

Local LLM Evaluation: Why Verifiability Widens Rather than Closes the Human Alignment Gap

Narendar Kumar Ale

University of the Cumberland
Email: [nale02820\[at\]ucumberland.edu](mailto:nale02820[at]ucumberland.edu)

Abstract: *Regulation pulls automated evaluation in two directions at once. Frameworks governing high-risk AI systems require that model outputs be evaluated, while data-protection law treats the evaluation of outputs containing personal data through a hosted third-party judge as processing that requires a formal agreement. For a large class of deployments the consequence is that evaluation is mandatory and a hosted judge is unavailable: the judge must run on hardware the organisation controls. What that constraint costs has not been measured. We evaluate five open-weight judges that fit on a single workstation against three independent expert raters on a public summarisation benchmark, using a reference-matched geometric test: one rater is held out, and the judge and that rater are both scored against the mean of the remainder, so both face an identical target. We report three findings. First, the gap between judge and human is nearly twice as large on the rubric with a checkable answer (+52.6 degrees) as on rubrics that are matters of taste (+27.5 degrees)- the inverse of what is reported for frontier judges, and a result that replicates across two independent cohorts to within 0.3 degrees. Second, these judges disagree with one another more than people do, by up to 54 degrees, with effective rank 4.59-4.85 out of 5: they are near-independent rather than sharing an axis. Third, because their errors are independent noise rather than correlated bias, ensembling beats the best single judge on three of four rubrics but not on the verifiable one, where averaging cannot recover a capability none of them has. One model of the six tested declined the task outright. The practical consequence is uncomfortable: the rubric a constrained deployment most needs a judge for- whether a claim is supported by its source- is the rubric on which locally-runnable judges perform worst.*

Keywords: LLM-as-a-judge, evaluation, alignment, open-weight models, groundedness, data protection, reproducibility

1. Introduction

The argument for this study is a conflict between two bodies of regulation, not a matter of preference.

- Evaluation obligations attach to providers and deployers of high-risk AI systems.
- Sending evaluation data containing personal information to a third-party judge constitutes processing and requires an agreement with that provider.

An organisation subject to both must evaluate without a hosted judge. The existing literature on judge quality does not address this case: it compares frontier judges with one another and with humans, on the implicit assumption that any judge may be called.

Contributions

- 1) A reference-matched geometric protocol in which a judge and a held-out human rater are scored against the same reference, making their angles comparable.
- 2) The first measurement of evaluation fidelity for judges constrained to locally-runnable hardware.
- 3) Evidence that verifiability **widens** rather than narrows the human-judge gap for such judges, inverting the frontier result.
- 4) Evidence that these judges are near-independent rather than sharing a blind-spot axis, with the consequence that ensembling helps where the frontier literature warns it amplifies error.
- 5) Released code, prompts, item identifiers and raw scores.

2. Related work

Judges and their limits. Language models are now the default raters for open-ended text [1]. They track human

raters closely where a rubric has a checkable answer and less closely on subjective judgement.

Consensus is not alignment. Judges may agree because they capture quality or because they share the same blind spots, and agreement statistics alone cannot separate those cases [2]. Mukherjee et al. address this by measuring the *shape* of disagreement rather than its size, treating each judge's scores as a vector and reporting the angle to human scores under a reference-matched comparison. **The geometric framing and the reference-matching protocol in Section 3 are taken from that work;** the contribution here is to apply them to a different axis- the hardware a judge is permitted to run on- and to report that two of its central findings invert under that constraint.

Reliability and validity of judges: Systematic evaluation across agreement, consistency and bias finds reliability is not sufficient for validity [3]. Judge cost at scale has also been examined directly [4].

Deterministic alternatives: Recent work replaces the judge with deterministic tests over what a system retrieved, cited and was permitted to access [5], and decomposes agent skill quality into interpretable signals [6, 7]. This bounds what a judge is needed for; it does not remove the need on subjective rubrics. Hallucination in agent trajectories has been benchmarked separately [8].

