

EMNLP 2026

The 2026 Conference on Empirical Methods in Natural Language Processing | Budapest, Hungary, October 24–29, 2026

The Geometry of LLM-as-Judge

Why Inter-LLM Consensus Is Not Human Alignment

Sourabrata Mukherjee

Hamna Hamna

Kalika Bali

Sunayana Sitaram



Microsoft Research, India

42

LLM judges

2

Indic benchmarks

8

languages

170,890

judge calls

87,618

human labels



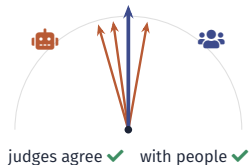
arXiv 2606.03043

When judges agree, what are they agreeing on?

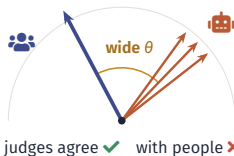
Judges agree with each other, far less with people

the lopsided pattern seen across multilingual meta-evaluations

Reading 1: shared insight



Reading 2: shared blind spot



Same numbers

Both readings give identical judge-judge agreement.

Agreement statistics alone cannot tell them apart.

Calderon et al. 2025; pattern in Hada et al. 2024, Watts et al. 2024

How we read this pattern decides how LLM evaluation gets done.

THE PROBLEM • WHY IT MATTERS

Misreading it has real costs



Panels trust it

Judge panels and ensembles treat agreement as reliability.



Humans get cut

Strong consensus is used to cut human annotation.



Errors compound

A shared blind spot is amplified by averaging, not cancelled.

WHAT JUDGES MISS • A REAL HINDI HEALTH QUESTION

घर में बिल्ली और कुत्ते जैसे पालतू जानवर हैं। क्या उनके संपर्क से एलर्जी हो सकती है?

"We keep cats and dogs at home. Can contact cause allergies? What everyday precautions should we take?"



Judges reward: HEPA filters, special vacuums, new furniture

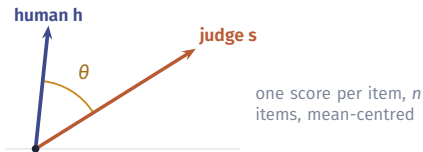


Raters reward: cheap household steps first

We need a test that sees the direction of agreement, not just its amount.

Measure the shape of disagreement

Scores as vectors



$$\theta = \arccos |r(\mathbf{s}, \mathbf{h})| \quad 1 - \cos^2 \theta = 1 - r^2$$

the second term is exactly the share of human variance the judge leaves unexplained

Four diagnostics, every judge

- 1 **Spread** σ_j / σ_H
does the judge use the human score range?
- 2 **Effective rank** r_{95}
how many independent directions its scores use
- 3 **Angle** θ
against a matched human floor and a permutation null
- 4 **Agreement triple** r_{LL}, r_{LH}, r_{HH}
always side by side, never one average

All four read observed scores only, never model internals.

Prediction: a shared blind spot means high r_{LL} , a wide θ , and no rescue from more judges.

Two benchmarks, 42 judges, one fair target



Real-world cultural QA

1,182 items · 8 languages · 3 domains
3 independent raters per item



Community-Health

405-item core · 4 subjective rubrics
Hindi, Kannada, Malayalam



42 LLM judges

15 base: closed APIs, open 12B–671B
27 trained variants of 3 open families

Hold one rater out

both face the same target; repeated for all 3 raters



TASK QUALITY, OBJECTIVE

averaged target

judges 58.5° · raters 61.7°

→ judges look closer to people than people are

matched target

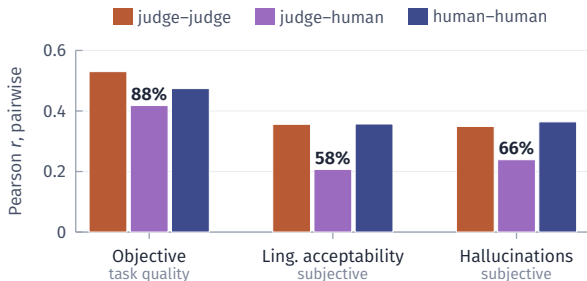
judges 60.7° · floor 56.4°

→ judges sit 4.3° outside the human floor

Every judge–human comparison in this talk uses the matched target.

Consensus without agreement

Three agreement legs, all pairwise



58–66%

of human agreement, on
subjective rubrics

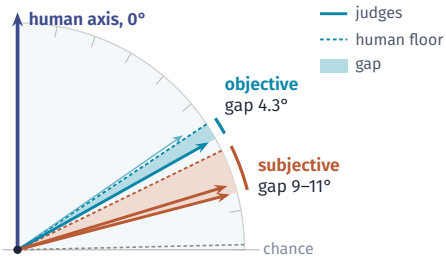
yet judge-judge equals
human-human: **0.356 vs 0.357**,
0.349 vs 0.364

Community-Health preferences:
 κ **0.445** among judges, **0.128** with
people

Judges reach human-level consensus on an axis people do not weight.

