

The Format is the Filter: Scoring Format Fixes What an LLM Judge Can Express

Lovedeep Gondara*, Jekaterina Novikova, Sagarika Suresh, Jack Ji, Jithin Pradeep
The Vanguard Group, Inc.*

*Correspondence: Lovedeep_Gondara@Vanguard.com

Abstract

When an LLM judge filters instruction-tuning data, its scoring format (binary verdict, 1–5 rating, 0–100 score, or comparison) is rarely justified or ablated. Holding the candidate pool, judge prompt, decoding and selection budget fixed, we score 15,406 UltraFeedback candidates under five formats with three judges, curate, fine-tune 66 models, and evaluate them with four judges. A 0–100 scale is worth only 7–12 equally used levels, two to three times a 1–5 scale rather than twenty, and a binary verdict leaves the top score tied in 77–78% of prompts, so most selections are made at random. These properties are set by the format almost independently of the judge. Kendall’s τ_b hides them: it ranks binary best while τ_a and top-1 agreement rank it last, because it discounts the ties a selector must break at random. Both comparative formats inherit a slot bias whose sign differs across judges. Downstream, curated data beat random selection under every evaluation judge, robust to length and markdown adjustment, but no ordering among formats is detectable. We propose four quantities that judge-based curation pipelines should report.

1 Introduction

A practitioner assembling an instruction-tuning corpus without an annotation budget typically asks an LLM judge to score every candidate, keeps the top fraction, and fine-tunes on what survives (Zheng et al., 2023; Chen et al., 2024; Sachdeva et al., 2026). How to ask for the judge’s opinion (ACCEPT/REJECT, a 1–5 rating, a 0–100 score, a comparison) is usually decided once, on grounds of convenience, and rarely ablated with the selector held fixed. Yet a curation pipeline consumes scores as an *ordering* and breaks ties at random,

so tie structure is a first-class property of a format: a binary verdict leaves every accepted candidate indistinguishable, and the choice at the frontier is made by chance.

We vary only the format across five designs, score a fixed pool with three judges, and follow one judge’s scores through curation and LoRA fine-tuning to 66 models. Our main contributions are as follows: (i) a fully crossed study of three judges $\times$ five formats, showing that the format, not the judge, determines how many ties the scores contain, a result that holds when any one judge is left out and when each format’s number of available levels is taken into account (§3.1); (ii) evidence that τ_b inverts the ordering a selection audit needs (§3.3); (iii) evidence that both comparative formats inherit a judge-specific slot bias, measured three ways (§3.4); and (iv) downstream results under four evaluation judges (§4).

Related work. Elicitation studies measure formats by agreement with human or reference ratings (Li et al., 2026; Novikova et al., 2018; Zhuang et al., 2024) or compare pointwise and pairwise protocols for bias (Tripathi et al., 2025), without following scores into training. Data-selection work does follow them, but with one elicitation protocol (Chen et al., 2024; Cui et al., 2023; Pang et al., 2025; Wettig et al., 2024). Deutsch et al. (2023) establish that ties distort Kendall-style meta-evaluation; we show the consequence for selection. Configuration beyond the judge model is a known variance source (Zhang, 2026; Don-Yehiya et al., 2026; Norman et al., 2026; Soumik, 2026). We add the controlled format ablation, carried through to trained models.

2 Method

Pool and reference. We draw 4,000 UltraFeedback prompts (Cui et al., 2023), yielding 15,406 multi-model responses after a 20–6,000 character filter ($k = 3.85$ per prompt); 300 further prompts

*© 2026 The Vanguard Group, Inc. All rights reserved.
This material is provided for informational purposes only and is not intended to be investment advice or a recommendation to take any particular investment action.

are held out. UltraFeedback’s GPT-4 scores are used as the *reference*. They never enter a prompt as numbers, the selection decision, or the training data, but they choose the anchor in the anchored format and the evaluation reference, so we report agreement with the GPT-4 reference and avoid the term accuracy.

Formats. The pointwise formats share one preamble (helpfulness, correctness, completeness, clarity) and differ only in the final instruction; the comparative formats use a parallel pairwise prompt with five verdicts mapped to $v \in \{+2, \dots, -2\}$ (Appendix A). **Binary** (ACCEPT/REJECT $\rightarrow \{0, 1\}$), **Likert-5** $((v - 1)/4)$ and **Numeric-100** $(v/100)$ are pointwise. **Anchored** compares the candidate (always Response A) with a fixed per-prompt anchor, the upper-middle candidate by GPT-4 score (Don-Yehiya et al., 2026). **Tournament** compares every pair of candidates within a prompt once. Each verdict v gives Response A a win probability $\sigma(v) = 1/(1 + e^{-v})$ (0.88 for a strong win, 0.5 for a tie), and a candidate’s score is its average win probability over its matches, a soft win rate. Decoding is greedy, with a 16-token verdict budget and reasoning disabled. Selection uses within-format ranks, which preserve ties, so each format’s tie structure passes unchanged into selection.

