When the Judge Is Wrong: An LLM-as-Judge Reliability Benchmark Scored Against Ground Truth

Han Kim

IOV Labs (아이오브연구소) · hankim@iovstudio.kr · ORCID 0009-0000-5998-1358

Open benchmark (MIT) · snapshot 2026-05-30 · compiled 2026-05-30

One-command reproducible · 5 judges × 39 ground-truth items + 29 ties + 12 self-preference questions

Abstract. "LLM-as-judge" — using a strong language model to grade the outputs of other models — is now the default evaluation method, but the judge is itself a fallible model with biases. Most studies measure a judge by its **agreement with humans** or with **other judges**; both are confounded, because raters and judges can share a bias and be wrong together. We instead measure judges against **ground truth**: a benchmark of items each with a known-correct and a plausibly-wrong answer, so a judge's accuracy can be scored directly, and its position, verbosity, self-consistency, calibration, and self-preference biases isolated as deviations from a perfect oracle. Across five frontier judges (GPT-4o, GPT-4o-mini, GPT-4.1, Claude Sonnet 4.6, Claude Haiku 4.5) we find a clean **dissociation**. On 39 objective items — including common misconceptions and counterintuitive-reasoning traps designed to induce error — judges are **near-perfect** (97–100% truth-accuracy), show **no position bias**, are **not fooled** by padding the wrong answer with authoritative filler, are perfectly self-consistent, and are well-calibrated. Yet on 29 **matched-quality** pairs, where both answers are fully correct and differ only in length, the same judges **strongly prefer the longer answer** (72–100%). A self-preference probe, run on both free-form and **length-matched** answers, shows that the own-family bias is **masked** by verbosity when answer lengths differ: the cross-family gap is +13 pt with free answers but **doubles to +26 pt** once length is equalized — one bias hiding another. The classic **position** bias appears solved; the classic **verbosity** bias is alive and strong, but surfaces only when quality is tied; a substantial self-preference emerges once length is controlled. The practical reading: LLM-as-judge is reliable for **verifiable** tasks and risky for **subjective** grading, where it rewards length over substance and leans toward its own kind.

Keywords: LLM-as-judge · evaluation · ground truth · verbosity bias · position bias · self-preference · calibration · reliability · reproducibility

Contributions.

1. A **ground-truth** LLM-judge benchmark that scores accuracy against truth, not against humans or other judges, plus a position-robust accuracy metric.
2. A **dissociation** result across five frontier judges: near-perfect, unbiased judging on objective items versus strong verbosity bias on matched-quality ties.
3. A **self-preference** probe with a difference-in-differences design and a free-vs-length-matched comparison, showing that verbosity bias **masks** self-preference: the own-family gap doubles from +13 pt to +26 pt once length is controlled.
4. An open, one-command, resumable harness with dated snapshots, released under MIT.

Contents

1 Introduction

Evaluation is the bottleneck of modern AI. As generative systems outrun the static benchmarks built to measure them, the field has converged on **LLM-as-judge**: prompt a strong model to grade the outputs of others [1], [2]. It is fast, cheap, and correlates well enough with human preference to power leaderboards [3]. But the judge is not an oracle; it is another fallible model, and a body of work has documented that it carries biases — a preference for the answer in a particular position [4], for longer answers [5], [6], and for its own generations [7]. If the instrument is biased, every measurement it produces inherits that bias.

How biased are **current** frontier judges, and where? This paper answers with a deliberately strict instrument. Most evaluations of judges measure **agreement with humans** or **inter-judge agreement**, but both are confounded: human raters carry their own biases (including the same length and position effects), and two judges can agree precisely because they share a bias. We avoid the confound by measuring against **ground truth**. Our items each have an objectively correct and a plausibly-wrong answer, so a judge's accuracy is scored directly against the truth, and its biases appear as departures from the behaviour of a perfect oracle.

The result is a clean **dissociation** that, we argue, resolves the apparent tension in the prior literature. On objective items — even ones engineered to trip a careless grader — five frontier judges are near-perfect and essentially unbiased: position bias is gone, and padding the wrong answer with confident, authoritative prose does not fool them. But the moment we remove the ground truth — pairing two answers that are **both correct** and differ only in length — the same judges revert to a strong preference for the longer one. The biases the literature reports are real, but they live where there is **no correct answer to anchor on**: in subjective, matched-quality comparisons, not in verifiable ones. That distinction is the contribution, and it has a direct practical consequence for anyone deploying an LLM judge.