Evaluation frameworks: Open frameworks implement faithfulness, relevancy and precision metrics for retrieval and agent systems [9, 10, 11, 12]. None report behaviour under a hardware constraint.

Benchmark: The summarisation benchmark used here provides three independent expert ratings per item across four rubrics [13].

Regulatory context: The obligation to evaluate high-risk systems arises under [14]; the treatment of third-party evaluation as processing requiring a formal agreement arises under [15].

3. Method

3.1 Scores as vectors

Each rater's scores over a fixed item set form a vector. Vectors are centred before comparison: without centring, two raters who both use the upper half of a five-point scale sit at a small angle purely because every entry is positive, reporting agreement that is an artefact of the scale.

The angle between centred vectors measures whether two raters rank items in the same direction- what correlation reports, but without conflating it with shared bias.

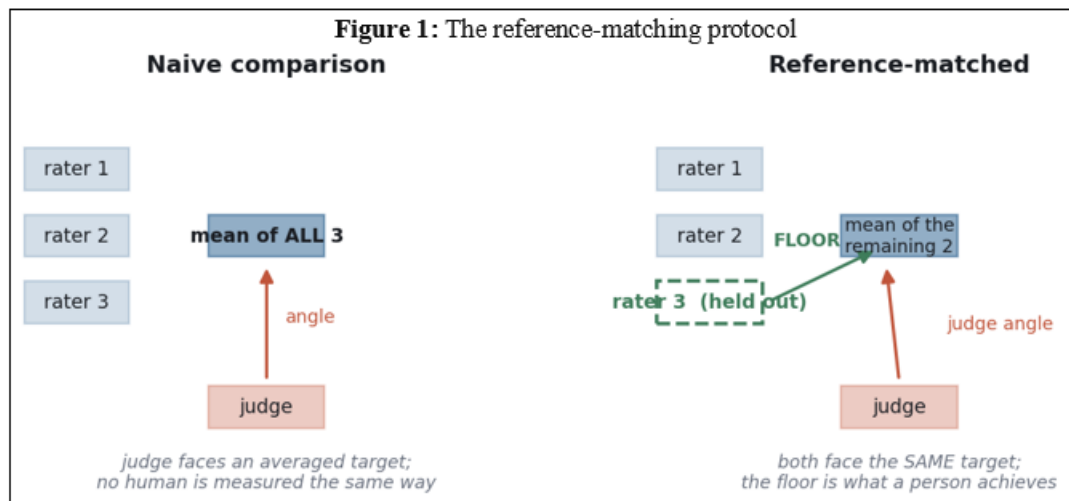


Figure 1: The reference-matching protocol. Left: the naive comparison, in which a judge is scored against the mean of all raters and no human faces the same target. Right: one rater is held out, and the judge and that rater are both scored against the mean of the remainder.

3.2 Reference matching

A judge compared against the mean of three raters faces an easier target than any single rater compared against one colleague, because averaging cancels individual noise. Reporting a judge-against-mean angle beside a rater-against-rater angle therefore overstates the judge.

Every comparison holds out one rater. The judge and that rater are both scored against the mean of the remaining two, and results are averaged over folds. The held-out rater's angle is the **floor**: what a competent human achieves under identical conditions. It is not zero, and it differs by rubric.

3.3 The chance baseline

The expected angle between two independent centred vectors is 90 degrees in any dimension; dimension changes the spread of that distribution, not its mean. The mean is therefore useless as a null. We report the 5th percentile of the null at the item count used- the angle shuffled scores reach by luck one time in twenty. At $n=179$ this is **82.8 degrees**.

This matters for small evaluations: at $n=40$ chance reaches 74.8 degrees against 85.3 degrees at $n=400$, so a judge measured on a small sample can appear aligned without being so.

3.4 Attainment

Raw angles are not comparable across rubrics because floors differ. Attainment is the fraction of the span from chance to the human floor that a judge covers: 1.0 means it reaches the floor, 0.0 means it is no better than shuffled scores.