Checkable answers close most of the gap

Angles from the human axis



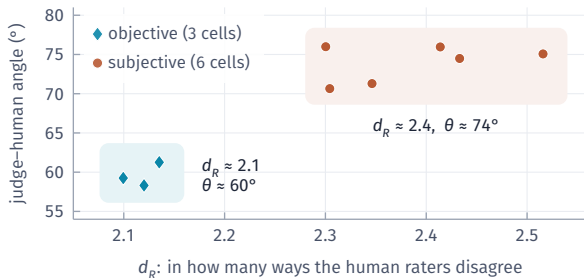
Matched angles, 5 judges, 3 raters

	judges	floor	gap	best
Objective task quality	60.7°	56.4°	4.3°	55.3° inside
Ling. accept. subjective	75.4°	64.4°	11.0°	67.3° outside
Hallucinations subjective	73.1°	63.8°	9.3°	70.8° outside
chance (permutation null) 88.6-88.7°				

Same judges, same items, same raters: only the rubric changes, and the gap with it.

Human raters alone predict the regime

Nine rubric × domain cells



Before any LLM runs

$$d_R = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2}$$

participation ratio of the centred
items × raters matrix

✓ $\rho = 0.68$, $p = 0.042$ across cells

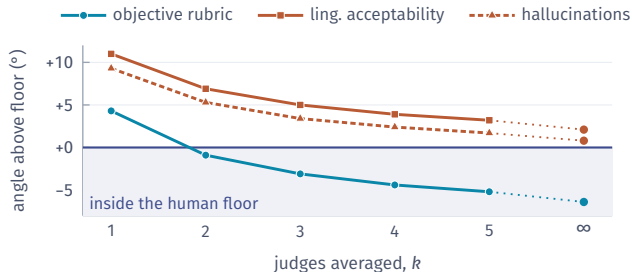
⌊ flat within a regime: $\rho = -0.03$

✗ rank vs alignment: $r = 0.12$

Rank is necessary; orientation is decisive. The humans tell you which regime you are in.

More judges converge on their shared axis

Averaging k judges, matched floor



One shared direction

86.6–89.0%

of judge variance lies in the top 3 principal components

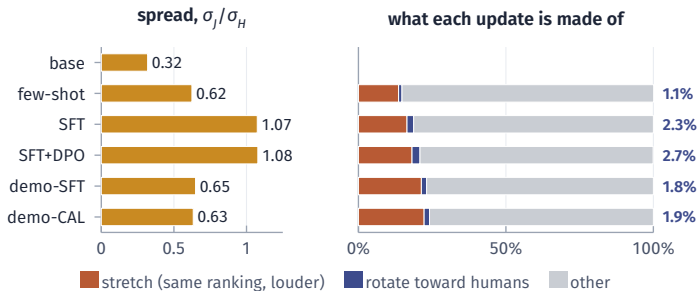
that axis sits **64–66°** from humans (subjective), **47.7°** (objective)

even $k \rightarrow \infty$ stays outside: **66.5°** vs 64.4°, **64.6°** vs 63.8°

Averaging reduces variance toward the judges' shared axis; it never finds a new one.

Training widens scores, barely turns them

Three open families, six stages



$\leq 3.5\%$

of any training update
turns a judge toward
people

spread 0.32 $\rightarrow$ 1.08, yet the
angle stays at 87–88°

best: human-rated demos
 $r = 0.184$, gpt-5.5 0.123

Training makes a judge louder, not more human.

The geometry survives every control



Not chance

15 of 15

DLQ judge × rubric angles beat all 500 shuffles, $p < 0.002$



Not a lucky judge

6 of 40

CH judges survive Benjamini-Hochberg, $q < .05$



Not the scale

$\leq 0.31^\circ$

angle shift after remapping to the human marginals



Not favouritism

$|\Delta| \leq 0.035$ SD

no family rates its own generator differently, $p \geq 0.58$



Not the item sample

$\leq 2.5^\circ$

angle agreement between two halves of the items



Not fixable by pooling

55 / 80 / 100%

human variance a ridge fit on all judges cannot recover

And we say what does not hold: Community-Health judges cannot be ranked by angle.

CONCLUSION

Consensus is not alignment

WHAT WE SHOWED

- 1 Judges reach human consensus, but only **58–66%** of human agreement
- 2 Checkable rubrics nearly close the gap; the best judge is inside the floor
- 3 A human-only dimensionality flags failing rubrics in advance
- 4 Ensembles and training do not turn judges toward people

BEFORE YOU TRUST JUDGE CONSENSUS

- ✓ **Measure** spread, rank and angle
- ✓ Build a **matched** human floor
- ✓ Test: beat chance, **reach** the floor
- ✓ Fail? Use **human-rated** examples

“ Consensus shows that judges share an axis;
geometry shows whether it is the human one. ”

Thank you

 Microsoft Research, India

Questions and discussion welcome

 connect.souro@gmail.com

 aka.ms/laj_geometry

 Code, prompts and data that reproduce every number



 Paper



 Website



 LinkedIn