Judges, curation and training. Three curation judges score the identical pool: deepseek-v4-flash (DeepSeek-AI et al., 2026), qwen3.5-9b (Yang et al., 2025), and nemotron-3-nano-30b-a3b; parse rates are ≥ 0.998 in all fifteen cells. Curation uses DeepSeek scores only. At retention $r \in \{10, 25, 50\}\%$ we keep the highest-scored response per prompt (ties broken at random), then the top r of these globally. RANDOM keeps one uniform response per prompt and a uniform fraction r of prompts; LENGTH applies the same policy to response length; FULL uses all candidates. Each subset fine-tunes Qwen3-4B (Yang et al., 2025) with LoRA (Hu et al., 2021) and three seeds: seven conditions at three rates plus FULL, a total of 66 models.

Evaluation. Each model answers the 300 held-out prompts; each response is compared with the highest-GPT-4-rated completion in both orders, scoring 1 when both orders agree on a win, 0 on a loss, and 0.5 otherwise. Four evaluation judges (deepseek-v4-pro, gemini-3.5-flash-lite,

grok-4.20, Olmo-3-7B) are analyzed separately. They resolve different fractions of comparisons into a win or loss, and every unresolved comparison scores 0.5, so a judge that resolves fewer pulls all win rates toward 0.5; levels are therefore comparable within a judge but not across judges. Per judge and rate we run a Friedman test (Friedman, 1937), paired seed-clustered bootstrap contrasts (10,000 resamples), and Holm correction (Holm, 1979), and report how many judges concur.

3 Scoring-Stage Results

3.1 Format Fixes Score Structure

For each diagnostic in Table 1 we decompose the sum of squares over the 15 format–judge cells into format, judge and residual shares (Table 2); with one observation per cell, the residual absorbs the interaction.

On structural diagnostics, format explains 97–99% and the judge under 0.5%. This is partly by construction: a K -level alphabet has tie probability at least $1/K$, and binary sits almost on its floor of 0.50. The informative question is how efficiently each judge uses its alphabet. Normalized by nominal capacity, D_2 is still 96% format (log-normalized, 85%): every judge uses binary and Likert-5 at 66–100% of capacity and numeric-100 at 7–10% (§3.2). The residual is not empty: tie probability above the $1/K$ floor is 54% format and 41% residual, so how a judge fills an alphabet is judge-and-format specific.

Reference agreement depends on the weakest judge. The judge accounts for 22–50% of reference agreement in the full grid, but leave-one-judge-out shows this is carried by Nemotron: with DeepSeek and Qwen alone, format explains 99.0–99.7% and the judge 0.1–0.4%. Given a format, DeepSeek and Qwen give nearly identical agreement, whereas Nemotron lowers agreement under every format. The result does not depend on the two comparative formats, whose implementations carry slot bias (§3.4): over the three pointwise formats alone, DeepSeek and Qwen give a format share of 98.7–99.5% for τ_a , Somers’ D and top-1 (93% for global Spearman; Appendix C, Table 7). With three judges we do not generalize this to a capability law.

Format	DeepSeek-V4-Flash							Qwen3.5-9B							Nemotron-30B-A3B						
	Tie	D_2	TT	Spr	τ_b	τ_a	T1	Tie	D_2	TT	Spr	τ_b	τ_a	T1	Tie	D_2	TT	Spr	τ_b	τ_a	T1
Binary	.501	2.0	.78	.636	.548	.222	.505	.511	2.0	.77	.603	.517	.210	.505	.578	1.7	.78	.517	.462	.156	.481
Likert-5	.240	4.2	.60	.685	.527	.344	.582	.304	3.3	.64	.680	.533	.337	.572	.261	3.8	.66	.427	.357	.224	.483
Numeric-100	.143	7.0	.45	.722	.515	.384	.607	.101	9.9	.40	.708	.502	.398	.616	.121	8.2	.47	.532	.401	.297	.537
Anchored	.365	2.7	.23	.412	.401	.315	.526	.259	3.9	.36	.397	.390	.308	.526	.317	3.2	.67	.207	.125	.086	.396
Tournament	.072	13.8	.05	.282	.336	.304	.544	.036	27.7	.10	.299	.335	.292	.536	.051	19.7	.04	.060	.036	.035	.378

