

Decomposing LLM-Judge–Human Agreement: How Much Is Inherited Surface Preference? *

Gaurav Kumar
Moveworks AI
gkumar@ucsd.edu

Abstract

LLM-as-judge evaluation is typically validated using judge–human agreement with standard benchmarks reporting agreement rates of roughly $\sim 80\text{--}85\%$. We show that this aggregate metric can be misleading. Most of that agreement comes from pairs where both judge and human prefer the longer, better-formatted response, rather than from agreement on response quality. On the remaining pairs, agreement drops to chance levels on MT-Bench, and below chance when surface cues mislead. We show this with two simple logistic-regression baselines on length, structure, and lexical features: one predicts the human’s pick, one predicts the judge’s. The pattern holds across closed and open judges and both benchmarks, and is strongest on the expert annotations the field treats as the quality reference. We suggest reporting the *inheritance share* alongside aggregate agreement: a simple reporting metric we defined that shows whether a judge agrees with humans for the right reasons.

1 Introduction

Using a strong LLM as an automatic judge of open-ended generation has become the default in model development and evaluation (Zheng et al., 2023; Liu et al., 2023; Chiang et al., 2024). It is validated by judge–human agreement: on MT-Bench, GPT-4-class judges agree with human annotators roughly 80% of the time, comparable to inter-human agreement (Zheng et al., 2023).

Length and formatting biases in LLM judges are well documented (Saito et al., 2023; Wu and Aji, 2023; Dubois et al., 2024). Humans show the same biases (Hosking et al., 2024; Watts et al., 2024), with LLM judges typically biased more on length (Dekoninck et al., 2025; MTalk-Bench Team, 2025). Taking these biases as established,

we ask instead: given that judges and humans share these biases, how much of the reported judge–human agreement is just the co-occurrence of shared surface preference, and how much is residual quality alignment? If much of the agreement is the former, the aggregate number measures correlated bias more than judge competence, and even a length-controlled judge (Dubois et al., 2024) does not solve the metric-level problem.

To make this quantitative, we train two trivial classifiers on the same surface features—length, structure, lexical—one targeting the human verdict, the other the judge’s. We decompose agreement into pairs both classifiers predict correctly (*surface-explained*) and the rest (*residual*). On MT-Bench, 85–88% of agreement sits in the surface-explained set, and on the residual it drops to chance.

Our main contribution is this decomposition and the resulting inheritance share $\pi_S \cdot \text{Agr}_S / \text{Agr}$ (detailed in Equation 2) proposed as a reporting metric. It quantifies how much of headline judge–human agreement is inherited from shared surface preferences rather than residual quality alignment. We find that the inherited component is largest on the expert MT-Bench annotations the field treats as the quality reference - the opposite of what annotator noise alone would predict - and smaller on crowdsourced Arena. We propose reporting the inheritance share alongside aggregate agreement: a simple diagnostic that can be computed from existing human labels.

2 Related Work

LLM-as-judge and its validation. MT-Bench and Chatbot Arena (Zheng et al., 2023) report $>80\%$ judge–human agreement, which the field has since used as the validity argument for the paradigm. Meta-evaluation work pushes back on this (Zeng et al., 2024; Bavaresco et al., 2025), and recent surveys cover the broader landscape (Gu

*Code is available at <https://github.com/gaurav5590/llm-judge-human-decomposition>.

et al., 2024; Li et al., 2024).

Judge-side biases. LLM judges exhibit position bias (Wang et al., 2024), self-preference (Panickssery et al., 2024), and broader cognitive biases and inconsistency (Koo et al., 2023; Stureborg et al., 2024). Verbosity and style biases on the judge side are extensively documented (Saito et al., 2023; Wu and Aji, 2023; Singhal et al., 2023), and Dubois et al. (2024) propose length-controlled regression for win rates.

Human-side bias and closest prior work. Humans show length and style biases too: Hosking et al. (2024) show human preference is not a reference, and MTalk-Bench Team (2025); Watts et al. (2024) report parallel human-judge length preferences. Closest to our work, POLYRATING (Dekoninck et al., 2025) fits a bias-aware rating system with shared surface-feature parameters on both human and LLM-judge data. They find that length bias is significant in both paradigms and larger for LLM judges. We take their parallel-bias finding as given. Our contribution is orthogonal: rather than estimating bias magnitudes in rating points, we decompose judge-human agreement into surface-explained and residual components (Eq. 2), propose the inheritance share as a reporting metric, and show it is largest on MT-Bench’s expert labels - a contrast not analyzed by Dekoninck et al. (2025). Hu et al. (2024) also propose a decomposition but of the pairwise win rate; we decompose the agreement metric itself.

