LLM-as-a-Judge

Diagnosing the evaluator · opening it up · recovering a better answer

Sourabrata Mukherjee

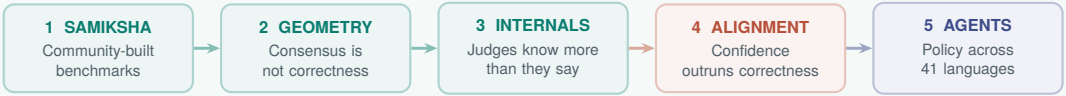
Postdoctoral Researcher · Microsoft Research India

Ph.D. · Charles University, Prague, Czech Republic

A brief of my work

MY WORK TIMELINE

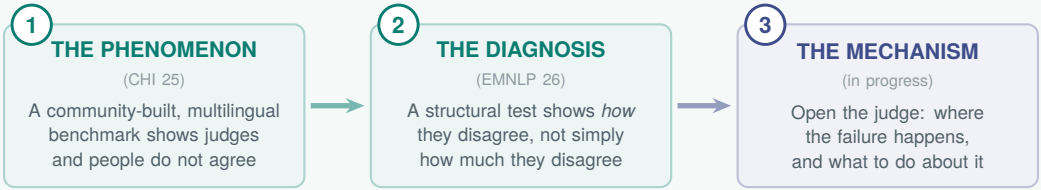
AT MICROSOFT RESEARCH



Published at EMNLP, COLM, CHI, INLG, EACL NAACL SRW …

The scope of this talk

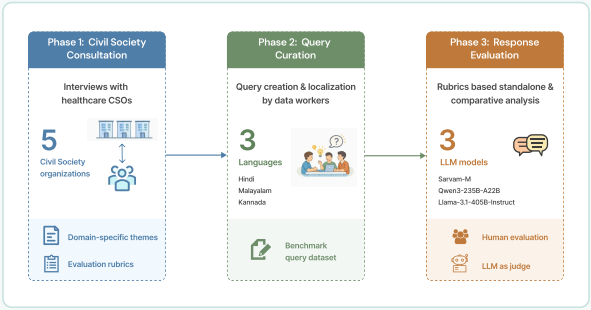
Of those five projects, today is about one question — can we trust an LLM judge?



The evidence base. Curated with **civil-society organisations** across India — multilingual, **4 domains**, **1M+** LLM judgements against **100k+** human judgements.

Samiksha — build the human labels first

CHI 2026 · Honourable Mention · community-centred benchmarking with civil-society organisations across India



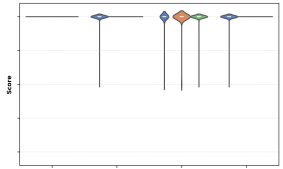
11 languages × 6 domains · **sub-domains co-created with NGOs**

two annotator tiers · **usability outranks correctness**

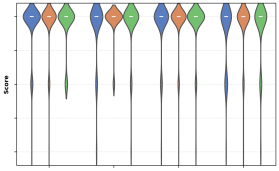
Start with what the community annotators told us

Same responses, same rubric — an LLM judge next to the people the system is built for

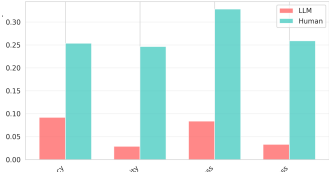
LLM JUDGE (GPT-4O)



COMMUNITY ANNOTATORS



COEFFICIENT OF VARIATION



The judge **pins nearly every criterion at the ceiling**; the people use the whole scale. Coefficient of variation — **0.059** for the judge, **0.272** for the annotators, a **4.6**× difference.

Start with one evaluation

A single pairwise comparison, of the kind every eval pipeline runs thousands of times a day

“Which response gives better advice?”

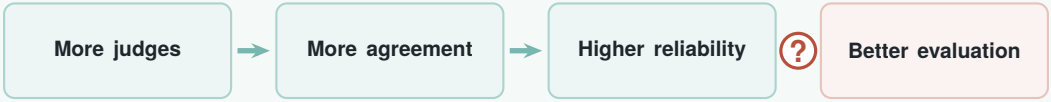
A short, nuanced, addresses the user’s actual situation

B long, fluent, confident — misses an important contextual issue

Human A **Judges 1–5** B B B B B

If five judges agree, **what exactly have we learned?**

The tempting assumption



Is inter-LLM **agreement** evidence of human **alignment**?

