson $r = 0.836$, Spearman $\rho = 0.817$, quadratic weighted $\kappa = 0.772$), Gemini 3 Flash scores systematically higher than GPT-5 mini by a median of $+1$ point ($p < 10^{-300}$, rank-biserial $r = +0.705$). This offset is present across all four axes: largest on grammar, information, and lexical ($+1$); smallest on coherence (0). All per-axis Wilcoxon tests are significant after Benjamini-Hochberg correction (Appendix Table 6).

Figures 6–7 show the agreement scatter and paired-difference distribution.

4.7 Scaling and Generalization Across Models

Across models from 3.8B to 228B parameters under the same minimal prompt, CR ranges from 0.026 (Mistral 7B) to 0.685 (GPT-5 mini). Among nine open-weight models with known sizes, $\log_{10}$ parameter count correlates positively with CR (Pearson $r = 0.78$, $p = 0.013$), but non-monotonic exceptions show that size alone is insufficient. Full scaling details are in Appendix E.

5 Mitigation Strategies & Results

We evaluate five strategies targeting different aspects of compression: quantile normalization, expected-score rescaling, asymmetry correction, auxiliary regression, and contrastive anchor prompting. Definitions and formulas are in Appendix G.

Table 3: Mitigation comparison across nine logprob-capable models unless noted. RU = range utilization; PA = pairwise accuracy.

Method	RU	ρ	PA	Notes
Baseline (argmax)	0.689	0.458	0.518	—
Expected score	0.579	0.490	0.649	$\Delta\rho = +0.032$
Asymmetry correction	0.579	0.490	0.649	$\alpha = 0$ selected
Auxiliary regression	0.919	0.496	0.702	supervised
Quantile uniform	0.922	0.458	0.518	rank preserved
Quantile beta	0.745	0.458	0.518	rank preserved
Contrastive anchors†	—	mixed	—	GPT: -0.110; Gemini: $+0.035$

†Evaluated on $n = 200$ samples per API model.

Results. Post-hoc corrections help modestly but do not solve compression. Quantile normalization restores range (RU: 0.689 → 0.922) while exactly preserving rank order (Bolstad et al., 2003). Expected-score rescaling improves rank correlation (ρ : 0.458 → 0.490) and pairwise accuracy (PA: 0.518 → 0.649) (Liu et al., 2023), but shrinks usable range (RU → 0.579) because the already-central logit distribution pulls expectations away from endpoints (Rafailov et al., 2023). Auxiliary regression gives the largest gains ($\rho \rightarrow 0.496$, PA → 0.702; Table 3) but requires supervised data. Contrastive anchor prompting is mixed on the two API models ($n = 200$ each): Gemini improves ($\Delta\rho = +0.035$) while GPT-5 mini degrades ($\Delta\rho = -0.110$), suggesting prompt engineering cannot reliably override deep distributional tendencies (Rafailov et al., 2023). The maximum rank gain ($\Delta\rho = +0.038$, Table 3) indicates that rank-order information is largely fixed at inference.

6 Discussion

6.1 The Compression Effect

Score compression is robust across all eleven models. GPT-5 mini and Gemini 3 Flash use only 62–69% of the ideal range and never produce boundary scores; CRs span 0.026–0.685 (Table 2, Section 4). Calibrated simulations show this is not a design artifact: equivalent linear scorers would produce hundreds of extreme scores (Section 4; Table 1). Compression is asymmetric and concave: pristine texts average 7.3–8.7 rather than 10 (Figure 8; Table 1), maximally degraded texts remain above the expected floor, and 34 of 44 axis×model regressions have significant quadratic terms (Appendix Table 9, Figure 12). Axis sensitivity reveals a psychometric blind spot: models penalise information loss and grammar more than discourse coherence (Appendix Figure 4; Appendix Table 6). Sentence shuffling destroys argumentative structure and entity transitions (Barzilay and Lapata, 2008), yet scores change less than for sentence-level perturbations, suggesting point-wise LLM evaluation overweights local token fluency relative to document-level reasoning (Lapata and Barzilay, 2005).

6.2 Absolute Scoring vs. Rank Ordering

High rank correlations (Spearman ρ up to 0.80; Table 2) coexist with severe absolute compression (CR as low as 0.026; Sections 4–4.7). Models perceive the quality gradient and often order degraded texts below intact ones, but their numeric translation is unreliable. Without relative anchors, they default toward the centre of the learned output distribution, consistent with reverse-KL mode seeking in RLHF (Rafailov et al., 2023). This matches evidence that LLMs are better pairwise rankers than absolute scorers (Zheng et al., 2023; Liu et al., 2023) and explains why post-hoc rank corrections are small ($\Delta\rho \leq 0.038$; Table 3): rank information is mostly present, but absolute calibration is lost and no monotonic rescaling can fully recover it (Appendix D).

6.3 Axis-Specific Sensitivity

Most models share an axis ordering: information deletion is steepest, while coherence or lexical disruption is shallowest (Appendix Table 6, Figure 9), with exceptions such as Llama 3.1 8B peaking on grammar. Low coherence sensitivity matters because sentence shuffling destroys argumentative flow and entity transitions (Barzilay and Lapata, 2008) while preserving surface features, again suggesting overweighting of sentence-level quality over global discourse structure (Lapata and Barzilay, 2005).

6.4 Systematic Inter-Model Bias

Despite high agreement ($r = 0.836$; Figure 6), Gemini 3 Flash scores one point higher than GPT-5 mini: a systematic offset (rank-biserial $r = 0.705$; Gemini higher on 85% of disagreements; Section 4.6, Figure 7). Provider scores are not interchangeable; a threshold calibrated on GPT will fail systematically on Gemini (Section 4.6; Table 2), analogous to inter-rater scale-use differences in human studies (Guilford, 1954).