Table 1: Scoring-stage diagnostics on the identical 15,406-candidate pool. *Tie*: probability two random candidates receive the same score; $D_2 = 1/\text{Tie}$: effective number of equally used levels. *TT*: fraction of prompts whose top score is tied, so that the selected response is drawn at random. *Spr*: global Spearman with the GPT-4 reference; τ_b , τ_a : mean within-prompt Kendall; *T1*: rate at which the top-scored response (one seeded tie-break; the exact expectation differs by at most 0.006) is a reference-top response; 28% of prompts have a tied reference top, so uniform selection scores 0.357. Pairs tied in the reference count in no τ numerator. Bold: best format per judge. Distinct-value counts and Somers’ D are in Appendix C, Table 5.

Diagnostic	All three judges			w/o Nem.
	Fmt %	Judge %	Res. %	Judge %
Tie probability	97.0	0.4	2.6	0.5
Distinct values	99.1	0.2	0.6	0.0
$D_2/\text{nominal levels}^\dagger$	96.0	0.3	3.8	0.0
Global Spearman	76.4	21.9	1.7	0.1
Within-prompt τ_b	55.6	36.0	8.5	0.4
Within-prompt τ_a	36.7	49.8	13.5	0.2
Top-1 agreement	38.9	49.2	11.9	0.1

Table 2: Variance shares over the format×judge grid. The last column drops Nemotron (DeepSeek and Qwen only); dropping either other judge leaves the judge share at 0.0–0.5% (structural) and 23–50% (reference agreement). † Pointwise and anchored formats only. Full leave-one-judge-out results in Appendix C, Table 6.

3.2 Nominal Granularity is Not Effective Granularity

Numeric-100 offers twenty times the levels of Likert-5, but judges emitted only 33–37 of its 101 values and placed 74.7%, 95.9% and 99.6% of their mass (Nemotron, Qwen, DeepSeek) on multiples of five. Counting distinct values overstates resolution, because a value used twice counts the same as one used nearly 3,800 times. The effective number of equally used levels, $D_\alpha = \exp(H_\alpha)$ for Rényi entropy H_α , corrects this: numeric-100 is worth 9.1–12.4 levels at $\alpha = 1$ and 7.0–9.9 at $\alpha = 2$, against 3.3–4.4 for Likert-5, two to three times the effective resolution rather than twenty. D_2 is exactly the inverse tie probability, so the tie column of Table 1 already reports it. It is also stable where counts are not: at a third of the pool, distinct values fall from 37 to 28.0 ± 2.0 while D_2 moves by about 0.1 s.d. or less (Appendix D, Table 10).

3.3 Kendall’s τ_b Misleads a Selection Audit

By τ_b , binary is the best format under DeepSeek and Nemotron and second under Qwen; by τ_a it is last under DeepSeek and Qwen. τ_a , Somers’ D and top-1 agreement all rank numeric-100 first, Likert-5 second and binary last under DeepSeek and Qwen, differing only on anchored versus tournament. τ_b (Kendall, 1945) divides by $\sqrt{(n_0 - n_1)(n_0 - n_2)}$, where n_1 counts judge-tied pairs, so coarse formats are divided by a smaller number. When the reference orders a pair strictly and the judge ties it, the selector breaks the tie at random, so the pair’s expected concordance is zero, which is how τ_a scores it and not how τ_b does. For example, with reference order $a > b > c$ and a binary judge accepting a, b , $\tau_a = 2/3$ and $\tau_b \approx 0.82$, yet the selector picks a only half the time. This is the common case here: binary leaves the top score tied in 77–78% of prompts (TT in Table 1), against 40–47% for numeric-100. The argument extends Deutsch et al. (2023) from meta-evaluation to selection.