Surface-cue meta-predictors and selective evaluation. Lehmann et al. (2025) fit a logistic-regression meta-predictor on surface cues for comparative QA, a different task; we adapt that methodology to LLM-as-judge and apply it to both the judge and the human side. Jung et al. (2025) and Radharapu et al. (2025) study confidence-based escalation and no-consensus items, settings where judge-human agreement is hardest to obtain.

3 Method

Data. We use two pairwise-preference benchmarks with human labels: **MT-Bench** human judgments (expert annotators) and a **Chatbot Arena** subset (crowdsourced), both from Zheng et al. (2023). We flatten conversations to assistant responses, deduplicate to unique (question, model pair) items, take the majority human label, and

drop ties. For Arena we subsample 1,000 pairs stratified by winner and length-difference quartile.

Not all pairs survive to analysis. First, pairs with no decisive human majority are dropped. Second, each pair is judged in both display orders; if a judge picks different winners across the two orders, the position-averaged verdict is a tie and the pair is dropped. Because judges differ in how often they flip, each judge yields a different number of usable pairs. This yields ~ 800 decisive MT-Bench pairs and ~ 700 – 900 decisive Arena pairs per judge (Appendix A).

Judges. We evaluate three judges: two closed (gpt-4o-mini-2024-07-18, gpt-4o-2024-11-20; OpenAI, 2024) and one open (Qwen2.5-7B-Instruct; Qwen Team, 2024), using the standard MT-Bench pairwise prompt (Appendix B) at temperature 0. Each pair is judged in both display orders. We take the position-averaged verdict and treat order-inconsistent pairs as ties. Verdicts parse cleanly (unparseable rate $\leq 1.6\%$ for every judge; Appendix C).

Surface features. From each response we compute twelve interpretable features grouped into three families: four *length* features (characters, tokens, words, sentences), five *structure* features (markdown headers, bullets, fenced code blocks, bold spans, paragraphs), and three *lexical* features (unique-word ratio, mean word length, question-response word overlap). Tokens use the cl100k_base tokenizer so counts are judge-independent.

For each pair (A, B) , we encode each feature as a signed difference $\Delta_f = f_A - f_B$ and, for length and structure features, additionally as a ratio $f_A/(f_A + f_B)$. The classifier sees only these Δ and ratio encodings. By construction this representation is *position-symmetric*: swapping A and B negates the feature vector and flips the prediction accordingly, so the baseline carries no positional preference of its own - in contrast to the judges.

The baseline model. Our baseline is deliberately the simplest competent classifier: ℓ_2 -regularized logistic regression on standardized features ($C=1$, no hyperparameter search; Pedregosa et al., 2011). We fit two instances with identical inputs but different targets: **Baseline-H** predicts the human verdict and **Baseline-J** predicts the judge verdict. Both use 5-fold stratified cross-validation, with out-of-fold predictions for every pair so the decomposition

Data	Judge	B-H	B-J	J-H	Flip
MT-B	4o-mini	0.804	0.868	0.854	0.141
	4o	0.803	0.865	0.852	0.119
	Qwen-7B	0.807	0.852	0.840	0.312
Arena	4o-mini	0.686	0.771	0.842	0.101
	4o	0.686	0.737	0.844	0.091
	Qwen-7B	0.716	0.745	0.830	0.259

Table 1: Surface baseline predicting humans (B-H) and judges (B-J) vs. judge–human agreement (J-H), and position-flip rate (Flip). Bootstrap 95% CIs have half-width ≈ 0.025 . B-H tracks B-J across all six cells.

uses no in-sample information.

All analyses are restricted to pairs decisive for both judge and human. We report bootstrap 95% confidence intervals over 1,000 resamples. Because the model is linear and deliberately untuned, any predictive power it shows is a lower bound on what surface cues afford.