Additional discussion of linearity, reliability, and calibration-related prior work is provided in Appendix F.

6.5 Implications

1. **Score calibration is necessary.** Raw LLM scores should not be treated as ground truth (Liu et al., 2023; Zheng et al., 2023); quantile normalization is a simple range-recovery method (Bolstad et al., 2003) (Section 5).
2. **Model-specific thresholds.** A 7 from GPT-5 mini is not equivalent to a 7 from Gemini 3 Flash (Section 4.6; Figure 7).
3. **Ordinal, not interval.** Score spacing is non-uniform (Stevens, 1971); effective endpoints differ from nominal ones (Table 1).
4. **Axis awareness.** LLMs are less sensitive to coherence-level quality variation (Appendix Figure 4; Appendix Table 6).
5. **Post-hoc corrections help modestly.** Maximum $\Delta\rho = +0.038$ (Table 3) shows rankings are largely fixed at inference time (Rafailov et al., 2023).
6. **Prompt-level interventions are unreliable.** Contrastive anchoring is inconsistent (Section 5) and not robust (Rafailov et al., 2023).
7. **Model scale matters noisily.** Larger models tend to compress less (Section 4.7), but exceptions exist (Table 2); size alone is insufficient.

7 Conclusion

Score compression is robust across eleven models (3.8B–228B parameters, multiple providers) and ∼99,000 evaluations (Sections 4–4.7). CRs range from 0.026 to 0.685 and correlate positively with model size ($r = 0.78$; Table 2, Section 4.7),

with notable exceptions. GPT-5 mini and Gemini 3 Flash compress the effective range to 62–69% of ideal (Section 4; Table 2).

Across five mitigations (Section 5), quantile normalization (Bolstad et al., 2003) restores range, expected-score rescaling (Liu et al., 2023) modestly improves rank correlation, auxiliary regression gives the best pairwise-accuracy gains, and contrastive anchor prompting (Zheng et al., 2023; Liu et al., 2023) is inconsistent; maximum $\Delta\rho = +0.038$ (Table 3) shows rankings are largely baked in at inference (Rafailov et al., 2023).

This is the first study to quantify distributional bias against an algebraically defined ground truth across eleven models, four psychometric axes (Attali and Burstein, 2006; McNamara et al., 2010), and five mitigations. The title captures the finding: no one gets an A—LLMs resist top marks and failing grades (Table 1), funneling scores into a safe middle range consistent with mode-seeking alignment training (Rafailov et al., 2023). Correcting this bias is essential for any application using LLM-generated quality ratings (Zheng et al., 2023; Liu et al., 2023).

Reproducibility. The complete codebase, degradation engine, scoring scripts, analysis pipeline, and all raw score data are publicly available at `https://anonymous.4open.science/r/Distributional-Bias-in-LLM-Text-Scoring-B8B1/README.md`. All experiments are deterministically reproducible via MD5-seeded randomness.

Limitations

First, although degradation axes are designed to be orthogonal (Section 3.2), residual correlations may remain (e.g., information deletion can affect coherence (Barzilay and Lapata, 2008)). Second, the corpus is English expository text from one source; other languages, genres, and domains require study (McNamara et al., 2010). Third, the broader model landscape, especially commercial models without logprob access, may vary further (Section 5; Appendix Table 4).

Fourth, the temperature asymmetry is a confound: GPT-5 mini does not support user-set temperature, while Gemini 3 Flash runs at 0.0. However, a temperature explanation predicts GPT should compress *more* than Gemini, opposite the observed CRs (0.685 vs. 0.620; Table 2). GPT's ICC remains excellent at 0.926 (Appendix Table 8), and majority-vote means are nearly identical (mean absolute deviation < 0.05). Future work should compare matched temperatures where APIs permit.

Fifth, logprob mitigations cover only the nine models exposing token probabilities, and contrastive prompting only two API models; controlled matching-model comparisons address this (Table 3). Sixth, the reference $S = 10(1 - \lambda)$ assumes linear quality degradation: a mathematical convenience, not a perceptual claim. Boundary-score absence (Section 4; Table 1) is independent of this assumption; under concave ground truth, our CRs would be understated, so the benchmark is conservative.

Seventh, the minimal scoring prompt measures intrinsic calibration but leaves chain-of-thought, decomposed evaluation, and fine-grained rubrics (Liu et al., 2023) for future work. The contrastive anchor experiment (Section 5) suggests the bias is not purely prompt-driven (Rafailov et al., 2023). Eighth, we study zero-shot general-purpose instruct models (Zheng et al., 2023); fine-tuning for evaluation may align score distributions with human ratings (Ke and Ng, 2019), but whether it eliminates or merely attenuates the bias, and at what cost to generality, remains open.

Ethical Considerations

Score compression affects high-stakes LLM deployment in essay grading, recruitment screening, and content moderation (Zheng et al., 2023). Naive thresholds are miscalibrated (Section 4 and Appendix D): requiring 9 for a scholarship or 8 for a resume screen can reject qualified candidates through ceiling compression (Table 1), while floor compression can shield poor submissions from remediation.

Inter-model bias compounds this (Section 4.6): Gemini 3 Flash scores identical texts about one point higher than GPT-5 mini (Figure 7; Table 2). Switching or mixing LLM evaluators can create systematic, opaque inequities (Guilford, 1954). Post-hoc mitigations (Section 5) only partly recover calibration and cannot retrieve signals the model missed (Appendix D). LLM scores are at best ordinal (Stevens, 1971); automated decisions require context-specific validation against domain-expert baselines (Ke and Ng, 2019).

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment*, 4(3):1–21.

