Model judges *and their biases*

When one language model grades another, how far the grade can be trusted, which distortions have been measured, and what statistics can and cannot repair.

MITWA LAB, MUMBAI

IN PLAIN WORDS

Checking whether a chatbot's answers are any good used to need people, who are slow and expensive. Many researchers now ask a second language model to do the grading, and this piece asks how far those grades can be trusted. On some tasks a strong model judge agrees with people about as often as people agree with each other. On others it does much worse, and a high agreement figure can hide large errors. Judges are swayed by which answer they read first, by length, and by writing that resembles their own, and the first of these is easy to correct. A small set of grades from people can keep a model judge honest, but a recent proof shows that those grades cannot be mostly replaced when the systems being judged are as capable as the judge.

Why one model is asked to grade another

Open-ended answers have no answer key. A summary, a piece of advice or a turn in a conversation can be good in several ways at once, and the traditional way to compare two systems is to show their outputs to people and ask which is better. That takes time and money and is hard to repeat. A common substitute is to give both answers to a strong language model, together with instructions describing a good answer, and to record its verdict. The literature calls this LLM-as-a-judge.

The main evidence for the method has been agreement with people, and the main evidence against it has been measured biases. The sections below take five questions in turn, setting the best case for the judge against the best case against it before saying how far the evidence goes. Our technical note 02, Eleven words, works through one of these biases in detail, so the section on length is short and points there.

Does a model judge agree with people?

The case for rests mainly on Zheng et al. (2023). Six chat models answered the 80 multi-turn questions of MT-bench, and the authors collected about 3,000 pairwise votes from 58 expert raters, most of them graduate students, alongside verdicts from GPT-4. Counting only votes that named a winner, GPT-4 agreed with a randomly chosen rater on 85% of first-turn comparisons, while two different raters agreed with each other on 81%. With ties kept in, both figures dropped to about two in three (66% and 63%), against a chance level of one in three rather than one in two. Agreement also depended on how far apart the two models were, climbing from about 70% for closely matched pairs to nearly 100% for pairs with a wide gap in win rate. An earlier study pointed the same way with a different measure. Chiang and Lee (2023) gave a model the same questionnaire as three English teachers, who each rated 200 human-written and 200 machine-generated stories, and found positive rank correlations with the teachers' average ratings on all four qualities rated, weak to moderate for three of them and strong only for relevance.

The case against comes from studies that changed the task, the raters or the measure. Wang et al. (2024) had three of the authors compare ChatGPT with Vicuna-13B on 80 questions. Measured against their majority vote, a single author matched it 71.7% of the time on average and GPT-4, used in the usual way, 52.7% (Cohen's kappa 0.54 and 0.24). The comparison favours the authors, since each one's vote is part of the majority. The sample is small and the raters were the authors, but the gap is wide. Koo et al. (2024) had 16 models judge one another's outputs and compared the resulting rankings with

those of six crowdworkers, who ranked the outputs of 13 models, using rank-biased overlap, a similarity score for ordered lists. Model and human rankings overlapped by 0.44 on average, against 0.54 among the humans. Bavaresco et al. (2025) tested 11 models on 20 datasets that already carried human annotations. Some properties, such as instruction following, were judged reliably, while toxicity and safety were among the worst, and on graded annotations every model tracked non-expert raters more closely than experts. Thakur et al. (2025) chose a task on which people almost never disagree, marking short answers to 400 trivia questions, and still found that raw agreement flattered weak judges. One judge agreed with the human marks more than 80% of the time yet scored only 59 on Scott's pi, a chance-corrected statistic here scaled to 100, and judges above 90% agreement could still be more than 10 points away from the human score for a model under test.

The strongest objection to the whole comparison is that people are not a clean reference. Hosking et al. (2024) asked crowdworkers to mark errors in answers whose tone had been made more or less assertive, then compared their marks with the authors' careful annotation of a subset. The crowdworkers missed more factual errors when the answer sounded sure of itself. Agreement with such raters can reward a judge for sharing their blind spots, a limit that Bavaresco et al. (2025) also acknowledge in their own design.

