

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

Sourabrata Mukherjee Hamna Hamna Kalika Bali Sunayana Sitaram

Microsoft Research

{t-somukherje, Sunayana.Sitaram}@microsoft.com

Abstract

LLMs-as-judges are now standard, yet judges agree strongly with one another while agreeing only weakly with humans. We test whether this is shared signal or shared bias by measuring four geometric quantities on the standard LLM-as-judge stack across four community-built Indic datasets, eight Indic languages, and 41 LLM judges: score spread, effective rank, principal angle to the human subspace, and stacked correlations among judges and humans, all with bootstrap intervals. On subjective rubrics judges use less than half the human score range ($\sigma_J/\sigma_H \approx 0.3\text{--}0.5$), their evaluation axis is nearly orthogonal to the human one and noticeably further from humans than humans are from each other ($87\text{--}89^\circ$ vs. $78\text{--}81^\circ$), and inter-LLM agreement runs well ahead of LLM human agreement ($r_{LL} \approx 0.35$ vs. $r_{LH} \approx 0.27\text{--}0.32$). On a rubric with a verifiable factual answer the same diagnostics fall back into the human range (axis 58.5° ; $r_{LH} = 0.519$). Fine-tuning and preference optimisation recover spread ($0.32 \rightarrow 1.08$) but barely move the axis (still $87\text{--}88^\circ$); only post-hoc calibration on a small human-anchored set lifts all four community-health rubrics together, putting a calibrated 24B Indic judge ($r = 0.184$) ahead of gpt-5.5 (0.123), yet still short of human reliability (HH $r = 0.474$ on the verifiable rubric). We argue that inter-LLM agreement should count as evidence of human alignment only when a direct geometric check on the judge’s score subspace passes; otherwise the consensus is consensus inside a collapsed subspace.

1 Introduction

LLMs are now the default raters for open-ended generation (Zheng et al., 2023; Chiang and yi Lee, 2023; Dubois et al., 2024). They track humans reasonably well when the rubric is close

to verification (factuality, constrained task completion), and much less well on the subjective, multilingual, culturally situated judgements that dominate real deployments (Wang et al., 2024; Liu et al., 2023; Hada et al., 2024; Watts et al., 2024; Calderon et al., 2025). The structural concern is not low agreement per se but a more specific asymmetry that recurs across recent meta-evaluations: LLM judges agree strongly with one another while agreeing only weakly with humans (Bowman et al., 2022; Schaeffer et al., 2023; Chu et al., 2025). Read as reliability, this supports current ensemble practice; read as shared bias, it overturns it; and the two readings are difficult to separate without a structural test.

The known failure modes (fluency and verbosity bias (Wang et al., 2024; Zheng et al., 2023), self-preference (Gilardi et al., 2023), cross-cultural degradation (Hada et al., 2024; Zhu et al., 2024; Hershcovich et al., 2022)) are consistent with one structural account: human evaluation occupies a multi-dimensional quality manifold, while LLM judges behave like low-rank projectors aligned with corpus-frequent, model-legible directions (Schütze, 1992; Arora et al., 2018; Raghu et al., 2017). The high inter-LLM correlations are then alignment of *projection matrices*: judges that share pre-training data and post-training objectives (Ouyang et al., 2022; Bai et al., 2022) covary even when both miss the human axis. Our claim is that LLM judges agree on subjective rubrics by collapsing onto a redundant subspace, while the same judges are rank-matched to humans on objective rubrics.

We evaluate this on real-world data to ensure representativeness. REAL-WORLD CULTURAL QA (Watts et al., 2024) contributes 8 Indic languages, three domains, three independent human raters, and rubrics spanning the

objective subjective spectrum; COMMUNITY-HEALTH (Hamna et al., 2026; Bhat et al., 2025) adds a community-curated Indic health benchmark with 14 base LLM judges and 27 trained, preference-optimised, or calibrated variants from three families (3B–405B). Beyond correlations and annotator agreement, we measure score-spread compression, effective rank, principal angles to the human subspace, null-space fraction, and paired item-bootstrap uncertainty across in-context demonstration, supervised fine-tuning, direct preference optimisation, and post-hoc demonstration-anchored calibration (Brown et al., 2020; Rafailov et al., 2023; Platt, 1999; Guo et al., 2017). The results are consistent. Across every subjective rubric LLMs agree more with one another than with humans, and the gap closes only on the single rubric with a verifiable factual answer; subjective-rubric judges sit far from the human subspace under heavy score compression, while the objective rubric rotates them back. Training restores the missing spread but barely moves the principal angle and degrades some rubrics; only post-hoc, demonstration-anchored calibration lifts all four community-health rubrics together, still short of human reliability. An explicit geometric check should therefore precede treating inter-LLM agreement as evidence of alignment with humans.

At the broader scale, our diagnostics show that the inter-human subspace is itself multi-dimensional and only moderately self-consistent on subjective rubrics (Sec. 4.1, 4.3; Figs. 5, 6), and a residual LLM human gap persists across every training stage (Sec. 4.8). For community-grounded, culturally situated, real-world subjective evaluation, even geometry-passing LLM ensembles do not remove the need for a human in the loop: nuanced, multi-perspective human judgement remains a critical signal.

2 A Geometric View of LLM-as-Judge: Rank-Deficient Projection

We formalise LLM-as-judge as a *rank-deficient projection operator* on a latent quality manifold (Belkin and Niyogi, 2003; Coifman and Lafon, 2006; Facco et al., 2017). The framework yields four quantitative predictions (σ -compression,

low effective rank, near-orthogonality of LLM and human subspaces, inter-LLM angular alignment) tested in Sec. 4, and a *boundary condition* (Eq. 3) that draws the objective-vs-subjective line.

Quality decomposition. Following work on disentangled semantic geometry (Schütze, 1992; Arora et al., 2018; Raghu et al., 2017), we model a response as $\mathbf{h} \in \mathbb{R}^D$ admitting an orthogonal decomposition

$$\mathbf{h} = \mathbf{h}_{\text{style}} + \mathbf{h}_{\text{fact}} + \mathbf{h}_{\text{cult}} + \boldsymbol{\epsilon}, \quad (1)$$

into mutually orthogonal subspaces $\mathcal{V}_{\text{style}}, \mathcal{V}_{\text{fact}}, \mathcal{V}_{\text{cult}}$. A human oracle scores via

$$f^*(\mathbf{h}) = \sum_{i \in \{\text{s,f,c}\}} w_i \|\mathbf{P}_i \mathbf{h}\|_2, \quad (2)$$

with *rubric-conditional* weights w_i : an objective rubric concentrates weight on w_{fact} , while subjective rubrics spread weight across w_{cult} and w_{fact} (Madaio et al., 2020; Hershovich et al., 2022).

Judge operator. An LLM judge induces an orthogonal projection $\mathbf{P}_{\mathcal{J}} = \mathbf{U}_{\mathcal{J}} \mathbf{U}_{\mathcal{J}}^{\top}$ (Golub and Van Loan, 2013; Meyer, 2000) with $\mathcal{S}_{\mathcal{J}} = \text{span}(\mathbf{u}_1, \dots, \mathbf{u}_k)$, $k \ll D$, and score functional $\text{Score}_{\mathcal{J}}(\mathbf{h}) = \|\mathbf{P}_{\mathcal{J}} \mathbf{h}\|_2^2$. If $\{\mathbf{u}_i\}$ is learned via next-token prediction on a fluency-dominated corpus (Brown et al., 2020; Kaplan et al., 2020; Radford et al., 2019; Bengio et al., 2003; Zipf, 1935), then for any rubric with $w_{\text{cult}}^R \gg w_{\text{style}}^R$, $\langle \mathbf{u}_i, \mathbf{h}_{\text{cult}} \rangle \approx 0$, so $\mathbf{h}_{\text{cult}} \in \mathcal{N}(\mathbf{P}_{\mathcal{J}})$: cultural content, being distributional outliers, projects minimally onto the fluency-aligned basis. The converse holds for rubrics with $w_{\text{fact}}^R \gg w_{\text{cult}}^R$, the regime Sec. 4.2 exploits.

Geometric signatures. The framework yields four item-level quantities measured in Sec. 4. (i) *Variance collapse*: $\sigma_{\mathcal{J}} \lesssim \sqrt{k/D} \sigma_H + |\text{bias}|$ predicts σ -compression $\sigma_{\mathcal{J}}/\sigma_H \ll 1$ on subjective rubrics. (ii) *Effective rank*: $r_{95}^{\mathcal{J}} = \min\{k : \sum_{i \leq k} \lambda_i / \sum_i \lambda_i \geq 0.95\} < r_{95}^H$ (Jolliffe and Cadima, 2016). (iii) *Subspace orthogonality*: the principal angle $\theta(\mathcal{S}_{\mathcal{J}}, \mathcal{S}_H)$ (Björck and Golub, 1973; Knyazev and Argentati, 2002) exceeds the inter-human floor $\theta(\mathcal{S}_{H_1}, \mathcal{S}_{H_2})$. (iv) *Inter-judge alignment*: if both judges project onto a shared low-rank subspace, $\theta(\mathcal{S}_A, \mathcal{S}_B) \ll \theta(\mathcal{S}_A, \mathcal{S}_H)$ (Ouyang

et al., 2022; Bai et al., 2022; Raghu et al., 2017), predicting LLM–LLM Pearson above LLM–human Pearson.

Boundary condition. Let d_R be the intrinsic dimensionality of the human-score manifold under rubric R :

$$\text{consensual blindness on } R \iff k_{\mathcal{J}} < d_R. \quad (3)$$

The four predictions hold in the high- d_R regime (subjective rubrics in both benchmarks) and are expected to fail in the low- d_R regime (REAL-WORLD CULTURAL QA factual rubric). The framework therefore predicts *which* judges fail, the *shape* of the failure, and *where* it vanishes, but says nothing about which training technique closes the gap, a dissociation Sec. 4.7 measures empirically. App. G reproduces the full rank-matching statement and the small-/large- d_R case analysis.