Yuma Asato, Kiyoaki Shirai, and Natthawut Kertkeidkachorn. 2026. Mitigating scoring bias in LLM-as-a-judge via random number generation. *arXiv preprint arXiv:2608.05726*. https://arxiv.org/abs/2608.05726.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Benjamin M. Bolstad, Rafael A. Irizarry, Magnus Åstrand, and Terence P. Speed. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193. https://doi.org/10.1093/bioinformatics/19.2.185.

Joy Paul Guilford. 1954. *Psychometric Methods*. McGraw-Hill, New York, 2nd edition.

Zihao Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 6300–6308. https://doi.org/10.24963/ijcai.2019/879.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203. https://aclanthology.org/2023.eamt-1.19.

Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 1085–1090.

Goran Glavaš and Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 63–68.

Dawei Li, Renliang Sun, et al. 2025. Preference leakage: A contamination problem in LLM-as-a-judge. *arXiv preprint arXiv:2502.01534*. https://arxiv.org/abs/2502.01534.

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*, pages 2511–2522. https://doi.org/10.18653/v1/2023.emnlp-main.153.

Danielle S. McNamara, Scott A. Crossley, and Philip M. McCarthy. 2010. Linguistic features of writing quality. *Written Communication*, 27(1):57–86.

Masanari Ohi, Masahiro Kaneko, Ryuto Koike, Mengsay Loem, and Naoaki Okazaki. 2024. Likelihood-based mitigation of evaluation bias in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12688–12701. https://aclanthology.org/2024.findings-acl.193.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 296–319. https://aclanthology.org/2025.ijcnlp-long.18/.

Stanley S. Stevens. 1971. Issues in psychophysical measurement. *Psychological Review*, 78(5):426–450. https://doi.org/10.1037/h0031324.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*. https://arxiv.org/abs/2405.01724.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu 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*, pages 9440–9450. https://aclanthology.org/2024.acl-long.511/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36. https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

A Mechanistic Background Moved from Related Work

Alignment training offers a plausible score-compression mechanism (Rafailov et al., 2023). In RLHF, the policy-gradient objective includes a KL

term that keeps the fine-tuned policy π_θ near the reference distribution π_{ref} (Rafailov et al., 2023):

$$\mathcal{L}(\theta) = \mathbb{E}[r(x, y)] - \beta\, \mathrm{KL}(\pi_\theta(y|x) \,\|\, \pi_{\text{ref}}(y|x)) \tag{3}$$

This regulariser penalises high-variance outputs and discourages extreme predictions far from the training mass (Rafailov et al., 2023). Because the reverse KL used in standard RLHF is mode-seeking, it concentrates probability on likely reference-policy outputs and can amplify any central tendency in π_{ref} (Rafailov et al., 2023). Since human preference datasets themselves rarely contain extreme ratings (Stevens, 1971), the interior of the scale becomes the safe aligned region. Rafailov et al. (2023) show that Direct Preference Optimization (DPO) implicitly reparameterises this objective; recent analysis suggests DPO can also contribute to distribution compression through mode-seeking behaviour (Rafailov et al., 2023). This logit-level effect can compound RLHF distributional regularisation, suppressing probability mass on extreme outputs (Rafailov et al., 2023). Whether it is the primary cause is open; our design documents the phenomenon rather than isolating mechanism (Section 3).

B Model Configurations

Table 4: Models and configurations. All eleven models evaluated at full factorial scale (9,000 samples each).

Model	Provider	Params	Access	Temp.	Logprobs
GPT-5 mini	OpenAI	—	API	n/a	No
Gemini 3 Flash	Google	—	API	0.0	No
GPT-OSS 120B	Fireworks	120B	API	0.0	Yes
MiniMax M2P1	Fireworks	228B	API	0.0	Yes
Gemma 2 9B	Local (Ollama)	9B	Local	0.0	Yes
Llama 3.1 8B	Local (Ollama)	8B	Local	0.0	Yes
Mistral 7B	Local (Ollama)	7B	Local	0.0	Yes
Phi-4 14B	Local (Ollama)	14B	Local	0.0	Yes
Phi-4 mini	Local (Ollama)	3.8B	Local	0.0	Yes
Qwen 2.5 14B	Local (Ollama)	14B	Local	0.0	Yes
Qwen 2.5 7B	Local (Ollama)	7B	Local	0.0	Yes

C Additional Analyses Moved from Main Text

The following analyses are reported here to reduce main-text length while preserving the full analysis.

C.1 Secondary Diagnostics

Secondary diagnostics confirm that score–degradation relationships are strictly monotonic but mildly concave, peaking in sensitivity at low degradation (Figure 5). Score variance exhibits a heteroscedastic "fan-out-then-in" pattern—peaking at intermediate degradation levels (Figure 14; Figure 15)—consistent with the hard ceiling and floor compression effects described above. Detailed non-linearity F-tests (Appendix Table 9), Kendall's τ rank correlations (Appendix Table 10), and variance diagnostics (Figure 12) are provided in Appendix I.

C.2 Mixed-Effects Model

The next two subsections focus on GPT-5 mini and Gemini 3 Flash—the two highest-performing frontier models—selected because they represent distinct API providers and both achieve CR > 0.60 and Spearman $\rho > 0.70$ (Table 2). The three-way specification requires exactly two model levels; inter-model agreement metrics are inherently pairwise.