3.4 Comparative Formats Inherit Slot Bias

Read each comparative verdict as a margin in favor of Response A. A genuine preference for the candidate follows the candidate, so its sign flips when the candidate moves from slot A to slot B; a preference for a slot does not move. Running 500 anchored candidates per judge in both orders therefore gives $v_{\text{fwd}} = q + b$ and $v_{\text{rev}} = -q + b$, separating the mean preference for the candidate q from the slot-A bias b (Table 4, Appendix B). The bias differs in sign on identical inputs: $b = -0.33 (\pm 0.09)$ for DeepSeek and Qwen, toward slot B, and $+0.89 [0.79, 0.99]$ for Nemotron. Self-comparisons confirm it without the additive model: the anchor is one of the candidates, so 4,000 queries compare a response

r		Contrast	DS-Pro	Gemini	Grok	Olmo
4/4	50	anch. > rand.	+0.093*	+0.086*	+0.084*	+0.049*
	50	Lik. > rand.	+0.099*	+0.087*	+0.110*	+0.054*
	50	num. > rand.	+0.080*	+0.088*	+0.091*	+0.038*
3/4	25	Lik. > rand.	+0.065*	+0.077*	+0.068*	+0.027
	25	num. > rand.	+0.061*	+0.094*	+0.072*	+0.009
	50	bin. > rand.	+0.048*	+0.086*	+0.082*	+0.026
	50	Lik. > len.	+0.062*	+0.079*	+0.078*	+0.013
	50	tourn. > rand.	+0.064*	+0.053	+0.058*	+0.047*
1/4	50	bin. < Lik.	-.052*	-.002	-.028	-.029
	50	bin. < num.	-.032*	-.003	-.009	-.013

Table 3: Downstream contrasts by the number of evaluation judges concurring (all 3/4 and 4/4 contrasts; two of the seven 1/4 contrasts). Cells are mean win-rate differences; *: $p_{\text{Holm}} < 0.05$ for that judge (seed-clustered bootstrap).

with itself, where any verdict other than TIE can only reflect position. DeepSeek and Qwen return TIE on 95.8% and 99.6% of these, Nemotron on 42.3%, with a mean verdict of +0.92 toward slot A. Because q is a mean preference rather than discrimination, we also correlate each item’s q_i with its GPT-4 gap to the anchor: 0.44 [0.37, 0.51] for DeepSeek, 0.37 [0.30, 0.44] for Qwen, and 0.10 [0.01, 0.19] for Nemotron, whose anchored scores are mostly position.

The tournament inherits the bias through scheduling: pairs were enumerated in listing order, so the first-listed candidate is always Response A and the last always Response B. Listing order is unrelated to quality (Spearman 0.05 with GPT-4 score), yet it drives selection in each judge’s direction. DeepSeek selects the last-listed response in 44.6% of four-candidate prompts, Nemotron the first-listed in 67.5%, against 25% without a position effect. The tournament’s low agreement (global Spearman 0.06–0.30) is therefore at least partly due to slot bias amplified by unbalanced scheduling.

4 Downstream Consequences

The omnibus Friedman test detects differences among the training conditions at every retention rate under every evaluation judge ($p \leq 0.017$). Every contrast that replicates across judges is a curated format beating a baseline (Table 3): three hold under all four evaluation judges and five more under three. A difference of +0.09 means winning against the reference on about nine more prompts in 100 than RANDOM. Against FULL, trained on all 15,406 candidates, a 400-example

curated subset ($r = 10\%$) is at least as good: 18 of 20 format×judge differences are positive (up to +0.059), though only 2 survive Holm correction, and RANDOM at 10% stays within 0.02 of FULL.

Shared judge taste. Curated students write 1,044–1,257 characters with markdown in 38–46% of responses, against 1,072–1,113 and 36–40% for RANDOM. Adding log length relative to the reference and a markdown indicator to a prompt-fixed-effects regression changes every curated-versus-RANDOM contrast by less than 0.02 (Appendix C, Table 8), so the gains are not a length or markdown effect.

Different data, similar students. The formats do not select the same examples. The Jaccard overlap between two formats’ DeepSeek-curated subsets averages 0.02 at $r = 10\%$, 0.06 at 25% and 0.15 at 50%, so each format trains on a largely disjoint set. That such different subsets yield statistically indistinguishable students suggests that, in this pool, many selections clear the bar that separates curated from random data, and the large scoring-stage differences are compressed by training.

Between formats. No format beats another under more than one evaluation judge. Binary does trail the finer formats in direction, below Likert-5 in all 12 judge×rate comparisons and below numeric-100 in 11 of 12, matching its weaker scoring-stage agreement. These gaps are too small for the design to resolve: with 300 prompts and three seeds it detects a difference of about 0.03–0.05 win-rate points with 80% power, and its 95% CIs rule out format differences larger than 0.064–0.113, depending on the judge. The missing ordering should therefore be read as undetected; it is not evidence that the formats are equivalent.