3 Experimental Setup

The study spans **244,000** LLM-judge scoring events over four community-built domain datasets, eight Indic languages, $\sim 3,600$ unique question-answer pairs, and **41** LLM judges (base plus fine-tuned, preference-optimised, and calibrated variants), evaluated against 4 human reference pools. All source questions and reference labels are community-authored; per-language demographics for the COMMUNITY-HEALTH pool are in App. A.

Datasets and languages. We use two community-built Indic benchmarks. COMMUNITY-HEALTH (Hamna et al., 2026; Bhat et al., 2025) is a medical-advisory benchmark with paired standalone and comparative human labels on the same items; it anchors the calibration rank-ordering analysis. The three REAL-WORLD CULTURAL QA domains (Watts et al., 2024), **health & wellness**, **finance**, and **everyday life**, supply three independent humans per item, giving a measurable human human ceiling and sweeping the rubric-objectivity axis from culturally-loaded subjective health text to a verifiable factual rubric (Joshi et al., 2020; Blasi et al., 2022; Hershovich et al., 2022). The eight unique languages (Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil, Telugu) span Indo-Aryan and Dravidian

families and include both well-represented and low-resource tails. Per-domain breakdowns are in Table 2.

Judges. 16 base LLMs cover closed-API frontier, open multilingual, and Indic-focused families ($\sim 12\text{B}$ – 400B), API-hosted or locally served; gpt-4o and gpt-4 (OpenAI, 2024; OpenAI et al., 2024) anchor both datasets. Three Indic-capable families (sarvam-24b, gemma-3-27b, aya-expanse-32b) are exercised under six training-side techniques: few-shot (Brown et al., 2020), LoRA SFT (Ouyang et al., 2022), DPO (Rafailov et al., 2023), SFT+DPO, adaptive-tune (SFT with 1–3 in-context demonstrations), and post-hoc demonstration-anchored calibration (Platt, 1999; Guo et al., 2017). All updates use LoRA/PEFT; SFT and DPO corpora are the COMMUNITY-HEALTH standalone and comparative labels respectively. Full enumeration, sizes, and hyper-parameters are in Tables 4, 2 and App. F.

Evaluation paradigms. Every (question, answer) pair is scored under two paradigms. **Standalone:** per-rubric ordinal scores on a single (question, answer) pair, with four subjective COMMUNITY-HEALTH rubrics (clarity, helpfulness, accuracy, conciseness) and three REAL-WORLD CULTURAL QA rubrics (factual-verifiable-ground-truth rubric, linguistic acceptability, hallucinations) (Chiang and yi Lee, 2023; Watts et al., 2024). **Comparative:** judges pick the better of two answers with a one-sentence rationale; COMMUNITY-HEALTH has rater-paired labels, the REAL-WORLD CULTURAL QA comparative split is excluded from analyses (single-LLM label-orientation artefact) but counted in event totals. The two paradigms separate *calibration* from *rank-ordering* failures (Zheng et al., 2023; Liu et al., 2023). Full rubric definitions and verbatim prompts are in App. C and App. E.

Measurement and inference. Per (judge, rubric, language, domain, eval type) slice we compute six item-level metrics: Pearson r across three agreement families ($\bar{r}_{LL}, \bar{r}_{LH}, \bar{r}_{HH}$); Cohen’s κ (Cohen, 1960) on comparative panels; quadratic-weighted κ on standalone ordinal scores; spread compression $\sigma_{\mathcal{J}}/\sigma_H$; effective rank r_{95} of the per-judge (item \times rubric) ma-

trix; and the principal angle $\theta(\mathcal{S}_J, \mathcal{S}_H)$ against a rubric-floor reference θ_{rubric} . For the training analysis (Sec. 4.7) we use a paired item bootstrap (Efron, 1979) with $B = 1000$ replicates resampling the 405 COMMUNITY-HEALTH item rows and reusing the same rows across (family, stage); we report 95% percentile CIs and *bootstrap support* (fraction of replicates moving in the predicted direction). With $n = 3$ trainable families the smallest one-sided Wilcoxon (Wilcoxon, 1945) p -value is 0.125, so Wilcoxon serves only as a directional cross-family summary alongside the bootstrap.

4 Empirical Analysis

We test the geometric account with four analyses. The pairwise-agreement analysis asks whether inter-LLM consensus overstates alignment with humans. The boundary analysis tests whether the same judge cohort behaves differently when the rubric is close to objective verification. The geometry analysis examines the shape of any disagreement: score spread, rank, angle, and null space. The training analysis asks which part of that geometry changes under few-shot prompting, fine-tuning, preference optimisation, and post-hoc calibration.

4.1 Inter-LLM consensus versus LLM–human agreement

On subjective rubrics, LLM judges are consistently closer to one another than to humans, but the magnitude is rubric- and benchmark-dependent rather than a single universal constant. The strongest collapse appears on COMMUNITY-HEALTH: the LLM-human/LLM-LLM ratio ranges from 0.41 on accuracy to 0.64 on completeness. REAL-WORLD CULTURAL QA’s subjective rubrics are less severe but preserve the direction: linguistic acceptability remains below the human-human ceiling, while hallucinations is a boundary case that nearly reaches it. The comparative COMMUNITY-HEALTH split sharpens the same conclusion in preference space: LLM-LLM $\kappa = 0.445$ versus LLM-human $\kappa = 0.128$, a $3.5\times$ gap.

Figure 1 visualises the contrast as an agreement-gap map: each row is one rubric \times dataset cell with LLM–human, LLM–LLM and the human–human ceiling plotted side by side, and segment length is the inter-LLM-minus-LLM–human gap.

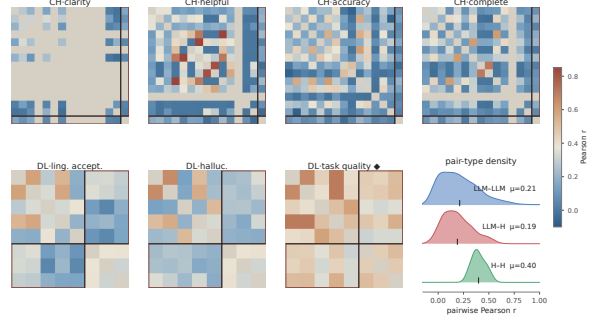


Figure 1: **Agreement-gap map.** Each row is one rubric \times dataset cell. Red=LLM–human; blue=LLM–LLM; black ticks=human–human ceiling. Segment length is the inter-LLM minus LLM–human gap. Subjective CH rubrics have the longest gaps; the objective factual rubric collapses the gap.

cohort	n	σ_J/σ_H	θ°	r	r_{ref}	Δ
(A) Rubric\timesdataset						
CH subj. mean	14	0.454	89.3	0.093	0.180	+0.087
DLQ LA	5	0.734	74.0	0.273	0.356	+0.083
DLQ HAL	5	0.784	71.6	0.316	0.349	+0.033
DLQ FR	5	0.687	58.5	0.519	0.530	+0.011
(B) Base-judge strata on CH						
FE	4	0.636	87.5	0.111	0.093	+0.018
OXL	4	0.305	88.0	0.073	0.093	−0.020
OM	4	0.264	89.3	0.075	0.093	−0.018
IF	2	0.642	89.6	0.080	0.093	−0.013
(C) Training stage on CH						
BASE	3	0.319	87.8	0.081	0.081	0.000
FS	3	0.624	88.3	0.122	0.081	+0.041
FT	3	1.074	87.1	0.129	0.081	+0.048
FT+PO	3	1.078	87.1	0.132	0.081	+0.051
AT	1	0.649	88.0	0.177	0.081	+0.096
CAL	1	0.633	88.1	0.184	0.081	+0.103
(D) CH top judges						
sarvam-m CAL v3	1	0.633	88.1	0.184	0.123	+0.062
sarvam-m FS	1	0.639	88.3	0.184	0.123	+0.061
grok-4	1	0.636	87.5	0.167	0.123	+0.044

Table 1: **Consolidated results grid.** r =Pearson vs. reference r_{ref} ; $\Delta = r - r_{\text{ref}}$ (inter-LLM in A, CH-pool in B, base in C, gpt-5.5 in D). Yellow : worst LLM–human gap. Blue : boundary-crossing objective rubric, the only stage lifting all four CH rubrics, and the best calibrated 24B Indic judge. Full column legend, code legend, and per-panel narrative in App. D.1.

4.2 The boundary is rubric-conditional

The strongest reading of consensual blindness would say that LLM judges are generically unreliable. REAL-WORLD CULTURAL QA falsifies that reading. With the same items, same five LLM judges, and same three humans, the factual rubric behaves differently from the subjective rubrics: LLM-human agreement reaches 0.519, almost identical to LLM-LLM agreement (0.530) and above the human-human mean (0.474). Figure 2 plots this as a phase portrait. The horizontal axis asks whether LLM-

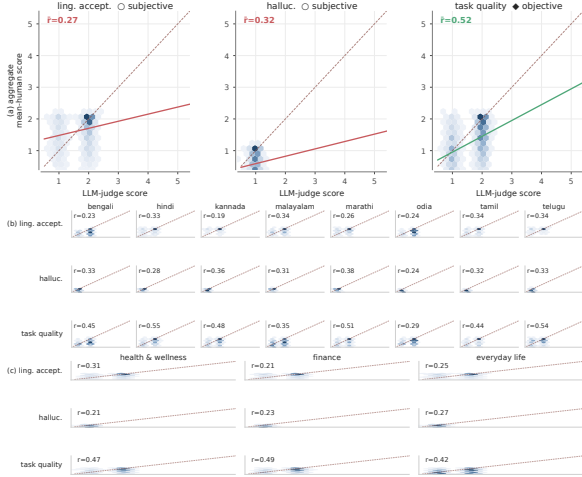


Figure 2: **Boundary phase portrait.** The shaded band is $\pm 15\%$ around human–human agreement. Subjective linguistic acceptability remains outside the band and below LLM–LLM consensus; hallucinations nearly reaches the band; the objective factual rubric enters the band and nearly reaches the LLM–LLM diagonal.

human reaches the human ceiling; the vertical axis asks whether it catches up to inter-LLM consensus. Linguistic acceptability sits in the blindness quadrant, hallucinations approaches the boundary, and the factual rubric sits in the rank-matched region.