Inheritance decomposition. We partition pairs into those the trivial baseline correctly predicts on both sides (*surface-explained*) and those it does not (*residual*), then ask what fraction of agreement lives in each. Let h_i, j_i be the human and judge verdicts on pair i and $\hat{h}_i, \hat{j}_i$ the corresponding out-of-fold baseline predictions. Define

$$S_i = \mathbb{1}[\hat{h}_i = h_i \text{ and } \hat{j}_i = j_i], \quad \pi_S = \frac{1}{N} \sum_i S_i. \quad (1)$$

Let $\text{Agr}(j, h) = \frac{1}{N} \sum_i \mathbb{1}[j_i = h_i]$ be the overall judge–human agreement rate, and let $\text{Agr}_S, \text{Agr}_R$ denote the same quantity restricted to pairs with $S_i = 1$ and $S_i = 0$, respectively. Then

$$\text{Agr}(j, h) = \pi_S \text{Agr}_S + (1 - \pi_S) \text{Agr}_R. \quad (2)$$

We call $\pi_S \cdot \text{Agr}_S / \text{Agr}$ the *inheritance share*: the fraction of total agreement that a trivial baseline already accounts for.

4 Results

4.1 The baseline predicts humans almost as well as judges

Table 1 reports the two baselines and judge–human agreement. On MT-Bench, the surface baseline predicts *human* verdicts at 80.3–80.7% across all three judges - only 4–6 pp below its accuracy at predicting the judge (B-J, 85.2–86.8%). A model that sees only length and format reproduces human preferences nearly as well as a frontier LLM

Data	Judge	π_S	Agr_S	Agr_R	Agr	Inh.%
MT-B	4o-mini	.745	.976	.498	.854	85.1
	4o	.755	.990	.425	.852	87.8
	Qwen-7B	.737	.992	.415	.840	87.0
Arena	4o-mini	.628	.991	.590	.842	74.0
	4o	.602	.985	.631	.844	70.3
	Qwen-7B	.616	.989	.576	.830	73.4

Table 2: Inheritance decomposition (Eq. 2). π_S : surface-explained fraction; $\text{Agr}_S/\text{Agr}_R$: agreement on explained/residual pairs; Inh.%: inheritance share $\pi_S \cdot \text{Agr}_S / \text{Agr}$. On MT-Bench the residual collapses to chance.

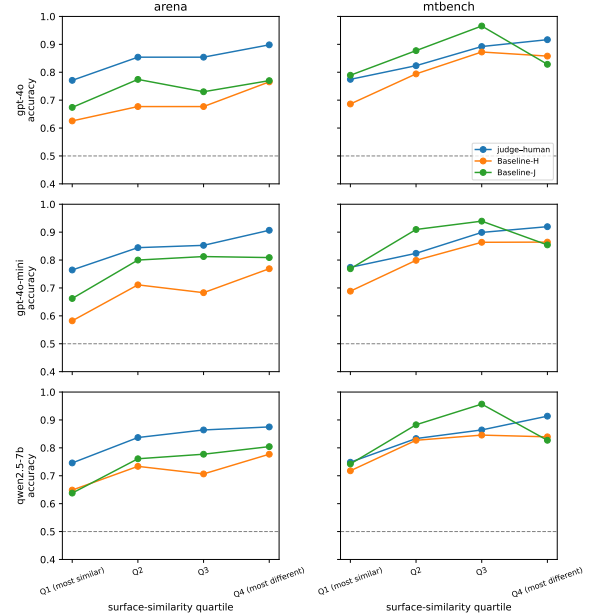


Figure 1: Accuracy by quartile of multi-feature surface distance (Q1 most similar, Q4 most different), per judge and dataset. Judge–human, Baseline-H, and Baseline-J move together; where surface cues are weak (Q1), agreement weakens too. Dashed line: chance.

agrees with those humans. We use this as input to the decomposition. On crowdsourced Arena the baseline is weaker (B-H 68.6–71.6%, still within 3–9 pp of B-J), with judge–human agreement further above B-H - a larger non-surface component than on MT-Bench.