5 Conclusion

Under a fixed decoding rule, the format sets a judge’s tie structure almost independently of the judge, and tie structure decides how often selection is random; comparative formats also import a judge-specific slot bias, and the format alone largely changes which examples are selected. We propose reporting the following: (1) the format verbatim, with the decoding rule; (2) tie probability (or D_2), with any distinct-value count given alongside the pool size; (3) for comparative formats, how

positions were swapped or balanced, and the estimated b ; and (4) the verdict parse rate. In practice, we recommend Likert-5 or numeric-100 over a binary verdict, swap and balance slots in comparative formats, and validate formats with τ_a , Somers’ D or top-1 agreement rather than τ_b .

Limitations

Decoding. All verdicts use greedy decoding. Probability-weighted scores (e.g., the expected Likert value under the verdict-token distribution, as in G-Eval; Liu et al., 2023) or repeated sampling would make binary and Likert-5 effectively continuous. Our structural results describe format together with greedy decoding, not format alone. **Variance decomposition.** There is one observation per cell, so the residual absorbs the format $\times$ judge interaction; three judges give the judge factor two degrees of freedom, and the judge share in reference agreement is driven by one judge (Table 2). **Reference.** Agreement is measured against UltraFeedback’s GPT-4 scores, not human labels. The same scores choose the anchor shown to the judge in the anchored format and the evaluation reference. **Downstream design.** Training data come from one curation judge (DeepSeek), so we do not test whether a less accurate curation judge yields worse students; one Nemotron-curated condition would. We have no untuned student or GPT-4-oracle selector baseline, so we cannot say whether better reference agreement yields better training data. Our length and markdown adjustment conditions on properties of the trained model’s own outputs, so it is a sensitivity check rather than a causal control. The LENGTH baseline is confounded with truncation (the 1,024-token cap removes 11.4–25.2% of its examples against 1.8–6.2% elsewhere). **Format-specific confounds.** Binary’s disadvantage is inseparable from its random tie-breaking. Anchored is intentionally unswapped, and 26% of its queries compare the anchor with itself, which assigns every anchor the same score. The tournament’s slot assignment follows listing order, so its selections, including the DeepSeek-curated training data, carry each judge’s slot bias; its soft win rate is computed within small slates ($\bar{k} = 3.85$), so a weak slate’s winner can score as high as a strong slate’s. **Scope.** We only use one candidate pool, one student, and one language.

References

- Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. 2024. Alpaga-sus: Training a better alpaca with fewer data. In *International Conference on Learning Representations*, volume 2024, pages 34767–34797.
- Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. 2023. [Ultrafeedback: Boosting language models with scaled ai feedback](#). Preprint, arXiv:2310.01377.
- DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyu Hou, Chenhao Xu, Chenze Shao, Chong Ruan, Conner Sun, and 300 others. 2026. [Deepseek-v4: Towards highly efficient million-token context intelligence](#). Preprint, arXiv:2606.19348.
- Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929.
- Shachar Don-Yehiya, Asaf Yehudai, Leshem Choshen, and Omri Abend. 2026. Mediocrity is the key for llm as a judge anchor selection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15491–15513.
- Milton Friedman. 1937. [The use of ranks to avoid the assumption of normality implicit in the analysis of variance](#). *Journal of the American Statistical Association*, 32(200):675–701.
- Sture Holm. 1979. [A simple sequentially rejective multiple test procedure](#). *Scandinavian Journal of Statistics*, 6(2):65–70.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). Preprint, arXiv:2106.09685.
- M. G. Kendall. 1945. [The treatment of ties in ranking problems](#). *Biometrika*, 33(3):239–251.
- Weiyue Li, Minda Zhao, Weixuan Dong, Jiahui Cai, Yuze Wei, Michael Pocreass, Yi Li, Wanyan Yuan, Xiaoyue Wang, Ruoyu Hou, et al. 2026. Grading scale impact on llm-as-a-judge: human-llm alignment is highest on 0–5 grading scale. *npj Artificial Intelligence*.
- 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.

Justin D Norman, Michael U Rivera, and D Alex Hughes. 2026. [Reliability without validity: A systematic, large-scale evaluation of llm-as-a-judge models across agreement, consistency, and bias](#). *Preprint*, arXiv:2606.19544.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. Rankme: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78.

Jinlong Pang, Jiaheng Wei, Ankit Parag Shah, Zhaowei Zhu, Yaxuan Wang, Chen Qian, Yang Liu, Yujia Bao, and Wei Wei. 2025. Improving data efficiency via curating llm-driven rating systems. In *International Conference on Learning Representations*, volume 2025, pages 100986–101016.

Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, Lichan Hong, Ed H Chi, James Caverlee, Julian McAuley, and Derek Cheng. 2026. How to train data-efficient llms. In *International Conference on Learning Representations*, volume 2026, pages 45491–45545.

Sadman Kabir Soumik. 2026. [Judging the judges: A systematic evaluation of bias mitigation strategies in llm-as-a-judge pipelines](#). *Preprint*, arXiv:2604.23178.

Tuhina Tripathi, Manya Wadhwa, Greg Durrett, and Scott Niekum. 2025. Pairwise or pointwise? evaluating feedback protocols for bias in llm-based evaluation. *arXiv preprint arXiv:2504.14716*.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. 2024. Qrating: Selecting high-quality data for training language models. *arXiv preprint arXiv:2402.09739*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

Xinran Zhang. 2026. How sensitive are safety benchmarks to judge configuration choices? In *International Conference on Intelligent Computing*, pages 173–184. Springer.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhaghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in neural information processing systems*, volume 36, pages 46595–46623.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. 2024. Beyond yes and no: Improving zero-shot llm rankers via scoring fine-grained relevance labels. In *Proceedings of the 2024 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies (volume 2: short papers)*, pages 358–370.

A Method Details

Prompts. Pointwise formats:

You are evaluating the quality of a response to an instruction. Consider helpfulness, correctness, completeness, and clarity.

Instruction: {instruction}

Response: {response}

{format line}

with format lines **Binary**: “Is this response of acceptable quality? Reply with exactly one word: ACCEPT or REJECT.” **Likert-5**: “Rate the response on a 1-5 scale (1=very poor, 2=poor, 3=acceptable, 4=good, 5=excellent). Reply with only the number.” **Numeric-100**: “Rate the response on a 0-100 scale, where 0 is unusable and 100 is flawless. Reply with only the number.” Comparative formats use “You are comparing two responses to the same instruction. Consider helpfulness, correctness, completeness, and clarity.” followed by the instruction, Response A, Response B, and “Which response is better? Reply with exactly one of: A++ (A much better), A+ (A slightly better), TIE, B+ (B slightly better), B++ (B much better).” Verdicts map to +2, +1, 0, −1, −2.

Anchored. The anchor is the candidate at index $\lfloor k/2 \rfloor$ after sorting the prompt’s candidates by GPT-4 score (mean anchor score 6.93 against 6.45 for all candidates, so $q < 0$ is expected); the candidate is always Response A, and the score is $(v + 2)/4$. Because the anchor is drawn from the candidates, 4,000 of the 15,406 queries compare the anchor with itself.

Tournament. All $\binom{k}{2} \leq 6$ within-prompt pairs are judged once, in the order produced by enumerating the prompt’s candidates in listing order, so candidate i is Response A against every later candidate. A verdict v gives the first response win probability $\sigma(v)$ and the second $1 - \sigma(v)$; a candidate’s score is its mean win probability over its matches.

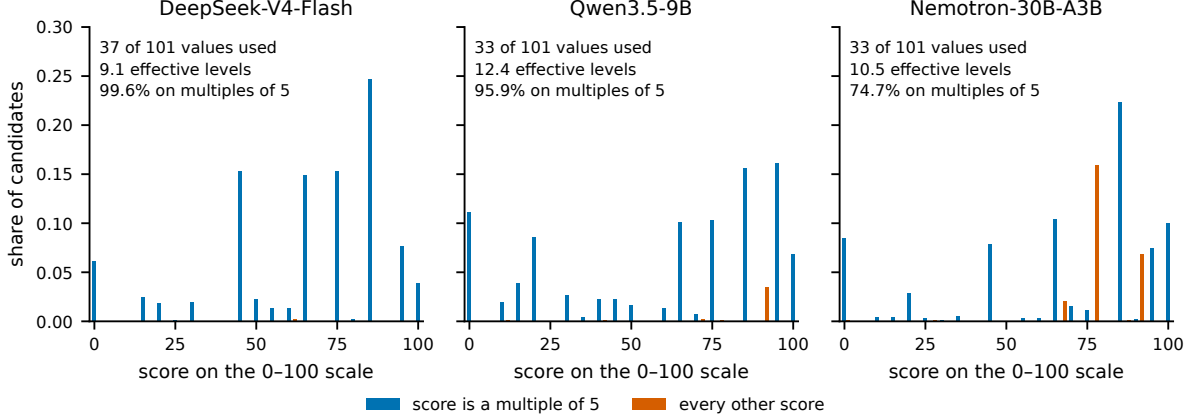


Figure 1: Numeric-100 score distributions, one panel per curation judge; blue bars are multiples of five. Nemotron departs most from the round-number grid (74.7% on multiples of five) yet has the lowest reference agreement.