We do not claim that LLM-as-judge fails in general. We claim that subjective rubrics with high intrinsic dimension expose a mismatch between the judge subspace and the human subspace, whereas objective rubrics can be rank-matched. Concrete items illustrating both regimes (factual recall and shared cultural knowledge on one side; healthcare access, accessibility, overmedicalisation, and emotional framing on the other) are listed in Tables 11 and 12 (App. H).

4.3 Geometric signature of the disagreement

The subjective-rubric failure is not only a correlation gap. It has a characteristic geometry: compressed score spread, low effective rank, high angle to the human subspace, and large null-space fraction. Table 1 summarizes the current geometry. On COMMUNITY-HEALTH, the median base judge uses less than half the human score spread (σ -ratio 0.45) and sits at 89.3° to the human subspace. On REAL-WORLD CULTURAL QA, the base LLMs have higher

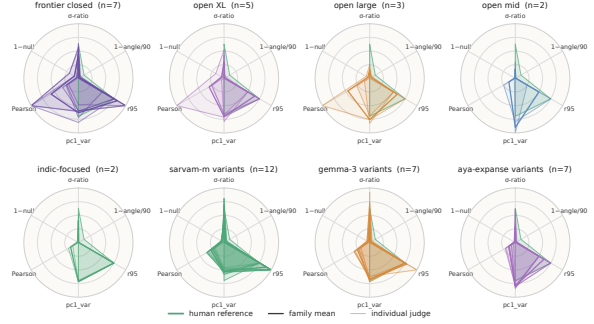


Figure 3: **Per-judge geometric signature.** Each judge is a glyph in $(\sigma$ -ratio, principal-angle) space. Colour gives Pearson alignment, marker size gives r_{95} , edge/shape gives judge family, and arrows trace trained variants.

Pearson because the benchmark includes more objective signal, but their stacked angle still sits around 88.1° while humans lie at 77.8 – 80.8° from each other. This is why the angle, not just Pearson, matters.

Figure 3 shows the full judge-level geometry: position encodes σ -ratio and angle, marker size encodes effective rank, colour encodes Pearson alignment, and arrows trace training trajectories. Training reliably moves judges rightward (spread) but not downward (angle).

The per-rubric geometry explains why the boundary analysis is not an accident (Fig. 4). The factual rubric does not simply raise Pearson; it changes the orientation of the judge vectors. The LLM mean angle on the factual rubric (58.5°) is inside the human-pair floor (61.7°), even though the score spread remains compressed (σ -ratio 0.69). Thus correlation alignment and spread calibration are separable: the objective rubric points in the right direction, but still under-uses the range.

4.4 Geometry across Indic languages

The remaining diagnostics use REAL-WORLD CULTURAL QA (DLQ), the only setup with three independent humans per item, so per-cell H–H references exist. Across the eight Indic languages, the LLM-mean subspace sits at 63 – 87° from human, with the high-resource languages near the inter-human band and the low-resource tail closer in (Bengali 63.5° , Malayalam 68.1°). σ -compression is universal (0.58–0.97), and the LL–LH gap is positive for 7/8 languages (largest on Bengali $+0.10$) but flips on Telugu (-0.06): consensual blindness

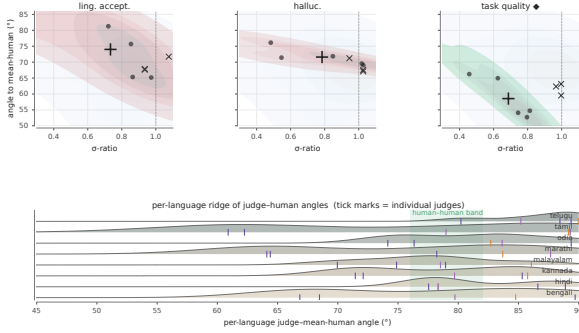


Figure 4: **Rubric-conditional geometry on DLQ.** Each panel is one rubric. Points=LLM judges; X=human-pair floors; diamond=LLM mean. The objective factual rubric rotates the LLM mean into the human floor while keeping σ below 1.

is a geometric, language-conditional offset, not a uniform constant (Fig. 5).

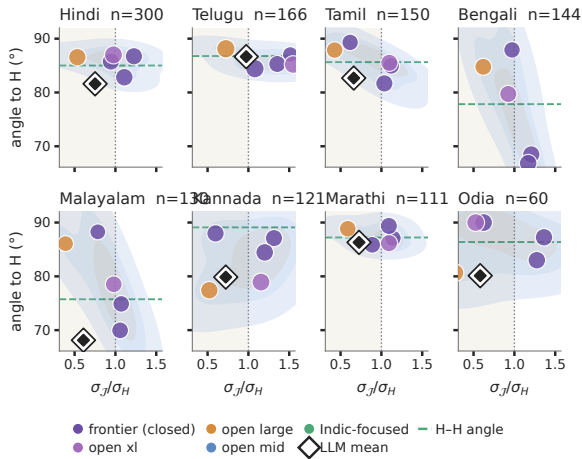


Figure 5: **Per-language DLQ geometric signature.** One facet per Indic language. Each panel plots every base-LLM judge in $(\sigma_J/\sigma_H, \text{angle to mean human})$ space; marker size encodes $|r|$ vs. mean human, colour encodes judge category, KDE contours render the judge density. Cream band marks the $\sigma < 1$ compression zone; dashed green line is the per-language inter-human angle. Open diamond is the LLM mean.

4.5 Geometry across domains

Across the three DLQ domains {health, finance, everyday-life}, the rubric-conditional boundary is reproduced inside every domain: on the objective TASK-QUALITY rubric (shaded), the LLM mean sits at 58–62°, at or below the inter-human angle (63°), and LH Pearson enters the H–H band (0.38–0.42 vs. 0.45); on the two subjective rubrics, the LLM mean stays

4–9° above the inter-human angle and LH falls clearly below H–H. σ -compression (0.67–0.82) is essentially domain-invariant. The objective–subjective boundary is therefore not a single-domain artefact (Fig. 6).

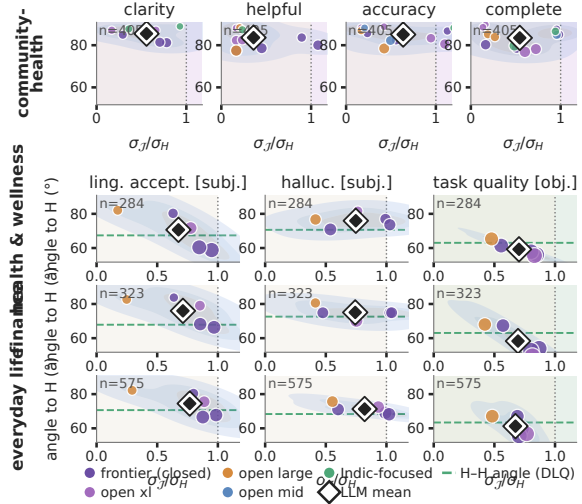


Figure 6: **Per-(domain, rubric) geometric signature.** *Top strip:* COMMUNITY-HEALTH (Samiksha; 14 base LLM judges over four subjective rubrics; single human pool, so no H–H reference line). *Bottom block:* DAILY-LIFE QA (Pariksha), rows: {health & wellness, finance, everyday life}; columns: {linguistic acceptability, hallucinations, task quality}. Same legend as Fig. 5. Pale-green background marks the single objective rubric (*task quality*); pale-lavender shades the COMMUNITY-HEALTH strip. The objective DLQ column collapses toward the inter-human angle on every row; the two subjective DLQ columns sit clearly above it; community-health judges sit at very high angles ($\geq 80^\circ$) under heavy σ -compression.

4.6 Subspace topology of the judge cohort

The principal-angle diagnostics so far ask, per judge, *how far is its evaluation axis from the human one?* A complementary structural question is whether judges all sit in the *same* direction or split into sub-clusters that happen to land equally far from humans, i.e., whether the apparent consensus is uniform or an average over distinct judge perspectives. We Z-score each judge’s per-rubric scores across the 405 COMMUNITY-HEALTH items (removing the score-spread component already characterised in Sec. 4.3–4.7) and compute the cosine similarity of every pair of 1620-d (item \times rubric) judge vectors; cosine then measures pure subspace

orientation.