4.2 Most agreement is the inherited surface term

Table 2 decomposes agreement via Eq. 2. On MT-Bench, $\pi_S \approx 0.74$ – 0.76 of pairs are surface-explained, agreement on them is $\approx 98\%$, and on the residual it *collapses to chance* ($\text{Agr}_R = 0.42$ – 0.50). The inheritance share thus accounts for **85–88%** of all judge–human agreement. Arena is

more tempered: fewer pairs are surface-explained ($\pi_S \approx 0.60$ – 0.63) and, crucially, residual agreement does *not* reach chance ($\text{Agr}_R = 0.58$ – 0.63). The inherited share still dominates (70–74%), but a genuine residual quality-alignment signal remains. The effect is therefore strongest exactly for the expert annotations treated as the quality reference – the opposite of what an annotator-noise account would predict. The inherited share stays high even on *objective* categories (math, coding, extraction), where extra words should not track quality (Appendix I), and is stable across filtering choices, including the aggressive position-flip drop (Appendix J). Additional robustness checks are in Appendices E, F, and G.

4.3 Agreement tracks surface separability and confidence

Figure 1 stratifies pairs by surface distance (the ℓ_2 norm of the standardized Δ -feature vector). All three curves rise together across quartiles, with judge–human agreement going from 0.75 on the most surface-similar pairs to 0.91 on the most surface-different. Agreement on the most-similar quartile does not hit 0.50; that drop only shows up on the *residual* set (Section 4.2), where the baseline is actively wrong. The point of the figure is that when surface cues are weak, human and judge agreement is weak too. Binning by baseline confidence (Figure 2) gives the same picture: agreement climbs from 0.70–0.80 in the least-confident bin to over 0.90 in the most-confident bin on both benchmarks. When the surface signal is strong, judge and human agree; when it is weak, they do not.

4.4 Humans and judges weight the same cues

The two baselines are not just accurate, they rely on similar features. Table 3 shows that token count and token ratio dominate the standardized coefficients for both targets, with the same sign and comparable magnitude. Humans and judges respond to the same length cues in the same direction, which is the mechanism the decomposition needs. Position-flip rates (Table 1) are low for the GPT judges (9–14%) but high for Qwen-7B (26–31%), consistent with smaller open judges being less order-robust (Wang et al., 2024). We position-average throughout to handle this.

4.5 Length causally moves judge verdicts

Tables 2 and 3 establish that both judges and humans rely on similar length-related cues. However,

feature	$ \bar{w}_H $	$ \bar{w}_J $
ratio_tokens	2.159	2.358
delta_chars	0.982	1.882
delta_tokens	1.178	1.752
delta_unique_word_ratio	1.010	1.409
delta_words	0.223	0.776
ratio_sentences	0.654	0.414
delta_sentences	0.641	0.354
delta_paragraph_count	0.362	0.578

Table 3: Top features by mean |standardized coefficient| for Baseline-H ($|\bar{w}_H|$) and Baseline-J ($|\bar{w}_J|$), averaged over judge $\times$ dataset. Length/verbosity dominates both, with the same sign.

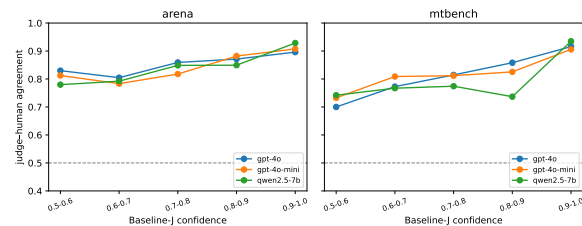


Figure 2: Judge–human agreement vs. Baseline-J confidence ($\max(p, 1-p)$), per dataset – one line per judge. Agreement increases monotonically with surface confidence on both benchmarks. Dashed line – chance.

these analyses are observational and do not show whether length itself influences judgments or simply correlates with response quality. To isolate the effect of length, we perform a verbosity-injection intervention.

We begin with the 198 MT-Bench pairs for which the judge originally preferred the *shorter* response. For each, we rewrite the longer, losing response with gpt-4o to roughly twice its length, instructing the model to add no new information and only rephrase or restate existing content. An independent model from a different family compared to all judges (claude-haiku-4-5) then verifies that no facts or claims were added. We retain the 133 rewrites that pass and grow in length, and re-judge each in both presentation orders (Appendix H).