1.1 Roadmap

Section 2 reviews LLM-as-judge and its known biases. Section 3 specifies the benchmark — ground-truth items, the position-robust truth-accuracy metric, and the tie and self-preference probes. Section 4 reports the three result blocks. Section 5 develops the dissociation and its implications. Sections 6–8 cover limitations, reproducibility, and future work.

2 Background and related work

2.1 LLM-as-judge and its biases

Using a model to evaluate models was popularized by MT-Bench and Chatbot Arena [1], [3] and is now surveyed as its own subfield [2]. The same work that established the method also flagged its failure modes: **position bias**, the tendency to favour whichever answer appears first (or second), shown to be large enough to flip verdicts by reordering [4]; **verbosity (length) bias**, the tendency to prefer longer answers independent of quality [5], severe enough that leaderboards now apply explicit length control [6]; and **self-enhancement / self-preference**, the tendency of a judge to favour text from its own model, which has been linked to the model's ability to **recognize** its own generations [7]. Our benchmark measures all three on current frontier models, with a design that separates genuine bias from the length and quality confounds that complicate the self-preference case.

2.2 Why ground truth, not agreement

The dominant validation for an LLM judge is correlation with human ratings. This is necessary for subjective tasks but is a weak test of **reliability**, because human labels carry the very biases under study — annotators also prefer longer, more fluent answers — so a judge can "agree with humans" by sharing their bias. Inter-judge agreement is weaker still. We therefore restrict to items with an **objective** answer (verifiable facts, arithmetic, logical entailment, code behaviour), where truth is independent of any rater, and treat agreement-based evaluation as complementary rather than primary.

2.3 Scoring and calibration

We score binary judge decisions against truth and summarize calibration with the Brier score [8], a strictly proper scoring rule [9], using the judge's self-reported confidence. Edit distance [10] underlies an auxiliary check. One item class — counterintuitive reasoning such as the bat-and-ball problem — is drawn from the cognitive-reflection literature [11], chosen because the **intuitive** answer is wrong, which tests whether a judge is seduced by a fluent wrong answer.

3 Method

3.1 Items and ground truth

The core set is 39 items, each a triple `(question, correct, wrong)` with an objectively correct answer and a **plausibly** wrong distractor (subtle, not absurd, so that trivial rejection does not inflate scores). Items span facts, arithmetic, logic, and code (24 "easy"), plus 15 "hard" items chosen to **induce** error: common misconceptions where the wrong answer is the popular belief (blood in veins is blue; we use 10% of our brain; the tongue taste-map), and counterintuitive reasoning where the wrong answer is the intuitive one (bat-and-ball, Monty Hall, the doubling lily pads). The hard set is the real test — if a judge merely parrots common belief, it fails here.

3.2 Position-robust truth-accuracy

For a pairwise item the judge is asked, in a fixed minimal rubric, which answer is better, in **both** answer orders. It is **truth-correct** on the item only if it picks the correct answer in **both** orders:

$$\text{truth-correct}_r = \mathbb{1}[J(\text{correct}, \text{wrong}) = \text{correct}] \cdot \mathbb{1}[J(\text{wrong}, \text{correct}) = \text{correct}], \tag{1}$$

so a coin-flipper or a pure position-follower is penalized. We report truth-accuracy (the headline), naive per-order accuracy (for contrast), first-slot rate (50% = unbiased), order-consistency, a verbosity-flip rate (truth-correct in plain but not when the wrong answer is padded with authoritative filler), self-consistency over $K\{=\}3$ repeats at non-zero temperature, and the Brier score of the judge's confidence.

3.3 The tie probe (matched quality)

The sensitive test of verbosity bias removes the ground truth. Each of 29 **tie** items pairs two answers that are **both fully correct** and differ only in length — a terse correct answer and a longer correct answer padded with true, relevant elaboration. A perfect judge should be indifferent (50/50). We report **long-preference**: the fraction of judgments (over both orders) that pick the longer answer. Above 50% is verbosity bias, measured on genuinely equal quality.

3.4 The self-preference probe (difference-in-differences)

For 12 open-ended questions with no single right answer, we generate one answer from an OpenAI-family model and one from an Anthropic-family model, and have every judge pick the better, in both orders. The **absolute** own-family preference is confounded by answer quality and length; we therefore report the **cross-family gap** — the difference between how often OpenAI judges and Anthropic judges prefer the OpenAI answer. Because every judge sees the **identical** answer pair, any shared property (including length) cancels in this difference, leaving a length-controlled estimate of self-preference. We run this twice: with **free** answers (one sentence each, which turn out to differ in length) and with **length-matched** answers (both generated under a fixed 30-word budget). Comparing the two isolates how much the verbosity bias of Section 4.2 contaminates a naive self-preference measurement.