Fine-tuning. Each model is trained with LoRA (rank 16, $\alpha = 32$, dropout 0.05) for 2 epochs, using AdamW with learning rate 2×10^{-4} , an effective batch size of 8 and bf16 precision. Training sequences are capped at 1,024 tokens. The cap truncates 4.3% of the full pool; across curated subsets it ranges from 1.8% (numeric-100) to 6.2% (tournament), but reaches 11.4–25.2% for LENGTH, which selects the longest responses by design (see Limitations). All training ran on a single NVIDIA RTX 5090. Curation and evaluation judges were queried through hosted APIs, except Olmo-3-7B-Instruct, which we ran locally with vLLM.

Statistics. Every format is evaluated on the same 300 prompts, a repeated-measures design, so we use the Friedman test per judge and rate, paired bootstrap contrasts resampling both prompts and seeds, and Holm correction across format pairs within each rate.

B Anchored Order-Swap Decomposition

Verdicts map to a slot-A-oriented margin: “A much better” $\rightarrow +2$, “A slightly better” $\rightarrow +1$, tie $\rightarrow 0$, “B slightly better” $\rightarrow -1$, “B much better” $\rightarrow -2$. The candidate’s preference travels with the candidate, so its sign flips when the candidate moves to slot B; a position preference stays with the slot. Hence $v_{\text{fwd}} = q + b$, $v_{\text{rev}} = -q + b$, $q = (v_{\text{fwd}} - v_{\text{rev}})/2$, $b = (v_{\text{fwd}} + v_{\text{rev}})/2$. The additive model treats the five verdicts as an interval scale and can be distorted near the clipped endpoints. The reverse pass uses a seeded sample of 500 candidates per judge, excluding self-

comparisons; 95% CIs below are from 10,000 bootstrap resamples of items.

Judge	q	b	$\rho(q_i, \Delta)$	Self-TIE
DeepSeek	-0.52 [-.65, -.39]	-0.33 [-.42, -.24]	0.44 [.37, .51]	95.8%
Qwen	-0.16 [-.26, -.07]	-0.33 [-.42, -.24]	0.37 [.30, .44]	99.6%
Nemotron	-0.07 [-.14, +.00]	+0.89 [+.79, +.99]	0.10 [.01, .19]	42.3%

Table 4: Order-swap decomposition of anchored verdicts ($n = 500$ per judge; 95% CIs below). q : preference for the candidate over the anchor net of position (< 0 : anchor preferred, as expected). b : slot bias (> 0 : favors Response A). $\rho(q_i, \Delta)$: per-item correlation of q_i with the GPT-4 score gap, i.e. whether preferences track quality. Self-TIE: TIE rate when the anchor is compared with itself ($n = 4,000$). DeepSeek and Qwen share a small slot-B bias but track quality; Nemotron is dominated by a slot-A bias.

DeepSeek and Qwen have indistinguishable slot bias, so the data do not support a monotone relation between capability and position bias; three judges could not establish one in any case. Prior reports range from at most 0.04 across frontier judges (Soumik, 2026) to above 0.10 in production judges (Norman et al., 2026).

C Additional Diagnostics

Table 5 gives distinct-value counts and Somers’ D ; Table 6 the leave-one-judge-out shares; Table 7 the pointwise-only decomposition; and Table 8 the length and markdown-adjusted contrasts.

Judge	Format	Vals	Somers' D
DeepSeek	Binary	2	.261
	Likert-5	5	.406
	Numeric-100	37	.457
	Anchored	5	.369
	Tournament	65	.364
Qwen	Binary	2	.247
	Likert-5	5	.398
	Numeric-100	33	.472
	Anchored	5	.361
	Tournament	66	.350
Nemotron	Binary	2	.187
	Likert-5	5	.266
	Numeric-100	33	.355
	Anchored	5	.098
	Tournament	56	.036

Table 5: Distinct values emitted and mean within-prompt Somers' D .

Judge share (%)	All	−DS	−Qwen	−Nem.
Tie probability	0.4	0.4	0.0	0.5
D_2 /nominal [†]	0.3	0.4	0.2	0.0
Distinct values	0.2	0.2	0.3	0.0
Global Spearman	21.9	22.8	23.7	0.1
τ_b	36.0	32.4	34.4	0.4
τ_a	49.8	47.0	50.4	0.2
Somers' D	49.1	46.1	49.4	0.2
Top-1 (expected)	49.0	46.2	49.4	0.3

Table 6: Leave-one-judge-out judge share of the sum of squares, computed from unrounded values; top-1 uses the exact expectation under random tie-breaking. Column headers name the judge removed. [†]Binary, Likert-5, numeric-100, anchored.