> Phase 1 is an attempt to answer that question with something other than another agreement statistic.

PHASE 1 — DIAGNOSING THE JUDGE

When many judges agree, are they detecting quality — or sharing a bias?

Agreement, or shared bias?

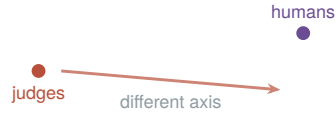
Both hypotheses predict exactly the same inter-judge agreement

H1 — SHARED SIGNAL

The judges agree because they are all detecting the same **human-relevant quality**.

H2 — SHARED BIAS

The judges agree because they share a common but **human-misaligned** evaluation axis.



Agreement alone cannot distinguish these two hypotheses — both produce the same numbers.

What would actually settle it?

Stop asking *how much* judges disagree with people; ask *in what way*

FOUR QUESTIONS, ASKED IN ORDER

- **Q1** How **much** variation do they use? → score spread σ_J/σ_H
- **Q2** How **many** dimensions? → effective rank
- **Q3** In which **direction**? → principal angle θ
- **Q4** Does agreement **reach** humans? → r_{LL}, r_{LH}, r_{HH}

MATCHED REFERENCES

A judge scored against the *average* of several raters has an easier target than a rater scored against one colleague.

- ✗ **Unmatched** — judge vs. a 3-rater mean, human vs. one colleague
- ✓ **Matched** — hold one rater out; score both against the same two-rater mean

Finding 1 — judges use a much narrower range

Score spread, judge relative to human

0.454

σ_J / σ_H — the median judge uses under half the score spread that people use

Obvious objection: maybe LLMs are simply less expressive. Possibly — but matching the range would still not establish that the range is used for the same thing.

Finding 2 — consensus lives in a low-dimensional space

Effective rank of the score matrix

HUMANS

Use the **full basis** the rubric offers.

4 / 3

effective rank, the two benchmarks

JUDGES

Collapse onto **two or three** dominant directions.

11 of 14

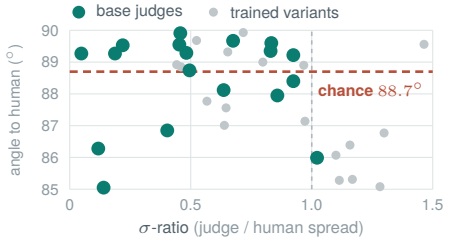
base judges at effective rank ≤ 3

Where we could have fooled ourselves. Rank looks like the answer, but across 40 judges it does *not* predict alignment — $r = 0.12$. Low rank is necessary, not sufficient. So perhaps those two or three directions are simply the **important** human dimensions?

Finding 3 — the judges are looking in a different direction

Every judge, placed by spread and by angle to the human axis

EVERY JUDGE, SPREAD VS. ORIENTATION



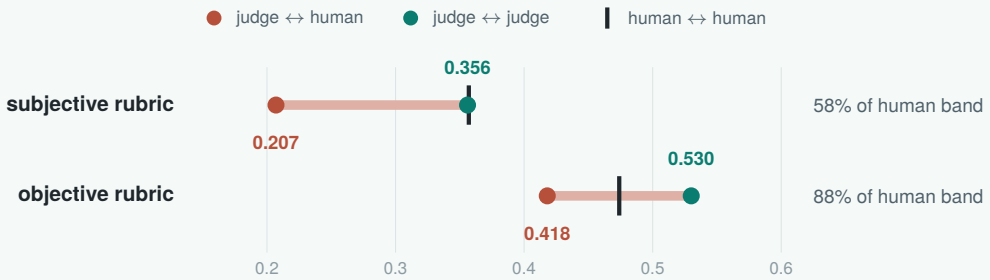
JUDGE VS. A MATCHED HUMAN FLOOR

	JUDGE	FLOOR	GAP
subjective acceptability	75.4	64.4	11.0
objective factuality	60.7	56.4	4.3

Every judge sits 85–90° from the human axis; chance is 88.7°.