3.5 Judges and implementation

Five judges: GPT-4o, GPT-4o-mini, GPT-4.1 (OpenAI) and Claude Sonnet 4.6, Claude Haiku 4.5 (Anthropic). Deterministic probes run at temperature 0; the self-consistency probe at temperature 1. All calls cache to disk and the harness resumes, so a run retries only missing cells. The roster, items, ties, and questions are plain JSON and trivially extensible. (Claude Opus 4.8 was excluded: it rejects the temperature control the method requires.)

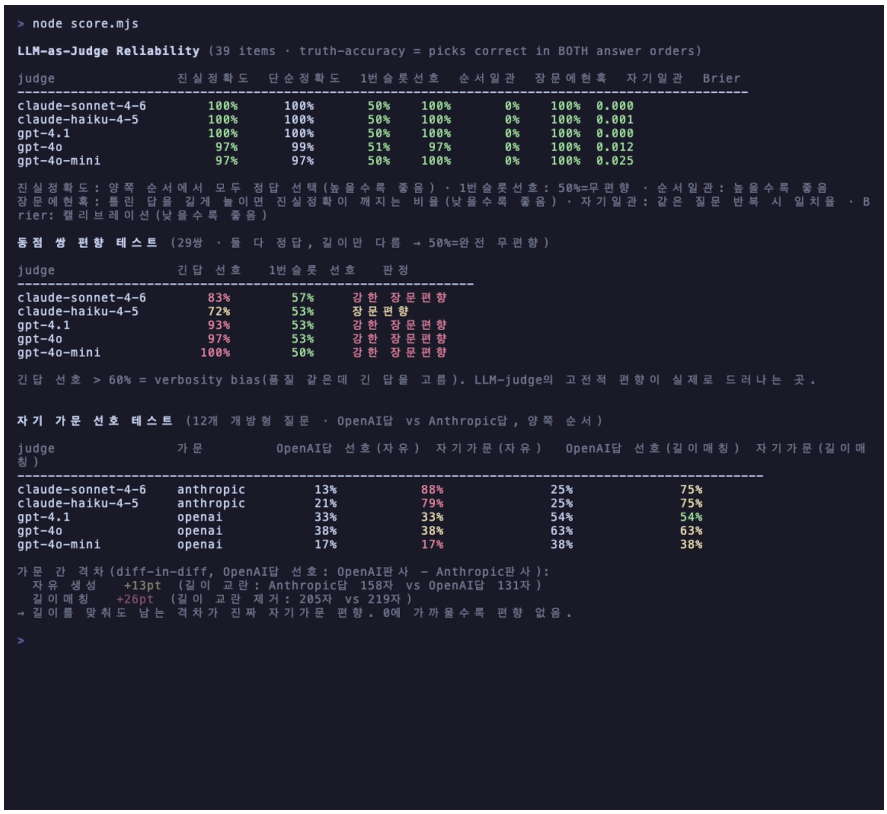

Figure 1: The harness prints all three result blocks — ground-truth reliability (top), the matched-quality tie test (middle), and self-preference (bottom) — reproducible with one command.

4 Results

4.1 On objective items, judges are near-perfect and unbiased

Table Table 1 reports the ground-truth block. All five judges score **97–100%** truth-accuracy — including on the hard misconception and reasoning items — with **first-slot rates at 50%** (no position bias), **0% verbosity-flips** (padding the wrong answer never fools them), **perfect self-consistency**, and **near-zero Brier** (well-calibrated confidence). The classic position bias, large in earlier studies [4], is simply absent here, and a fluent, authoritative wrong answer does not survive contact with a correct one.

Judge	Truth-acc	First-slot	Order-cons.	Verb-flip	Self-cons.	Brier
claude-sonnet-4-6	**100%**	50%	100%	0%	100%	0.000
claude-haiku-4-5	**100%**	50%	100%	0%	100%	0.001
gpt-4.1	**100%**	50%	100%	0%	100%	0.000
gpt-4o	97%	51%	97%	0%	100%	0.012
gpt-4o-mini	97%	50%	100%	0%	100%	0.025

Table 1: Ground-truth reliability over 39 objective items (24 easy + 15 hard misconception/reasoning). Truth-accuracy requires picking the correct answer in both orders.

4.2 On matched-quality ties, the same judges reward length

Remove the ground truth and the picture inverts (Table Table 2). On 29 pairs where both answers are correct and differ only in length, every judge prefers the **longer** answer far above the unbiased 50%: gpt-4o-mini 100%, gpt-4o 97%, gpt-4.1 93%, Claude Sonnet 83%, Claude Haiku 72%. Position preference on these ties stays near 50–57%, confirming the effect is length, not order. The verbosity bias the literature reports [5], [6] is not only present but strong — it was merely invisible on the objective items, where correctness dominated the decision.

