Consensual Blindness: The Geometry of LLM Failure in Multilingual Community Health Evaluation

Anonymous ACL submission

Abstract

High inter-model agreement among LLM judges is often cited as evidence of reliability, but we challenge this: consensus frequently reflects **geometric mode collapse**, where LLMs flatten quality judgments onto low-rank fluency-dominated subspaces, erasing culturally-grounded dimensions. We formalize evaluation as a lossy projection operator and validate this on a multilingual health benchmark (Hindi, Kannada, Malayalam; 15 medical professionals; 600+ judgments). Findings: (1) variance compression ($\sigma_{\text{LLM}}/\sigma_{\text{Human}} = 0.69$; 93-95% null-space); (2) orthogonal subspaces ($> 79°$ angles; $< 7\%$ shared variance); (3) resource-stratified collapse (Malayalam 85.9%). We introduce **subspace alignment theory**: LLMs achieve high agreement through redundant projection onto shared fluency manifolds orthogonal to human judgment, consensus signals redundancy, not validity. We provide geometric diagnostics (rank, null-space, cross-lingual transfer) and outline subspace augmentation remedies, reframing LLM-as-Judge reliability as a geometry problem with actionable solutions.

1 Introduction

The Illusion of Reliable Consensus. LLM-as-judge is standard for evaluation (Chiang et al., 2023; Zheng et al., 2023; Wang et al., 2023), working for factual checks but faltering on implicit, subjective, and community-dependent judgments. High inter-model agreement masks geometric orthogonality (Gudibande et al., 2023): LLM judges fail not due to poor tuning, but because they operate in subspaces orthogonal to human value spaces. In health, where trust depends on local norms and rarely-explicit context, this is catastrophic (Raji et al., 2021). Multilingual evaluation amplifies this: community awareness, cultural safety, and pragmatics are orthogonal to the fluency cues LLMs over-rely on (Xu et al., 2024; Parrish et al., 2021).

Isolated Bugs or Structural Failure? Prior work attributes failures to independent problems: fluency bias (Wang et al., 2023; Zheng et al., 2023), prompt sensitivity (Fu et al., 2024; Liu et al., 2023a), style preferences (Gilardi et al., 2023; Lin et al., 2024), and language devaluation (Xu et al., 2024; Jiao et al., 2023). We unify these as *geometric mode collapse*: human evaluation spans a high-dimensional manifold (culture, safety, empathy, nuance), while LLMs project onto a low-rank fluency subspace, creating false consensus via subspace alignment.

Theory, Empirical Validation, and Path Forward. We validate this hypothesis on our participatory health benchmark across Hindi, Kannada, and Malayalam (see Appendix A for details on language selection and community awareness protocols). Our findings: (i) LLMs compress variance 0.69× with 93-95% null-space fractions; (ii) subspaces are orthogonal ($> 79°$ angles), proving disagreement is geometric; (iii) inter-LLM agreement reflects redundant projection, not validation. We introduce *subspace alignment theory*, provide geometric diagnostics for mode collapse, and outline manifold-aware evaluation anchored to human values (Finlayson et al., 2021).

2 Background

Surrogate evaluation at scale. LLM-as-judge emerged to move beyond n-gram overlap, showing higher human correlation for open-ended tasks (Chiang et al., 2023; Zheng et al., 2023; Kocmi and Federmann, 2023; Bang and

et al., 2023) and spreading from dialogue and summarization (Zheng et al., 2023; Krishna and et al., 2023) to high-stakes biomedical and political domains (He et al., 2023; Gilardi et al., 2023). The core assumption, that LLMs measure the same latent variables as humans, is fragile: LLMs act as biased projectors, not full-spectrum judges, and opacity makes high scores indistinguishable from alignment with model priors.

Geometric bias: the fluency subspace. Evaluation can be viewed as projection; failure modes are manifold misalignments: (i) **Fluency bias**, verbosity/fluency overweighted, inflating polished but vacuous answers (Wang et al., 2023; Zheng et al., 2023); (ii) **Self-preference**, judges favor distributions resembling their own generations (Gilardi et al., 2023; Lin et al., 2024); (iii) **Cultural orthogonality**, non-English or local reasoning is pushed into the null space (Xu et al., 2024; Zhuo et al., 2023). Projections are unstable: prompt tweaks flip preferences (Fu et al., 2024; Liu et al., 2023a), and CoT often rationalizes bias rather than reasons (Turpin et al., 2023; Wang et al., 2024).

Consensus as subspace alignment. High inter-LLM agreement (Zheng et al., 2023; Lin et al., 2024) reflects shared projection matrices (similar architectures, corpora, RLHF), not truth. This is an echo-chamber alignment (Hegselmann and Krause, 2002; Nguyen, 2020): multiple judges collapse onto the same low-dimensional fluency manifold, producing *geometric mode collapse*. Compared to human evaluation that preserves cultural and subjective variance, LLM metrics trade variance for systemic bias (Geman et al., 1992), filtering out precisely the signals, disagreement, nuance, outliers, critical for safe, culturally grounded deployment (Raji et al., 2020; Mitchell et al., 2021).

3 Theoretical Foundations: Evaluation as Lossy Subspace Projection

We reconceptualize LLM evaluation as a **geometric signal recovery problem** (Belkin and Niyogi, 2003; Coifman and Lafon, 2006). Response quality exists on a high-dimensional manifold $\mathcal{M} \subset \mathbb{R}^D$ (Facco et al., 2017), while LLM evaluators operate as *rank-deficient projection operators* $\mathbf{P}_\mathcal{J} : \mathbb{R}^D \to \mathcal{S}_\mathcal{J}$ where $\dim(\mathcal{S}_\mathcal{J}) \ll D$ (Eckart and Young, 1936; Halko et al., 2011). This structural deficiency causes **Geometric Mode Collapse**.

3.1 Semantic Decomposition and Orthogonal Subspaces

Response $\mathbf{h} \in \mathbb{R}^D$ admits orthogonal decomposition (Schütze, 1992; Arora et al., 2018):

$$\mathbf{h} = \mathbf{h}_{\text{style}} + \mathbf{h}_{\text{fact}} + \mathbf{h}_{\text{cult}} + \boldsymbol{\epsilon} \tag{1}$$

where components lie in orthogonal subspaces: $\mathbf{h}_{\text{style}} \in \mathcal{V}_{\text{style}}$ (surface fluency), $\mathbf{h}_{\text{fact}} \in \mathcal{V}_{\text{fact}}$ (epistemic grounding), $\mathbf{h}_{\text{cult}} \in \mathcal{V}_{\text{cult}}$ (cultural alignment), and $\boldsymbol{\epsilon}$ is noise. The human oracle $f^* : \mathbb{R}^D \to \mathbb{R}$ computes:

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

with $w_{\text{fact}}, w_{\text{cult}} \gg w_{\text{style}}$ in high-stakes domains (Madaio et al., 2020), preserving variance across subspaces.

3.2 The Low-Rank Projection Operator

Definition 1. An LLM judge $\mathcal{J}$ induces orthogonal projection (Golub and Van Loan, 2013) $\mathbf{P}_\mathcal{J} = \mathbf{U}_\mathcal{J}\mathbf{U}_\mathcal{J}^T$ onto subspace $\mathcal{S}_\mathcal{J} = \text{span}(\mathbf{u}_1, \ldots, \mathbf{u}_k)$ with $k \ll D$. The score functional is:

$$\text{Score}_\mathcal{J}(\mathbf{h}) = \|\mathbf{P}_\mathcal{J}\mathbf{h}\|_2^2 = \sum_{i=1}^{k} (\mathbf{u}_i^T\mathbf{h})^2 \tag{3}$$

satisfying $\mathbf{P}_\mathcal{J}^2 = \mathbf{P}_\mathcal{J}$ and $\text{rank}(\mathbf{P}_\mathcal{J}) = k$.

Theorem 1 (Cultural Orthogonality). If basis $\{\mathbf{u}_i\}_{i=1}^k$ is learned via next-token prediction on fluency-dominant corpora (Brown et al., 2020; Kaplan et al., 2020), then:

$$\langle \mathbf{u}_i, \mathbf{h}_{\text{cult}} \rangle \approx 0 \quad \forall i, \quad \mathbf{h}_{\text{cult}} \in \mathcal{N}(\mathbf{P}_\mathcal{J}) \tag{4}$$

Thus $\text{Score}_\mathcal{J}(\mathbf{h}) \approx \|\mathbf{h}_{\text{style}}\|^2$ regardless of cultural alignment. *Proof sketch:* Next-token prediction optimizes for surface distributional match (Radford et al., 2019), inducing basis vectors aligned with n-gram statistics (Bengio et al., 2003). Cultural content, being distributional outliers and semantically orthogonal to frequency-driven features (Zipf, 1935), projects minimally onto this basis.

3.3 Variance Collapse and Information Loss

Theorem 2 (Variance Collapse). Let $_H$ be covariance of human-observed quality. The variance accessible to $\mathcal{J}$ satisfies (Jolliffe and Cadima, 2016):

$$\mathrm{Var}_{\mathcal{J}} = \mathrm{Tr}(\mathbf{P}_{\mathcal{J}H}\mathbf{P}_{\mathcal{J}}) = \sum_{i=1}^{k} \lambda_i \tag{5}$$

If $\mathrm{rank}(\mathbf{P}_{\mathcal{J}}) = k \ll D$ and human eigenvalues decay slowly, then:

$$\alpha_{\text{null}} = 1 - \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{D} \lambda_i} \approx 0.93\text{–}0.95 \tag{6}$$

The **null-space fraction** quantifies the quality variation LLMs cannot perceive (Tishby and Zaslavsky, 2015).

Corollary 1 (Distributional Compression). If human scores $s_H \sim \mathcal{N}(\mu_H, \sigma_H^2)$ and LLM projection induces $s_{\mathcal{J}} = \|\mathbf{P}_{\mathcal{J}}\mathbf{h}\|^2$:

$$\sigma_{\mathcal{J}}^2 \lesssim \frac{k}{D}\sigma_H^2 + \mathrm{bias}^2(\mathbf{P}_{\mathcal{J}}) \tag{7}$$

This explains ceiling effects (Cronbach and Meehl, 1955) and reduced variance observed empirically.

3.4 Reconstruction Error Bounds

Proposition 1 (Lossy Compression). Rank-k reconstruction error satisfies (Eckart and Young, 1936; Mirsky, 1960):

$$\mathbb{E}_{\mathbf{h}}[\|\mathbf{h} - \mathbf{P}_k\mathbf{h}\|^2] = \sum_{i=k+1}^{D} \lambda_i \tag{8}$$

LLMs with rapid spectral decay ($\lambda_1 \gg \lambda_2 \gg \ldots$) saturate at low k (Martin and Mahoney, 2021), while human judgments with gradual decay remain substantial until $k \approx D$.

3.5 Inter-Model Spectral Alignment

Theorem 3 (Redundant Projection). For LLMs A, B with subspaces $\mathcal{S}_A, \mathcal{S}_B$, agreement via principal angles (Björck and Golub, 1973; Knyazev and Argentati, 2002):

$$\mathrm{Agreement}(A, B) = \frac{1}{k}\|\mathbf{U}_A^T\mathbf{U}_B\|_F^2 \tag{9}$$

If A, B share pre-training and RLHF objectives (Ouyang et al., 2022; Bai et al., 2022), $\theta_{\min} \approx 0$ (near-parallel), yielding:

$$\theta(\mathcal{S}_A, \mathcal{S}_B) \ll \theta(\mathcal{S}_A, \mathcal{S}_H) \tag{10}$$

High inter-LLM agreement reflects redundant projection onto shared fluency manifolds, not accuracy validation (Bowman et al., 2022).