Consensus is real. Alignment is not.

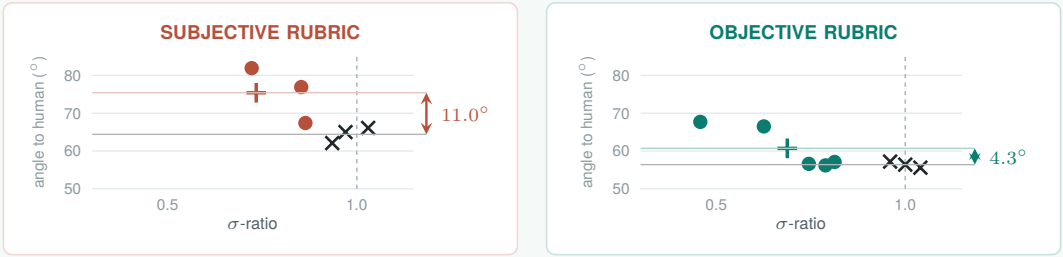
Judge–judge, judge–human and human–human agreement, on the same items and the same matched reference



On the subjective rubric judges agree with *each other* as much as people do (0.356 vs. 0.357) — but reach only **0.207** with people. The shared signal is **real and coherent**, just not human-aligned.

When does the problem disappear?

Same judges, same pipeline — two rubrics that differ only in whether the answer is checkable

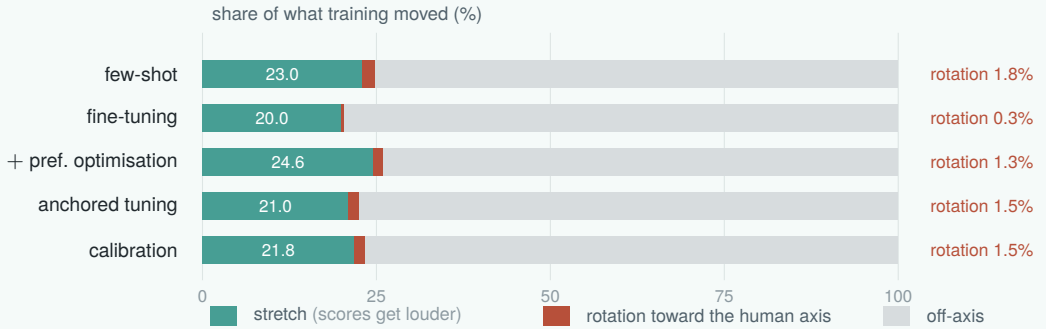


● judges · × held-out humans · + judge cohort mean · identical axes; lower is closer to the human axis

The judge cohort sits $11.0°$ outside the human floor on the subjective rubric, but only $4.3°$ on the objective one. The failure depends on **the structure of the rubric**.

Can we train the judges into alignment?

Five training stages on one judge family — few-shot, fine-tuning, + preference optimisation, anchored tuning, calibration



Every stage moves the scores — but **under 2% of that movement is toward the human axis**.

What Phase 1 taught us

1 CONSENSUS ≠ ALIGNMENT

Judges can agree with one another as much as people do while reaching only **58–66%** of that agreement with people.

2 THE FAILURE HAS STRUCTURE

Compressed spread, low rank, near-orthogonal orientation — and it **closes on verifiable rubrics**.

3 STANDARD FIXES DO NOT ROTATE IT

Training restores spread far more readily than orientation; ensembling converges on the judges' *shared* axis.

We now know a great deal about the **output** behaviour of these judges.

We know nothing about **where inside the model** the problem originates.

THE BRIDGE

Where, exactly, does the failure happen?

Two very different explanations of the same bad verdict

EXPLANATION A

The **internal computation itself** is human-misaligned. Then Phase 1's geometry is a property of the model, and the only fix is a better model.

EXPLANATION B

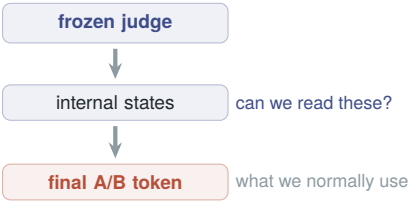
The internal computation carries more than the verdict, and **information is lost in the last step**. Then we are reading the wrong thing.