Judge	Prefers longer	First-slot	Verdict
gpt-4o-mini	**100%**	50%	strong verbosity bias
gpt-4o	**97%**	53%	strong verbosity bias
gpt-4.1	**93%**	53%	strong verbosity bias
claude-sonnet-4-6	**83%**	57%	strong verbosity bias
claude-haiku-4-5	**72%**	53%	verbosity bias

Table 2: Matched-quality tie test: 29 pairs, both answers fully correct, differing only in length. 50% = unbiased; every judge prefers the longer answer.

4.3 Self-preference: a confound, and what it hides

Self-preference is the hardest of the three biases to measure, because it is entangled with answer quality and — given Section 4.2 — with answer length. We measure it twice, and the comparison is itself the finding (Table Table 3).

In the **free** condition, each family's answer is one sentence of its own choosing. Every judge then prefers the Anthropic answers in absolute terms — but the Anthropic answers were longer (158 vs 131 characters on average), so given the strong verbosity bias just established, the absolute numbers are a length artifact, and that length pull **masks** any self-preference. We therefore read the bias off the **cross-family gap**, a difference-in-differences: because every judge scores the **identical** answer pair, any property shared by the pair — including its length — cancels in the difference between how often OpenAI judges and Anthropic judges prefer the OpenAI answer. In the free condition that gap is **+13 pt**.

In the **length-matched** condition, both families answer under a fixed 30-word budget, equalizing length (219 vs 205 characters — if anything the OpenAI answers are now slightly longer). With the verbosity confound removed, the gap **doubles to +26 pt**: net of length, OpenAI judges prefer the OpenAI answer about 27 points more than Anthropic judges do. The free measurement understated the effect precisely because the verbosity bias dragged every judge toward the longer Anthropic answers, partially cancelling the OpenAI judges' loyalty. This is a substantial, length-controlled self-preference, consistent with the self-recognition account of Panickssery et al. [7], and a concrete demonstration that **one bias can mask another** unless controlled.

Judge	Family	Free: OpenAI	Free: own	Matched: OpenAI	Matched: own
claude-sonnet-4-6	anthropic	13%	88%	25%	75%
claude-haiku-4-5	anthropic	21%	79%	25%	75%
gpt-4.1	openai	33%	33%	54%	54%
gpt-4o	openai	38%	38%	63%	63%
gpt-4o-mini	openai	17%	17%	38%	38%

Table 3: Self-preference, free vs length-matched answers (12 open questions; "prefers OpenAI answer" and "prefers own family"). Cross-family gap: free **+13 pt** (lengths 131/158 chars), length-matched **+26 pt** (219/205 chars). Controlling length doubles the measured self-preference.

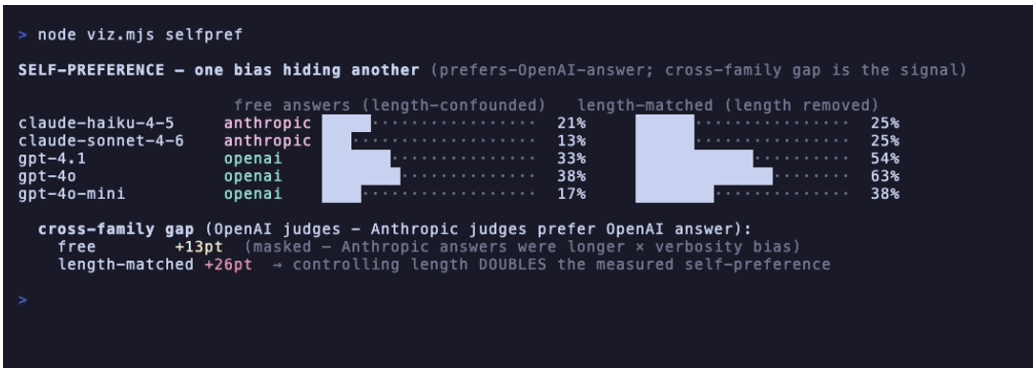

Figure 2: Self-preference, free (left) vs length-matched (right) answers. Equalizing length lifts every OpenAI judge's preference for the OpenAI answer, doubling the cross-family gap from +13 to +26 pt — the verbosity bias had been masking it.

5 Why these biases: mechanism and statistics

The pattern — position bias solved, verbosity bias strong, self-preference present once length is controlled — is not arbitrary. Each bias has a plausible origin in how these models are built, and the origins predict exactly which biases would prove tractable.