On agreement, then, the evidence is mixed. The figure of over 80% is real, and Zheng et al. (2023) found a similar 87% against crowd votes on Chatbot Arena, but both figures count only votes that named a winner and use a raw agreement rate. On other tasks and measures the same kind of judge does markedly worse. Because these studies did not measure one quantity, their numbers cannot be placed on one scale. Read together, they suggest that agreement has to be measured afresh for each task, which is what Bavaresco et al. (2025) recommend.

Does the order of the two answers decide the verdict?

Here the two sides agree on the facts. Zheng et al. (2023) built pairs of similar answers by sampling one model twice, and showed each pair in both orders. GPT-4 gave the same verdict both ways in 65% of cases. Claude-v1 managed it in 23.8% and favoured whichever answer came first in 75%. Wang et al. (2024) showed how far this can move a headline result. With ChatGPT as the judge, Vicuna-13B beat ChatGPT on 2.5% of the 80 questions when its answer was shown first, and on 82.5% when it was shown second. GPT-4 was steadier without being immune, giving Vicuna 51.3% in one order and 23.8% in the other. Both papers found the effect largest when the two answers were close in quality. Koo et al. (2024) report that their human raters were also swayed by order, though less often than the Vicuna judge they set beside them.

The defence is that this bias is cheap to remove. Zheng et al. (2023) suggest judging every pair in both orders and declaring a winner only when the two verdicts match. Wang et al. (2024) averaged over both orders and had the judge set out its reasons before giving a score. On their 80 questions this lifted GPT-4's agreement with the authors' majority from 52.7% to 62.5%, still below the 71.7% of a single author.

Order bias is large, has been found by several groups and has a known remedy that costs one extra judgement per pair, so a judged result that does not say whether both orders were used is incomplete.

Does a judge prefer longer answers?

Saito et al. (2023) found that in their setting GPT-4 favoured the longer of two answers more than human raters did. Zheng et al. (2023) padded 23 answers by adding a reworded copy of the list each one contained. GPT-3.5 and Claude-v1 preferred the padded version in 91.3% of cases, GPT-4 in 8.7%. Our note 02 sets out what such a taste for length does to the score of a deliberately short reply, and why it has to be estimated before that score can be read, so we leave the subject there.

Does a judge prefer its own writing?

On this question the evidence is suggestive and the mechanism is disputed. Zheng et al. (2023) saw GPT-4 give itself a win rate 10% higher than human raters gave it, and Claude-v1 25% higher, but GPT-3.5 did not favour itself, and the authors judged their data too limited to decide. Panickssery et al. (2024) worked with news summaries. Without any extra training, GPT-4 told its own summaries apart from those of two other models and of people 73.5% of the time. When the authors trained GPT-3.5 and Llama 2 further to raise or lower that ability, each model's preference for its own summaries rose and fell with it, and across the further-trained versions the two were linearly correlated. The study did not control for the true quality of the summaries, and its authors describe the causal account as supported rather than settled.

Wataoka et al. (2024) used Chatbot Arena pairs that people had already voted on, so each verdict could be set against a human one. Of eight judges, GPT-4 showed the largest gap. When people preferred GPT-4's answer, GPT-4 agreed 94.5% of the time, and when people preferred the other answer it agreed only 42.5% of the time. Three open models (stablelm-tuned-alpha-7b, dolly-v2-12b and oasst-pythia-12b) leaned the other way and rated their own answers below the human verdict. For six open models the authors could also measure perplexity, which is low when a text is unsurprising to the model reading it. Most of these judges rewarded low-perplexity answers more than people did, whoever had written them. That points to familiarity rather than recognition of the self, although GPT-4, the judge with the largest gap, could not be included in that analysis.

Several studies find that judges favour text resembling what they would write themselves, and whether that is a preference for the self or for the familiar remains open. In either case a judge from the same family as one of the systems under test is a party to the comparison.

Can a few human labels correct a biased judge?

The most ambitious defence of model judges is statistical. If a judge is biased, its bias can be measured on a small set of items that people have also labelled, and subtracted. Prediction-powered inference, introduced by Angelopoulos et al. (2023), puts this on a formal footing. It combines many machine predictions with a few gold-standard labels and returns confidence intervals that stay valid whatever errors the machine makes, and that shrink as its predictions improve. Boyeau et al. (2025) applied it to model evaluation. Using a few hundred of 16,000 Chatbot Arena votes over 20 models as human labels, with GPT-4o-mini as judge, the corrected estimates were as precise as the human labels alone would have been with 20 to 25% more of them, and with other judges the gain ranged from 20 to 35%. In their image classification and protein experiments it was about 50%.