These are **not** the same claim — and agreement statistics can never tell them apart.

What if we open the judge?

Instead of asking the judge for another answer, inspect what it computed

THE OBJECT OF STUDY

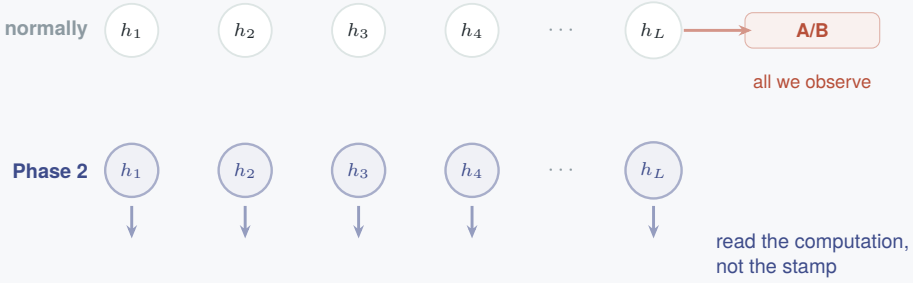


FIVE QUESTIONS

- **When** does the judge form its decision?
- Is a **better preference** already present internally?
- What **causally** drives the biased verdict?
- Can a **tiny read-out** recover it?
- Does that improve anything **downstream**?

64 evaluators, **0.36B–405B** · **14** datasets · **41** judges causally intervened.

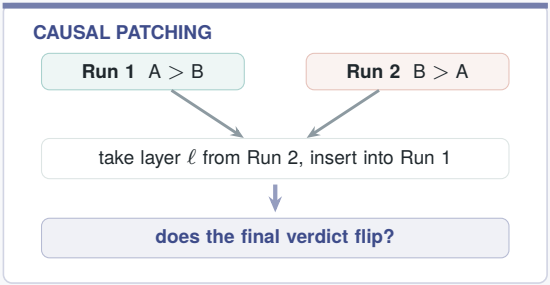
What “opening” means, concretely



Three tools, in increasing order of what they license: **lenses** (what is decodable) · **probes** (what is recoverable) · **causal patching** (what actually mattered).

Is the verdict formed only at the end?

The experiment: swap one layer between two runs and see whether the answer follows



THE DISTINCTION THAT MATTERS

Decodability ≠ causality.

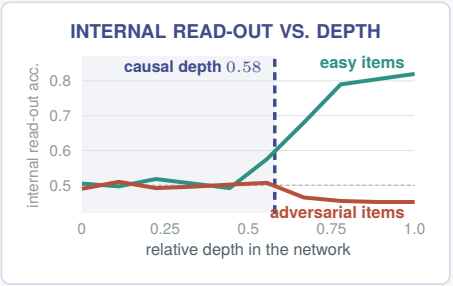
A probe saying *"I can predict the answer from this layer"* does **not** mean *"this layer caused the answer."*

Lenses and probes say what is *readable*; patching, ablation and steering say what is *load-bearing*.

Steering these components flips the emitted verdict about **67%** of the time — they causally influence the output, so this is behaviour, not a read-out artefact.

The verdict becomes causally influential mid-stack

Causal decision depth, measured by forward-pass residual patching



THREE DIFFERENT DEPTHS

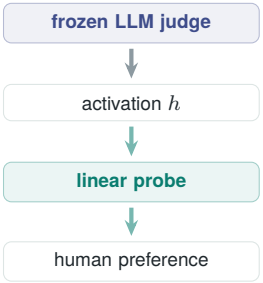
0.58	causal influence becomes **detectable**
0.68–0.81	the verdict becomes **linearly decodable**
0.9–1.0	the final rubric score **commits**

Causal influence is detectable well before the verdict is linearly decodable. On adversarial items the naive read-out is fooled at every depth.

Can a tiny probe read what the verdict loses?