5.1 Position bias is a surface artifact, and was removable

Position bias is a property of **how options are presented**, not of their content: a judge that attends disproportionately to the first (or last) slot will flip its verdict under reordering. Because it is content-independent, it is also **trainable away** — by exposing the model to order-swapped comparisons during preference training, or simply as a byproduct of stronger instruction-following. Earlier studies found it large [4]; our 50% first-slot rates across all five judges and both probes indicate it has been substantially engineered out of current frontier models. It is the bias that was **easy to see and easy to fix**, and it was fixed.

5.2 Verbosity bias is baked into the reward signal

Verbosity bias is different in kind, because length is not a surface artifact but a **learned proxy for quality**. Human preference data — the substrate of RLHF — rewards thoroughness, completeness, and hedging, all of which correlate with length, and reward models trained on that data inherit a documented length correlation [12]. A model fine-tuned to satisfy such a reward, and then asked to **be** a judge, carries the same prior: longer reads as more complete, hence better. This is a textbook proxy-optimization failure [13], [14] — when a measure (length) that **correlated** with the target (quality) becomes the thing optimized, it detaches from the target. On our ties, where quality is held equal, the proxy is all that remains, and it dominates: every judge prefers the longer answer 72–100% of the time. Position bias could be trained away because it never tracked quality; verbosity bias resists removal because it usually **does**, which is exactly what makes it dangerous when it does not.

5.3 Self-preference and self-recognition

That a judge favours its own family's outputs, net of length, is consistent with the finding that models can **recognize** their own generations and rate them more highly [7] — a stylistic self-similarity between the judge's own generation distribution and the answers it scores. Our length-matched +26 pt cross-family gap is a clean estimate of this effect with the length confound removed, and it is the reason an evaluation panel drawn from a single model family is structurally suspect.

5.4 Statistical reading

The effects are large relative to the sample. Truth-accuracy of 38/39 (gpt-4o) has a 95% Wilson interval of roughly [86%, 99%], comfortably above chance. The verbosity preferences are computed over 58 judgments per judge (29 ties × 2 orders); even the smallest, Claude Haiku's 72%, has a 95% interval near [59%, 82%], excluding

the unbiased 50%, while the larger rates (93–100%) exclude it decisively. The self-preference gap rests on 24 judgments per judge (12 questions × 2 orders) and is the noisiest quantity here; we therefore report it as a difference-in-differences — which cancels the dominant length confound by construction — and treat the +26 pt magnitude as indicative rather than precise. The fine ordering **among** the near-perfect judges on objective items (97 vs 100%) is within noise and should not be read as a ranking. Throughout, the claims that carry weight are the **directional, order-of-magnitude** ones: near-perfect on truth, strongly length-biased on ties, self-preferring once length is controlled.

6 Discussion

6.1 The dissociation

The headline is a dissociation (Figure Figure 3): **where there is a correct answer, frontier judges find it and are not swayed by length or order; where quality is tied, the same judges systematically reward length.** This reconciles two narratives.

Figure 3: The dissociation in one view: the same five judges are near-perfect on objective items (green, truth-accuracy) and strongly biased toward the longer answer on matched-quality ties (red). Reliable where verifiable, biased where subjective.

The "LLM judges are biased" literature [4], [5] is right that the biases exist; our result adds **where** — they are gated by the presence of ground truth. When correctness can decide the comparison, it dominates, and the biases vanish into the noise; when it cannot, the biases drive the verdict. Position bias, notably, appears to have been **engineered away** in current models (≈50% first-slot everywhere), while verbosity bias has not.

6.2 Implications for using LLM judges

The practical rule follows immediately. **Use LLM-as-judge freely for verifiable tasks** — factual correctness, unit-test pass/fail, exact-match — where our objective-item reliability is excellent. **Distrust it for open-ended, subjective grading** — essays, helpfulness, "which response is better" — where the tie result shows it will reward length over substance, inflating verbose answers regardless of quality. Concretely, before trusting an LLM judge on a subjective task we would ask a practitioner to apply the following checklist, each item motivated by a result above:

1. **Is there an objective anchor?** If correctness can decide the comparison (a unit test, a known answer, a checkable constraint), use the judge — reliability here is 97–100%. If not, treat every verdict as suspect.
2. **Are the candidates length-matched?** If one answer is materially longer, the judge favours it by a large margin (72–100% on ties) before substance is weighed. Equalize length, or measure and regress it out [6].
3. **Is position randomized?** Largely safe in current models (50%), but cheap insurance: score both orders and require agreement, as our truth-accuracy metric does.