Inter-judge subspace similarity (community-health)

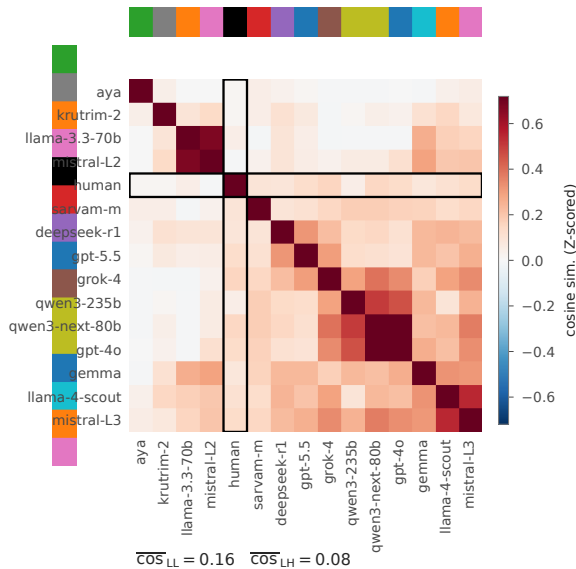


Figure 7: **Inter-judge subspace similarity (community-health)**. Pairwise cosine similarity between per-rubric Z-scored item \times rubric vectors of the 14 base LLM judges and the human reference, hierarchically reordered (Ward on $1 - \cos$). Side strips colour rows/columns by model family; the human row/column is outlined. The human cells are uniformly pale: no LLM judge aligns with the human axis after spread is removed. $\overline{\cos}_{LL} = 0.16$ vs. $\overline{\cos}_{LH} = 0.08$ (factor of two), and the deepest red is within-family (qwen3-235b \leftrightarrow qwen3-next-80b; llama-4-scout \leftrightarrow mistral-L3).

Three structural features in Fig. 7 are invisible to per-judge angle plots. (i) Within-family covariance is the strongest orientation signal (cosines 0.45–0.65 on qwen3, llama/mistral pairs vs. a 0.16 across-family mean): shared pre-/post-training data dominates over scale or access. (ii) The cohort splits into two macro-blocks, a tight frontier+open block (deepseek, gpt-5.5, gpt-4o, grok-4, qwen3, gemma, llama-4-scout, mistral-L3; intra 0.27) and a looser mid-tier multilingual block (aya, krutrim-2, llama-3.3-70b, mistral-L2; intra 0.18), with across-block cosines at 0.08; reported as a single **LL** aggregate, these average to a consensus that no judge actually holds. (iii) The human row is uniformly pale even after Z-scoring, so the collapse is neither a scale artefact nor a single-judge artefact but a property of the LLM-shared subspace itself.

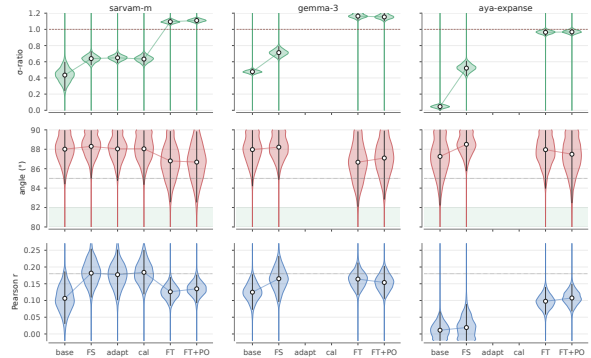


Figure 8: **Stage-wise bootstrap progression**. Left: stage trajectories in $(\sigma\text{-ratio}, \theta)$ with 95% CIs. Right: Pearson r with 95% CIs. Training reliably increases spread, but angle CIs overlap.

4.7 Training restores spread before orientation

We now ask which part of the geometry training can change. On COMMUNITY-HEALTH, we follow sarvam-m 24B, gemma-3 27B, and aya-expanse 32B through their available stages – base, few-shot, fine-tune, fine-tune+pref-optim, and for sarvam-m, adaptive-tune and calibrated – selecting the best in-stage judge per family by stacked Pearson and running paired item-bootstrap CIs. σ -ratio rises strongly under fine-tune and fine-tune+pref-optim ($0.32 \rightarrow 1.08$), but the angle axis barely moves: stage means remain near $87\text{--}88^\circ$ with overlapping CIs (Fig. 8). The aggregate also masks a per-rubric sign reversal: fine-tune/pref-optim over-spreads accuracy ($\sigma\text{-ratio} \approx 2.2$) and degrades completeness ($\Delta r = -0.055$); only calibration lifts all four rubrics together, placing the best calibrated 24B Indic judge at $r = 0.184$ – above the gpt-5.5 anchor (0.123) yet well below the human ceiling.

4.8 Decomposing the training delta: stretch, rotate, residual

The bootstrap view above describes what training *outputs*, but not *where each unit of delta goes* in geometry space. For every (family, stage) we decompose the per-rubric score-change vector $\Delta s = s_{\text{stage}} - s_{\text{base}}$ over the 405 items into three orthogonal energy components: **stretch** along the base axis u_b (amplifies what base already ranks), **rotate** \rightarrow **human** along the human axis after removing u_b (the only direction that can shrink the principal angle), and **off-axis residual** (item-rank reshuffling

that aligns with neither).

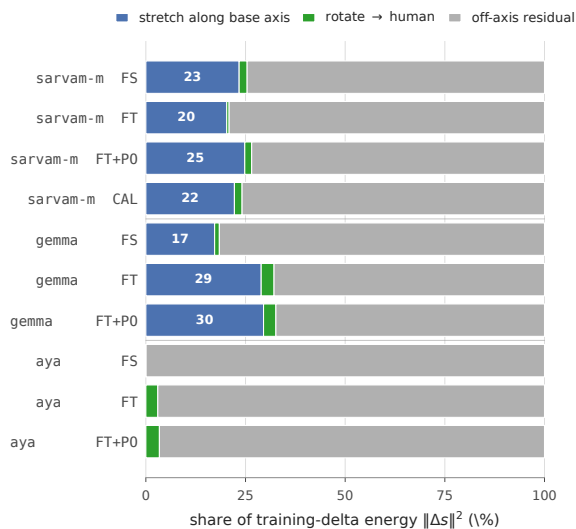


Figure 9: **Where each unit of training delta goes.** Per-(family, stage) decomposition of $\|\Delta s\|^2$ into stretch along the base axis (blue), rotation toward the human-orthogonal axis (green), and off-axis residual (grey); rubric-mean. Rotation toward human is $\leq 3\%$ across all eleven stage cells; the structured part of Δs is almost entirely stretch.

Two findings make the orientation result mechanistic rather than just observed. First, **rotation toward the human axis is uniformly tiny**: rotate \rightarrow human is $\leq 3\%$ of $\|\Delta s\|^2$ in every cell (stage means 1.1%/2.3%/2.8%/1.9% for FS/FT/FT+PO/CAL). Whatever spread training restores, almost none of the delta lies along the only direction that could reduce the principal angle. Second, the structured part of Δs is dominated by **stretch** (17–30% on sarvam-m and gemma, $\leq 0.1\%$ on aya), and tracks the stage’s observed σ -ratio gain: variance added by training is mostly a louder version of the base ranking. The remaining ~ 70 –99% is item-rank reshuffling that aligns with neither anchor; on aya, $\sim 97\%$ of $\|\Delta s\|^2$ is reshuffling and almost none is stretch, which is why aya gains less stacked Pearson than sarvam-m at the same stage despite similar σ -ratio movement. Training acts on judges as *amplify-then-reshuffle*; the orientation axis is essentially never moved.

Item-level walkthroughs that ground both regimes, objective rank-matching and subjective subspace mismatch, are in App.H (Tables 11–12).

5 Discussion

The four analyses converge on one finding: subjective-rubric LLM–human agreement sits well below inter-LLM consensus, the gap has a distinctive geometric shape (spread compression, rank loss, near-orthogonal angles to the human subspace), the gap closes on the one rubric with a verifiable factual answer, and training restores spread before orientation while masking per-rubric harm. We draw three implications.

Subjective inter-LLM consensus is perspective collapse, not alignment. On REAL-WORLD CULTURAL QA’s subjective rubrics, the three human raters agree at $r_{HH} \approx 0.36$; the LLM cohort reaches the same consensus level ($r_{LL} \approx 0.35$) but agrees with humans at only $r_{LH} \approx 0.27$ –0.32. The inter-human floor is low because reasonable raters legitimately differ on culturally embedded items; the LLMs match that ceiling not by recovering this perspective variance but by collapsing onto a shared low-rank surrogate (effective rank ≤ 3 against a human band that uses the full rubric basis). The healthcare items in Table 12 (App. H, rows SUBJ-1–SUBJ-4) make the collapse concrete: across complaint procedure, surgical access, household allergy management, and post-Caesarean delivery, the LLM cohort consistently scores responses on medical-completeness while humans weight institutional vulnerability, affordability, low-cost everyday precautions, and emotional reassurance. Reporting **LL**, **LH**, and **HH** separately, rather than a single aggregate, makes this collapse legible and lets an evaluator choose between a consensus signal and a perspective-preserving one.

Training corrects spread, not orientation, and aggregate metrics hide per-rubric harm. Fine-tuning and FT+PO move $\sigma_{\mathcal{J}}/\sigma_H$ from 0.32 to 1.08, but the principal angle stays at 87–88°; the delta-decomposition (Sec. 4.8) explains why – rotate \rightarrow human is $\leq 3\%$ of $\|\Delta s\|^2$ across every stage. The same stages degrade completeness while over-spreading accuracy (Wolpert, 1996); only calibration lifts all four rubrics, with the calibrated 24B Indic judge ($r=0.184$) above gpt-5.5 (0.123).

6 Limitations

The study covers Indic LLM-as-judge over four community datasets, eight languages, and 41 judges. Diagnostics assume a fixed rubric basis; free-form judgements would need an embedding step. Per-cell H–H values outside the three-human DLQ subset are estimates.