The model

$$S_{ijk} \sim \lambda \times \text{Axis} \times \text{Model} + (1 + \lambda \mid \text{Article}) \tag{4}$$

converged with random intercepts and slopes. Key findings (full coefficients in Appendix Table 5):

- Degradation level: large negative main effect ($\beta = -4.76$, $z = -45.6$, $p < 10^{-300}$).
- Model intercept (Gemini vs. GPT): non-significant ($\beta = +0.07$, $p = 0.25$), indicating comparable baseline scoring at $\lambda = 0$. The significant $+1$ point median offset (Section 4.6) emerges from interaction terms as degradation increases.
- All axis main effects significant; grammar shows the largest intercept reduction ($\beta = -1.49$, $z = -25.1$).
- All three-way Level $\times$ Axis $\times$ Model interactions significant. Largest: Level $\times$ Lexical $\times$ Gemini ($\beta = +1.93$, $z = +11.2$), confirming Gemini is substantially less sensitive to lexical degradation (Appendix Table 6).
- Random effects: article-level intercept variance $= 0.54$; slope variance $= 0.49$ (Appendix Table 5).

C.3 Intra-Condition Reliability

ICC(1,1) across three repetitions per condition indicates high consistency across all eleven models (Appendix Table 8). GPT-5 mini (ICC $= 0.926$) and Gemini 3 Flash (ICC $= 0.905$) are highest; ten of eleven models achieve ICC > 0.70. The exception is MiniMax M2P1 (ICC $= 0.433$). Grammar yields the highest reliability for the top four models; Information or Lexical degradation is most reliable for the rest (Figure 11; Appendix Table 8).

C.4 Effect Sizes

Within-model effect sizes for the undegraded-vs.-maximally-degraded contrast are extremely large: Cliff's $\delta = 0.986$ (GPT) and 0.984 (Gemini), with Cohen's $d = 4.65$ and 5.33 (Table 1; Table 2). Between-model effect sizes (Gemini − GPT) peak at $\lambda = 0.2$ (Cohen's $d = 0.81$, rank-biserial $r = 0.79$; Figure 6; Figure 7).

C.5 Covariates and Confounds

Article length. A z-scored length covariate is non-significant for GPT-5 mini ($\beta = -0.058$, $p = 0.21$) but small and significant for Gemini 3 Flash ($\beta = -0.151$, $p = 0.002$), with negligible effect size. This is consistent with prior work noting that verbosity signals can modulate LLM evaluation outputs (Zheng et al., 2023).

Category effects. Kruskal-Wallis tests on undegraded scores reveal significant category differences for both models (GPT: $H = 202.6$, $p < 10^{-42}$; Gemini: $H = 377.1$, $p < 10^{-80}$). Both models score Science & Technology highest (GPT: 8.44; Gemini: 8.73) and Politics & Society lowest (GPT: 7.58; Gemini: 7.88), despite all articles being expert-authored pieces of comparable editorial quality (Figure 3; Figure 8).

C.6 Multiple Testing Correction

Benjamini-Hochberg FDR correction across all 49 p-values leaves 47 significant at $\alpha = 0.05$. The two exceptions are the article-length covariate for GPT-5 mini ($p_{\text{adj}} = 0.21$; Appendix C.1) and Mistral 7B's lexical slope ($p_{\text{adj}} = 0.49$; Appendix Table 6).

D Calibration Recovery

We test post-hoc correctability by splitting 150 articles 80/20 (training: 120 articles, 7,200 samples; test: 30 articles, 1,800 samples) and fitting affine, four-parameter sigmoid, and isotonic calibrations on training data.

Affine calibration reduces RMSE by 25% for GPT-5 mini ($2.51 \rightarrow 1.88$) and 10% for Gemini 3 Flash ($1.94 \rightarrow 1.74$) (Table 7). Sigmoid and isotonic variants add only marginal improvement, showing that compression is approximately linear and that residual RMSE of ∼1.7–1.9 points is an irreducible floor from imperfect rank discrimination—no monotonic transform can recover it (Liu et al., 2023). Rank correlations are unchanged by all calibrations, as expected (Figure 13).

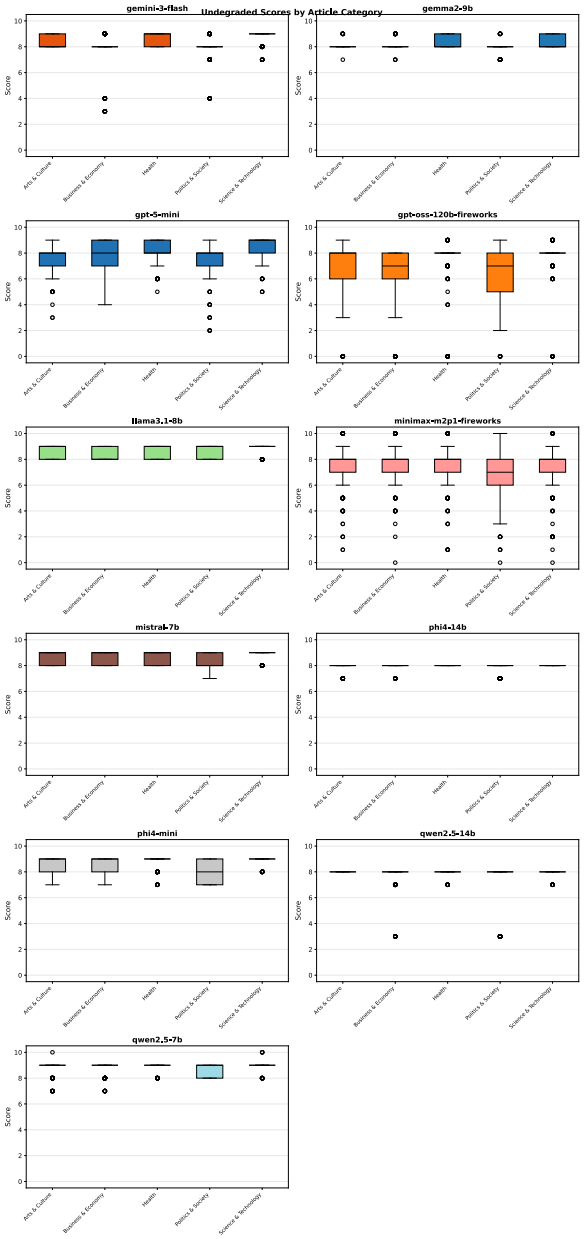


Figure 3: Undegraded scores by article category. Both models score Science & Technology highest and Politics & Society lowest despite comparable editorial quality across categories.