3.5 Cohort geometry

For a set of judges we report mean pairwise angle and effective rank — the exponential of the entropy of the normalised singular values of the score matrix. A cohort measuring one thing has effective rank near 1 however many judges it contains.

3.6 Exclusion

A judge whose replies cannot be parsed for at least half the items is excluded and reported. Carrying it is not an option: item sets are intersected across judges, so a judge with zero coverage empties the intersection and drives every angle to exactly 90 degrees- an artefact that resembles a finding.

4. Experimental Setup

Benchmark: A public summarisation benchmark with three independent expert ratings per item on four rubrics.

Consistency asks whether the summary's claims are

supported by the source and has a checkable answer; **coherence**, **fluency** and **relevance** are judgements of quality.

Items: 200 sampled stratified across 16 summarisation systems, so the item set spans a range of quality; 179 retained after intersecting coverage.

Judges: Six open-weight models runnable on a single 24 GB workstation, served by a local inference server. All four rubrics are requested in one call. No data leaves the machine.

Token budget: Generation is capped at 400 tokens. An earlier run capped at 60 truncated judges that append justifications, losing 26 per cent of one judge's items. That loss is **not random**: verbose replies cluster on items the judge finds ambiguous, so a tight cap discards the hard items and biases the surviving scores toward easy ones.

Missing scores: A reply not yielding all four rubrics is recorded as missing and the item dropped from every vector, never imputed. Substituting a middle value would pull judges toward the centre of the scale and manufacture agreement that was never expressed.

Determinism: Sampling temperature is zero. We measured what this actually buys: three judges were re-run three times over 40 items, under a fixed seed and under varying seeds.

Every score was identical in every condition- 100 per cent reproducibility, zero variance, across all four rubrics.

This corrects a claim commonly made in passing, including in an earlier draft of this work: at temperature zero on a locally-hosted model, a judge IS reproducible, and the seed is irrelevant because greedy decoding does not sample. The mechanisms usually cited for judge drift- floating-point non-associativity in batched kernels, expert routing- did not move a single score at this scale.

The reproducibility argument against judges therefore does not apply to local deployment in the form it is usually stated. It applies to HOSTED judges, where the weights behind an endpoint can change without notice and the caller cannot detect it. That distinction favours local deployment, and is reported here because it runs against this paper's own framing.

5. Results

5.1 Excluded judge

One of six models returned no parseable score on any item, replying that the task was subjective and outside its ability. It is excluded from the geometry and reported here. **For a constrained deployment, a model that will not grade is a different failure from one that grades badly:** it is unavailable, not inaccurate.

Table 1: Reference-matched angles (n=179, chance = 82.8 degrees)

Rubric	Kind	Human floor	Judge	Angle	Gap	Attainment
consistency	verifiable	21.4	gemma3:4b	61	39.6	36%
			qwen2.5vl:7b	67.8	46.4	24%
			mistral	77	55.6	10%
			qwen2.5-coder:7b	77.7	56.3	8%
			llama3.2	86.8	65.4	0%
coherence	subjective	43.8	qwen2.5-coder:7b	65.9	22.1	43%
			mistral	69.5	25.7	34%
			gemma3:4b	70.3	26.5	32%
			qwen2.5vl:7b	71.2	27.4	30%
			llama3.2	82.3	38.5	1%
fluency	subjective	38.8	gemma3:4b	62.9	24.1	45%
			qwen2.5vl:7b	67.6	28.8	35%
			mistral	75.2	36.4	17%
			qwen2.5-coder:7b	75.2	36.5	17%
			llama3.2	86.2	47.4	0%
relevance	subjective	56.4	qwen2.5vl:7b	70.9	14.5	45%
			gemma3:4b	72.1	15.6	41%
			qwen2.5-coder:7b	75.9	19.5	26%
			mistral	79.7	23.2	12%
			llama3.2	82.2	25.8	2%

Mean gap: subjective +27.5 degrees, verifiable +52.6 degrees.

An independent earlier run with a different three-judge cohort at n=118 gave +27.8 and +52.7 — a replication to within 0.3 degrees.