Because content is held constant by construction, any change in verdict is attributable to length and formatting alone. Table 4 reports the fraction of pairs for which the judge now prefers the padded response, averaged over both display orders: gpt-4o-mini flips on 43%, Qwen2.5-7B on 28%, and gpt-4o on 16%, with all three confidence intervals excluding zero. Increasing verbosity alone thus reverses a substantial fraction of decisions while leaving content unchanged, and the effect is strongest for the weakest-performing

Judge	n	Flip	95% CI
gpt-4o-mini	91	.429	[.330, .538]
gpt-4o	91	.165	[.099, .242]
Qwen2.5-7B	76	.276	[.184, .382]

Table 4: Verbosity-injection results on MT-Bench. Flip is the fraction of decisive re-judged pairs for which the judge switches to the padded response. Responses are expanded to $\approx 2.2\times$ length without changing content (133 verified rewrites).

Judge	Objective		Subjective	
	Agr_R	Inh.%	Agr_R	Inh.%
gpt-4o-mini	.659	76.9	.384	89.8
gpt-4o	.578	81.7	.342	90.5
Qwen2.5-7B	.467	81.2	.407	88.9

Table 5: Inheritance decomposition pooled over *objective* (math, coding, extraction) vs. *subjective* (writing, roleplay, humanities) MT-Bench categories. The inherited share stays $\geq 77\%$ on objective tasks, where extra length cannot make a near-binary answer more correct; Per-category detail in Appendix I.

judge.

4.6 Length behaves as bias even on objective tasks

If length reflects genuine quality rather than bias, its effect should be smaller on tasks where longer responses are unlikely to be more correct. MT-Bench lets us test this by separating objective tasks (math, coding, extraction) from subjective ones (writing, roleplay, humanities), where more detailed responses may legitimately improve quality.

We recompute the decomposition for each group while keeping the global baseline fixed (Table 5). The inherited share is 8–13 pp lower on pooled objective categories, suggesting that length carries some genuine signal there. However, it remains at least 77%, indicating that most of the agreement between judges and humans is still explained by length even where verbosity is unlikely to improve correctness. Since each category contains relatively few examples ($n \approx 80\text{--}120$), we focus on the pooled comparison rather than individual categories (Appendix I).

5 Discussion

What agreement measures. A single judge–human agreement number mixes two regimes: surface-explained pairs, where judges and humans agree trivially, and the residual, where genuine

quality alignment would have to live. Reporting one aggregate hides that the headline is carried by the surface-explained regime, and that the residual where actual quality adjudication is needed can sit close to a chance probability.

An audit of the agreement metric. We propose reporting the *inheritance share* $\pi_S \cdot \text{Agr}_S / \text{Agr}$ alongside aggregate agreement, and/or stratifying by surface separability (e.g. residual agreement Agr_R). This requires only the human labels already collected and a trivial baseline; it costs nothing extra and surfaces whether a judge agrees with humans for the right reasons. It is complementary to length-controlled estimators (Dubois et al., 2024), which debias the score, and to bias-aware rating systems (Dekoninck et al., 2025), which estimate bias magnitudes on ratings; we directly audit the *agreement metric itself*. We do not claim the inheritance share is sufficient but it is a near-free signal currently missing from standard reports.

6 Conclusion

We split judge–human agreement into two parts using a simple surface baseline: pairs the baseline gets right on both sides, and the rest. On MT-Bench, 85–88% of agreement falls in the first part. Our results suggest that headline judge–human agreement largely reflects shared surface preferences rather than shared quality judgments, motivating the inheritance share as a complementary reporting metric. The inheritance share is a cheap way to see this in any evaluation with human labels.

Limitations

We test turn-1, English, pairwise preference on two benchmarks and three judges; the numbers may differ for other tasks, languages, judges, and pointwise scoring. The split into surface-explained and residual depends on the features we use; a richer set (e.g. semantic embeddings) would likely move more pairs into the surface-explained bucket, so our numbers are a lower bound on what surface cues can capture. We treat human labels as ground truth even though annotators have biases, make errors, and sometimes legitimately disagree (Plank, 2022; Hosking et al., 2024; Bavaresco et al., 2025)—our claim is that this “ground truth” shares the judge’s surface preference, not that either is correct. The Arena subset is 1,000 pairs to keep costs down, so the confidence intervals there are wider.