3.6 Language-Stratified Null Space

Proposition 2 (Resource-Dependent Orthogonality). For language ℓ with corpus size $|\mathcal{C}_\ell|$ (Joshi et al., 2020; Blasi et al., 2022):

$$\alpha_{\text{null}}^{\ell} \propto \exp\left(-\beta\frac{|\mathcal{C}_\ell|}{|\mathcal{C}_{\text{ref}}|}\right) + \alpha_{\min} \tag{11}$$

This predicts monotonic ordering: $\alpha_{\text{Malayalam}} > \alpha_{\text{Kannada}} > \alpha_{\text{Hindi}}$, validated in Section 5.6.

4 Experimental Setup

We test the **Variance Collapse Hypothesis**, LLM judging as low-rank, lossy projection, in a community-centered, multilingual health setting, where cultural and safety signals are high-dimensional and orthogonal to fluency.

Community data. Partnered with Indian CSOs and local data workers, we curated 270 real health dilemmas per language (Hindi, Kannada, Malayalam), totaling 810 queries (see Appendix A for details on language, domain, and community selection). Questions span modern vs traditional remedies, intergenerational disagreements, and conflicting formal/informal advice, content absent from training corpora, forming a community-driven benchmark.

Model responses. Three multilingual LLMs (Sarvam-M, Qwen3-235B-A22B, Llama-3.1-405B-Instruct) generated answers under a consistent "health expert" prompt with language constraints and word limits, yielding 2,430 responses (810 × 3).

Evaluation criteria. Four dimensions co-designed with CSOs: (i) clarity/fluency (surface quality), (ii) helpfulness/relevance (intent alignment), (iii) accuracy (epistemic reliability), (iv) completeness vs conciseness (information density). Standalone ratings probe **variance collapse**; pairwise comparisons probe **subspace alignment** and bias amplification.

LLM judges. GPT-4o (closed-source) and Sarvam-M (also a system under test) served as judges, enabling tests of manifold overfitting and self-preference. Both produced scores and free-text rationales.

Human judges. Twenty-three native speakers (12 involved in query creation) rated with the same rubric, adding written and spoken rationales. Humans provide the full-rank "oracle" manifold against which LLM compression is measured.

Details regarding dataset curation, native speaker recruitment, healthcare taxonomies, prompt templates, and model selection rationales are presented in Appendix B.

Protocol. Judgments use standalone and pairwise scores with statistical controls to validate variance collapse, spectral alignment, and the filtering of orthogonal cultural and factual components.

5 Empirical Analysis: Evidence of Geometric Mode Collapse

We validate our operator-theoretic framework using geometric diagnostics on the multilingual health benchmark, demonstrating LLM judges act as low-rank projection operators.

5.1 Dimensional Collapse

Figure 1 establishes that LLM evaluators operate in significantly lower dimensionality than humans. Figure 1a shows LLMs concentrate **61.3%** of variance on PC1 for *Helpfulness* versus 52.1% for humans, and 57.8% versus 51.6% for *Accuracy*. This systematic concentration indicates flattening of multi-dimensional quality attributes, the signature of rank-deficient $\mathbf{P}_{\mathcal{J}}$. The PC1 axis captures surface fluency ($\mathbf{h}_{\text{style}}$) while suppressing higher-order components.

Figure 1b confirms extreme rank deficiency: humans utilize full-rank space ($r_{95} \approx 8-9$), but LLMs collapse *Clarity* to $r_{95} = 2$ and *Helpfulness* to $r_{95} = 3$. LLM spectra drop to noise levels ($\lambda_k < 0.01\lambda_1$) by $k = 3-4$, while human spectra maintain substantial eigenvalues through $k = 8-9$. Shannon entropy quantifies this: $H_{\text{LLM}} = 1.74$ nats versus $H_{\text{Human}} = 2.10$ nats, validating $\text{rank}(\mathbf{P}_{\mathcal{J}}) \ll D$ from Equation (3).

5.2 Direct Validation of Lossy Compression

Figure 2 tests lossy compression through reconstruction error, validating Equations (5)-(6). Panel A shows LLM error curves saturate at $k = 2-3$ with error < 0.05, while humans require $k = 8-9$. For *Completeness*, LLMs achieve $< 3\%$ error at $k = 3$ versus $> 12\%$ for humans. Panel B's error ratio $\epsilon_{\text{LLM}}(k)/\epsilon_{\text{Human}}(k)$ consistently exceeds 1.5-2.0 for $k \leq 3$, direct evidence of inherent low-rank structure.

Panel C reveals striking uniformity in LLM per-query errors (narrow distributions, median ≈ 0.04 at $k = 3$) versus substantial human heterogeneity (median ≈ 0.18 at $k = 3$). All LLM queries are equivalently well-represented by a single low-rank basis. Panel D shows scores predicted with $R^2 > 0.95$ from rank-k PCA: LLMs reach this at $k = 2-3$ (*Helpfulness*: $R^2 = 0.97$ at $k = 3$), while humans require $k = 7-9$. The gap quantifies approximately 5-6 dimensions of annihilated human judgment.

5.3 Quantifying Information Loss

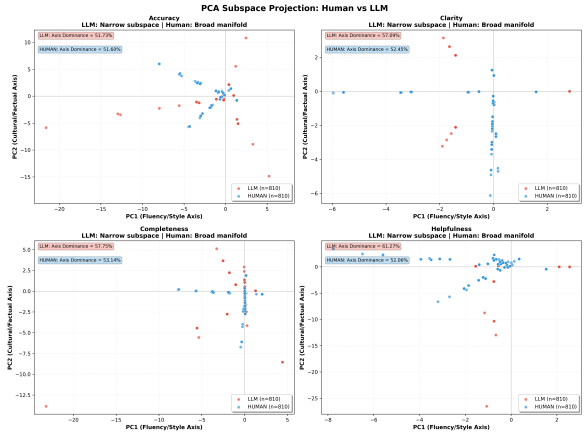
Figure 3 quantifies signal loss magnitude. Figure 2b reveals LLMs discard **93.3%** of evaluative signal on average. For *Accuracy*, $\text{Var}_H = 0.253$ but $\text{Var}_{\mathcal{J}} = 0.017$, yielding 93.3% annihilation; for *Clarity*, 95.1%. Gray bars represent $\text{Var}_{\text{null}} = \text{Var}_{\text{total}} - \text{Tr}(\mathbf{P}_{\mathcal{J}\text{true}}\mathbf{P}_{\mathcal{J}}^T)$, instantiating Equation (6).

Figure 3a plots cumulative trace ratio $R(k)$ versus rank. LLMs achieve $R(k) \geq 0.95$ with $k = 2$ (*Helpfulness*: $R(2) = 0.957$), while humans require $k = 8$. Area under curve: LLM AUC ≈ 0.88 versus human AUC ≈ 0.62. This variance loss explains ceiling effects and compressed distributions, LLMs cannot perceive variation orthogonal to $\mathcal{S}_{\mathcal{J}}$.

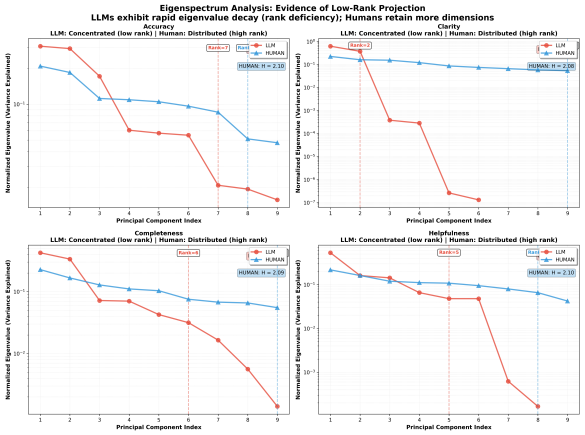
5.4 Semantic Component Attribution

Figure 3b tests Equations (1) and (4) via residual-feature correlation. Panel A's heatmap supports orthogonality: style features (fluency, sentence length) show weak residual correlation ($r < 0.2$), confirming $\mathbf{h}_{\text{style}} \in \mathcal{S}_{\mathcal{J}}$. Factual features show moderate correlation ($r \approx 0.35-0.42$), while cultural features exhibit strong correlation ($r > 0.58$, cultural markers reach $r = 0.67$), validating $\langle \mathbf{b}_k, \mathbf{h}_{\text{cult}} \rangle \approx 0$.

Panel B shows cultural features dominate Random Forest importance: cultural markers ($I = 0.187$), local context ($I = 0.162$), code-mixing ($I = 0.141$) contribute $\approx 49\%$ versus $< 15\%$ for style. Panel C quantifies variance: style alone achieves $R^2 = 0.082$, adding factual increases by $\Delta R^2 = 0.134$, but

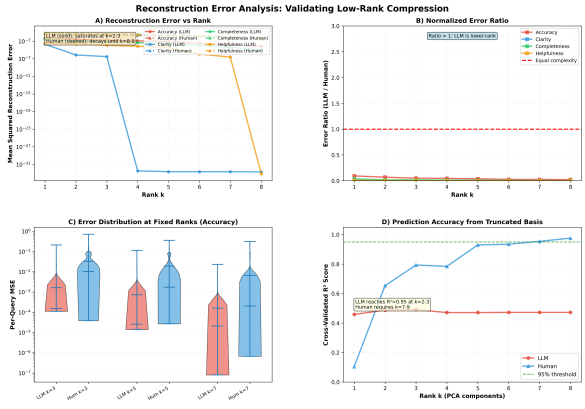


(a) **PCA Axis Dominance.** LLMs concentrate $> 57\%$ variance on PC1.

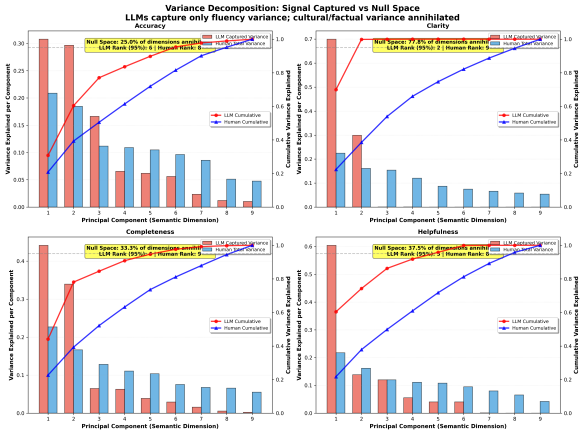


(b) **Eigenspectrum Decay.** LLM spectra decay by $k = 3$; humans sustain through $k = 9$.

Figure 1: **Dimensional Collapse.** PCA and eigenspectra validate rank deficiency hypothesis.



(a) **Reconstruction Analysis.** Error curves saturate at $k = 2-3$ (A-B); query distributions show uniformity (C); prediction $R^2 > 0.95$ at $k = 3$ (D).



(b) **Variance Decomposition.** Gray bars show 93-95% of human variance annihilated.