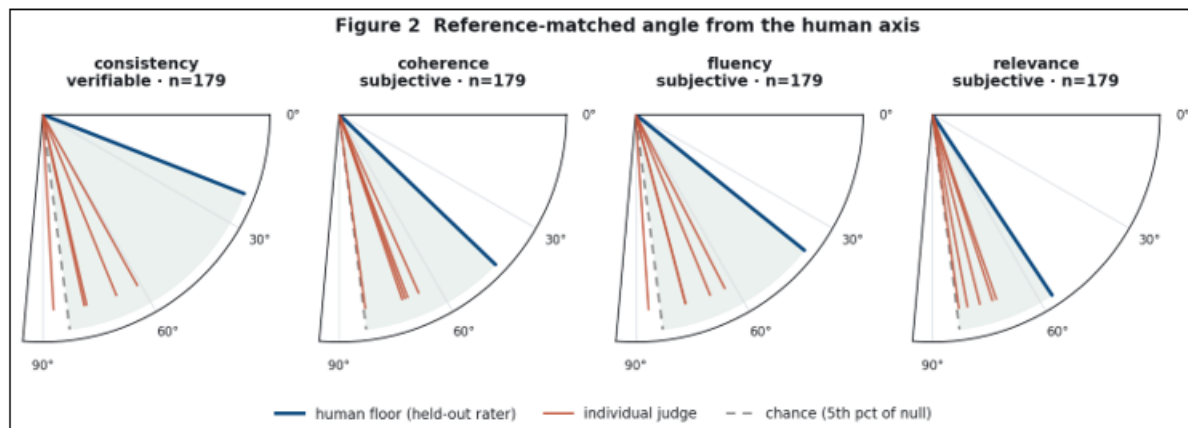


Figure 2: Reference-matched angle from the human axis, per rubric. The solid line is the human floor, thin lines are individual judges, the dashed line is the 5th percentile of the null.

5.2 Verifiability widens the gap

Human floors vary sharply by rubric: **21.4 degrees on consistency**, where the answer is checkable and people converge, against **56.4 degrees on relevance**, where they do not. Judges do not track that variation. On consistency the best judge sits 39.6 degrees beyond a floor of 21.4; on relevance the best sits 14.5 beyond a floor of 56.4.

The mechanism is that **verifiability raises the human bar faster than it raises the judge's capability**. People check facts consistently. Models of this size largely do not, and the gap is the difference.

(Figure 2, Figure 3)