Thus affine correction is practical and cheap, but the RMSE floor shows compression is not merely an offset: models lose discriminative information that rescaling cannot restore (Section 6).

E Full Scaling and Generalization Analysis

Models span 3.8B (Phi-4 mini) to 228B (MiniMax M2P1) parameters and use the identical minimal prompt (Appendix Table 4).

E.1 Compression Across Models

CRs range from 0.026 (Mistral 7B) to 0.685 (GPT-5 mini) (Table 2). Sub-10B models all have CR < 0.30; models above 100B all have CR ≥ 0.42 (Table 2).

E.2 Relationship Between Model Size and Compression

Among nine open-weight models with known sizes, $\log_{10}(\text{parameters})$ correlates positively with CR (Pearson $r = 0.78$, $p = 0.013$; Table 2). The trend is noisy ($n = 9$) and non-monotonic: MiniMax M2P1 (228B) has lower CR (0.42) than Qwen 2.5 14B (0.45), and Gemma 2 9B reaches Spearman $\rho = 0.70$ despite CR 0.30. Thus size alone does not determine compression; architecture, training data, and RLHF procedures (Rafailov et al., 2023) likely contribute. We treat this as a trend, not a scaling law.

F Additional Discussion Details

F.1 Linearity as Approximation

The linear reference $S = 10(1 - \lambda)$ is interpretable despite concave dose-response curves (Figure 5). Boundary absence is model-free (Table 1); affine, sigmoid, and isotonic calibrations converge to similar curves (Appendix D, Figure 13); and if true perceptual quality is concave, our CRs understate compression, making the benchmark conservative.

F.2 Reliability

Both API models have ICC > 0.90, and ten of eleven models exceed 0.70 (Appendix Table 8). Compression is therefore a stable property of the scoring function, not noise, and GPT-5 mini's fixed sampling behaviour does not add material variance relative to temperature-0.0 models (Appendix Table 8).

F.3 Relationship to Prior Work on Calibration

Supervised AES studies show task-specific fine-tuned models can match human score distributions (Ke and Ng, 2019); rubric or binary-like LLM evaluation reports smaller gaps between absolute scoring and relative ranking (Liu et al., 2023). These results are compatible with ours because they use specialist models or constrained scales requiring supervision or prompt engineering (Ke and Ng, 2019; Liu et al., 2023). We study intrinsic zero-shot calibration of general-purpose instruct models, the common LLM-as-judge setting (Zheng et al., 2023). That compression can be reduced under ideal conditions (Ke and Ng, 2019) makes its persistence in uncalibrated deployment more important.

G Mitigation Formulas

G.1 Strategy Definitions

We evaluate five strategies targeting different aspects of compression; the corresponding formulas are given below:

- **Quantile Normalization:** maps the empirical distribution to Uniform$[0, 10]$ or Beta$(2, 2)$ by quantile matching (Bolstad et al., 2003), preserving rank order.
- **Expected-Score Rescaling:** replaces argmax with the probability-weighted expected score $\mathbb{E}[S]$ (Liu et al., 2023).
- **Asymmetry Correction:** adjusts $\mathbb{E}[S]$ by a skewness penalty (Liu et al., 2023); cross-validation selects $\alpha = 0.0$, reducing it to expected-score rescaling.
- **Auxiliary Regression:** trains supervised regressors on logprob-derived features to predict proxy ground truth.
- **Contrastive Anchor Prompting:** prepends explicit scale anchors (e.g., "0 = unreadable garbage") to the prompt (Zheng et al., 2023; Liu et al., 2023).

G.2 Formula Details

Expected-Score Rescaling. For models returning token-level log-probabilities, the argmax score is replaced by the probability-weighted expected value (Liu et al., 2023):

$$\hat{S} = \mathbb{E}[S] = \sum_{k=0}^{10} k \cdot p(k) \tag{5}$$

where $p(k)$ is the softmax probability of score token k.

Asymmetry Correction. An extension of expected-score rescaling (Liu et al., 2023) correcting for distribution skewness (Stevens, 1971):

$$\hat{S}_{\text{adj}} = \mathbb{E}[S] - \alpha \cdot \text{skewness}(p) \tag{6}$$

with α tuned via five-fold cross-validation. In our experiments, the optimal α was 0.0 for all models, reducing this method entirely to expected-score rescaling (Liu et al., 2023).

H Mixed-Effects Model: Key Coefficients

Table 5 reports all fixed-effect coefficients from the three-way model described in Section 4.5. Reference levels are the Grammar axis and GPT-5 mini. All three-way Level × Axis × Model interactions were significant ($p < 0.001$); the largest is reported in the main text (Section 4.5). Random-effects estimates are listed at the bottom of the table.

Table 5: Fixed-effect coefficients for the three-way linear mixed-effects model. "—" indicates the value is subsumed in the reference level intercept. All terms marked $p < 0.001$ were significant after Benjamini-Hochberg correction.