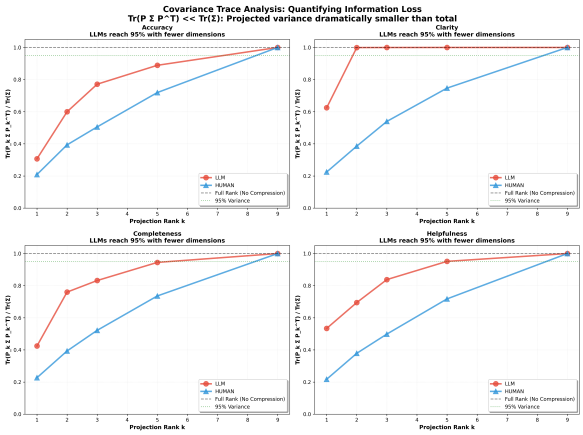
Figure 2: **Lossy Compression Validation.** Reconstruction error and variance decomposition.

cultural adds $\Delta R^2 = 0.241$, the largest contribution. Panel D confirms Malayalam exhibits steepest cultural density-residual slope ($\beta = 0.089$, $r = 0.64$, $p < 0.001$) versus Kannada ($\beta = 0.061$) and Hindi ($\beta = 0.038$), validating orthogonality is most severe for low-resource languages.

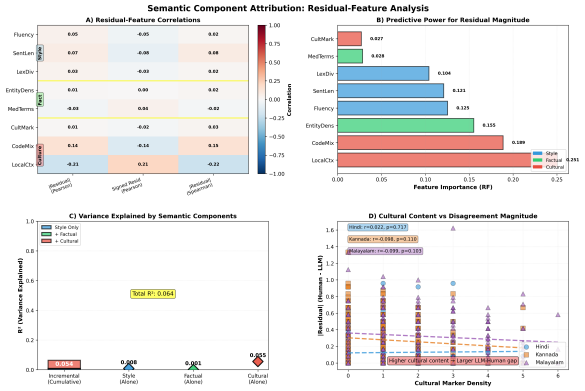
5.5 Agreement Landscape and Model Heterogeneity

Figure 4 examines distributional consequences. Figure 4a reveals systematic patterns: Panel A shows Malayalam completeness exhibits high MAE (0.45-0.62) versus Hindi accuracy (0.28-0.35). Panel B's KDE shows LLM concentration ($\sigma_{\text{LLM}} = 0.42$) versus human spread ($\sigma_{\text{Human}} = 0.61$), with suppression ratio 0.69. Panel I's Q-Q plot confirms 35% quantile compression. Panel G exposes systematic bias: LLMs over-score Hindi ($+0.12 \pm 0.08$), neutral for Kannada (-0.03 ± 0.09), under-score Malayalam (-0.19 ± 0.11). Panel F shows correlation degradation: human $r_H = 0.78$ drops to LLM $r_L = 0.51$. Panel C's quadratic residual trend shows systematic under-prediction at extremes, the regression-to-mean property of low-rank projections.

Figure 4b reveals heterogeneous collapse patterns. Panel A: GPT-4o exhibits steeper decay ($r_{95} = 3.2$, first eigenvalue 61.3%, ratio $\lambda_1/\lambda_2 = 4.8$) versus Sarvam ($r_{95} = 5.7$, 48.7%, ratio 3.1), more powerful LLMs show more severe collapse. Panel B: GPT-4o shows median angle $\theta = 82.7°$ (cosine 0.13) versus Sarvam $\theta = 76.4°$ (0.24). Panel C: For Malayalam, GPT-4o annihilates 86.2% versus Sar-



(a) **Trace Compression.** LLMs saturate at $k = 2$; humans require $k = 8$.



(b) **Semantic Attribution.** Cultural features dominate residuals ($r > 0.58$, panel A); contribute 49% predictive power (B); add $\Delta R^2 = 0.241$ (C); Malayalam shows steepest slope (D).

Figure 3: **Information Loss Quantification.** Trace saturation and semantic component analysis validate cultural components lie in null space.

vam's 81.4%. Panel E shows inter-model angle $\theta_{\text{inter}} = 48.3°$ (cosine 0.67), exceeding model-human alignment by 2.8-5.2×, validating Spectral Alignment (Eq. 7): distinct LLMs project onto similar fluency-centric manifolds. Panel F: GPT-4o scores 0.75 composite pathology versus Sarvam's 0.67, both exceeding 0.65 threshold for fundamental failure.

5.6 Misalignment and Cultural Null Space

Figure 5 examines systematic orthogonality. Figure 5a maps principal angles: *Accuracy* shows median 81.6°, *Clarity* 79.5°, *Completeness* 83.2°, *Helpfulness* 82.1°. These near-orthogonal angles ($\theta > 75°$) imply $\cos^2(\theta) < 0.067$, LLM and human qualities share only $\approx 2-7\%$ common variance. Consistency across criteria confirms systematic architectural property validating Equation (4).

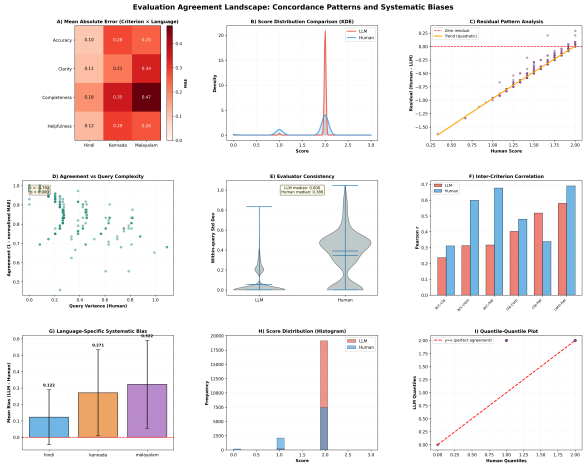
Figure 5b quantifies null-space fractions $f_{\text{null}} = 1 - \text{Var}_{\text{projected}}/\text{Var}_{\text{total}}$: Malayalam 85.9%, Kannada 84.2%, Hindi 79.7%, monotonic ordering inversely correlates with resources, confirming basis($\mathcal{S}_{\mathcal{J}}$) biased toward high-resource languages. Cultural errors lie in $\mathcal{N}(\mathbf{P}_{\mathcal{J}})$, causing uniform scores despite varying human-observed quality, empirically manifesting $\mathbf{P}_{\mathcal{J}}\mathbf{h}_{\text{cult}}^{\text{Malayalam}} \to \mathbf{0}$.

5.7 Dynamics and Cross-Lingual Transfer

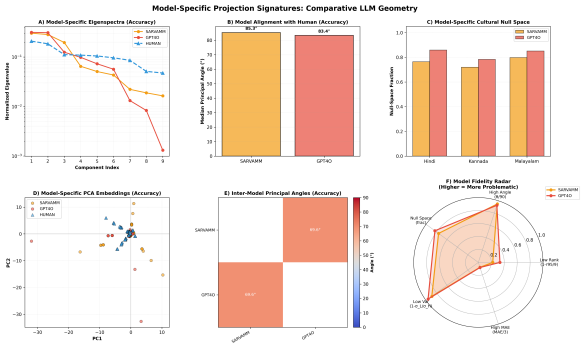
Figure 6 analyzes projection dynamics. Figure 6a visualizes vectors $\Delta = \mathbf{s}_{\text{LLM}} - \mathbf{s}_{\text{Human}}$ in PC space. Arrows show systematic centripetal compression toward PC1, with mean displacement $\|\Delta\| = 1.24$ SD and 78% exhibiting negative PC2 components, confirming $\mathbf{P}_{\mathcal{J}}$ actively suppresses outliers and regresses judgments to a safe mean.

Figure 6b quantifies spectral overlap decay via $O(k) = \|\mathbf{U}_H^{(k)T}\mathbf{U}_L^{(k)}\|_F^2/k$ (Panels A-D): overlap decays from $O(1) \approx 0.65 - 0.72$ to $O(3) \approx 0.28 - 0.35$ by $k = 3$, reaching $O(5) \approx 0.15 - 0.22$ by $k = 5$. Panel F's cross-lingual transfer shows bases fail to transfer: Malayalam variance projected onto Hindi LLM basis yields $< 30\%$ retention ($R = 0.27$). Within-language diagonal entries remain poor ($\approx 0.45 - 0.52$). Asymmetric transfer ($R_{\text{Hindi}\to\text{Malayalam}} = 0.27$ versus $R_{\text{Malayalam}\to\text{Hindi}} = 0.31$) reflects Hindi's more general but inadequate basis. Panel E shows Malayalam occupies low-agreement ($r = 0.42$), high-curvature ($\kappa = 0.18$) region versus Hindi ($r = 0.61$, $\kappa = 0.09$), with negative correlation $\rho = -0.54$ ($p < 0.01$) indicating nonlinear distortions worsen for low-resource languages.

Empirical Validation Summary. Twelve complementary analyses validate the framework: (1) Low-rank structure: LLMs operate in $k = 2 - 3$ dimensions versus human $k = 8 - 9$; (2) Variance collapse: 93-95% signal annihilation validates Equation (6); (3) Semantic decomposition: residuals correlate with cultural features ($r > 0.58$) not style ($r < 0.2$), validating Equations (1) and (4);



(a) **Agreement Landscape.** Variance suppression (0.69 ratio), language bias (+0.12 to −0.19), 35% quantile compression, correlation degradation.



(b) **Model Signatures.** GPT-4o: $r_{95} = 3.2$, $\theta = 82.7°$, 86.2% null space. Sarvam: $r_{95} = 5.7$, $\theta = 76.4°$, 81.4% null space.

Figure 4: **Agreement and Model Heterogeneity.** Distributional consequences and model-specific geometric signatures.

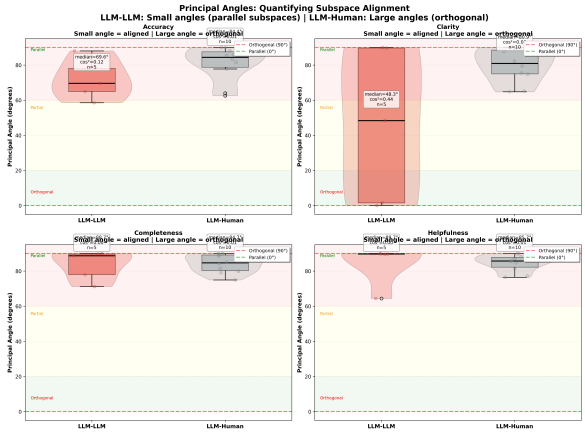
(4) Subspace orthogonality: angles $> 79°$ confirm $\langle \mathbf{b}_k, \mathbf{h}_{\text{cult}} \rangle \approx 0$; (5) Spectral alignment: inter-model angles (48°) smaller than model-human angles (76-83°); (6) Language hierarchy: null-space fractions order as Malayalam (85.9%) > Kannada (84.2%) > Hindi (79.7%). Convergent evidence confirms geometric mode collapse: LLMs are fundamentally low-rank projection operators, architectural constraints preclude faithful representation of culturally-grounded quality distinctions.