Diagnostic	All three judges		DeepSeek + Qwen	
	Fmt %	Judge %	Fmt %	Judge %
Tie probability	97.5	0.4	98.0	0.1
Global Spearman	9.9	80.2	93.4	4.6
τ_b	7.7	80.8	52.9	19.8
τ_a	70.3	27.5	99.4	0.0
Somers' D	71.5	26.3	99.5	0.0
Top-1 (expected)	58.0	36.4	98.7	0.2

Table 7: Variance shares over the three pointwise formats only (binary, Likert-5, numeric-100), which have no slot-assignment confound.

$r = 50\%$	DS-Pro	Gemini	Grok	Olmo
Binary	.048/.047	.086/.083	.082/.080	.026/.025
Likert-5	.099/.094	.087/.083	.110/.106	.054/.052
Numeric-100	.080/.081	.088/.086	.091/.090	.038/.038
Anchored	.093/.080	.086/.083	.084/.078	.049/.044
Tournament	.064/.051	.053/.049	.058/.052	.047/.042

Table 8: Curated minus RANDOM win rate at $r = 50\%$, unadjusted / adjusted for log response length relative to the reference and a markdown indicator (OLS with prompt fixed effects over seed-level outcomes). The markdown indicator is 1 if a response contains a heading, a bullet or numbered list item, bold text, or a code fence. The largest change is 0.013 (at $r = 25\%$ it is smaller); dropping numbered lists, adding tables, using headings and bold only, or using a log count of markdown elements gives 0.009–0.014.

D Effective Granularity

For a score distribution p , the Rényi entropy of order α is $H_\alpha = \frac{1}{1-\alpha} \log \sum_i p_i^\alpha$ and $D_\alpha = \exp(H_\alpha)$ is the number of equally used levels with the same entropy. D_1 is the exponentiated Shannon entropy; $D_2 = 1 / \sum_i p_i^2$ is the inverse tie probability, and $D_2 \leq D_1$. A distinct-value count weights a score emitted twice the same as one emitted 3,791 times: DeepSeek's four most frequent numeric-100 scores (85, 75, 45, 65) carry 70.2% of the mass. Figure 1 shows the numeric-100 distributions, and Table 9 lists nominal, emitted and effective levels for every pointwise format.

Judge	Format	Nom.	Dist.	D_1	D_2
DeepSeek	Binary	2	2	2.00	1.99
	Likert-5	5	5	4.44	4.17
	Anchored	5	5	3.17	2.74
	Numeric-100	101	37	9.07	6.98
Qwen	Binary	2	2	1.98	1.96
	Likert-5	5	5	3.76	3.28
	Anchored	5	5	4.34	3.86
	Numeric-100	101	33	12.39	9.94
Nemotron	Binary	2	2	1.85	1.73
	Likert-5	5	5	4.27	3.84
	Anchored	5	5	3.65	3.16
	Numeric-100	101	33	10.55	8.23

Table 9: Nominal levels, distinct values emitted, and effective levels.

Binary and Likert-5 use close to their nominal capacity. Anchored uses fewer effective levels than Likert-5 under DeepSeek and Nemotron, consistent with a verdict distribution concentrated by the fixed anchor and slot bias; under Qwen the order reverses (D_1 4.34 vs. 3.76).

	DeepSeek	Qwen	Nemotron
Distinct	28.0 ± 2.0	27.8 ± 1.5	27.8 ± 1.4
shift	-9.0	-5.2	-5.2
D_1	$9.05 \pm .09$	$12.36 \pm .12$	$10.52 \pm .11$
shift	$-.02$	$-.03$	$-.03$
D_2	$6.98 \pm .08$	$9.93 \pm .11$	$8.22 \pm .10$
shift	$-.00$	$-.01$	$-.01$

Table 10: Numeric-100 measures at $n/3$ (mean $\pm$ s.d., 2,000 resamples without replacement) and their shift from the full-pool value.

Subsampling to a third of the pool (Table 10) removes 16–24% of the distinct-value count and collapses all three judges to 27.8–28.0, erasing the full-pool difference (37 vs. 33). D_1 and D_2 shift by at most 0.3% of their value and preserve the judge ordering; the D_2 shift is about 0.1 s.d. or less and the D_1 shift within 0.3 s.d.