Term	$\hat{\beta}$	z	p
Fixed effects			
Degradation level (λ)	-4.76	-45.6	$< 10^{-300}$
Model: Gemini vs. GPT (intercept)	$+0.07$	—	0.25 (n.s.)
Axis: Grammar (intercept reduction)	-1.49	-25.1	< 0.001
Axis: Coherence	—	—	significant
Axis: Information	—	—	significant
Axis: Lexical	—	—	significant
Level × Lexical × Gemini	$+1.93$	$+11.2$	< 0.001
All other Level × Axis × Model	—	—	all < 0.001
Random effects (article level)			
Intercept variance	0.54		
Slope variance	0.49		

I Extended Results Tables

Table 6: Regression slopes (β_1) of score on degradation level. All $p < 0.001$ except Mistral 7B lexical ($p = 0.49$, shown in gray).

Model	Grammar	Coherence	Information	Lexical	Mean
GPT-5 mini	-6.82	-4.78	-7.87	-6.38	-6.46
Gemini 3 Flash	-6.81	-5.58	-7.05	-5.26	-6.18
GPT-OSS 120B	-5.16	-3.36	-6.08	-4.55	-4.79
Qwen 2.5 14B	-5.97	-1.21	-6.64	-4.46	-4.57
MiniMax M2P1	-5.59	-2.04	-5.56	-3.73	-4.23
Gemma 2 9B	-4.77	-2.02	-3.63	-1.69	-3.03
Phi-4 14B	-2.77	-0.47	-3.09	-0.90	-1.81
Llama 3.1 8B	-3.70	-0.69	-1.25	-0.65	-1.57
Phi-4 mini	-1.12	-0.47	-1.35	-0.23	-0.79
Qwen 2.5 7B	-0.75	-0.54	-0.84	-0.22	-0.59
Mistral 7B	-0.41	-0.28	-0.41	-0.03	-0.28

Table 7: Calibration recovery on held-out test set. RMSE computed against $S^* = 10(1 - \lambda)$.

Model	Calib.	RMSE	ΔRMSE	Kendall τ	Spearman ρ
GPT-5 mini	Raw	2.51	—	0.629	0.741
	Affine	1.88	−25%	0.629	0.741
	Sigmoid	1.86	−26%	0.629	0.741
	Isotonic	1.85	−26%	0.629	0.741
Gemini 3 Flash	Raw	1.94	—	0.684	0.782
	Affine	1.74	−10%	0.684	0.782
	Sigmoid	1.74	−10%	0.684	0.782
	Isotonic	1.73	−11%	0.684	0.782

Table 8: Intra-condition reliability ICC(1,1) by axis.

Model	Grammar	Coherence	Info	Lexical	Overall
GPT-5 mini	0.963	0.850	0.925	0.897	**0.926**
Gemini 3 Flash	0.950	0.844	0.913	0.890	**0.905**
GPT-OSS 120B	0.914	0.566	0.777	0.718	0.751
Qwen 2.5 14B	0.845	0.605	0.856	0.744	0.849
MiniMax M2P1	0.394	0.181	0.422	0.359	0.433
Gemma 2 9B	0.891	0.823	0.852	0.873	0.888
Phi-4 14B	0.765	0.702	0.811	0.794	0.826
Llama 3.1 8B	0.716	0.689	0.711	0.767	0.805
Phi-4 mini	0.619	0.734	0.759	0.778	0.719
Qwen 2.5 7B	0.722	0.612	0.740	0.807	0.722
Mistral 7B	0.717	0.663	0.726	0.880	0.745

Table 9: F-statistic for quadratic term. Bold = significant ($F > 3.84$, $p < 0.05$); "—" = $F < 1$.

Model	Grammar	Coherence	Info	Lexical
GPT-5 mini	**3,556**	**110**	**1,640**	**1,335**
Gemini 3 Flash	**1,757**	**463**	**192**	**382**
GPT-OSS 120B	**880**	**49**	**120**	**317**
Qwen 2.5 14B	**1,683**	—	**86**	—
MiniMax M2P1	**254**	**17**	**21**	**29**
Gemma 2 9B	**73**	**25**	**63**	**33**
Phi-4 14B	**63**	—	**279**	—
Llama 3.1 8B	**96**	**13**	7	10
Phi-4 mini	3	—	**105**	8
Qwen 2.5 7B	2	—	**44**	—
Mistral 7B	**14**	—	**45**	**19**

Table 10: Kendall's τ between degradation level and score. All $p < 10^{-4}$ except Mistral 7B lexical ($\tau = -0.005$, $p = 0.49$).

Model	Grammar	Coherence	Info	Lexical
GPT-5 mini	-0.733	-0.568	-0.827	-0.675
Gemini 3 Flash	-0.818	-0.636	-0.818	-0.661
GPT-OSS 120B	-0.673	-0.408	-0.700	-0.528
Qwen 2.5 14B	-0.708	-0.382	-0.761	-0.559
MiniMax M2P1	-0.487	-0.270	-0.516	-0.420
Gemma 2 9B	-0.768	-0.585	-0.747	-0.524
Phi-4 14B	-0.700	-0.264	-0.716	-0.375
Llama 3.1 8B	-0.605	-0.358	-0.506	-0.315
Phi-4 mini	-0.265	-0.136	-0.348	-0.058
Qwen 2.5 7B	-0.351	-0.267	-0.399	-0.128
Mistral 7B	-0.172	-0.129	-0.181	-0.005