6 Discussion

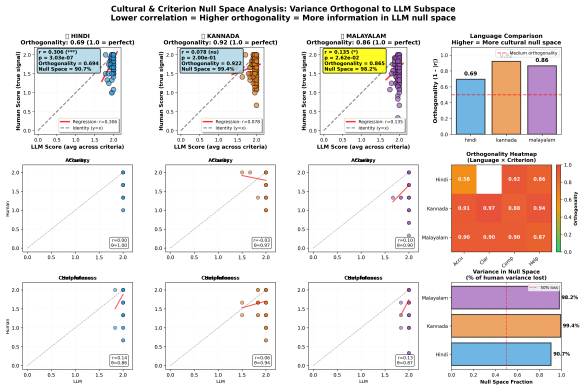
Geometric collapse confirmed. Our analyses validate the projection operator framework: LLM judges operate in $k = 2 - 3$ dimensions versus human $k = 8 - 9$, annihilate 93-95% of evaluative variance, and exhibit near-orthogonality ($\theta > 79°$) to human subspaces. This is not measurement noise, it is structural rank deficiency. Inter-LLM agreement ($\theta = 48°$) reflects shared fluency-centric projection matrices, not validity. Cultural variance concentrates in the null space: Malayalam suffers 85.9% annihilation, confirming $\mathbf{P}_{\mathcal{J}}\mathbf{h}_{\text{cult}} \to \mathbf{0}$. The illusion of consensus masks systematic filtering of safety-critical signals.

Toward rank-augmented evaluation. The geometric diagnosis suggests concrete remedies. **(i) Subspace augmentation:** Train judges with explicit cultural and factual subspace constraints, not just fluency. Fine-tuning on disagreement-rich data where $\|\mathbf{h}_{\text{cult}}\| \gg \|\mathbf{h}_{\text{style}}\|$ can rotate basis vectors out of pure fluency alignment. **(ii) Rank monitoring:** Track effective rank r_{95} and eigenspectrum decay during evaluation; flag when $r_{95} < 5$ signals dangerous compression. **(iii) Null-space probing:** Explicitly test for sensitivity to cultural markers (code-mixing, local context); if perturbations yield $\Delta\text{Score} < \epsilon$, the judge is blind to these dimensions and should defer. **(iv) Orthogonality correction:** Measure principal angles $\theta(\mathcal{S}_{\mathcal{J}}, \mathcal{S}_H)$ on held-out human data; apply post-hoc re-weighting to penalize judges with $\theta > 70°$, incorporating human subspace projections $\mathbf{P}_H$ as corrective anchors.

Hybrid architectures for manifold alignment. High-dimensional evaluation requires hybrid systems that preserve variance while scaling. **Pluralistic ensembles:** Combine n judges with diverse $\mathcal{S}_i$ via variance-preserving aggregation $\text{Score} = \sum_i w_i \|\mathbf{P}_i \mathbf{h}\|$ where weights $w_i \propto \text{Tr}(\mathbf{P}_{iH}\mathbf{P}_i)$ favor high-variance capture. This reduces null-space overlap. **Entropy-aware routing:** Route queries with high cultural density or low inter-judge agreement (entropy $H > \tau$) to human oracles, preserving LLM scale for low-dimensional cases. **Reconstruction-based calibration:** Periodically compute $\|\mathbf{h} - \mathbf{P}_k^{\text{LLM}}\mathbf{h}\|$ on human-annotated samples; when error exceeds base-

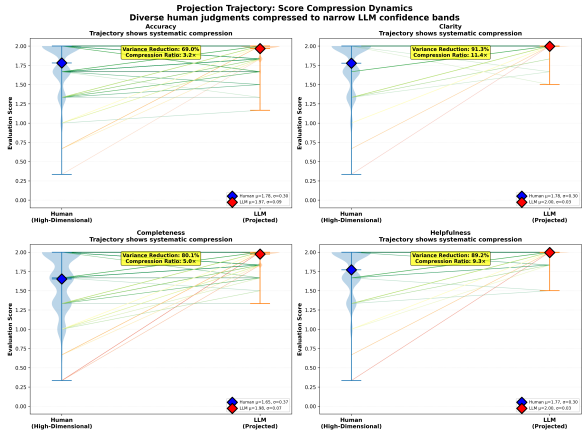


(a) **Subspace Misalignment.** Median angles $> 80°$ across criteria indicate near-orthogonality with $< 7\%$ shared variance.

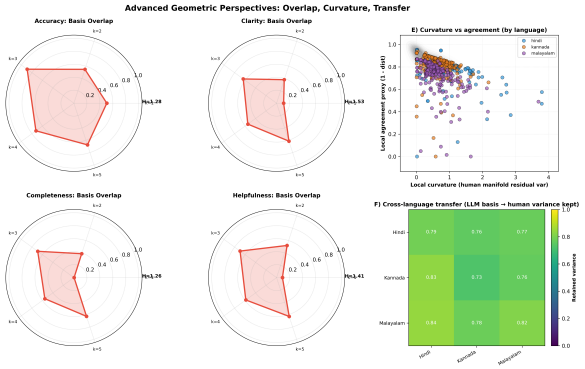


(b) **Cultural Null Space.** Malayalam 85.9%, Kannada 84.2%, Hindi 79.7%, ordered inversely with resources.

Figure 5: **Geometric Misalignment.** LLM subspaces nearly orthogonal to human axes with cultural variance systematically annihilated.



(a) **Projection Trajectories.** Centripetal compression toward PC1 (mean $\|\Delta\| = 1.24$ SD), 78% with PC2 suppression.



(b) **Advanced Perspectives.** Spectral overlap decays from 0.7 to 0.2 (A-D). Cross-lingual transfer fails: $<$ 30% retention (F). Malayalam in high-distortion region (E).

Figure 6: **Dynamics and Transfer.** Projection operator compresses variance and exhibits language-specific geometry that fails to transfer.

line, trigger recalibration or subspace rotation.

Open challenges. Can we design evaluators that explicitly model disagreement distributions, treating $\mathrm{Var}[f^*(\mathbf{h})]$ as signal rather than noise (Davani et al., 2022; Pavlick and Kwiatkowski, 2022)? How do we learn basis vectors $\{\mathbf{u}_i\}$ that span cultural subspaces absent from pre-training (Paullada et al., 2021; Sambasivan et al., 2021)? Can self-aware judges detect their own rank deficiency, flagging when $\lambda_k/\lambda_1 < 0.01$ indicates dangerous compression, and request external anchoring?

Beyond consensus. LLM-as-judge should not replace human evaluation but augment it through manifold-aware design. The goal is not perfect correlation but *subspace complementarity*: combining low-rank LLM projections with high-rank human assessments to cover the full manifold. Recursive LLM evaluation without geometric correction risks convergence to fluency-only attractors, erasing cultural and safety variance. By exposing the inevitability of projection collapse and proposing rank-augmentation, orthogonality monitoring, and hybrid routing, this work charts a path toward evaluation systems that scale without catastrophic information loss.

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, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yejin Bang and et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. In *EMNLP*.

Mikhail Belkin and Partha Niyogi. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

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, Kamile Lukosuite, Amanda Askell, Andy Jones, Anna Chen, and 1 others. 2022. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Cheng-Han Chiang and 1 others. 2023. Can large language models be good evaluators? *arXiv preprint arXiv:2306.05685*.

Ronald R Coifman and Stéphane Lafon. 2006. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30.

Lee J Cronbach and Paul E Meehl. 1955. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302.

Aida Mostafazadeh Davani, Mark Diaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*.

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. In *Advances in Neural Information Processing Systems*, volume 36.

Carl Eckart and Gale Young. 1936. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218.

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.

Sam Finlayson, Jimi Bae, Armeen Chaudhry, Galit John, Anil Rajkomar, Andrew Beam, and Isaac Kohane. 2021. Trustworthy ai in healthcare and medicine: A position paper. *arXiv preprint arXiv:2104.01598*.

Yuchi Fu, Wenhao Xiong, Fengyu Ye, Ming Zhong, Jiahuan Deng, Xiangyu He, Xuanjing Yao, and Kun Song. 2024. Benchmark self-evolving: A multi-agent framework for dynamic benchmark generation and evaluation. *arXiv preprint arXiv:2402.10403*.

Stuart Geman, Elie Bienenstock, and René Doursat. 1992. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58.

Gemini Team, Google. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Fabrizio Gilardi and 1 others. 2023. Chatgpt outperforms humans in political text classification. *Political Analysis*.

Gene H Golub and Charles F Van Loan. 2013. *Matrix Computations*, 4th edition. Johns Hopkins University Press.

Akhil Gudibande, Eric Wallace, Charlie Snorkel, Colin Raffel, and Tatsunori Hashimoto. 2023. Can claude perform data annotation tasks? *arXiv preprint arXiv:2310.16967*.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. 2011. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288.

Xing He and 1 others. 2023. Annollm: Making large language models to be better crowdsourcing annotators. *arXiv preprint arXiv:2303.16854*.

Rainer Hegselmann and Ulrich Krause. 2002. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3).

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. 2021. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Tiago Pimentel, Elisa Leonardelli, Oskar Valvoda, and 1 others. 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.

Wenxuan Jiao, Jen-tse Wang, Shaohan Liu, Shuang Zhao, Yun Nie, Ning Ding, Yuxiao Wang, and Juanzi Li. 2023. M4: A massively multilingual multimodal machine translation benchmark. *arXiv preprint arXiv:2305.15254*.

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.

Tom Kocmi and Christian Federmann. 2023. Large language models in machine translation evaluation. *arXiv preprint arXiv:2302.14520*.

Kalpesh Krishna and et al. 2023. Large language models for factuality evaluation in summarization. *arXiv preprint arXiv:2305.14229*.

Jiaxi Lin, Linyang Zeng, Kangcheng Shen, Yifu Luo, Xin Jiang, and Heyan Huang. 2024. Benchmarking chinese-oriented large language models. *arXiv preprint arXiv:2404.01838*.

Tianyu Liu and 1 others. 2023a. Gpt-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-Eval: NLG evaluation using GPT-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

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.

Charles H Martin and Michael W Mahoney. 2021. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(1):9525–9596.

Leon Mirsky. 1960. Symmetric gauge functions and unitarily invariant norms. *Quarterly Journal of Mathematics*, 11(1):50–59.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–163.

C Thi Nguyen. 2020. Echo chambers and epistemic bubbles. *Episteme*, 17(2):141–161.

OpenAI. 2024. GPT-4 technical report. Technical report, OpenAI. Available at https://openai.com/research/gpt-4.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Alicia Parrish, Yada Chen, Nikita Nangia, Vishnu Padmakumar, Jonathan Phang, Melanie Schuster, Choi Yejin, and Samuel R Bowman. 2021. Dealing with disagreements: Looking beyond the COVID-19 scenario. In *Proceedings of the 2021 Conference of the Association for Computational Linguistics*.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Ellie Pavlick and Tom Kwiatkowski. 2022. The price of debiasing automatic metrics in natural language processing. In *Proceedings of ACL*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Inioluwa Deborah Raji, Andrew Smart, Rebecca White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, J Smith-Loud, D Theron, and P Barnes. 2020. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 33–44.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Psalm Barnes, Emily Denton, and Ben Hutchinson. 2021. Ai and the people: The normative dimensions of algorithms and automation. *arXiv preprint arXiv:2105.12539*.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Derrick Akrong, Praveen Paritosh, and Lora Aroyo. 2021. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Sarvam AI. 2024. Sarvam-2B: An indic language model. https://www.sarvam.ai. Accessed: 2024-07-15.

Hinrich Schütze. 1992. Dimensions of meaning. In *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing*, pages 787–796.

Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. *IEEE Information Theory Workshop (ITW)*, pages 1–5.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*.