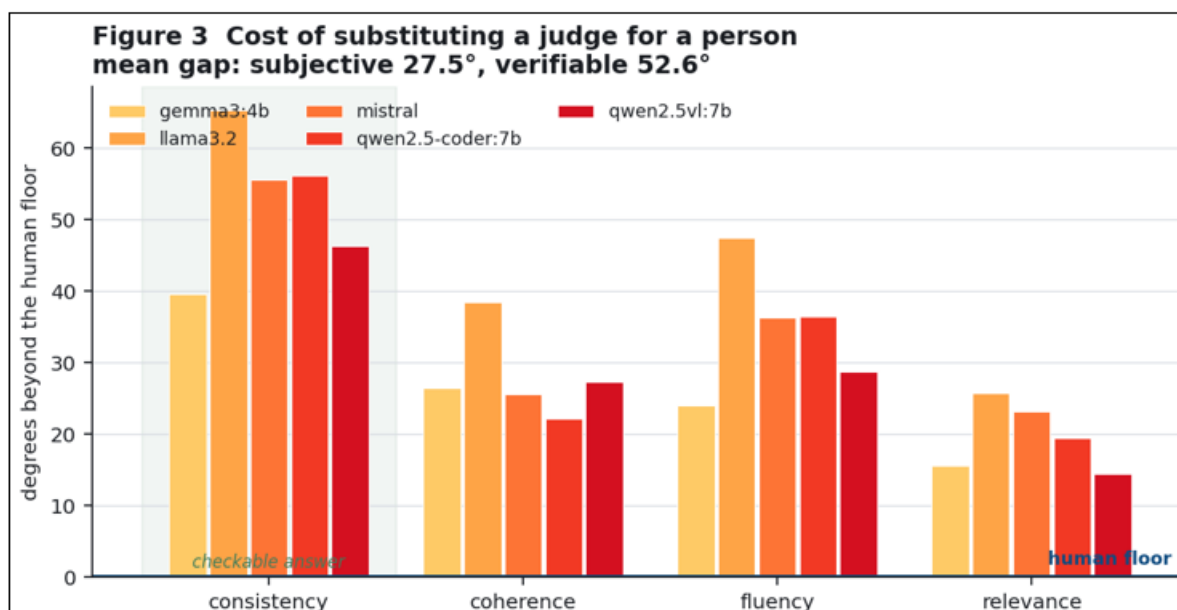


Figure 3: Degrees beyond the human floor, by judge and rubric. The shaded column is the rubric with a checkable answer

Table 2- The flattery effect

Comparing a judge against the mean of all three raters, rather than holding one out, improves its reported angle by **0.03 to 1.41 degrees, mean 0.59**.

This is smaller than reported for frontier judges. The likely reason is that judge error here dominates the noise that averaging removes: when a judge sits 40 degrees from the floor, a degree of reference flattery is not what decides the conclusion. **The correction matters most precisely where judges are good enough for it to matter.**

Table 3: Cohort geometry

Rubric	judge-judge	human-human	Difference	Effective rank (of 5)
consistency	78.6	24.5	54	4.85
fluency	80.6	43.8	36.9	4.74
coherence	65.7	49	16.7	4.59
relevance	75.3	61.8	13.6	4.78

These judges disagree with one another more than people do, on every rubric.

Effective rank near 5 out of 5 means near-independence, not a shared axis.

(Figure 4)

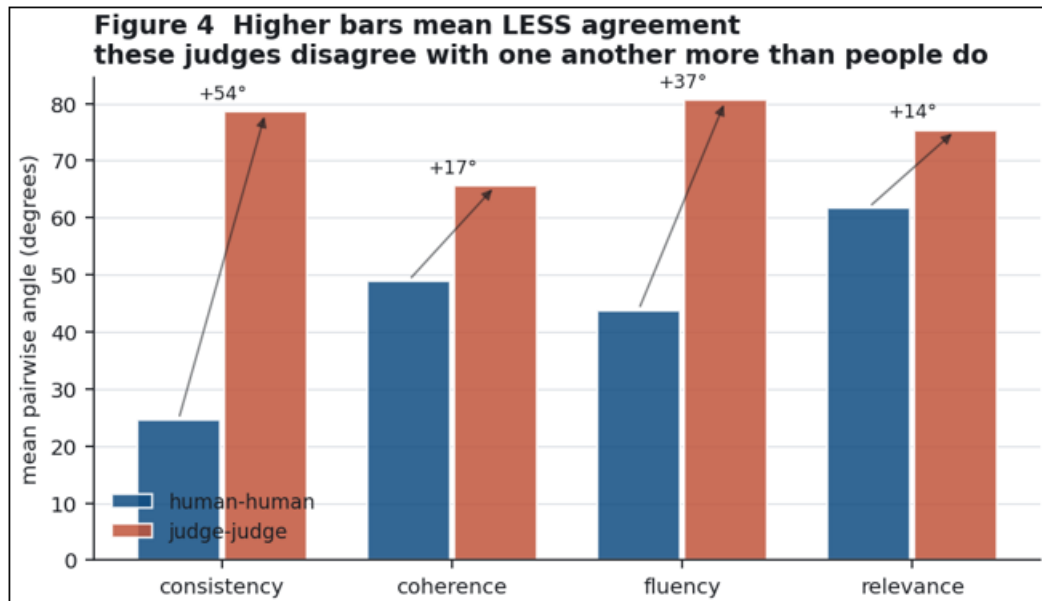


Figure 4: Mean pairwise angle within the judge cohort against mean pairwise angle among human raters. Higher bars indicate less agreement.