References

- Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, and 1 others. 2025. [LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anas-tasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. [Chatbot arena: An open platform for evaluating LLMs by human preference](#). In *International Conference on Machine Learning (ICML)*.
- Jasper Dekoninck, Maximilian Baader, and Martin Vechev. 2025. [POLYRATING: A cost-effective and bias-aware rating system for LLM evaluation](#). In *The Thirteenth International Conference on Learning Representations (ICLR)*.
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. [Length-controlled AlpacaEval: A simple way to debias automatic evaluators](#). *arXiv preprint arXiv:2404.04475*.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. [A survey on LLM-as-a-judge](#). *arXiv preprint arXiv:2411.15594*.
- Tom Hosking, Phil Blunsom, and Max Bartolo. 2024. [Human feedback is not gold standard](#). In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Zhengyu Hu, Linxin Song, Jieyu Zhang, Zheyuan Xiao, Tianfu Wang, Zhengyu Chen, Athanasios V. Vasilakos, and Hui Xiong. 2024. [Explaining length bias in LLM-based preference evaluations](#). *arXiv preprint arXiv:2407.01085*.
- Jaehun Jung and 1 others. 2025. Trust or escalate: LLM judges with provable guarantees via selective evaluation. *arXiv preprint*.
- Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. [Benchmarking cognitive biases in large language models as evaluators](#). *arXiv preprint arXiv:2309.17012*.
- Jan Lehmann and 1 others. 2025. Knowing the facts but choosing the shortcut: Surface cues in comparative question answering. *arXiv preprint*.
- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2024. [From generation to judgment: Opportunities and challenges of LLM-as-a-judge](#). *arXiv preprint arXiv:2411.16594*.
- 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 EMNLP*.
- MTalk-Bench Team. 2025. MTalk-Bench: Evaluating speech-to-speech models with human and LLM judges. *arXiv preprint*.
- OpenAI. 2024. [GPT-4o system card](#). *arXiv preprint arXiv:2410.21276*.
- Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. [LLM evaluators recognize and favor their own generations](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Barbara Plank. 2022. [The “problem” of human label variation: On ground truth in data, modeling and evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Qwen Team. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Bhaktipriya Radharapu and 1 others. 2025. Arbiters of ambivalence: Challenges of using LLM judges on no-consensus tasks. *arXiv preprint*.
- Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. [Verbosity bias in preference labeling by large language models](#). *arXiv preprint arXiv:2310.10076*.
- Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2023. [A long way to go: Investigating length correlations in RLHF](#). *arXiv preprint arXiv:2310.03716*.
- Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. [Large language models are inconsistent and biased evaluators](#). *arXiv preprint arXiv:2405.01724*.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, 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 (ACL)*.
- Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. 2024. [PARIKSHA: A large-scale investigation of human-LLM evaluator agreement on multilingual and multi-cultural data](#). *arXiv preprint arXiv:2406.15053*.

Minghao Wu and Alham Fikri Aji. 2023. [Style over substance: Evaluation biases for large language models](#). *arXiv preprint arXiv:2307.03025*.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. [Evaluating large language models at evaluating instruction following](#). In *International Conference on Learning Representations (ICLR)*.

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](#). In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

A Data and Filtering

Both datasets are reduced to a normalized pair schema (question, answer_a, answer_b, human_winner). MT-Bench uses the turn-1 human-judgment split; conversations are flattened to assistant text. We drop tie and tie (bothbad) labels and pairs with a tied human vote. Arena is subsampled to 1,000 pairs stratified by winner $\times$ length-difference quartile (seed 42), English-only. Decisive-pair counts per cell range from 649 to 905.

B Judge Prompt

We use the standard MT-Bench pairwise system prompt:

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants. . . Avoid any position biases. . . Do not allow the length of the responses to influence your evaluation. . . After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.

The user turn presents the question and the two answers; we parse the last bracketed token as the verdict.

C Additional Results

Verdict unparseable rates are uniformly low: gpt-4o-mini 0.0%, gpt-4o 0.4–1.6%, Qwen2.5-7B 0.0–0.1%. Per-cell decisive-pair counts (n) appear in Table 1’s caption range and in our released CSVs.

D Position Bias by Judge

Table 6 reports position-flip rates out of all human-decisive pairs (the same denominator used throughout the paper). GPT judges flip on 10–18% of pairs; Qwen flips on 31–36%. All analyses use position-averaged verdicts; inconsistent pairs are treated as ties and excluded.