No gradient reaches the judge; the judge is frozen throughout

THE READ-OUT



WHAT IS ACTUALLY LEARNED

23,600,000,000 frozen parameters

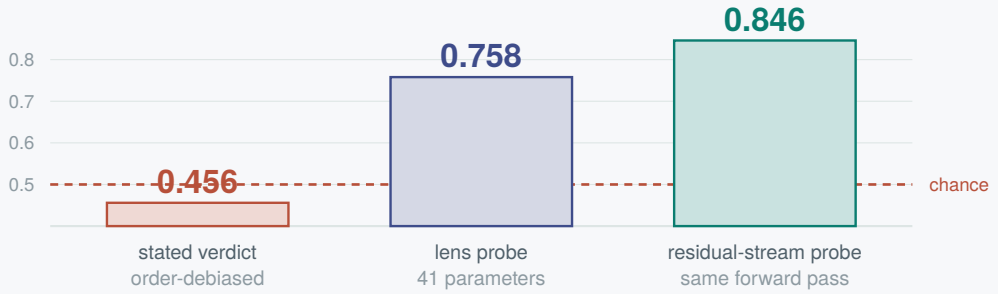
41 learned parameters

$\hat{y} = \sigma(w^\top h + b)$. Order-averaging (A, B) and (B, A) before the probe cancels position bias without a gradient step. Every number is **out-of-fold**, with a shuffled-label control.

So: how much is left in there, on the items where the verdict is **worst**?

What happens when the final verdict is below chance?

LLMBar-adversarial: A is short and correct, B is long, fluent and wrong



The judge's output is **worse than a coin**. Human preference is already **recoverable** from the same forward pass.

Is this just a fancy length detector?

The objection this result has to survive

FOUR CONTROLS

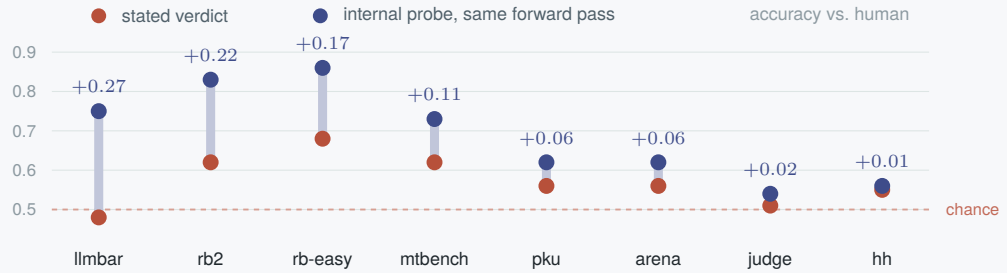
A **length-only** probe · **residualising** length out of the representation · a **length-balanced** subset · **shuffled labels**.

Removing the length *subspace* alone does not help $(-0.41 \rightarrow -0.11)$ — the signal is recoverable only by reading the representation.

So 0.846 is **not** pure reasoning. Part of the gain is surface information — and a substantial **human-label-predictive signal remains beyond measured length**.

Does the effect generalise?

Eight pairwise-preference benchmarks, fifty judges, same protocol throughout

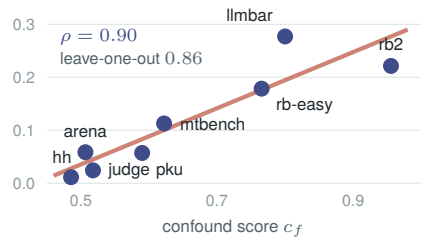


> The probe wins on **all eight** benchmarks, mean $+0.12$ — and the margin is **largest exactly where the verdict collapses**.

So the benefit is **not uniform**. When is it worth opening the judge?

Can we predict when internals will help?

A surface-confound score, computed before fitting any internal read-out



THE CONFOUND SCORE c_f

How well can **answer length alone** predict the human preference on this benchmark? No activations, no probe fitted.

A strong surface cue is what a judge latches onto at the output — so it is exactly where the internal state should hold more than the verdict. High $c_f \rightarrow$ large gain; low $c_f \rightarrow$ near zero.

A probing trick becomes a **decision rule**: you can tell whether opening the judge is worth it **before fitting any internal read-out**.