References

- Sanjeev Arora, Rong Ge, Ankur Moitra, and Yonatan Halpern. 2018. A linear algebraic structure of word senses, with applications to polysemy. In *Transactions of the Association for Computational Linguistics*, volume 6, pages 483–495.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. [Training a helpful and harmless assistant with reinforcement learning from human feedback](#).
- Mikhail Belkin and Partha Niyogi. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396.
- Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.
- Gayatri Bhat, Hamna, Sourabrata Mukherjee, Manan Uppadhyay, Brintha Chandrasekaran, Monali Shelar, Faisal Lalani, Evan Heafield, Kavitha K, Vivek Seshadri, Manu Chopra, Divya Siddarth, Kalika Bali, and Sunayana Sitaram. 2025. [Samiksha: What large-scale, community-driven evaluation reveals about language models for india](#).
- Åke Björck and Gene H Golub. 1973. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594.
- Damián Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world’s languages. *arXiv preprint arXiv:2110.06733*.
- Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiušė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, and 27 others. 2022. [Measuring progress on scalable oversight for large language models](#).
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners.
- Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. [The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16051–16081, Vienna, Austria. Association for Computational Linguistics.
- Cheng-Han Chiang and Hung yi Lee. 2023. [Can large language models be an alternative to human evaluations?](#)
- Tianzhe Chu, Jiaxin Zhang, Zhenyu Liao, Qiuyu Ren, Tahsin Saffat, Zitong Yang, Yi Ma, and Yue Zhang. 2025. Gauss eval: Human-llm judge consistency analysis. *GAUSS Blogs*. <https://gauss-math.ai/eval.html>.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. In *Educational and Psychological Measurement*.
- Ronald R Coifman and Stéphane Lafon. 2006. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30.
- Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2024. [AlpacaFarm: A simulation framework for methods that learn from human feedback](#).
- Bradley Efron. 1979. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*.
- Elena Facco, Maria d’Errico, Alex Rodriguez, and Alessandro Laio. 2017. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):1–8.
- Gemini Team, Google. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. [Chatgpt outperforms crowd workers for text-annotation tasks](#). *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

- Gene H Golub and Charles F Van Loan. 2013. *Matrix Computations*, 4th edition. Johns Hopkins University Press.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. *The llama 3 herd of models*. *Preprint*, arXiv:2407.21783.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *ICML*.
- Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2024. *Are large language model-based evaluators the solution to scaling up multilingual evaluation?* In *Findings of the Association for Computational Linguistics: EAACL 2024*, pages 1051–1070.
- Hamna Hamna, Gayatri Bhat, Sourabrata Mukherjee, Faisal M. Lalani, Evan Hadfield, Divya Siddarth, Kalika Bali, and Sunayana Sitaram. 2026. *Building benchmarks from the ground up: Community-centered evaluation of LLMs in healthcare chatbot settings*. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA. Association for Computing Machinery.
- Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. *Challenges and strategies in cross-cultural NLP*. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.
- Ian T Jolliffe and Jorge Cadima. 2016. *Principal Component Analysis*, 2nd edition. Springer.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. *arXiv preprint arXiv:2004.09095*.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Andrew V Knyazev and Merico E Argentati. 2002. Principal angles between subspaces in an a -based scalar product: Algorithms and perturbation estimates. *SIAM Journal on Scientific Computing*, 23(6):2008–2040.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. *G-eval: NLG evaluation using gpt-4 with better human alignment*. pages 2511–2522.
- Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.
- Carl D Meyer. 2000. *Matrix analysis and applied linear algebra*. SIAM.
- OpenAI. 2024. Gpt-4o technical report. <https://openai.com/research/gpt-4o>.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. *Gpt-4 technical report*.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. *Training language models to follow instructions with human feedback*.
- John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*.
- Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. *Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability*. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Sarvam AI. 2024. Sarvam-2B: An indic language model. <https://www.sarvam.ai>. Accessed: 2024-07-15.

- Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. [Are emergent abilities of large language models a mirage?](#) In *Advances in Neural Information Processing Systems*, volume 36, pages 55565–55581. Curran Associates, Inc.
- Hinrich Schütze. 1992. [Dimensions of meaning](#). In *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing*, pages 787–796.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. [Large language models are not fair evaluators](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.
- Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. 2024. [PARIKSHA: A large-scale investigation of human-LLM evaluator agreement on multilingual and multi-cultural data](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7900–7932, Miami, Florida, USA. Association for Computational Linguistics.
- Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*.
- David H Wolpert. 1996. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.
- Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. [Multilingual machine translation with large language models: Empirical results and analysis](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.
- George Kingsley Zipf. 1935. *The Psycho-Biology of Language*. Houghton Mifflin.

A Per-domain breakdown of datasets, languages, generators, and judges

Per-domain coverage (referenced from Sec. 3). Table 2 lists languages, item counts, and rubrics per domain; Table 3 lists answer generators and judge variants per domain. Trained judge variants apply only to COMMUNITY-HEALTH.

Domain dataset	Languages	Qs	QA	Rubrics (eval type)
COMMUNITY-HEALTH (Hamna et al., 2026; Bhat et al., 2025)	Hi; Kn; Ml	243	1,215	clarity, helpfulness, accuracy, completeness (STANDALONE, {0, 1, 2}); A/B/NS (COMPARATIVE)
DLQ: HEALTH & WELLNESS (Watts et al., 2024)	8 Indic [†]	432	432	factual rubric ({0, 1, 2}, <i>obj.</i>); linguistic acceptability ({0, 1, 2}); hallucinations ({0, 1}) (STANDALONE)
DLQ: FINANCE (Watts et al., 2024)	8 Indic [†]	432	432	same as above
DLQ: EVERYDAY LIFE (Watts et al., 2024)	8 Indic [†]	432	432	same as above
Totals	8 unique langs	1,539	2,511	4 domain datasets; >240,000 scoring events

Table 2: **Per-domain breakdown (datasets, languages, items, rubrics)**. *Hi/Kn/Ml* = Hindi/Kannada/Malayalam. [†]DLQ languages: Bengali, Hindi, Tamil, Telugu, Marathi, Kannada, Malayalam, Odia. Counts: COMMUNITY-HEALTH is 81 questions \times 3 languages \times 5 answer generators.

Field	community-health
Answer generators	\approx 15 generator models combined across COMMUNITY-HEALTH and REAL-WORLD CULTURAL QA (Hamna et al., 2026; Bhat et al., 2025; Watts et al., 2024)
Judges, <i>base</i> (14)	deepseek-r1; gpt-5.5; gpt-4o; gpt-4; grok-4; mistral-large-3; mistral-large-2; llama-4-scout-17b; llama-3.3-70b; qwen3-235b; qwen3-next-80b; krutrim-2-instruct; aya-expanse-32b; gemma-3-27b; sarvam-m-24b
Judges, <i>improved sarvam-m</i> (9)	few-shot; fine-tune; pref-optim; fine-tune+pref-optim; adaptive-tune-v1; adaptive-tune-v1-inf3; adaptive-tune-v2; calibrated-v2; calibrated-v3
Judges, <i>improved gemma-3</i> (4)	few-shot; fine-tune; pref-optim; fine-tune+pref-optim
Judges, <i>improved aya-expanse</i> (4)	few-shot; fine-tune; pref-optim; fine-tune+pref-optim
Field	DLQ: health & wellness / finance / everyday life (identical roster across all three)
Answer generators	\approx 15 generator models combined across COMMUNITY-HEALTH and REAL-WORLD CULTURAL QA (Hamna et al., 2026; Bhat et al., 2025; Watts et al., 2024)
Judges, <i>base</i> (5)	gpt-4o; gpt-4; gemma-3-27b; qwen3-235b; llama-3.3-70b
Judges, <i>improved</i>	none (trained variants not applied to DLQ)

Table 3: **Answer generators and judge variants per domain**. COMMUNITY-HEALTH totals 14+27=41 judge variants; the three DLQ domains share the same 5 base judges. Combined: 46 **distinct judge variants** across the study.

Human reference panels. COMMUNITY-HEALTH: 15 native-speaker medical evaluators (5 Hindi, 6 Kannada, 4 Malayalam) from 5 partner community organisations; sparse per-rater overlap, so merged into one HUMAN-COMMUNITY-HEALTH reference. REAL-WORLD CULTURAL

QA: three independent humans (h_1, h_2, h_3) per item per (domain, language), giving a measurable H–H ceiling. All source questions and human reference labels are community-curated and human-authored.

B Judge models

Table 4 lists the 16 base LLM judges; Table 5 enumerates the full roster of 41 LLM judges (the 13 non-trainable bases, the 3 trainable family bases, and 25 trained variants over three Indic-capable families), plus the 4 human reference pools used as anchors. gpt-4o and gpt-4 act as cross-dataset anchors.

Model	Family / orientation	Size	Access
gpt-5.5	GPT (OpenAI), general	API	API
gpt-4o	GPT (OpenAI), general (OpenAI, 2024)	API	API
gpt-4	GPT (OpenAI), general (OpenAI et al., 2024)	API	API
grok-4	Grok (xAI), general	API	API
deepseek-r1	DeepSeek, reasoning	API	API
qwen3-235b	Qwen (Yang et al., 2025), multilingual	235B	local
qwen3-next-80b	Qwen (Yang et al., 2025), multilingual	80B	local
llama-3.3-70b	Llama (Grattafiori et al., 2024; Touvron et al., 2023), general	70B	local
llama-3.1-405b	Llama (Grattafiori et al., 2024), general	405B	API
llama-4-scout-17b	Llama, general	17B	local
mistral-large-2	Mistral, general	123B	API
mistral-large-3	Mistral, general	123B	API
gemma-3-27b	Gemma (Gemini Team, Google, 2023), multilingual	27B	local
aya-expanse-32b	Aya (Cohere), multilingual	32B	local
sarvam-m-24b	Sarvam (Sarvam AI, 2024), Indic-focused	24B	local
krutrim-2-instruct	Krutrim, Indic-focused	12B	local

Table 4: Base LLM judges. Size shows reported active-parameter count where disclosed; API otherwise. Per-domain coverage in Table 2; trained variants in App. F.