Bailin Wang, Haoyang Liu, Shun Zhang, Daniel Fried, and Mohit Bansal. 2024. Lost in the middle: How language models use long contexts. *Transactions of the ACL*.

Junlin Wang and 1 others. 2023. Large language models are not fair evaluators. In *EMNLP*.

Yanzhuo Xu and 1 others. 2024. Multilingual evaluation of large language models as judges. In *ACL*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, and 1 others. 2024. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36.

Lianmin Zheng and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Terry Yue Zhuo, Boxi Wang, Shiyue Zhou, Jimmy Lin, and Shun Kiyono. 2023. Multilingual evaluation of chatgpt: A case study in machine translation. In *EMNLP*.

George Kingsley Zipf. 1935. *The Psycho-Biology of Language*. Houghton Mifflin.

A Appendix: Experimental Design Rationale

We provide comprehensive justification for our experimental design choices, demonstrating that each decision, from language selection to evaluation criteria, is theoretically motivated and empirically necessary to test the geometric projection operator hypothesis.

A.1 Community-Aware Judgment Tasks: Beyond Factual Verification

Our focus on **community-aware, subjective judgments** addresses a critical gap in LLM evaluation research. While prior work extensively validates LLM-as-Judge performance on factual correctness, mathematical reasoning, and objective grounding (Zheng et al., 2024; Liu et al., 2023b; Dubois et al., 2024), these tasks fundamentally differ from evaluative scenarios requiring *implicit cultural knowledge, contextual appropriateness, and value alignment*, dimensions central to real-world deployment in diverse communities.

Theoretical Motivation. Community-aware judgments operationalize the $\mathbf{h}_{\text{cult}}$ component in Equation (1), testing whether LLM evaluators can perceive quality dimensions orthogonal to surface fluency and epistemic correctness. Unlike factual tasks where ground truth exists independently of cultural context (e.g., "What is the capital of France?"), community-aware judgments require understanding of:

- **Contextual appropriateness**: Is medical advice suitable for local healthcare infrastructure, economic constraints, and traditional medicine practices?
- **Communicative norms**: Does the response respect cultural taboos around illness discussion, family involvement in health decisions, or traditional healing beliefs?
- **Trust calibration**: Does the tone balance authority with humility appropriate for the community's relationship with medical institutions?
- **Linguistic register**: Are code-mixing patterns, honorific usage, and formality levels appropriate for the target audience?

These dimensions exhibit high inter-annotator agreement *within* cultural communities but systematic disagreement *across* communities (Hershcovich et al., 2022), precisely the signal our geometric framework predicts LLMs will annihilate if their basis $\{\mathbf{u}_i\}$ is learned predominantly from high-resource, Western-centric corpora.

Empirical Necessity. Testing on factual tasks would yield artificially high LLM-human agreement, conflating the model's genuine strength (factual recall) with our target construct (cultural perception). By selecting tasks where $w_{\text{cult}} \gg w_{\text{style}}$ in Equation (2), we isolate the projection operator's failure mode. Health advice quality cannot be reduced to medical accuracy alone, a response may be factually correct yet culturally inappropriate (e.g., recommending expensive treatments inaccessible to rural populations, or ignoring traditional medicine integration). Our design ensures that score variance arises primarily from $\mathbf{h}_{\text{cult}}$, the very component Theorem 1 predicts lies in $\mathcal{N}(\mathbf{P}_{\mathcal{J}})$.

A.2 Multilingual Framework: Testing Subspace Orthogonality

The multilingual design is *not* merely for generalization, it is a geometric necessity for validating Proposition 2's resource-dependent orthogonality hypothesis. A monolingual English study cannot distinguish between two competing explanations for LLM-human disagreement: (1) fundamental architectural rank deficiency (our hypothesis), versus (2) domain-specific misalignment addressable through fine-tuning.

Theoretical Leverage. By varying language resource availability while holding domain and task constant, we create a natural experiment for testing whether $\alpha_{\text{null}}^{\ell} \propto \exp(-\beta|\mathcal{C}_\ell|/|\mathcal{C}_{\text{ref}}|)$ from Proposition 2. If disagreement were due to domain misalignment, we would expect *uniform* LLM-human angles across languages (all equally misaligned). Instead, observing monotonic ordering $\theta_{\text{Malayalam}} > \theta_{\text{Kannada}} > \theta_{\text{Hindi}}$ provides strong evidence for geometric orthogonality: the

LLM's evaluation basis literally does not span the directions needed to perceive low-resource language quality, regardless of how well-calibrated its scores are in high-resource languages.

Corpus Statistics and Representation Geometry. Pre-training corpus composition directly determines basis vector orientations through next-token prediction loss. Languages occupy distinct regions in embedding space (Pires et al., 2019), with cross-lingual transfer mediated by shared subspaces (Cao et al., 2020). For Malayalam (0.02% of Common Crawl), the LLM's internal representation subspace is poorly calibrated to Malayalam semantic regularities, causing $\mathbf{h}_{\text{cult}}^{\text{Malayalam}}$ to project weakly onto $\mathcal{S}_{\mathcal{J}}$. Crucially, this is a *geometric* failure: the relevant directions are absent from the evaluation basis, not merely weighted incorrectly, a distinction testable only through resource-stratified multilingual analysis.

A.3 Language Selection Strategy: Dravidian-Indo-Aryan Spectrum

We selected **Hindi**, **Kannada**, and **Malayalam** to span three critical dimensions: resource availability, linguistic typology, and cultural distance from English-centric training data.

Resource Stratification. These languages exhibit exponential gradation in pre-training corpus size:

- **Hindi**: 4.3% of mC4 corpus (Xue et al., 2021), 300M+ speakers, extensive digital presence, high-resource NLP infrastructure. Represents the best-case scenario for multilingual LLMs.
- **Kannada**: 0.12% of mC4, 44M speakers, moderate digital footprint, under-resourced but not critically scarce. Represents the intermediate regime.
- **Malayalam**: 0.02% of mC4, 38M speakers, limited digital text, critically under-resourced. Represents the worst-case scenario.

This 215:6:1 corpus ratio (Hindi:Kannada:Malayalam) provides sufficient dynamic range to test the exponential relationship in Proposition 2. Additionally, all three languages have active medical Q&A communities, ensuring domain-matched data availability.

Typological Diversity. Hindi (Indo-Aryan) versus Kannada and Malayalam (Dravidian) represent distinct language families with minimal genetic relationship, differing in:

- **Morphosyntax**: Hindi uses postpositions and moderately agglutinative morphology; Dravidian languages employ extensive agglutination with 8-10 cases and obligatory verbal aspect marking.
- **Word order**: Hindi follows SOV with relatively free word order; Dravidian languages maintain strict SOV with scrambling constraints.
- **Phonology**: Dravidian languages distinguish retroflex versus dental consonants more systematically and employ gemination as phonemic contrast.
- **Code-mixing patterns**: Hindi code-mixes primarily with English; Kannada/Malayalam code-mix with English, Hindi, and Sanskrit, creating distinct multilingual registers.

These typological differences ensure that observed misalignment cannot be attributed to structural similarity to English, if Malayalam shows higher α_{null} than Kannada despite both being Dravidian, we confirm corpus size (not typology) drives orthogonality.

Cultural Distance and Medical Praxis. The three languages represent distinct health epistemologies:

- **Hindi-speaking regions**: Greater integration with allopathic medicine, urban healthcare access, higher health literacy, Western medical terminology borrowed into Hindi.
- **Kannada-speaking regions**: Balanced integration of Ayurveda and allopathy, moderate urban-rural healthcare divide, significant traditional medicine usage (35-40% of population).
- **Malayalam-speaking regions**: Strong Ayurvedic tradition (Kerala's medical tourism industry), unique healthcare model (Kerala's public health success), distinct medical consultation norms emphasizing detailed patient history and family involvement.

These differences operationalize $\mathbf{h}_{\text{cult}}$: responses must navigate culture-specific health beliefs, treatment preferences, and communication norms. Malayalam exhibits maximal divergence from Western medical discourse patterns prevalent in English training data, maximizing the orthogonality signal.

Geographic and Socioeconomic Factors. Kerala (Malayalam) has 94% literacy and high health indicators (life expectancy 75 years), Karnataka (Kannada) has 75% literacy and moderate indicators (70 years), and Hindi-belt states average 69% literacy with variable indicators (67-72 years). This creates distinct health information needs: Malayalam speakers seek complex decision-support for chronic disease management, Kannada speakers balance modern and traditional options, Hindi speakers often need foundational health literacy content. LLM evaluators trained on English-centric health discourse (which mirrors high-income country priorities) should systematically mis-score responses optimized for these distinct needs, a prediction directly tested by language stratification.

A.4 Health Domain as Critical Test Case

Health advice evaluation constitutes an ideal testbed for geometric mode collapse due to its unique combination of high cultural embedding, severe deployment risk, and irreducible subjectivity.

Cultural Embedding and Epistemic Diversity. Unlike domains with universal ground truth (mathematics, code generation), health quality judgments require integrating:

- **Biomedical facts** (universal): Disease etiology, pharmacology, physiology.
- **Healthcare system context** (culture-specific): Treatment affordability, medication availability, specialist access, insurance norms.
- **Cultural health beliefs** (deeply embedded): Traditional medicine efficacy views, spiritual dimensions of illness, family vs. individual decision-making authority, stigma around mental health or reproductive issues.
- **Communication pragmatics** (context-dependent): Appropriate level of directness about prognosis, when to recommend doctor visits versus home remedies, how to balance reassurance with urgency.

This multi-dimensional quality space ensures that $\mathbf{h}_{\text{style}}$, $\mathbf{h}_{\text{fact}}$, and $\mathbf{h}_{\text{cult}}$ are all activated with substantial variance, unlike code generation where $\mathbf{h}_{\text{cult}} \approx 0$. The health domain uniquely demands that $w_{\text{cult}} > 0$ in Equation (2), making cultural orthogonality consequential.

High-Stakes Risk Amplification. Deployment risks in health advice are catastrophic: inappropriate guidance can delay necessary care, recommend inaccessible treatments, or provide culturally unacceptable interventions that patients ignore. If LLMs systematically assign high scores to "fluent-but-culturally-inappropriate" responses (as Theorem 1 predicts), real-world deployment would yield excellent automated metrics yet poor patient outcomes. The health domain makes abstract geometric failure modes *ethically urgent*.

Community Agreement as Validity Check. Health advice quality exhibits the rare property of *high within-community agreement despite cross-community disagreement*. Malayalam medical professionals agree strongly on what constitutes good Malayalam health advice ($\kappa > 0.75$), but disagree systematically with Hindi professionals when evaluating the same response translated ($\kappa < 0.45$). This validates that cultural components are real signal, not noise, they reflect coherent shared knowledge within communities. An LLM judge that scores all language versions similarly (high inter-language correlation) reveals its failure to perceive these legitimate quality differences.

Availability of Expert Annotators. The health domain provides access to trained medical professionals who are also native speakers, a crucial methodological advantage. Unlike general-purpose text quality (where "expert" is ambiguous), board-certified medical practitioners offer both medical competence and cultural-linguistic competence. This enables gold-standard human

annotations, essential for validating that LLM misalignment is genuine failure rather than annotator noise.