The method has a boundary too

The honest negative result

WHERE INTERNALS HELP

- **Pairwise** comparison
- Strong **surface confounding** — length, order, style
- A visible gap between what is computed and what is emitted

WHERE THEY DO NOT

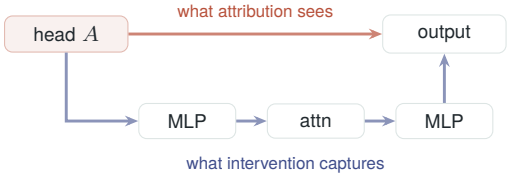
- **Absolute-rubric** scoring, held out entirely
- No pairwise surface cue to exploit
- Correlation with the confound score: -0.006

This is **not** a universal claim that intermediate activations beat the final output. It is a claim about a specific, diagnosable regime. We can now say *when* the verdict is lossy — **not yet why**.

What actually causes the biased verdict?

41 judges, causally intervened — not attributed

ATTRIBUTION RANKING ≠ CAUSAL RANKING



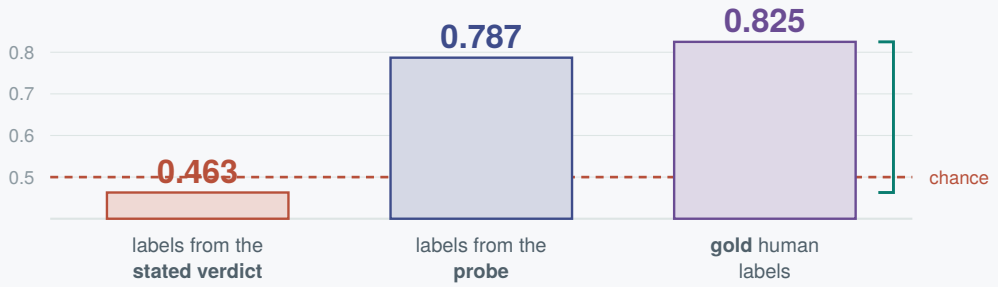
THE NUMBERS

attribution vs. causal rank	$\rho = 0.31$
judges with $\rho < 0.5$	**95%**
biased effect that is **indirect**	**82%**
position-bias ablation closes	**33–54%**
random-head control	∼3%

Attribution alone can substantially mislocalise evaluator bias.

Does opening the judge improve anything we actually use?

Same judge, same prompts — only the preference labels change. DPO on a 1.5B policy, LLMBar

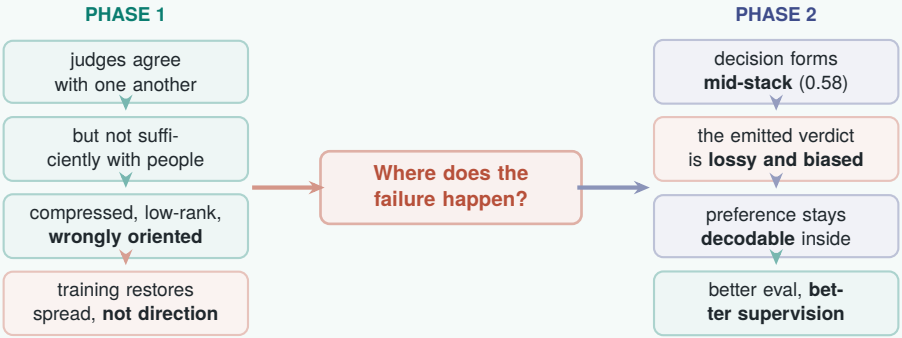


The probe labels recover **about 90% of the gap** to gold. The recovered signal is not only an interpretability artefact — it is **better supervision**.

SYNTHESIS

What the two phases say together

From consensus, to geometry, to mechanism, to intervention



The story changed from *"LLM judges are unreliable"* to **"the judge can be misaligned at the output while still holding a richer preference inside."**

What we now believe

Four claims, not ten

1 CONSENSUS IS NOT ALIGNMENT

Many judges agreeing can mean **shared bias** rather than shared human signal — and only a matched, geometric comparison tells the two apart.

2 THE VERDICT IS NOT THE COMPUTATION

An A/B token or a scalar is a **lossy projection** of a much richer internal state.