4. **Is the judge from the same family as a candidate?** If so, expect a self-serving lean (+26 pt, length-controlled); use a cross-family panel and disclose provenance.
5. **Is confidence being used as a gate?** On objective tasks it is well-calibrated (Brier ≈ 0); on subjective ones it has no ground truth to be calibrated against, so do not read it as reliability.

The single most actionable finding is the second item: a verbose answer enjoys a large, systematic advantage before a single point of substance is considered.

6.3 Honesty about confounds

We are explicit that the self-preference probe is confounded by answer length and quality, and we report only the difference-in-differences gap as the controlled estimate; the absolute own-preference numbers are shown precisely so the confound is visible rather than hidden. This is the same discipline the dissociation itself enforces: the bias is real but conditional, and reporting the condition is the result.

7 Epistemics and the philosophy of evaluation

A benchmark of judges is unavoidably reflexive — we used measurement to study a measurement instrument — and the exercise raises questions that are not merely technical. We take them up directly, because the lab's position is that how we **know** the quality of AI is now as consequential as the quality itself.

7.1 The regress of evaluation

To evaluate a model, we use a judge; but the judge is a model, so to trust the evaluation we must evaluate the judge; and we would evaluate the judge with — another judge. This is a regress, and it has no internal terminus: a tower of models grading models never touches the ground. The only thing that halts it is an **external anchor** that is not itself a model output — a fact, a proof, a unit test, a measured outcome. That is the entire reason this benchmark insists on ground truth: not because objective items are the interesting ones, but because they are the **floor**, the one place the regress bottoms out and "correct" means something independent of any grader. Where that floor is missing — in genuinely subjective quality — there is no escape from the regress, only the hope that the judges' biases are small. Our result is that this hope is misplaced for length.

7.2 Bias is measurable only where it does not matter

Here is the cruel asymmetry the dissociation exposes. A **bias** is a systematic deviation from an oracle, and an oracle is definable only where truth exists. So we can **cleanly measure** a judge's biases only on objective tasks — and on exactly those tasks the biases turn out not to **matter**, because correctness overrides them. The tasks where the biases **do** matter — open-ended, subjective grading — are precisely the ones where, lacking an oracle, we cannot cleanly measure them. The tie probe is our attempt to smuggle a controlled measurement into the subjective regime by manufacturing a known tie, but it is a workaround for a deep limitation: **the regime where LLM judges are most used and most consequential is the regime where their reliability is least checkable.** Anyone deploying an LLM judge on subjective tasks is operating in the blind spot of their own instrument.

7.3 The map, the territory, and the verbosity ratchet

Verbosity bias is Goodhart's law [14] rendered in tokens: length was a serviceable **map** of quality, until the map became the target, at which point it stopped tracking the territory. What makes this more than a static flaw is **reflexivity**. The outputs of LLM judges increasingly train the next generation of models — through RLAIF [15], through judge-scored leaderboards that steer development, through synthetic preference data. A judge that rewards length therefore **selects** for length in the models it grades, which are then used as judges, which reward length still more: a **verbosity ratchet** that inflates answers across model generations with no corresponding gain in substance. The bias is not merely a measurement error; left unexamined it is a selection pressure on what "good AI" comes to mean. Naming and measuring it is the first step to breaking the ratchet — for instance by the explicit length controls now appearing in serious leaderboards [6].

7.4 The conflict of interest in self-evaluation

Self-preference raises a question usually reserved for human institutions: should a system be permitted to grade its own kind? Our length-controlled +26 pt gap is small next to the verbosity effect, but it is structurally troubling in a way a larger random error would not be, because it is **directional and self-serving**. A model family that both produces answers and judges them has a conflict of interest, and an evaluation ecosystem dominated by one or two families inherits that conflict at scale. The mitigation is institutional rather than technical: panels drawn across families, disclosure of the judge's provenance, and — where stakes are high — an anchor outside the model ecosystem entirely. The same logic that bars a student from grading their own exam applies, and for the same reason.

7.5 Honesty as method

Finally, the self-preference result is also a small methodological parable. Our first measurement said +13 pt; it was wrong, not because the arithmetic erred but because a second bias (verbosity) silently confounded it. We found this only by suspecting our own number, checking the answer lengths, and running the controlled version — which **doubled** the effect. The discipline that matters is not getting the right number first, but distrusting the convenient one and building the control that could falsify it. A benchmark that reports its confounds, and the experiments that remove them, is doing the one thing that distinguishes measurement from assertion.