C Rubrics and scoring scales

Per-dataset rubric definitions and ordinal scales (Sec. 3). The COMMUNITY-HEALTH comparative paradigm returns one of $\{A, B, \text{NOT SURE}\}$ per item with a one-sentence English rationale. Verbatim prompts in App. E.

D Additional figures

D.1 Consolidated results grid: column key and codes

Compact legend for Table 1. Per-panel narrative summary in Table 8.

E Evaluation prompts

Two prompts: STANDALONE multi-rubric and COMPARATIVE pairwise. Both are language-agnostic; the judge evaluates within the answer’s language and writes the explanation in English.

Standalone Multi-Rubric Scoring Prompt

Task. Evaluate the answer to the question on four rubrics.

Audience. Internal evaluator. Do not browse. Use only general knowledge and the given text.

Language. The (question, answer) pair may be in Hindi, Kannada, Malayalam, Bengali, Tamil, Telugu, Marathi, or Odia. Judge *within* the answer’s language. Write the explanation in concise English.

Rubrics (one rating per rubric: Yes / Somewhat / No):

- i.* **Clarity & fluency:** is the language clear, grammatically correct, easy to understand?
- ii.* **Helpfulness & relevance:** does the answer directly address the question and provide useful informa-

Variant ID	Recipe	Display
sarvamm	base	sarvam-m (24b)
sarvamm_fs	few-shot	sarvam-m (fewshot)
sarvamm_ft_v1	fine-tune	sarvam-m (FT v1)
sarvamm_ft_v2	fine-tune	sarvam-m (FT v2)
sarvamm_ft_v3	fine-tune	sarvam-m (FT v3)
sarvamm_po_v2	pref-optim	sarvam-m (PO v2)
sarvamm_ft_po_v1	FT + PO	sarvam-m (FT+PO v1)
sarvamm_ft_po_v2	FT + PO	sarvam-m (FT+PO v2)
sarvamm_ft_v3_demo1	adaptive-tune	sarvam-m (AT v1)
sarvamm_ft_v3_demo2	adaptive-tune	sarvam-m (AT v2)
sarvamm_ft_v3_demo1_infer3	adaptive-tune	sarvam-m (AT v1-inf3)
sarvamm_po_v3_cal_2demo	calibrated	sarvam-m (CAL v2)
sarvamm_po_v3_cal_3demo	calibrated	sarvam-m (CAL v3)
<hr/>		
gemma-3-27b-it	base	gemma-3 (27b)
gemma-3-27b-it_fs	few-shot	gemma-3 (fewshot)
gemma-3-27b-it_ft_v1	fine-tune	gemma-3 (FT v1)
gemma-3-27b-it_ft_v1_b128	fine-tune	gemma-3 (FT v1-b128)
gemma-3-27b-it_ft_v2	fine-tune	gemma-3 (FT v2)
gemma-3-27b-it_po_v1	pref-optim	gemma-3 (PO v1)
gemma-3-27b-it_po_v2	pref-optim	gemma-3 (PO v2)
gemma-3-27b-it_ft_po_v2	FT + PO	gemma-3 (FT+PO v2)
<hr/>		
aya-expanse-32b	base	aya-expanse (32b)
aya-expanse-32b_fs	few-shot	aya-expanse (fewshot)
aya-expanse-32b_ft_v1	fine-tune	aya-expanse (FT v1)
aya-expanse-32b_ft_v2	fine-tune	aya-expanse (FT v2)
aya-expanse-32b_po_v1	pref-optim	aya-expanse (PO v1)
aya-expanse-32b_po_v2	pref-optim	aya-expanse (PO v2)
aya-expanse-32b_ft_po_v1	FT + PO	aya-expanse (FT+PO v1)
aya-expanse-32b_ft_po_v2	FT + PO	aya-expanse (FT+PO v2)
<hr/>		
h_ch	native medical eval	CH human reference
h1	native annotator	DLQ human h_1
h2	native annotator	DLQ human h_2
h3	native annotator	DLQ human h_3

Table 5: The 41 LLM judges plus 4 human reference pools: 16 base LLMs from Table 4 (the 13 non-trainable bases plus the three trainable family bases listed here), and 25 trained variants across three open Indic-capable families (SARVAM-M, GEMMA-3-27B-IT, AYA-EXPANSE-32B). Recipe definitions and hyper-parameters in App. F.

Dataset	Rubric	Scale	Definition
COMMUNITY-HEALTH	clarity & fluency	{0, 1, 2}	language clear, grammatical, easy to understand
COMMUNITY-HEALTH	helpfulness & relevance	{0, 1, 2}	directly addresses the question with useful info
COMMUNITY-HEALTH	accuracy	{0, 1, 2}	trustworthy under general medical knowledge
COMMUNITY-HEALTH	completeness vs. concis.	{0, 1, 2}	sufficient detail without verbosity
<hr/>			
REAL-WORLD CULTURAL QA	factual rubric (<i>obj.</i>)	{0, 1, 2}	satisfies task constraints; verifiable ground truth
REAL-WORLD CULTURAL QA	linguistic acceptability	{0, 1, 2}	native-speaker grammaticality and naturalness
REAL-WORLD CULTURAL QA	hallucinations	{0, 1}	binary presence of unsupported content

Table 6: Rubric definitions and ordinal scales. Yes= 2 / Somewhat= 1 / No= 0 on COMMUNITY-HEALTH; factual rubric is the single objective baseline in this study.

tion?
<i>iii.</i> Accuracy (general perception): based on general knowledge, does the information seem trustworthy and factually correct?
<i>iv.</i> Completeness & conciseness: does the answer provide sufficient detail without being verbose or repetitive?
Scoring. Yes= 2, Somewhat= 1, No= 0. Overall = sum of the four.
Output schema (strict, single minified JSON, no prose, no code fences):
{

Symbol / code	Meaning
n	judges/humans in cohort
$\sigma_{\mathcal{J}}/\sigma_H$	score-spread ratio (judge vs. human)
θ°	principal angle to human subspace
r	Pearson vs. reference r_{ref}
$\Delta = r - r_{\text{ref}}$	shortfall below the panel reference
$r_{95}^{\mathcal{J}}$	effective rank (95% variance); range {2.0, 2.18, 2.3, 2.81, 3.0, 3.5, 3.57, 3.96, 3.99, 4.0}
CH / DLQ	community-health / real-world cultural QA
LA / HAL / FR	linguistic acceptability, hallucinations (subj.); factual rubric (obj.)
FE / OXL / OM / IF	frontier-closed / open XL-large / open mid / Indic-focused strata
BASE / FS / FT / FT+PO / AT / CAL	base, few-shot, fine-tune, FT+pref-optim, adaptive-tune, calibrated (AT, CAL: sarvam-m only)

Table 7: Column key and code legend for the consolidated results grid.

Panel	One-line narrative
(A)	Δ shrinks monotonically with rubric objectivity (0.087 \rightarrow 0.011); DLQ FR $r = 0.519$ enters HH band (0.474).
(B)	no base stratum closes more than 0.018 of the CH gap.
(C)	$\sigma_{\mathcal{J}}/\sigma_H$ rises 0.319 \rightarrow 1.078 across FT/FT+PO; θ° stays at 87–88°; largest Δ from CAL.
(D)	best calibrated 24B Indic judge surpasses gpt-5.5; absolute r remains below the DLQ HH floor.

Table 8: Per-panel narrative of Table 1.

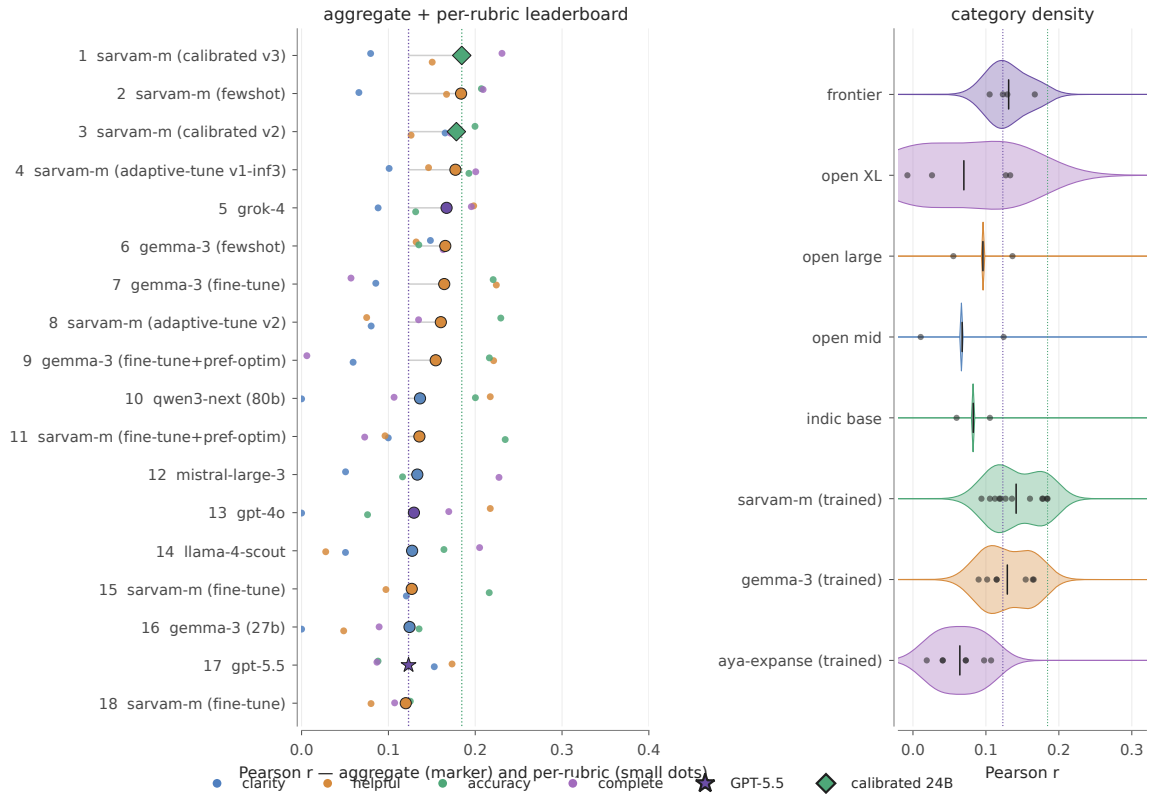


Figure 10: **Judge ranking.** Top-20 judges on stacked COMMUNITY-HEALTH Pearson r ; red line = gpt-5.5 (0.123), green line = best calibrated 24B judge (0.184).

```
"clarity":{"rating":"Yes|Somewhat|No","explanation":"..."},
```