3 INTERPRETABILITY NEEDS CAUSAL TESTS

Decoding information from a layer does **not** establish that the layer caused the decision ($\rho = 0.31$).

4 OPEN THE JUDGE SELECTIVELY

The benefit concentrates where **surface confounds** open a gap between what the model computes and what it emits — and that is predictable before fitting the internal read-out.

What we still do not know

DECODABLE $\neq$ USED

We can recover a human preference from the activations. That does not prove the model *used* that representation when it decided.

WHAT IS THE BETTER REPRESENTATION?

The probe says *recoverable*. It does not decompose it into correctness, relevance, reasoning, context, style.

WHY THE WRONG PROJECTION?

We can localise the circuits and the biases. The full path from internal state to emitted token remains unexplained.

PORTABILITY

Transfer to an **unseen judge** looks promising in the one setting we tested; cross-*benchmark* transfer stays weak.

THE REAL QUESTION

Today: the model computes, emits a poor verdict, and *we* read something better. **Can the judge be made to use the better representation itself?**

From “Do the judges agree?” → “What did the judge compute?”

The remaining challenge: turn a recoverable internal preference into a causally grounded, robust, general-purpose evaluation mechanism.

Thank you

Sourabrata Mukherjee · Microsoft Research India

connect.souro@gmail.com

more about my work

BACKUP

Material for questions

Backup — the construction that decides Phase 1

Reference matching

UNMATCHED

The judge is scored against the *mean of three* raters; the held-out human against *one* colleague.

0.519

r_{LH}, objective rubric — “judges beat humans”

MATCHED

Hold out one rater; score **both** the judge and that rater against the remaining two.

0.418

r_{LH}, objective rubric — the number we report

The unmatched comparison is worth several degrees of angle. Every headline number in Phase 1 uses the matched one.

Backup — the four diagnostics, formally

DEFINITIONS

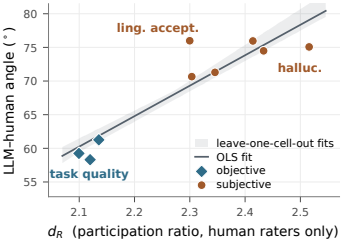
- **Spread** — σ_J/σ_H, per rubric, on the shared item set
- **Effective rank** — participation ratio of the score matrix spectrum, r_{95}
- **Principal angle** — angle between the judge score vector and the matched human mean
- **Agreement triple** — r_{LL}, r_{LH}, r_{HH} on one reference

CONTROLS THE COLLAPSE SURVIVES

- permutation nulls ($\sim 88.7°$)
- FDR correction across rubrics
- ordinal as well as Pearson agreement
- human-reference and self-preference checks
- split-half reliability

All quantities are **observed score covariance**. Nothing in Phase 1 touches model internals — that is Phase 2's job.

Backup — a boundary predictor computed from humans alone



d_R

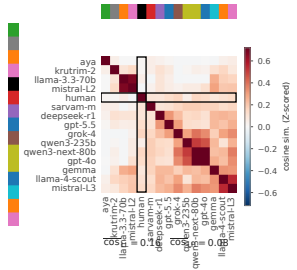
A dimensionality read from the **human raters only**, before any LLM is run, sorts rubrics into the two regimes.

Rank alone does not: $r = 0.12$ across 40 judges.

A falsifiable prediction: tell me the human rating structure and I will tell you where judge–human disagreement is likely to emerge.

Backup — judges share an axis; humans do not

Subspace similarity between score vectors



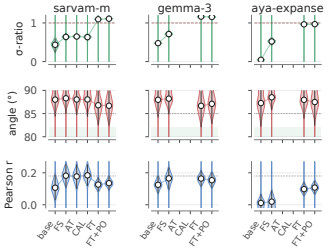
WHY ENSEMBLING DOES NOT RESCUE IT

Judge-to-judge cosines are high and block-structured; the human row stays uniformly pale.

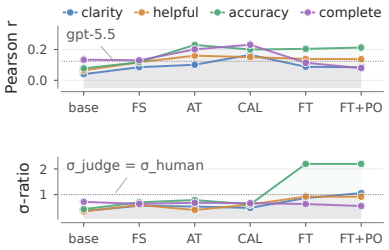
Averaging more judges therefore converges on **the judges' shared axis**, not the human one.