E Flip-Pair Sensitivity

We drop position-inconsistent pairs as ties throughout the main analysis. Table 7 checks what changes if we count them as disagreements instead. For GPT judges the shift is small (−8 to −12 pp); for Qwen, which flips on 26–31% of pairs, inclusive agreement falls to 0.59–0.61. Either way the main

Dataset	Judge	Flip rate	Consistent
Arena	gpt-4o-mini	10.5%	89.5%
Arena	gpt-4o	13.5%	86.5%
Arena	Qwen2.5-7B	30.6%	69.4%
MT-Bench	gpt-4o-mini	14.5%	85.5%
MT-Bench	gpt-4o	18.1%	81.9%
MT-Bench	Qwen2.5-7B	36.4%	63.6%

Table 6: Position-flip rates per judge and dataset, out of all human-decisive pairs. Qwen2.5-7B is much less robust to display order than the GPT judges. These are the same Flip% values shown in Table 7.

results hold and Qwen’s agreement with humans is lower than the strict figures suggest.

Data	Judge	Flip%	Strict	Incl.
Arena	gpt-4o-mini	10.5	.842	.757
Arena	gpt-4o	13.5	.844	.764
Arena	Qwen2.5-7B	30.6	.830	.612
MT-B	gpt-4o-mini	14.5	.854	.732
MT-B	gpt-4o	18.1	.852	.750
MT-B	Qwen2.5-7B	36.4	.840	.588

Table 7: Sensitivity to treating position-inconsistent pairs as disagreements. Strict J–H is the figure used throughout the paper (flips dropped); inclusive counts them as judge–human disagreements. The gap is small for GPT judges and large for Qwen2.5-7B, confirming the strict figures are conservative.

F The Residual Collapse Is Not Mechanical

One concern is that the residual pairs are just hard and any two signals would agree at chance on them. We check this by splitting pairs into three groups: **S** (both baselines correct, same as the surface-explained set), **CW** (both baselines predict the same winner but are wrong), and **D** (baselines predict different winners).

If the collapse were mechanical, both CW and D should be near chance. Table 8 shows they are not: CW pairs on MT-Bench give J–H agreement of 0.33–0.38, *below chance*, while D pairs give 0.65–0.79. When surface cues confidently point the wrong way, judge and human actually anti-correlate. The D result (4–9% of pairs) suggests a small genuine quality signal when surface cues are ambiguous.

G Feature-Set Robustness

We rerun the decomposition with three feature sets. The ranking (MT-Bench > Arena across all judges)

Data	Judge	S	CW	D
Arena	gpt-4o-mini	.991	.530	.828
Arena	gpt-4o	.985	.551	.835
MT-B	gpt-4o-mini	.976	.357	.730
MT-B	gpt-4o	.990	.378	.654

Table 8: J-H agreement by group (Qwen follows the same pattern). **S**: both baselines right. **CW**: both predict the same winner but are wrong. **D**: baselines disagree. CW is *below* chance on MT-Bench—the collapse is not just “hard pairs.” D is above chance, suggesting a small real quality signal.

holds for all three. The inherited share rises slightly with more features. Length-only already gives 67–82% – so the numbers we report in the main paper are a lower bound.

Data	Judge	length	l+struct	full
Arena	mini	.724	.723	.740
Arena	4o	.673	.679	.703
MT-B	mini	.786	.839	.851
MT-B	4o	.806	.845	.878

Table 9: Inherited share under three feature sets (Qwen follows the same pattern). *length*: 6 features. *l+struct*: adds sentence, paragraph, and formatting features (13 total). *full*: all 16 symmetric features. The MT-Bench > Arena ranking is stable; numbers reported in the main paper (full) are a lower bound.

H Verbosity Experiments

This appendix details the intervention experiments of Section 4.5.

Sampling. We select MT-Bench decisive pairs where a majority of judges unanimously preferred the *shorter* response. This gives 198 pairs in which the initial verdict goes against length. The shorter response is preferred in only 22–27% of decisive pairs, so the 198 pairs cover nearly all such cases. In all of them, the losing response is longer than the winning response (median 260 vs. 68 tokens).

Padding. We use gpt-4o-2024-11-20 at temperature 0 to rewrite each losing response to roughly twice its original length. The model is instructed not to add information, and can only rephrase or expand existing sentences, restructure the response into bullets, or add summaries of existing content.