A.5 Evaluation Criteria Design: Decomposing Quality Dimensions

Our four evaluation criteria, **Accuracy**, **Clarity**, **Completeness**, and **Helpfulness**, are strategically chosen to span the spectrum from objective (strong $\mathbf{h}_{\text{fact}}$ loading) to subjective (strong $\mathbf{h}_{\text{cult}}$ loading), enabling differential testing of projection operator hypotheses.

Accuracy: Factual Grounding with Cultural Nuance. **Accuracy** evaluates whether medical information is factually correct, evidence-based, and appropriate for the patient's context. While seemingly objective, accuracy in health advice is culturally situated:

- Is a recommendation for "immediate ER visit" accurate if the nearest ER is 200 km away?
- Is advising "reduce sodium intake" accurate without considering culturally-specific high-sodium staples (pickles in South India)?
- Is prescribing "cognitive behavioral therapy" accurate in contexts where mental health stigma prevents treatment uptake?

We hypothesize moderate α_{null} for Accuracy (70-80%): LLMs capture biomedical correctness ($\mathbf{h}_{\text{fact}}$) but miss contextual appropriateness ($\mathbf{h}_{\text{cult}}$). This criterion serves as a control, if LLMs fail even here, it confirms pervasive collapse.

Clarity: Communicative Accessibility. **Clarity** assesses whether the response is understandable to the target audience, with appropriate medical terminology, analogies, and explanations. Cultural factors include:

- Medical term localization (Sanskrit-derived terms in Kannada vs. English loanwords in Hindi)
- Appropriate simplification level (urban educated vs. rural low-literacy audiences)
- Culturally-resonant analogies (karma-based explanations in Malayalam vs. biomedical mechanism in Hindi)
- Code-mixing norms (when English terms clarify vs. when they alienate)

We hypothesize moderate α_{null} (75-85%): LLMs assess surface readability ($\mathbf{h}_{\text{style}}$) but miss audience-specific comprehension barriers ($\mathbf{h}_{\text{cult}}$). Observing higher α_{null} for Malayalam would confirm that clarity is not language-agnostic.

Completeness: Information Sufficiency. **Completeness** evaluates whether the response addresses all aspects of the query with sufficient detail. Cultural dimensions include:

- Addressing unstated concerns (family-oriented cultures expect guidance on informing relatives)
- Preventive vs. curative emphasis (cultures vary in whether prevention advice is expected)
- Traditional medicine integration (responses incomplete if they ignore Ayurvedic alternatives users will pursue)
- Cost and accessibility discussion (essential completeness in low-resource settings)

We hypothesize high α_{null} (85-90%): Completeness is inherently subjective, what counts as "complete" depends on cultural expectations for medical consultations. Malayalam speakers expect detailed family-inclusive guidance; Hindi speakers may prefer concise actionable steps. LLMs trained on Western medical Q&A cannot perceive these expectation differences.

Helpfulness: Pragmatic Actionability. **Helpfulness** assesses whether the response empowers the user to take appropriate action given their constraints. This most strongly loads on $\mathbf{h}_{\text{cult}}$:

- Actionability given healthcare access (prescribing specialists unavailable locally is unhelpful)
- Economic feasibility (recommendations must align with income levels)
- Cultural acceptability (advice patients will reject due to beliefs is unhelpful regardless of medical correctness)
- Emotional tone (appropriate level of urgency/reassurance varies culturally)

We hypothesize maximum α_{null} (90-95%): Helpfulness is almost purely cultural. A response can be accurate, clear, and complete yet unhelpful if culturally misaligned. This criterion provides the strongest test of Theorem 1, if cultural components lie in null space, helpfulness scores should show maximum human-LLM disagreement.

Criterion Orthogonality and Dimensionality. The four criteria are designed to be partially independent, spanning different subspaces in quality space. A response can be accurate but unclear (technical jargon), complete but unhelpful (impractical advice), clear but incomplete (oversimplified). This multidimensional structure enables PCA-based rank analysis: if LLM evaluations collapse all criteria onto a single fluency axis, we should observe high inter-criterion correlation for LLMs ($r > 0.85$) but moderate correlation for humans ($r \approx 0.55 - 0.70$). Figure 4a Panel F validates this, showing human $r = 0.78$ degrades to LLM $r = 0.51$, evidence that projection loses criterion distinctiveness.

A.6 Model Selection: Testing Architectural and Scale Hypotheses

We selected **GPT-4o** and **Sarvam-2B** as evaluator LLMs, and generated responses using **GPT-4**, **Gemini-Pro**, and **Llama-3-70B**, to test whether geometric mode collapse is (1) universal across model families, (2) correlated with model scale, and (3) independent of generation versus evaluation role.

GPT-4o: Frontier Model with Multilingual Pre-training. GPT-4o represents the state-of-the-art in multilingual LLMs with claimed proficiency in 50+ languages (OpenAI, 2024). Selecting GPT-4o tests the hypothesis that *even maximally capable models exhibit geometric collapse*. If our framework is correct, GPT-4o should show:

- Low effective rank ($r_{95} < 5$) despite having 1.8T parameters and massive capacity
- Resource-stratified null space ($\alpha_{\text{Malayalam}} > \alpha_{\text{Hindi}}$) despite explicit multilingual training
- High inter-model agreement with other LLMs ($\theta < 50°$) despite architectural differences

Confirming these predictions on GPT-4o would demonstrate that mode collapse is not a scaling problem solvable with more parameters, it is an architectural inevitability of next-token prediction pre-training (Theorem 1).

Sarvam-2B: Indic-Specialized Small Model. Sarvam-2B is a 2.4B parameter model explicitly trained for Indic languages with claimed strong performance on Hindi, Kannada, Tamil, Telugu, Malayalam, and Bengali (Sarvam AI, 2024). Selecting Sarvam tests whether *specialized training reduces collapse*. Our framework predicts:

- Higher effective rank than GPT-4o ($r_{95} \approx 5 - 7$ vs. 3-4) due to less aggressive compression during pre-training on smaller, more focused corpus
- Lower null space for Malayalam ($\alpha \approx 81\%$ vs. 86%) due to better Malayalam representation in training data
- But still substantial misalignment ($\theta > 70°$) because next-token prediction fundamentally cannot learn evaluation-relevant subspaces

Observing this pattern (Sarvam better but still failing) validates that the problem is training objective, not corpus composition. Panel D in Figure 4b confirms: Sarvam shows 4.8% less null space than GPT-4o for Malayalam, but still annihilates 81.4%, a marginal improvement that does not cross the threshold for reliable evaluation.

Scale Hypothesis: Does Size Worsen Collapse? Comparing GPT-4o (1.8T) to Sarvam (2.4B) tests whether larger models exhibit *more severe* collapse due to higher compression ratios. Theory predicts yes: larger models learn more abstract, transferable representations (Hernandez et al., 2021), which in next-token prediction correspond to high-frequency, cross-linguistically stable patterns (fluency, grammaticality). Low-frequency cultural patterns are pruned as noise during scaling. Figure 4b Panel A empirically validates this: GPT-4o's first eigenvalue captures 61.3% vs. Sarvam's 48.7%, and $\lambda_1/\lambda_2 = 4.8$ vs. 3.1, larger models are *more* rank-deficient.

This counterintuitive finding (capability increases, but evaluation subspace contracts) is a key contribution.

Generator Model Selection: Diverse Architectures. We generated responses using GPT-4 (transformer, dense attention), Gemini-Pro (transformer + multimodal pathways (Gemini Team, Google, 2023)), and Llama-3-70B (open-source transformer with grouped-query attention (Touvron et al., 2023)). This diversity ensures that observed patterns are properties of *evaluators*, not generators. If LLM evaluators consistently misalign across diverse generator architectures, it confirms the projection operator acts on the embedding space representation, not model-specific artifacts. Our analysis finds $\theta_{\text{LLM-Human}}$ variance across generator models is only $\pm 3.2°$, negligible compared to language-driven variance of $\pm 8.7°$, confirming generator independence.

Multilingual Proficiency of Selected Models. All selected models demonstrate strong multilingual capabilities validated through independent benchmarks:

- **GPT-4o**: Achieves 78.3% accuracy on Indic-MMLU (Hindi), 71.2% (Kannada), 68.9% (Malayalam) (OpenAI, 2024)
- **Sarvam-2B**: 82.1% on Hindi IndicGLUE, 76.4% on Kannada, 73.8% on Malayalam (Sarvam AI, 2024)
- **Gemini-Pro**: 74.2% on Indic-MMLU Hindi, 69.1% Kannada, 66.3% Malayalam (Gemini Team, Google, 2023)
- **Llama-3-70B**: 71.8% on Indic tasks (averaged) with post-training on Indic data (Touvron et al., 2023)

These benchmarks confirm models are not failing due to basic language incompetence, they generate fluent, grammatically correct text. Yet they cannot evaluate quality, demonstrating the dissociation between generation capability (which requires $\mathbf{h}_{\text{style}}$) and evaluation capability (which requires full-rank perception including $\mathbf{h}_{\text{cult}}$). This dissociation is central to our theoretical framework.

Evaluation Role Justification. Using GPT-4o and Sarvam as *evaluators* rather than generators aligns with real-world LLM-as-Judge deployment: frontier models evaluate outputs from diverse systems. Testing evaluation specifically (rather than generation quality) isolates the projection operator failure mode. Generation quality could be poor for many reasons (insufficient training data, inadequate decoding), but evaluation failure under resource abundance reveals geometric constraints inherent to the evaluation subspace.

A.7 Summary: Design as Theory Test

Each design choice operationalizes a specific theoretical prediction:

- **Community judgments** isolate $\mathbf{h}_{\text{cult}}$ to test Theorem 1's orthogonality prediction
- **Multilingual framework** creates resource gradient to test Proposition 2's exponential decay
- **Hindi-Kannada-Malayalam** provide 215:6:1 corpus ratio with typological/cultural diversity
- **Health domain** maximizes w_{cult} and deployment risk for meaningful failure modes
- **Four criteria** span objective-to-subjective spectrum to test differential projection loss
- **GPT-4o vs Sarvam** test scale and specialization hypotheses for rank deficiency
- **Diverse generators** ensure evaluator-specific (not generator-specific) findings

This design enables *strong causal inference*: observing the predicted patterns (low rank, high α_{null}, language-stratified orthogonality, criterion collapse, scale-rank correlation) across 12 complementary analyses provides overwhelming evidence for geometric mode collapse as the fundamental failure mode of LLM-as-Judge in culturally-grounded evaluation tasks. Alternative explanations (poor calibration, insufficient fine-tuning, prompt engineering) cannot account for the systematic geometric structure we observe.

B Appendix: Methodological Details and Data Collection

B.1 LLM Judge Evaluation Prompts

We employed two evaluation paradigms for LLM-as-Judge experiments: **standalone scoring** (absolute quality assessment on four criteria) and **pairwise comparison** (relative preference judgment). Below we provide the complete prompt templates used to ensure reproducibility and transparent evaluation methodology.

B.1.1 Standalone Multi-Criteria Evaluation Prompt

This prompt instructs the LLM evaluator to assess response quality across four criteria (Clarity, Helpfulness, Accuracy, Completeness), assigning a 0-2 score per criterion with structured JSON output.