Dorner et al. (2025) asked how large such gains can become. Their main result concerns scores that are simply right or wrong. If the judge's agreement with the true scores is no higher than the score of the model it is judging, then no unbiased method can reduce the number of human labels needed by more than half, however many judge labels are added. In experiments on MMLU and MT-bench, savings above a factor of two appeared only when a leading model judged a much weaker one. They also prove that an agreement rate places no useful limit on bias. A judge that agrees with the truth on a share r of items can be off by as much as 1 minus r in either direction, so separating leading models that differ by a few percentage points would, without further assumptions, need agreement above 99% for every model judged.

The two views are compatible. Prediction-powered inference makes a judged score honest, with an interval that means what it says, and the theorem says that at the frontier, where the systems under test may be as capable as the judge, honesty is most of what correction buys, and at least half of the human labelling remains, often more in the experiments of Dorner et al. (2025). Statistics can remove the bias from a judged score once people have labelled a sample, but they cannot make those people unnecessary.

What is not yet known

- How model judges behave on long conversations with emotional content, where human raters may disagree more and none of the studies reviewed here measured agreement.
- Whether a judge's biases stay the same when its provider updates it, which matters for any comparison repeated over months.
- Which human reference a judge should be validated against when crowd and expert raters disagree, and how the human raters themselves should be validated.
- Whether self-preference comes from recognising one's own text or from familiarity, tested on a closed model large enough to show the effect.
- How far graded judge scores, rather than right-or-wrong ones, can push savings past a factor of two at the frontier, which Dorner et al. (2025) bound only in a weaker form, and what happens when the true scores themselves are graded.

References

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. (2023). Prediction-powered inference. Science, 382(6671), 669–674. https://arxiv.org/abs/2301.09633

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., et al. (2025). LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. Proceedings of ACL 2025. https://arxiv.org/abs/2406.18403

Boyeau, P., Angelopoulos, A. N., Li, T., Yosef, N., Malik, J., and Jordan, M. I. (2025). AutoEval done right: Using synthetic data for model evaluation. Proceedings of the 42nd International

Conference on Machine Learning, PMLR 267. https://arxiv.org/abs/2403.07008

Chiang, C.-H., and Lee, H. (2023). Can large language models be an alternative to human evaluations? Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 15607–15631. https://arxiv.org/abs/2305.01937

Dorner, F. E., Nastl, V. Y., and Hardt, M. (2025). Limits to scalable evaluation at the frontier: LLM as judge won't beat twice the data. ICLR 2025. https://arxiv.org/abs/2410.13341

Hosking, T., Blunsom, P., and Bartolo, M. (2024). Human feedback is not gold standard. ICLR 2024. https://arxiv.org/abs/2309.16349

Koo, R., Lee, M., Raheja, V., Park, J., Kim, Z. M., and Kang, D. (2024). Benchmarking cognitive biases in large language models as evaluators. Findings of the Association for Computational Linguistics: ACL 2024, 517–545. https://arxiv.org/abs/2309.17012

Panickssery, A., Bowman, S. R., and Feng, S. (2024). LLM evaluators recognize and favor their own generations. Advances in Neural Information Processing Systems 37. https://arxiv.org/abs/2404.13076

Saito, K., Wachi, A., Wataoka, K., and Akimoto, Y. (2023). Verbosity bias in preference labeling by large language models. arXiv:2310.10076. https://arxiv.org/abs/2310.10076

Thakur, A. S., Choudhary, K., Ramayapally, V. S., Vaidyanathan, S., and Hupkes, D. (2025). Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM2 2025), 404–430. https://arxiv.org/abs/2406.12624

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., et al. (2024). Large language models are not fair evaluators. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 9440–9450. https://arxiv.org/abs/2305.17926

Wataoka, K., Takahashi, T., and Ri, R. (2024). Self-preference bias in LLM-as-a-judge. NeurIPS 2024 Safe Generative AI Workshop. https://arxiv.org/abs/2410.21819

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems 36 (Datasets and Benchmarks Track). https://arxiv.org/abs/2306.05685