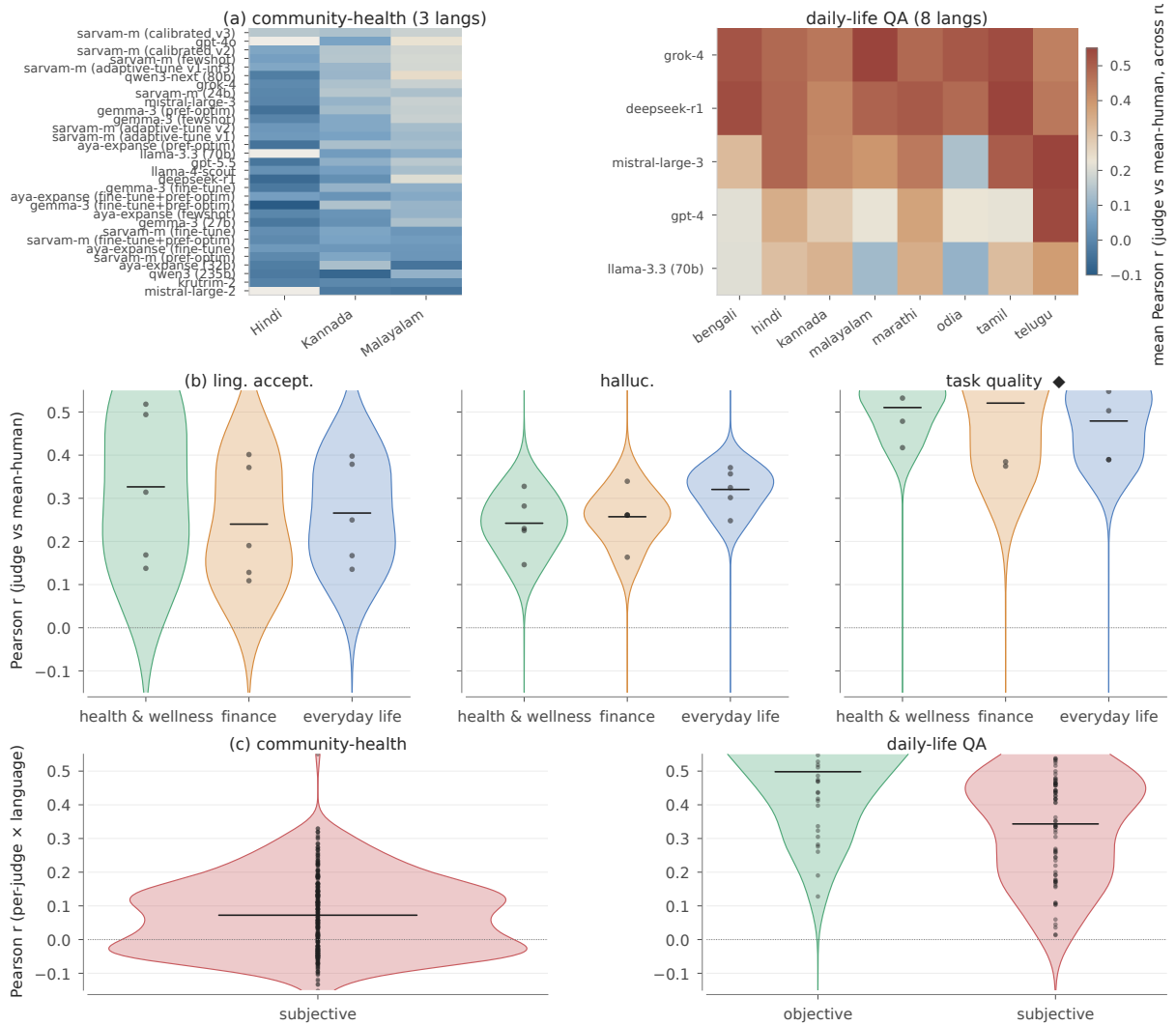


Figure 11: **Per-language, per-domain, and rubric-kind agreement across the four datasets.** (a) Per-language mean Pearson r between each judge and the mean-human reference, ordered by judge row mean; COMMUNITY-HEALTH (3 languages) on the left, REAL-WORLD CULTURAL QA (8 languages) on the right. (b) Per-domain Pearson distributions, one panel per DLQ rubric; the objective factual rubric sits highest in every domain. (c) Objective vs. subjective rubric-kind density per dataset; COMMUNITY-HEALTH has no objective rubric, while REAL-WORLD CULTURAL QA separates the two by ≈ 0.16 in median Pearson r .

```

"helpfulness":{"rating":"Yes|Somewhat|No","explanation":"..."},
"accuracy":{"rating":"Yes|Somewhat|No","explanation":"..."},
"completeness":{"rating":"Yes|Somewhat|No","explanation":"..."},
"scores":{"clarity":int,"helpfulness":int,"accuracy":int,
          "completeness":int,"overall":int}
}
Question: {{question}}
Answer: {{answer}}
Return the JSON now.

```

On REAL-WORLD CULTURAL QA the same template substitutes three rubrics: **factual rubric**, **linguistic acceptability** ($\{0, 1, 2\}$), **hallucinations** ($\{0, 1\}$).

Comparative (Pairwise Preference) Prompt

Task. Compare two answers to the same question and pick the better one overall.

Audience. Internal evaluator. Use only the question, the two answers, and general world knowledge; no external browsing.

Languages. The (question, A, B) triple may be in Hindi, Kannada, or Malayalam. Judge within the answers’ language; write the explanation in concise English (≤ 40 words).

Holistic criteria. Clarity & fluency, helpfulness & relevance, factual accuracy / plausibility, completeness (no critical omissions), and safety. Prefer the answer that better serves the user overall; do not reward hallucinated or unsafe content.

Decision rule. Choose A or B if one answer is even slightly better overall. Choose **Not sure** only if both are essentially equal or both are unintelligible/empty.

Output (strict, single JSON, exact keys):

```
{
  "winner": "A" | "B" | "Not sure",
  "explanation": "<concise justification <=40 words>"
}
```

Question: {{question}}

Answer A: {{answer_a}}

Answer B: {{answer_b}}

Return only the JSON now.

F Judge variant recipes

Six recipes applied across three open base families on COMMUNITY-HEALTH. Per-family roster in Table 2; common controls: temperature 0.2, top- $p=0.9$, max-output 256 tokens, training seed 42, bootstrap seed 20260525, single A100-80GB.

Recipe	Applied to	What & key hyper-parameters
few-shot	all three families	$k = 3$ stratified per (language \times rubric) demos; frozen weights; same decoding as base.
fine-tune (FT)	all three families	LoRA rank 32, $\alpha = 64$, target $\{q, k, v, o, \text{gate}, \text{up}, \text{down}\}$; AdamW LR 2×10^{-4} , batch 16, 3 epochs, cosine warm-up 3%, bf16; train on 80% of CH (Q \times A) \times rubric, language-stratified, held-out 20%.
pref-optim (PO)	all three families	DPO (Rafailov et al., 2023) on $\approx 4,500$ A/B/NS comparative triples; $\beta=0.1$, LR 5×10^{-6} , batch 8, 1 epoch, LoRA as above.
FT + PO	all three families	stage-wise: FT checkpoint \rightarrow DPO (1 epoch); adapter re-merged before DPO. Dominant trainable signal in our repair analysis.
adaptive-tune (AT)	sarvam-m only	FT reused at inference with top- k nearest gold-scored demos retrieved by sentence-embedding cosine; variants: AT-v1 ($k=1$), AT-v1-inf3 (train $k=1$, infer $k=3$), AT-v2 (only demos with judge-human agreement ≥ 0.6).
calibrated (CAL)	sarvam-m only	post-hoc rubric-conditional Platt / temperature scaling (Platt, 1999; Guo et al., 2017) on FT+PO; CAL-v2 (per-rubric T + global Platt), CAL-v3 (per-rubric and per-language Platt); L-BFGS on NLL; no base-weight updates.

Table 9: Trained judge variant recipes. All variants share the base-judge JSON schema and decoding settings; calibration sets are anchored to a small human-rated demo pool.

G Boundary condition and scope of the framework

Extends Sec. 2. Let d_R be the intrinsic dimensionality of the human-score manifold under rubric R and k_J the judge cohort’s effective rank. The four geometric predictions of Sec. 2 are gated by

$$\text{consensual blindness on } R \iff k_J < d_R.$$

Regime	Rubrics in this study	Predicted behaviour
small- d_R (objective)	DLQ factual rubric	oracle concentrates on \mathbf{h}_{fact} ; $\mathbf{P}_{\mathcal{J}}$ and f^* share a subspace; LLM-human agreement enters the H-H band (Sec. 4.2).
large- d_R (subjective)	CH all four; DLQ LA, HAL	$w_{\text{cult}}^R > 0$; $\mathbf{h}_{\text{cult}} \in \mathcal{N}(\mathbf{P}_{\mathcal{J}})$; the four signatures of Sec. 2 (σ -compression, low rank, θ° gap, inter-judge clustering) are quantitatively confirmed.