J Additional Figures

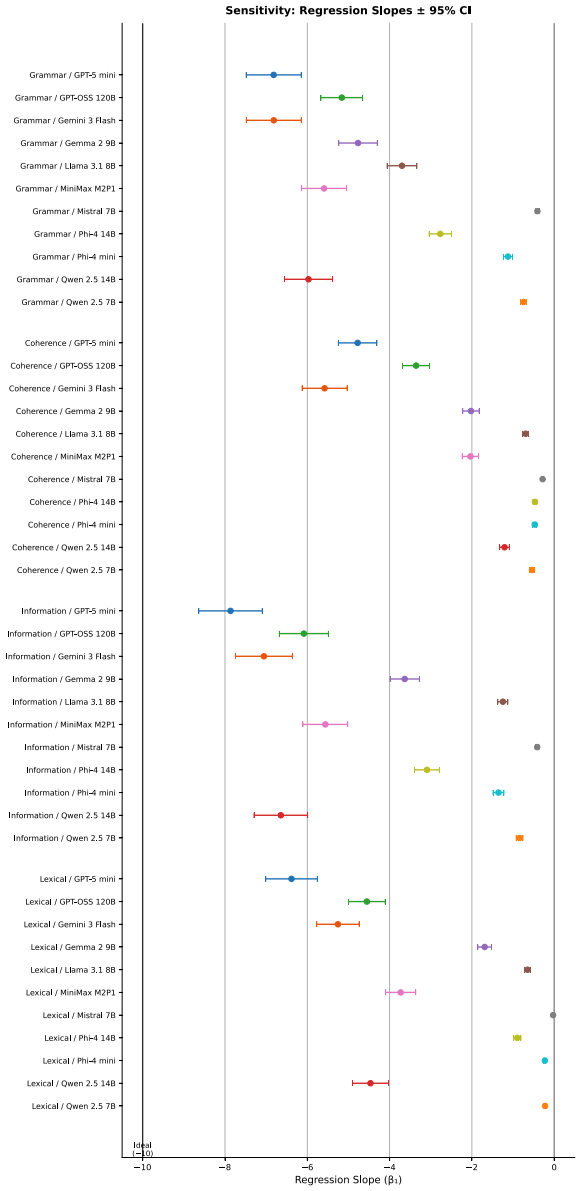


Figure 4: Forest plot of regression slopes (β_1) with 95% CIs across all axis×model combinations. All slopes significantly negative except Mistral 7B lexical.

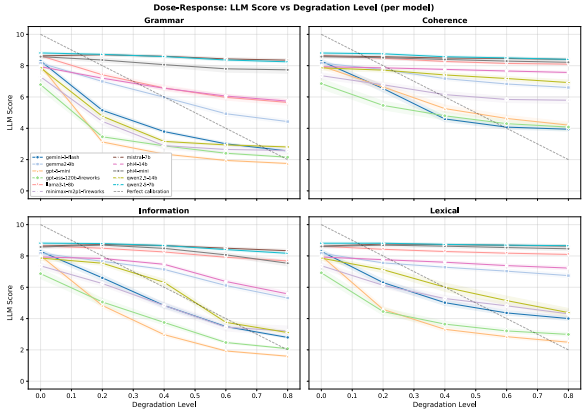


Figure 5: Dose-response curves: mean score vs. degradation level per axis. Shaded bands = 95% block-bootstrap CIs (5,000 iterations). Dashed line = naive linear reference. Concave shape reflects diminishing sensitivity at higher degradation.

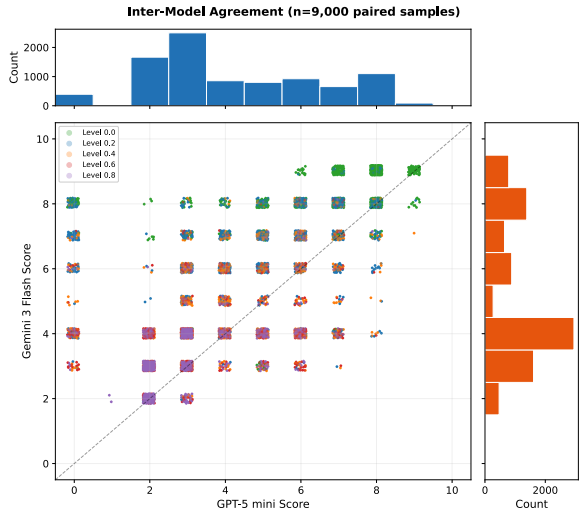


Figure 6: Inter-model scatter ($n = 9{,}000$ paired samples), jittered and colored by degradation level. Points above the identity line indicate Gemini scored higher.

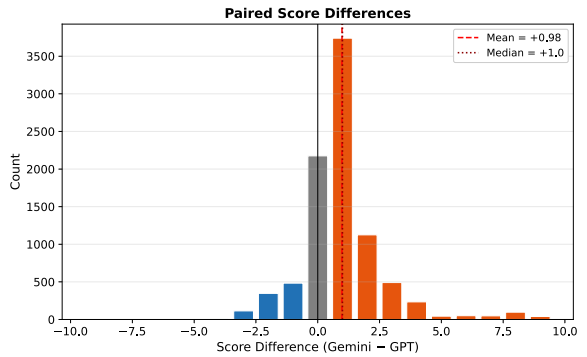


Figure 7: Paired score differences (Gemini − GPT): right-shifted distribution with median $+1$, mean $+0.89$.

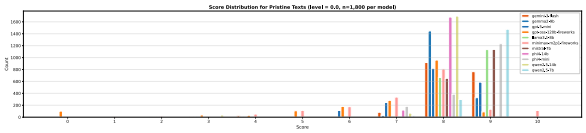


Figure 8: Score distributions for pristine ($\lambda = 0.0$) texts only ($n = 1{,}800$ per model). All models concentrate scores in a narrow top-of-scale band; most never assign 10 to expert-authored articles.