5.3 Ensembling

Rubric	Best single gap	Ensemble gap	Outcome
relevance	+14.5	+10.7	beats best single
fluency	+24.1	+23.5	beats best single
coherence	+22.1	+20.3	beats best single
consistency	+39.6	+40.7	does not beat best single

Ensembling helps on the three subjective rubrics and fails on the verifiable one. This follows directly from Table 3: independent errors average out, so a cohort spanning five directions gains from combination. On consistency no member has the capability, and averaging five judges who cannot verify a claim does not produce one that can. (Figure 5)

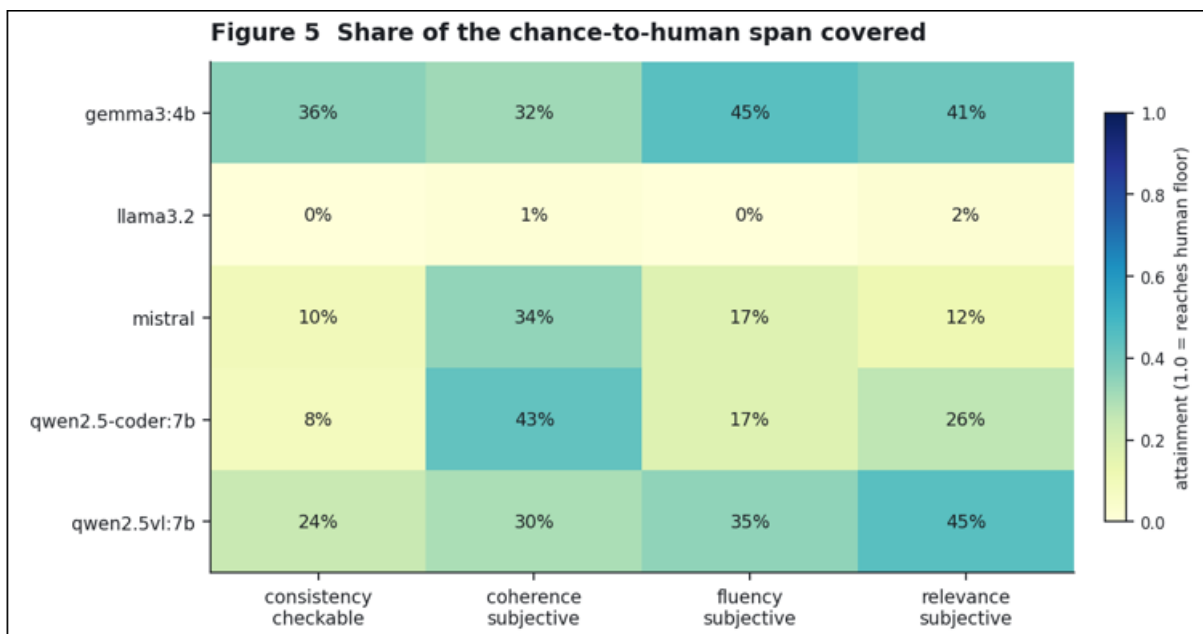


Figure 5: Attainment: the share of the span from chance to the human floor that each judge covers.

6. Discussion

The constraint bites hardest where it matters most. A deployment that cannot call a hosted judge is usually one handling sensitive data, and the evaluation such a deployment most needs is groundedness: is this claim supported by the source. That is the consistency rubric, and

it is where locally-runable judges are worst- mean gap +52.6 degrees against a human floor of 21.4.

Frontier results do not transfer. Two published patterns invert here. Verifiability closes the gap for frontier judges and widens it for these. Frontier judges share an axis, so ensembling amplifies a common blind spot; these judges are

near-independent, so ensembling helps- except where none has the capability at all.