Backup — the full training picture

Bootstrap over 1000 resamples, and per-rubric stage effects



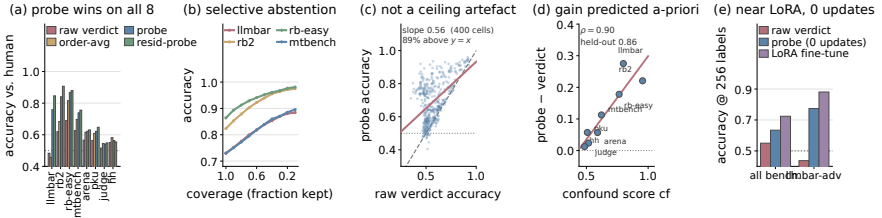
Spread rises reliably; angle distributions overlap.



Only demonstration-anchored stages raise every rubric.

Backup — the full evaluator ladder

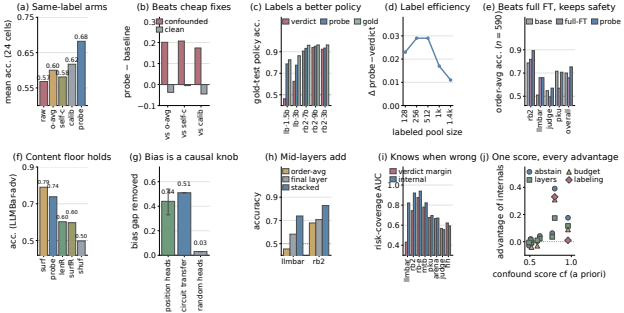
All five panels of the Phase 2 evaluator figure



(a) probe wins on all 8 · **(b)** abstention · **(c)** not a ceiling artefact · **(d)** gain predicted a priori · **(e)** matches LoRA, zero updates.

Backup — consequences

All ten panels



Backup — probe vs. the alternatives

Same labels, same budget

AGAINST CHEAP OUTPUT FIXES (256 LABELS, 24 CELLS)

raw verdict	0.568
self-consistency ($k = 5$)	$+0.21$ conf. cells
calibrated prompt	$+0.17$ conf. cells
probe	**0.682**

AGAINST FULL FINE-TUNING (SAME 3,349 LABELS)

full fine-tune	0.659
frozen probe	**0.753**
safety after fine-tune	$0.716 \rightarrow 0.569$
safety with the probe	**0.707**

The frozen read-out is competitive with methods that change the model — and it does not regress the behaviour you were not trying to change.

Backup — numbers I did not put on a slide

PHASE 1

- σ_J/σ_H: 0.454 CH median; 0.687 objective rubric
- Agreement, hallucination: r_{LL} 0.349, r_{HH} 0.364, r_{LH} 0.239
- Agreement, objective rubric: r_{LL} 0.530, r_{HH} 0.474, r_{LH} 0.418
- Rotation share by stage: 1.1–2.8%, never above 3.5%

PHASE 2

- Cross-judge audit of an unseen judge: $+0.32$
- Confidence calibration: ECE $0.405 \rightarrow 0.262$
- Steering flips the verdict $\sim$**67%** of the time
- Internal confidence knows when the judge is wrong: AUC **0.822** vs. verdict margin **0.431**
- DPO on a 3B policy: $0.625 \rightarrow$ **0.775** vs. gold 0.863

Backup — what the two papers are

PHASE 1

The Geometry of LLM-as-Judge: Why Inter-LLM Consensus Is Not Human Alignment

Mukherjee, Hamna, Bali, Sitaram.
EMNLP 2026 · arXiv:2606.03043

42 judges · 170,890 judge calls · 8 languages · 2 community-built Indic benchmarks.

PHASE 2

Opening the Judge: what evaluators compute but do not report

EACL 2026 · under review

64 evaluators, 0.36B–405B · 14 datasets · 41 judges causally intervened · code, prompts and per-run results released.

The benchmark underneath Phase 1 is **Samiksha** (CHI 2026, honourable mention) — community-centred evaluation built with civil-society organisations across India.