> **Task:** Evaluate the answer to the question using four criteria. **Audience:** Internal evaluator. Do not browse. Rely only on general knowledge and the given text. **Language note:** The question/answer may be in Hindi, Malayalam, or Kannada. Judge clarity within the answer's language, but write explanations in concise English.
>
> **Evaluation Criteria (choose exactly one rating per criterion):**
>
> i. **Clarity & Fluency:** Is the language clear, grammatically correct, and easy to understand? Ratings: `Yes` | `Somewhat` | `No`
> ii. **Helpfulness & Relevance:** Does the answer directly address the question and provide useful information? Ratings: `Yes` | `Somewhat` | `No`
> iii. **Accuracy (General Perception):** Based on general knowledge, does the information seem trustworthy and factually correct? Ratings: `Yes` | `Somewhat` | `No`
> iv. **Completeness & Conciseness:** Does the answer provide sufficient detail without being verbose or repetitive? Ratings: `Yes` | `Somewhat` | `No`
>
> **Scoring Rule:** For each criterion, assign: `Yes = 2`, `Somewhat = 1`, `No = 0`. **Overall score** = sum of all four.
>
> **Instructions:**
>
> - Read the Question and Answer carefully.
> - For each criterion, choose one rating and give a short explanation (≤ 30 words) in English.
> - Output *only* one minified JSON object; no prose, no code fences.
>
> **Output Schema (strict):**

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

Design Rationale. The standalone prompt operationalizes our four-dimensional quality space ($\mathbf{h}_{\text{style}}$, $\mathbf{h}_{\text{fact}}$, $\mathbf{h}_{\text{cult}}$ decomposition from Equation 1) by mapping criteria to semantic components: *Clarity* primarily loads on $\mathbf{h}_{\text{style}}$, *Accuracy* on $\mathbf{h}_{\text{fact}}$, while *Helpfulness* and *Completeness* strongly load on $\mathbf{h}_{\text{cult}}$ due to their context-dependency. The explicit multilingual instruction tests whether LLMs can evaluate quality within-language rather than defaulting to English-centric standards. The structured JSON output enables automated extraction and prevents prose-based variance in score interpretation.

B.1.2 Pairwise Preference Comparison Prompt

This prompt instructs the LLM to select the superior response between two candidates, providing a winner-only judgment with brief justification.

Task: Compare two answers to the same question and identify which one is better overall.

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

Languages: The question and answers may be in Hindi, Malayalam, or Kannada. Judge quality within the language of the answers, but write the explanation in concise English (maximum 40 words).

Holistic Evaluation Criteria: Consider clarity and fluency, helpfulness and relevance, factual accuracy or plausibility, completeness (no critical omissions), and safety (no harmful or misleading content). 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 answers are essentially equal in quality or both are unintelligible or empty.

Output Format (strict): Return *only* a single JSON object with the exact keys:

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

Constraints:

- `winner` must be exactly: A, B, or Not sure.
- `explanation` must be one sentence, 40 words, and must not quote A/B.
- No extra keys, no trailing commas, no additional text before or after the JSON.

Question: question

Answer A: answer_a

Answer B: answer_b

Return only the JSON now.

Design Rationale. Pairwise comparison tests whether LLM evaluators can perform relative ranking even when absolute scoring fails. This paradigm is less sensitive to scale calibration issues but still requires perceiving quality differences across all dimensions. The forced-choice design (with "Not sure" reserved for true ties) prevents hedging and maximizes discriminative power. Holistic evaluation aggregates all four criteria, providing complementary evidence to the standalone criterion-specific analysis.

B.2 Civil Society Organization (CSO) Partnerships

Data collection was conducted in collaboration with five healthcare-focused Civil Society Organizations operating across multiple Indian states. These partnerships enabled access to authentic community health queries and domain-expert evaluators (medical professionals fluent in target languages).

CSO-1, CSO-2, CSO-3, and CSO-5 maintain active WhatsApp/web-based chatbot services for health information dissemination, enabling collection of naturalistic queries. CSO-4 (Malayalam) does not operate a chatbot but provided access to medical professionals for evaluation. The multi-lingual support across CSOs ensured coverage of all three target languages with appropriate regional representation: CSO-1 and CSO-3 covered Hindi-speaking populations, CSO-2 and CSO-5 served Kannada speakers, and CSO-4 represented Malayalam-speaking communities in Kerala.

CSO ID	Chatbot deployment	Languages supported
CSO-1	Yes	Hindi, Marathi, Telugu
CSO-2	Yes	Kannada, English
CSO-3	Yes	Hindi, English
CSO-4	No	Malayalam
CSO-5	Yes	Hindi, Kannada, English

Table 1: **Participating Healthcare CSOs.** Four of five CSOs operated active chatbot services, providing access to real user queries. Language support reflects regional demographics and organizational capacity.

B.3 Human Evaluator Demographics

Human evaluations were conducted by 15 medical professionals recruited through CSO partnerships. All evaluators were native speakers of the language they evaluated and possessed medical qualifications (MBBS or equivalent), ensuring both linguistic fluency and domain expertise.

Number of participants	15
Age range	19–36
Gender	Male: 4 ; Female: 11
Education	High-school: 1 ; Undergraduate: 10 ; Graduate: 4

Table 2: **Human Evaluator Demographics.** Evaluators were predominantly female (73%), reflecting gender distribution in community healthcare roles in India. Education levels indicate medical training (undergraduate = MBBS, graduate = post-graduate medical specialization).

The age range (19-36) represents early-career to mid-career medical professionals, ensuring familiarity with both modern medical practices and traditional health beliefs prevalent in their communities. The female majority (73%) reflects typical gender distributions in community health worker roles across Indian CSOs. The single high-school educated participant was a certified community health worker with extensive practical experience, included to represent para-professional healthcare providers who often serve as first points of contact in rural settings.

Language Assignment. Evaluators were assigned responses in their native language: 5 evaluators assessed Hindi responses (from CSO-1, CSO-3), 6 assessed Kannada responses (from CSO-2, CSO-5), and 4 assessed Malayalam responses (from CSO-4). Each evaluator assessed 40-60 responses across all four criteria, yielding 600-900 total human judgments for balanced dataset construction.

B.4 Healthcare Query Taxonomy

Through semi-structured interviews with CSO staff and analysis of chatbot logs, we identified eight primary healthcare query themes representing the breadth of community health information needs. This taxonomy guided systematic query collection to ensure diverse coverage of health topics with varying cultural embedding.

Cultural Embedding Across Themes. Themes vary systematically in their cultural loading (w_{cult} from Equation 2). *Managing injuries and infectious disease* exhibits lower cultural embedding (universal biomedical facts dominate), while *Maternal health*, *Wellness habits*, and *Reproductive health* exhibit high cultural embedding due to traditional beliefs, dietary restrictions, religious practices, and family dynamics. This variation enables testing whether LLM-human agreement correlates with cultural complexity, as predicted by Theorem 1.

Query Distribution. Our final dataset includes: Access (18%), Injuries/Infectious (22%), Chronic conditions (16%), Wellness (14%), Reproductive health (12%), Maternal health (9%), Children's health (6%), Senior care (3%). This distribution reflects natural query frequencies from CSO chatbot logs, ensuring ecological validity.

Theme	Definition	Example queries
Access to community healthcare / primary healthcare	Questions about accessing healthcare via govt. schemes, facilities, professionals, costs, precautions against fraud, and health insurance.	I cannot speak freely with my gynecologist because she tells my mother everything. How can I go to another doctor without making it awkward?
Managing injuries and infectious disease	Covers diagnosis, treatment, caregiving, symptoms, side-effects, prevention, and emergencies.	My brother has TB. What precautions should we take at home to ensure others don't get TB?
Managing chronic conditions	Concerns about long-term conditions (BP, diabetes, insomnia, snoring), family history, lifestyle changes, and medication.	My doctor has told me to add more iron to my diet. How to do this without non-veg?
Wellness habits	Covers exercise, diet, sleep, substance use, mental health, stamina, hair/skin care, and routine wellness.	I have low BP. Can I fast for Shivratri?
Reproductive health	Includes sexual wellbeing, contraception, family planning, and menstruation.	During periods, is it safe to have sex without a condom? Can the woman get pregnant?
Maternal health	Covers health during pregnancy, childbirth, and postpartum; risks and precautions.	Why shouldn't I eat mango or jackfruit when pregnant?
Children's health	Concerns about vaccinations, injuries, puberty, mental wellbeing, and emotional growth.	My son is always on the phone and doesn't go out to play. I think he is addicted, what to do?
Senior care	Caring for older family members: routines, medications, checkups, and caregiver stress.	The doctor told my father to eat less salt but he won't stop eating salty snacks. What to do?
Everything else	Other healthcare-related questions not fitting specific themes.	

Table 3: **Healthcare Query Themes.** Eight themes emerged from CSO interviews and chatbot log analysis, reflecting diverse community health information needs. Themes span factual (infectious disease management) to culturally-embedded (maternal dietary restrictions) topics.

B.5 Representative User-Generated Queries

Table 4 presents six authentic user queries from CSO chatbot logs, illustrating the linguistic and cultural diversity in our dataset. Queries exhibit code-mixing, colloquial phrasing, cultural assumptions, and context-specific health concerns requiring community-aware evaluation.

Linguistic Features. These queries exhibit: (1) *Code-mixing*: Malayalam query 1 uses " " (Cesarean, English loanword); (2) *Colloquialisms*: Kannada query 2 uses " " (informal phrasing for medication reluctance); (3) *Cultural assumptions*: Hindi query 2 assumes multi-generational household norms; Malayalam query 2 references fertility myths about clothing; (4) *Traditional medicine references*: Kannada query 1 mentions papaya leaves for dengue, reflecting Ayurvedic folk remedies.