Verification. We use claude-haiku-4-5 to check each padded response for substantive additions. The verifier is from a different model family

than the three judges and checks for new facts, statistics, entities, examples, claims, or recommendations. Rephrasing, added length, and repeated summaries are not counted as additions. Of the 198 pairs, 133 pass this check and also exceed a $1.3\times$ length ratio. The remaining pairs are dropped. The verified pairs have a mean length ratio of $2.47\times$ (median $2.22\times$).

Re-judging. We re-judge each verified pair (original winner vs. padded loser) with all three judges in both display orders at temperature 0. Table 4 reports the position-averaged flip rate. Padding also increases position instability: 32–51% of pairs become position-inconsistent after padding, compared with 14–36% at baseline. These pairs are excluded when computing the flip rate.

I Per-Category Detail

Table 10 gives the per-category inheritance share underlying the pooled objective/subjective contrast of Section 4.6 (Table 5). We recompute the decomposition within each of MT-Bench’s eight categories, keeping the global out-of-fold baseline fixed and stratifying only the evaluation. We report reasoning and stem but exclude them from the objective/subjective pooling since neither fits cleanly into either group. The inherited share exceeds 76% in every objective category for all judges. extraction is lowest (62–78%). The 100% share for stem is based on only $n \approx 15$ residual examples. Given the small per-category sample sizes ($n \approx 80$ –120), we focus on the pooled contrast in the main text.

Category	Type	mini	4o	Qwen
math	obj	86.3	83.8	85.7
coding	obj	83.3	85.7	80.3
extraction	obj	61.9	76.5	78.3
reasoning	mix	78.3	82.0	87.0
stem	mix	96.1	100.0	100.0
writing	subj	89.7	92.3	91.9
roleplay	subj	83.8	84.3	84.7
humanities	subj	96.7	95.8	92.2

Table 10: Per-category inheritance share (%) on MT-Bench, with one column per judge (mini = gpt-4o-mini). *obj*: objective; *subj*: subjective; *mix*: reasoning and stem, reported but not pooled.

J Filtering Sensitivity and Funnel

Our main analysis drops human ties and position-inconsistent judge verdicts. Table 11 shows the

number of pairs removed at each step. The position-inconsistent rate is 11–18% for the GPT judges and 36% for Qwen2.5-7B on MT-Bench, making filtering more consequential for the open judge.

Data	Judge	Dec.	Drop _{pos}	Final
Arena	gpt-4o-mini	1000	105 (10.5%)	899
Arena	gpt-4o	1000	135 (13.5%)	905
Arena	Qwen2.5-7B	1000	306 (30.6%)	737
MT-B	gpt-4o-mini	927	134 (14.5%)	795
MT-B	gpt-4o	927	168 (18.1%)	816
MT-B	Qwen2.5-7B	927	337 (36.4%)	649

Table 11: Filtering funnel. Dec.: human-decisive pairs entering per-judge analysis. Drop_{pos}: pairs dropped for position inconsistency. Final: pairs decisive for both judge and human. MT-Bench starts from 927 pairs after dropping tied human votes (from 961 unique turn-1 pairs).

Table 12 tests whether the results depend on this filtering. We compare three regimes: *current*, which drops position-inconsistent pairs; *incl.*, which keeps them and treats each as a judge–human disagreement; and *AB*, which uses a single display order with no filtering. The agreement rate changes substantially across regimes. For Qwen2.5-7B on MT-Bench, for example, it drops from 0.84 under *current* to 0.60 under *incl.*. The inheritance share is more stable, changing by at most ≈ 8 pp across all cells.

Data	Judge	current	incl.	AB
Arena	gpt-4o-mini	74.0	74.2	73.1
Arena	gpt-4o	70.3	69.2	67.7
Arena	Qwen2.5-7B	73.4	68.5	66.1
MT-B	gpt-4o-mini	85.1	88.1	85.5
MT-B	gpt-4o	87.8	86.8	86.0
MT-B	Qwen2.5-7B	87.0	81.7	78.9

Table 12: Inheritance share (%) under three filtering regimes. *current*: main-paper policy. *incl.*: position-inconsistent pairs are kept and counted as judge–human disagreements. *AB*: single display order with no filtering.