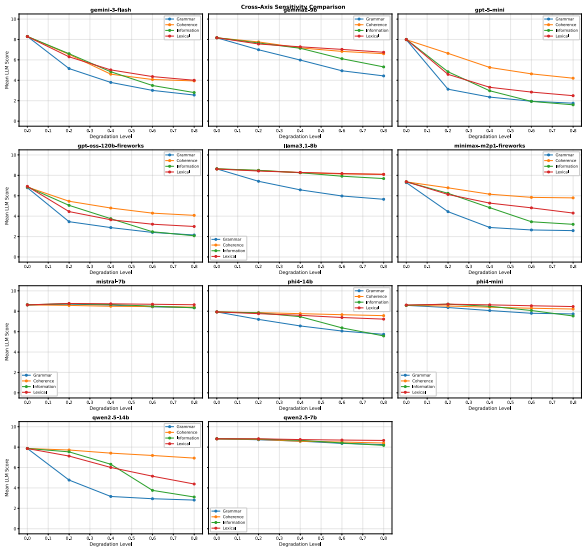


Figure 9: Cross-axis sensitivity per model. Both models are most sensitive to information deletion; GPT-5 mini least to coherence, Gemini 3 Flash least to lexical degradation.

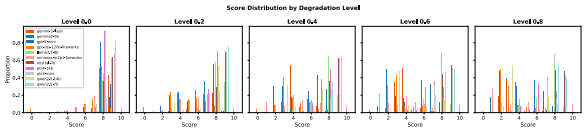


Figure 10: Score distributions by degradation level. At $\lambda = 0.0$, scores concentrate at 8–9; by $\lambda = 0.8$, at 2–4.

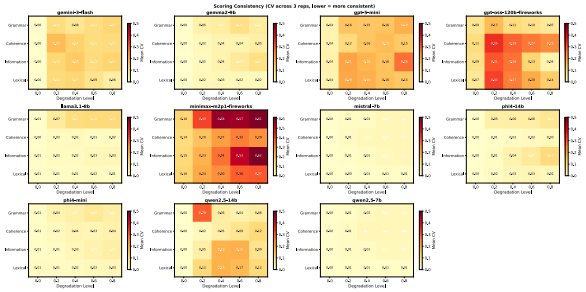


Figure 11: Scoring consistency heatmap: mean CV across three repetitions by axis and degradation level. Lower (lighter) = more consistent.

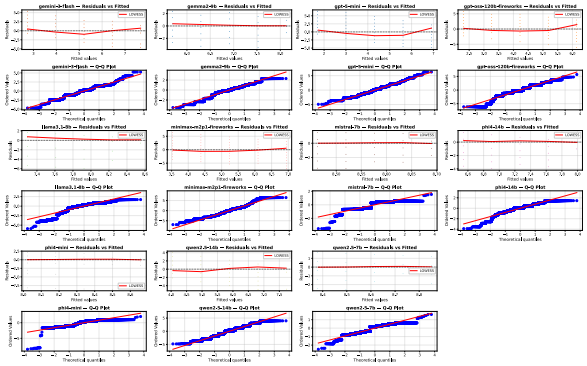


Figure 12: Residual diagnostics from linear regression. Left: residuals vs. fitted with LOWESS smoother, showing systematic curvature. Right: Q-Q plots showing heavy-tailed discrete residuals.

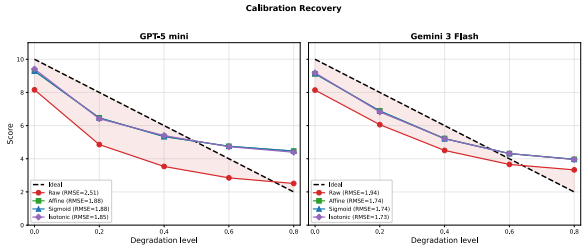


Figure 13: Dose-response curves before and after calibration. Affine, sigmoid, and isotonic calibrations converge to nearly identical curves; residual error is irreducible rank-order noise.

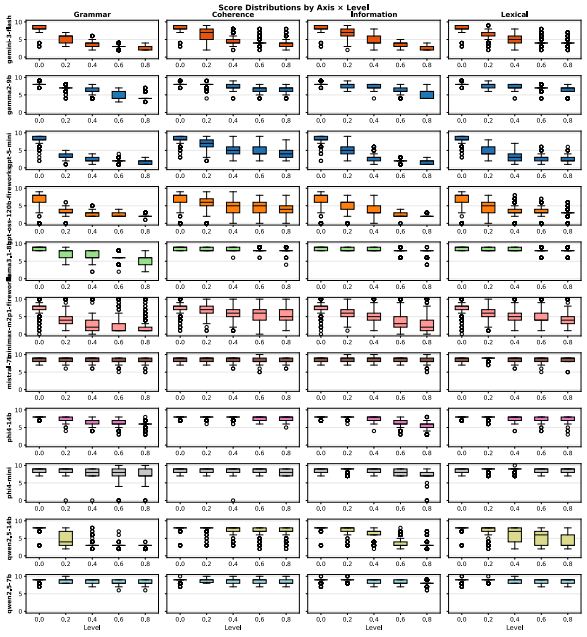


Figure 14: Score distributions by axis and level. IQR is narrowest at highest and lowest degradation levels, reflecting ceiling and floor compression.

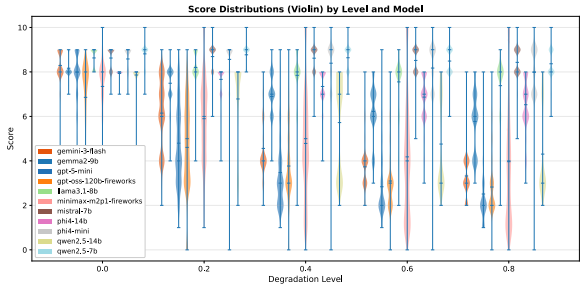


Figure 15: Violin plots by degradation level and model. Bimodal shape at $\lambda = 0.2$ reflects the onset of degradation effects.