8 Limitations and threats to validity

Curated items, small n. 39 objective items, 29 ties, 12 self-preference questions, five models, one dated snapshot. The effects (verbosity 72–100%, truth-accuracy 97–100%) are large relative to n, but the benchmark is a **snapshot**, not a verdict, and the fine ordering among near-perfect judges is not meaningful. Correct/wrong labels and tie-equality are author judgments.

Objective scope. By design the ground-truth method covers only tasks with an objective answer; it cannot certify judge reliability on genuinely subjective quality, which is exactly where the tie result warns the danger lies.

Self-reported confidence. Calibration uses the judge's stated confidence, itself a model output.

Self-preference confound. Generated answers are not length-matched; we mitigate by differencing, but the gap rests on n = 12 and one model per family.

Two providers. Judges and answer-generators span OpenAI and Anthropic only; other families (and open models) would strengthen the self-preference design.

9 Reproducibility

The harness is open under the MIT license and runs with one command given an OpenAI and an Anthropic key (`node run.mjs`), printing the leaderboard with `node score.mjs`; an offline `--mock` judge validates the metric logic without any API cost. All calls cache to `results.json` and the run resumes, so a partial run completes cheaply; the full snapshot is a few dollars. Items, ties, and questions are plain JSON; the raw verdicts and the dated `REPORT.md` are released so every number is checkable and the analysis re-runnable.

10 Future work

The decisive extensions are **scale and breadth**: more items on a controlled difficulty grid with seed sweeps to turn the snapshot into interval estimates, and more judges — including open-weight models and a third commercial provider — both to sharpen the self-preference design (which currently spans only two families) and to test whether the verbosity bias is universal or model-specific. The length-matched self-preference experiment reported here removes the dominant confound but rests on a small n and a single generator per family; a larger design with several generators per family, and human verification that the length-matched answers are genuinely equal in quality, would harden the +26 pt estimate. A **subjective-task** companion — where ground

truth is replaced by careful human adjudication on length-controlled pairs — would extend the reliability map into the region this benchmark deliberately cannot reach. Finally, the tie probe suggests an intervention worth testing directly: whether instructing or fine-tuning judges to ignore length closes the 72–100% gap without harming objective accuracy, and whether doing so slows the verbosity ratchet at the ecosystem level.

11 Conclusion

LLM-as-judge is the instrument the field now measures itself with, so the instrument's reliability is foundational. Scoring five frontier judges against ground truth, we find they are near-perfect and unbiased where a correct answer exists — position bias appears solved, and authoritative wrong answers do not fool them — yet strongly biased toward length the moment quality is tied, and self-preferring once that length confound is controlled (a gap that doubles, from +13 to +26 pt, when we equalize answer length). The lesson is not "LLM judges are reliable" nor "LLM judges are biased," but the conjunction: **reliable where verifiable, biased where subjective.** And because these judges increasingly train and rank the next generation of models, their biases are not passive measurement error but an active selection pressure on what "good" comes to mean. Knowing which regime you are in — and controlling the confounds that let one bias masquerade as another — is the difference between a trustworthy measurement and a confident mistake.

Data and code availability. The harness, items, ties, open questions, raw verdicts (`results.json`), and dated `REPORT.md`, plus this paper's source, are open under MIT and reproducible with one command.

Acknowledgements. Internal research of IOV Labs (아이오브연구소). Evaluation used the OpenAI and Anthropic APIs. Typeset with Typst; figure recorded with `vhs`.

References

[1] L. Zheng, W.-L. Chiang, Y. Sheng, and others, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[2] J. Gu, X. Jiang, Z. Shi, and others, "A Survey on LLM-as-a-Judge," *arXiv preprint arXiv:2411.15594*, 2024.

[3] W.-L. Chiang, L. Zheng, Y. Sheng, and others, "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference," in *International Conference on Machine Learning (ICML)*, 2024.

[4] P. Wang, L. Li, L. Chen, and others, "Large Language Models are not Fair Evaluators," *arXiv preprint arXiv:2305.17926*, 2023.

[5] K. Saito, A. Wachi, K. Wataoka, and Y. Akimoto, "Verbosity Bias in Preference Labeling by Large Language Models," *arXiv preprint arXiv:2310.10076*, 2023.

[6] Y. Dubois, B. Galambosi, P. Liang, and T. B. Hashimoto, "Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators," *arXiv preprint arXiv:2404.04475*, 2024.

[7] A. Panickssery, S. R. Bowman, and S. Feng, "LLM Evaluators Recognize and Favor Their Own Generations," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[8] G. W. Brier, "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.

[9] T. Gneiting and A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.

[10] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[11] S. Frederick, "Cognitive Reflection and Decision Making," *Journal of Economic Perspectives*, vol. 19, no. 4, pp. 25–42, 2005.