What follows for practice. On subjective rubrics a local cohort recovers a useful fraction of the human span, and ensembling is worth its cost. On verifiable rubrics the honest recommendation is not a better local judge but a **deterministic check**: whether a claim traces to the source is a question arithmetic and retrieval can answer without a model, reproducibly and at no cost. The judge should be reserved for what deterministic checks cannot see.

Refusal is a deployment property. One model declined the task. Under a hardware constraint the available cohort is small, and a member that will not grade removes a fifth of it.

7. Limitations

- One benchmark, one language, one task (news summarisation).
- Five judges after exclusion. Cohort geometry is indicative, not a survey.
- Three raters per item is the minimum for reference matching; more folds would tighten the floor estimate.
- No frontier judge is measured here. Comparison with published frontier figures is indicative only, since those were obtained under a different protocol and on different data.
- Prompt sensitivity is not ablated; a single prompt is used throughout.
- Model size is confounded with family and training recipe.

Declaration of interests

This work was conducted independently, on personally-owned hardware, using publicly available data. It received no funding or resources from the author's employer.

Reproducibility

Code, prompts, item identifiers and raw judge scores are released. The benchmark is public. No data leaves the machine running the evaluation, which is the condition the study is about.

src/geometry.py method: angles, reference matching, chance, effective rank

src/run_judges.py local scoring

src/analyse.py reference-matched analysis, exclusion

src/figures.py figures, drawn only from results.json

Figures

- 1) The reference-matching protocol, naive versus matched.
- 2) Reference-matched angle from the human axis, per rubric.
- 3) Cost of substituting a judge for a person.
- 4) Judge-judge against human-human agreement.
- 5) Attainment heatmap, judge by rubric.

References

- [1] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). *Judging LLM-as-a-Judge with MT-Bench and

Chatbot Arena.* Advances in Neural Information Processing Systems 36.

- [2] Mukherjee, S., Hamna, H., Bali, K., and Sitaram, S. (2026). *The Geometry of LLM-as-Judge: Why Inter-LLM Consensus Is Not Human Alignment.* arXiv:2606.03043.
- [3] Norman, J. D., Rivera, M. U., and Hughes, D. A. (2026). *Reliability without Validity: A Systematic, Large-Scale Evaluation of LLM-as-a-Judge Models Across Agreement, Consistency, and Bias.* arXiv:2606.19544.
- [4] Saha, A., Wagde, A., and Kveton, B. (2026). *LLM-as-Judge on a Budget.* arXiv:2602.15481.
- [5] Flynt, J. (2026). *GroundEval: A Deterministic Replacement for LLM-as-Judge in Stateful Agent Evaluation.* arXiv:2606.22737.
- [6] Han, J., Li, Q., Peng, Z., Wu, H., Liu, H., Shan, D., Wang, G., Deng, H., and Liu, N. (2026). *SkillEval: Decomposing Agent Skill Quality into Interpretable Signals.* arXiv:2608.06891.
- [7] Qin, S., Yin, L., and Wang, X. (2026). *Grounded Checklist Partial Credit for Agent Skill Trajectories.* arXiv:2608.27487.
- [8] Zhang, W., Sun, Y., Huang, P., Pu, J., Lin, H., and Song, D. (2025). *MIRAGE-Bench: LLM Agent is Hallucinating and Where to Find Them.* arXiv:2507.21017.
- [9] RAGAS: evaluation framework for retrieval-augmented generation. <https://github.com/explodinggradients/ragas>
- [10] DeepEval: evaluation framework for large language model applications. <https://github.com/confident-ai/deepeval>
- [11] Inspect: a framework for large language model evaluations. UK AI Security Institute. <https://inspect.aisi.org.uk/>
- [12] Giskard: open-source evaluation and testing for AI and LLM systems. <https://github.com/Giskard-AI/giskard-oss>
- [13] Fabbri, A. R., Kryscinski, W., McCann, B., Xiong, C., Socher, R., and Radev, D. (2021). *SummEval: Re-evaluating Summarization Evaluation.* Transactions of the Association for Computational Linguistics, 9, 391-409.
- [14] Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).
- [15] Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation), Article 28.