Table 10: Boundary condition: which rubrics fall in which regime and what the framework predicts there. The framework predicts *which* judges fail, the *shape* of failure, and *where* it vanishes; it does *not* predict which repair technique closes the gap, a dissociation Sec. 4.7 measures.

H Qualitative study: objective vs. subjective rubrics

This appendix grounds the quantitative geometry results in concrete items drawn from the evaluation set. We separate items into two regimes:

Objective rubrics. Items whose target answer is fact-based, verifiable, or grounded in widely shared cultural knowledge. Phrasing may vary across responders, but the underlying information is stable, so evaluation reduces to factual consistency and concept matching. Human and LLM judges converge sharply on these items (Table 11).

Subjective rubrics. Items that require contextual interpretation, reasoning, advisory guidance, or sensitivity to the user’s situation. There is no single fixed answer; evaluation depends on relevance, completeness, practical applicability, and contextual appropriateness rather than exact factual match. Healthcare items in particular carry emotional, financial, and access-related context that LLM judges tend to overlook even when responses are medically accurate (Table 12).

Rubric dimensions assessed. Responses are evaluated along four axes: (i) *task quality / factuality* (accuracy, expected details, direct address of the question); (ii) *linguistic acceptability* (grammar, word choice, spelling); (iii) *clarity* (well-formed, interpretable phrasing); and (iv) *helpfulness and local relevance* (practical usefulness and fit to the user’s situational or local context). The objective regime is dominated by axis (i); the subjective regime engages all four, and the failures in Table 12 concentrate on axis (iv).

Why this matters for the geometry. The objective items (Table 11, rows OBJ-1–OBJ-4) sit on a low-dimensional subspace that LLM judges already cover: the correctness axis is shared, and inter-LLM consensus is genuine consensus with humans. The subjective items (Table 12, rows SUBJ-1–SUBJ-4) require additional axes (access, vulnerability, affordability, reassurance) that LLM judges underweight or do not represent at all. This is the qualitative analogue of the principal-angle gap (Sec. 4): inter-LLM agreement is preserved because all judges collapse onto the same medical-completeness direction, while humans evaluate along directions that are orthogonal to it.

I Use of LLMs in Our Research Process

We used large language models to assist in polishing the writing of this paper and for exploring related work. However, all substantive analyses, evaluations, and interpretations were performed and validated manually by the authors to ensure accuracy and accountability.

ID	Lang	Question (script + English gloss)	Why human and LLM agree
OBJ-1	Pa	ਪੰਜਾਬ ਵਿੱਚ ਰੋਇਲ ਸਿਟੀ ਕਿਸ ਨੂੰ ਕਿਹਾ ਜਾਂਦਾ ਹੈ? (<i>What is called the Royal City in Punjab?</i>)	Factual recall with a single canonical answer (<i>Patiala</i>). The LLM rationale notes that the response satisfies all aspects of the prompt with no deviation; humans assign the same full-correct score. Exact agreement.
OBJ-2	Te	తెలుగు సంప్రదాయంలో ఆడపిల్లల తల్లితండ్రులు పిల్లల మొదటి జన్మకు ఎందుకు డబ్బు చెల్లిస్తారు? (<i>In Telugu tradition, why do the parents of a girl pay for the first childbirth?</i>)	Cultural practice grounded in widely shared understanding (e.g., <i>Janmantara Kanuka</i>). Responses vary in phrasing but capture the core intent; humans explicitly accept the “gist” as sufficient, and LLM judges score accordingly. High agreement.
OBJ-3	Hi	डेबिट कार्ड और क्रेडिट कार्ड में क्या अंतर है? (<i>What is the difference between a debit card and a credit card?</i>)	Conceptual but well-defined knowledge: debit draws from the user’s account; credit is borrowed and repaid with interest. Whenever responses articulate this contrast, both humans and LLMs rate them correct. Consistent agreement.
OBJ-4	Pa	ਮੀਰੀ ਪੀਰੀ ਦੀਆਂ ਤਲਵਾਰਾਂ ਕਿਸ ਗੁਰੂ ਸਾਹਿਬਾਨ ਨੇ ਧਾਰਨ ਕੀਤੀਆਂ ਸੀ ਅਤੇ ਇਸ ਨੂੰ ਧਾਰਨ ਕਰਨ ਦਾ ਕੀ ਉਦੇਸ਼ ਸੀ? (<i>Which Guru wore the swords of Miri and Piri, and what was their purpose?</i>)	Historical–religious fact: Guru Hargobind Sahib Ji; <i>Miri</i> (temporal power) and <i>Piri</i> (spiritual power). No ambiguity, no subjective interpretation; humans and LLMs agree completely.

Table 11: **Objective rubric, qualitative examples.** For fact-based or canonically known items, LLM judges converge with humans on near-identical scores. Referenced from Sec. 1 and Sec. 4 (boundary analysis).

ID	Lang	Theme	Question (script + English gloss)	What the LLM emphasises vs. what humans need
SUBJ-1	Kn	Healthcare access, vulnerability, and power dynamics	<p>ನಮ್ಮನೆ ಹತ್ತ ಇರೋ ಪ್ರಾಥಮಿಕ ಆರೋಗ್ಯ ಕೇಂದ್ರದಲ್ಲಿ ತೋರ್ಸೋಕೆ ಹೋದ್ರೆ ಎಲ್ಲಾದಕ್ಕೂ ದುಡ್ಡು ಕೇಳ್ತಾರೆ ನಾನ್ ಇವ್ ಮೇಲೆ ದೂರು ಕೊಡ್ಬೇಕು ಅಂದ್ರೆ ಏನ್ ಮಾಡ್ಬೇಕು?</p> <p><i>(If the nearby primary health centre asks for money to be seen, how do I file a complaint against them?)</i></p>	<p>LLM: detailed procedural guidance on filing complaints. Human-aligned: acknowledges fear of retaliation, dependence on the same centre, and recommends supportive pathways (community health workers, local officials, collective reporting). The question is about institutional trust and social hierarchy, not paperwork.</p>
SUBJ-2	MI	Medical recommendations vs. real-world accessibility	<p>എന്റെ ഒമ്പത് വയസ്സുള്ള മോൾ വി- ണ് നെറ്റി മുറിഞ്ഞു, ഒരുപാട് വലിയ പാട് വരാതെ സ്റ്റിച്ച് ചെയ്യാൻ ഏത് വിഭാഗത്തിലാണ് കുട്ടിയെ കാണി- ക്കേണ്ടത്?</p> <p><i>(My nine-year-old daughter cut her forehead in a fall; which department should I take her to so it can be stitched without a large scar?)</i></p>	<p>LLM: narrowly recommends a plastic surgeon. Human-aligned: balances clinical optimality with accessibility and affordability; many cuts are handled effectively by emergency physicians, paediatric surgeons, or general surgeons if treated promptly. The user's real concern is healing and minimal scarring, not access to specialist cosmetic care.</p>
SUBJ-3	Hi	Overmedicalisation of everyday living	<p>घर में बिल्ली और कुत्ते जैसे पालतू जानवर हैं। क्या उनके संपर्क से बच्चों और बड़ों को एलर्जी हो सकती है? रोज़मर्रा की जिंदगी में क्या एहतियात रखनी चाहिए?</p> <p><i>(We keep pets like cats and dogs at home. Can contact cause allergies in children and adults? What everyday precautions should we take?)</i></p>	<p>LLM: foregrounds HEPA filters, specialised vacuum cleaners, furniture replacement. Human-aligned: starts from low-cost, household-level precautions (cleanliness, handwashing, ventilation, symptom monitoring) and escalates only if needed. The user wants to keep living with pets, not to medicalise their home.</p>
SUBJ-4	MI	Emotional reassurance and risk framing in maternal health	<p>6 വർഷം മുമ്പ് ഞാൻ എന്റെ മകളെ പ്രസവിച്ചത് സിസേറിയൻ ആയിരു- ന്നു. ഇനി ഒരു കുട്ടിക്ക് ശ്രമിക്കേണ്ട സമയം ആയെന്ന എല്ലാവരും പറ- ഞ്ഞു തുടങ്ങി. ഞങ്ങളും അതെ തീ- രുമാനത്തിലാണ്. എന്നാലും എനി- ക്ക് ഇനി സുഖ പ്രസവം ആവാൻ ആണ് ആഗ്രഹിക്കുന്നത്. അതിനു എന്തെങ്കിലും ബുദ്ധിമുട്ടുകൾ ഉണ്ടാ- വുമോ? എന്തൊക്കെയാണ് അതി- നു വേണ്ടി മുൻകരുതേണ്ടത്?</p> <p><i>(My first delivery 6 years ago was a Caesarean. Family is asking us to try for a second child and we are deciding the same; I want a normal delivery this time. Are there difficulties? What precautions should I take?)</i></p>	<p>LLM: foregrounds uterine rupture, foetal distress, and worst-case complications under VBAC. Human-aligned: first validates the desire for a normal delivery, notes that many VBACs succeed under proper care, then discusses risks proportionately. The user is asking for reassurance and confidence under family pressure, not a complications checklist.</p>

Table 12: **Subjective rubric, healthcare qualitative examples.** On items that carry institutional, financial, or emotional context, LLM judges score on medical-completeness while humans weight practical accessibility, vulnerability, and reassurance. These items concentrate the principal-angle gap reported in Sec. 4. Referenced from Sec. 1 and Sec. 5.