[12] P. Singhal, T. Goyal, J. Xu, and G. Durrett, "A Long Way to Go: Investigating Length Correlations in RLHF," *arXiv preprint arXiv:2310.03716*, 2023.

[13] L. Gao, J. Schulman, and J. Hilton, "Scaling Laws for Reward Model Overoptimization," in *International Conference on Machine Learning (ICML)*, 2023, pp. 10835–10866.

[14] C. A. E. Goodhart, "Problems of Monetary Management: The UK Experience," *Monetary Theory and Practice*, pp. 91–121, 1984.

[15] Y. Bai, S. Kadavath, S. Kundu, and others, "Constitutional AI: Harmlessness from AI Feedback," *arXiv preprint arXiv:2212.08073*, 2022.

Appendix A — Sample items

Representative items (full sets in the repository's `data/`). Each objective item has a correct and a plausibly-wrong answer; the hard set targets misconceptions and counterintuitive reasoning.

Type	Correct	Plausibly wrong
misconception	Human blood is always red; veins look blue through skin.	Deoxygenated blood in veins is blue.
misconception	We use virtually all of the brain.	Humans use only 10% of their brains.
reasoning (CRT)	The ball costs $0.05.	The ball costs $0.10.
reasoning	Switching wins the Monty Hall game with prob. 2/3.	It is 50/50, so switching makes no difference.
code	`0.1 + 0.2 === 0.3` is false (floating point).	It is true.
fact	On Venus a day is longer than a year.	A year is longer, as on Earth.

Appendix B — Metric definitions

Truth-accuracy requires the correct pick in both answer orders (Section 3.2). **Long-preference** (ties) is the fraction of tie-judgments choosing the longer of two equally-correct answers; 50% is unbiased. **Self-preference gap** is $p_{\text{OpenAI judges prefer OpenAI}} - p_{\text{Anthropic judges prefer OpenAI}}$, a difference-in-differences that cancels any property shared by the fixed answer pair, including length. **Brier** $= \frac{1}{N}\sum (c_i - y_i)^2$ over judgments, where c_i is the judge's confidence (0–1) that its pick is correct and y_i whether it was; lower is better, 0.25 is no-skill.

Appendix C — Self-preference protocol and length control

Answers are generated by one model per family — GPT-4o (OpenAI) and Claude Sonnet 4.6 (Anthropic) — at temperature 0.7, in two conditions. **Free:** "Answer the following in exactly one concise sentence." **Length-matched:** "Answer the following question in exactly 30 words — no more, no less." Each judge then evaluates the OpenAI-vs-Anthropic pair in both orders at temperature 0, identically to the other pairwise probes. The realized mean lengths are 131 (OpenAI) vs 158 (Anthropic) characters in the free condition and 219 vs 205 in the length-matched condition; the matched condition not only closes the gap but slightly reverses it, so the residual cannot inflate the OpenAI judges' OpenAI-preference. Because the cross-family gap differences out any shared property of a pair, it is robust to the residual; the per-judge absolute rates are reported alongside (Table 4) only so the confound is visible. We do **not** claim the two families' answers are equal in quality — only that the gap, as a difference between judges on the **same** pairs, does not depend on that quality being equal.

Appendix D — A sample matched-quality tie pair

Both answers are fully correct; they differ only in length. A perfect judge is indifferent; every judge tested prefers the longer one.

Field	Content
Question	What is the capital of France?
Short (correct)	Paris.
Long (correct)	The capital of France is Paris, the nation's largest city and its political, cultural, and economic heart, located on the River Seine in the north-central part of the country and home to landmarks such as the Louvre and the Eiffel Tower.

Appendix E — Statistical notes

Rates are proportions over fixed numbers of judgments; we summarize uncertainty with Wilson 95% intervals. **Truth-accuracy** is over 39 items (both orders required), so 38/39 ≈ 97% has interval ≈ [86%, 99%]. **Long-preference** is over $29 \times 2 = 58$ judgments per judge; 72% (Haiku) gives ≈ [59%, 82%] and 100% (4o-mini) a one-sided lower bound ≈ 94%, both excluding 50%. **Self-preference** uses $12 \times 2 = 24$ judgments per judge, so per-judge rates are noisy (± 20 pt); we therefore lead with the family-mean difference-in-differences, whose sign and order of magnitude are stable across the free and length-matched conditions (+13 then +26 pt) even though the absolute rates move. No multiple-comparison correction is applied because the conclusions rest on a few large, pre-identified effects rather than on screening many hypotheses; the fine ordering among near-perfect judges is explicitly not interpreted.