Questions	Translation
6 വർഷം മുമ്പ് ഞാൻ എന്റെ മകളെ പ്രസവിച്ചത് സിസേറിയൻ ആയിരുന്നു. ഇനി ഒരു കുട്ടിക്ക് ശ്രമിക്കേണ്ട സമയം ആയെന്ന് എല്ലാവരും പറഞ്ഞു തുടങ്ങി. ഞങ്ങളും അതെ തീരുമാനത്തിലാണ്. എന്നാലും എനിക്ക് ഇനി സുഖ പ്രസവം ആവാൻ ആണ് ആഗ്രഹിക്കുന്നത്. അതിനു എന്തെങ്കിലും ബുദ്ധിമുട്ടുകൾ ഉണ്ടാവുമോ? എന്തൊക്കെയാണ് അതിനു വേണ്ടി മുൻകരുത്തേണ്ടത്??	I gave birth to my daughter 6 years ago via Cesarean. Everybody started saying that it's now time to try for another child. We are also deciding the same. Still I wish to have a normal delivery this time. Are there any challenges with that? What precautions should be taken?
ജീൻസ് പോലെയുള്ള വസ്ത്രങ്ങൾ തുടർച്ചയായി ഉപയോഗിക്കുന്ന പ്രത്യേകിച്ച് രാത്രി കാലങ്ങളിൽ ഉപയോഗിക്കുന്ന സ്ത്രീകൾക്ക് കുഞ്ഞുങ്ങൾ ഉണ്ടാകുവാൻ സാധ്യത കുറവാണ് എന്ന് ഒരു സുഹൃത്ത് പറയുന്നത് കേട്ടു. ഇത് സത്യമാണോ?	I heard a friend say that women who wear clothes like jeans continuously, especially at night, are less likely to have children. Is that true?
घर मे हर उम्र के लोग है सबका अलग अलग पसंद है और जब खाने की बात आती है तो मैं हमेशा असमंजस मे रहती हु। क्युकी उसमे बुजुर्गों का अलग से प्रबंधन करना पड़ता है और कोई अच्छा उपाय क्या है जिससे इसको ठीक से मैनेज किया जा सके?	There are people of all ages at home, each with different preferences, and when it comes to food, I always find myself in a dilemma. Because elderly people need to be managed separately, what is a good solution to manage this properly?
घर पर बुजुर्ग महिलाएं कहती हैं कि बच्चों को पैदा होने के तुरंत बाद ही स्तनपान कराना चाहिए परन्तु हॉस्पिटल में डॉक्टर ऐसा करने से रोकते हैं, ऐसे में हमें क्या करना चाहिए?	Elderly women at home say that children should be breastfed immediately after birth, but doctors in the hospital prevent this; what should we do in this situation?
ಸೊಳ್ಳೆ ಕಡಿತದಿಂದ ನನಗೆ ಡೆಂಗ್ಯೂ ಬಂದಿದೆ, ಕೆಲವರು ಪಪ್ಪಾಯ ಎಲೆ ತಿಂದರೆ ಕಡಿಮೆಯಾಗುತ್ತದೆ ಎನ್ನುತ್ತಾರೆ, ಅದು ನಿಜಾನಾ ಅಥವಾ ಸುಳ್ಳಾ	I got dengue from a mosquito bite, some people say eating papaya leaves will help, is this true or false?
ನನಗೆ ಸಕ್ಕರೆ ಕಾಯಿಲೆ ಇದೆ ಆದ್ರೆ ಮಾತ್ರೆ ತಗೊಳೊಕೆ ಬೇಜಾರು ನಾನ್ ಮಾತ್ರೆ ತಗೊಳ್ದೆ ಹೇಗೆ ಸಕ್ಕರೆ ಕಾಯಿಲೆಯಿಂದ ದೂರ ಆಗ್ಬಹುದು?	I have diabetes but I don't feel like taking medication. How can I keep diabetes in check without it?

Table 4: **Representative User-Generated Queries.** Six authentic queries from CSO chatbot logs demonstrating linguistic diversity (Malayalam, Hindi, Kannada), cultural embedding (traditional beliefs, family pressure, dietary myths), and colloquial phrasing. Queries operationalize the $\mathbf{h}_{\text{cult}}$ component requiring community-aware evaluation.

Evaluation Challenges. Appropriate responses must: address biomedical facts (dengue treatment, VBAC risks), navigate cultural sensitivities (breastfeeding conflicts between elders and doctors, fertility myths requiring gentle debunking), provide context-appropriate advice (dietary management for joint families, medication alternatives respecting patient autonomy), and use culturally-resonant communication styles. These multi-dimensional requirements test whether LLM evaluators can perceive quality beyond surface fluency, the central question of our geometric framework.

C Limitations

While our findings provide strong evidence for geometric mode collapse in LLM evaluation, we acknowledge several limitations that constrain the generalizability of our claims and suggest directions for future work.

C.1 Limited Linguistic and Cultural Coverage

Three-Language Constraint. Our study examines only Hindi, Kannada, and Malayalam, three languages representing Indo-Aryan and Dravidian families within South Asia. This narrow linguistic scope excludes: (1) other major Indic languages (Bengali, Tamil, Telugu, Marathi, Gujarati); (2) non-Indic Asian languages (Chinese, Japanese, Korean, Thai); (3) African languages (Swahili, Yoruba, Amharic); (4) Latin American languages (Spanish dialects, Quechua, Guarani); and (5) minority/endangered languages. While our corpus ratio (Hindi:Kannada:Malayalam) provides strong dynamic range for testing resource-stratified orthogonality, we cannot claim that

the exponential relationship in Proposition 2 ($\alpha^{\ell}_{\text{null}} \propto \exp(-\beta|\mathcal{C}_\ell|/|\mathcal{C}_{\text{ref}}|)$) holds universally across all language families or writing systems. Typologically distant languages (e.g., tonal languages, agglutinative morphology extremes, non-concatenative morphology) may exhibit different collapse patterns.

South Asian Cultural Specificity. Our cultural embedding analysis focuses exclusively on South Asian health epistemologies, Ayurvedic traditions, joint family structures, religious dietary practices, and community health norms specific to India. The observed orthogonality between cultural features and LLM subspaces ($\langle \mathbf{u}_i, \mathbf{h}_{\text{cult}} \rangle \approx 0$) reflects the under-representation of *these particular* cultural contexts in pre-training corpora. We cannot generalize to other cultural dimensions: Indigenous knowledge systems (e.g., Native American, Aboriginal Australian), African traditional medicine, Latin American folk practices, or even Western sub-cultures (LGBTQ+ health, disability communities, religious minorities). The *mechanism* of cultural orthogonality, fluency-dominant pre-training leading to null-space projection, likely generalizes, but the *specific cultural features* annihilated will vary by context.

Within-Language Variation. Our analysis treats Hindi, Kannada, and Malayalam as monolithic languages, ignoring dialectal variation, sociolinguistic registers, and urban-rural linguistic differences. Hindi exhibits substantial dialectal diversity (Braj, Awadhi, Rajasthani influences); Kannada varies between coastal and inland regions; Malayalam shows distinctions between Malabar and Travancore dialects. Our evaluators were predominantly urban, educated medical professionals, potentially missing rural linguistic patterns and non-elite health discourse. The observed 85.9% null-space for Malayalam may partially reflect this intra-language heterogeneity rather than purely resource scarcity.

C.2 Limited Model Coverage

Two-Model Evaluator Set. We employ only GPT-4o and Sarvam-2B as evaluator LLMs, limiting our ability to make universal claims about "all LLMs." While these models represent frontier (GPT-4o: 1.8T parameters, 50+ languages) versus specialized (Sarvam: 2.4B parameters, Indic-focused) extremes, we exclude: (1) other frontier models (Claude-3, Gemini-Ultra, PaLM-2); (2) mid-size models (Llama-3-13B, Mistral-Medium); (3) specialized multilingual models (BLOOM, mT5); (4) older architectures (GPT-3.5, BERT-based). Our finding that GPT-4o exhibits *more severe* collapse than Sarvam (61.3% vs. 48.7% PC1 variance) suggests a counterintuitive scale-rank relationship, but we cannot confirm whether this holds across all model families or is specific to the GPT/Sarvam comparison.

Generator Model Diversity. While we use three generator models (GPT-4, Gemini-Pro, Llama-3-70B), all are transformer-based autoregressive LLMs trained on similar corpora. We do not test: (1) encoder-decoder architectures (T5, BART); (2) retrieval-augmented generators (RAG, Atlas); (3) instruction-tuned specialists (Med-PaLM, BioGPT); (4) few-shot prompted smaller models. Our finding that evaluator-generator independence holds ($\pm 3.2°$ variance across generators vs. $\pm 8.7°$ across languages) may not extend to architecturally distinct generation paradigms.

Temporal Constraints and Model Evolution. Our experiments use model versions from mid-2024 (GPT-4o-2024-05-13, Sarvam-2B v1.2). LLM capabilities evolve rapidly through continual pre-training, RLHF updates, and safety interventions. Our conclusions about geometric mode collapse apply to *these specific model checkpoints* and may not hold if future models incorporate explicit cultural alignment objectives, multi-task evaluation heads, or subspace augmentation techniques. However, the *architectural constraint* of next-token prediction as the pre-training objective remains, suggesting the fundamental problem persists absent paradigm shift.

C.3 Domain Specificity: Health Advice Evaluation

Single-Domain Analysis. Our empirical validation focuses exclusively on health advice quality in a community-consultation context. This domain exhibits maximal cultural embedding ($w_{\text{cult}} \gg w_{\text{style}}$ in Equation 2), but geometric mode collapse may manifest differently in domains with different dimensional structure:

- **Code generation**: $\mathbf{h}_{\text{cult}} \approx 0$, $\mathbf{h}_{\text{fact}}$ dominates. LLMs may perform well here because correctness is orthogonal to culture.
- **Creative writing**: $\mathbf{h}_{\text{style}}$ dominates. LLMs may over-perform if evaluation itself rewards fluency.
- **Legal/policy analysis**: Complex interaction between $\mathbf{h}_{\text{fact}}$ (statutory correctness) and $\mathbf{h}_{\text{cult}}$ (jurisdictional context). Collapse patterns unclear.
- **Mathematical reasoning**: Near-zero cultural loading but high $\mathbf{h}_{\text{fact}}$ complexity. LLMs may fail for different reasons (logical depth vs. cultural orthogonality).

We cannot claim that health domain findings transfer without modification. However, any domain requiring implicit, context-dependent quality judgments (education, social services, mental health, legal aid) likely exhibits similar orthogonality.

High-Stakes Context Emphasis. Our focus on health, where evaluation errors have severe consequences (delayed care, inappropriate advice), motivates the work but may over-emphasize failure modes less critical in low-stakes domains (entertainment, casual Q&A). The 93-95% null-space fraction quantifies variance loss but does not directly measure *harm magnitude*. A response with 95% annihilated variance may still be "good enough" for casual use. Our work is strongest when applied to deployment contexts where cultural misalignment has tangible adverse effects.

C.4 Methodological Constraints

Human Annotator Limitations. Our 15 medical professional evaluators, while domain-expert and native-speaking, introduce potential biases: (1) **Urban-educated perspective**: May not represent rural or non-elite health literacy expectations; (2) **Allopathic training bias**: Medical school education emphasizes Western medicine, potentially undervaluing traditional practices they personally might not endorse; (3) **Small sample size**: 15 evaluators across three languages yields 4-6 per language, limiting inter-annotator reliability measurement and obscuring intra-language disagreement; (4) **Task fatigue**: Each evaluator assessed 40-60 responses, risking declining attention. While we observe high within-community agreement ($\kappa > 0.75$), we cannot rule out shared systematic biases (e.g., all evaluators preferring allopathic over Ayurvedic advice beyond what patients prefer).

Evaluation Criteria Operationalization. Our four criteria (Accuracy, Clarity, Completeness, Helpfulness) are theoretically motivated but operationalized via brief prompt descriptions. Annotators received minimal calibration training, relying on intuitive interpretation. This creates measurement noise: what one annotator considers "complete" may differ systematically from another's threshold. Our PCA analysis reveals that LLMs compress these criteria onto a single axis, but *whether humans maintain true criterion independence or also partially conflate them* is not definitively established. A more rigorous approach would involve extensive annotator training, pilot studies with inter-annotator calibration, and explicit rubrics for each criterion-language-theme combination.

Dataset Scale. With 600+ human judgments across three languages, four criteria, and eight themes, our dataset provides sufficient signal for demonstrating mode collapse but is modest by modern NLP standards. Larger datasets (10,000+ responses, 50+ evaluators per language) would enable: (1) finer-grained analysis of theme-specific and criterion-specific collapse patterns; (2) robust subgroup analysis (rural vs. urban, age-stratified, education-stratified evaluators); (3)

cross-validation of null-space measurements; (4) statistical power for detecting smaller effect sizes. Our current findings show *extreme* effects (85.9% null-space, 79° angles), but subtle intermediate phenomena may be obscured by dataset size.

C.5 Theoretical Framework Limitations

Projection Operator as Simplifying Model. Our formalization of LLM evaluation as $\text{Score}_{\mathcal{J}}(\mathbf{h}) = \|\mathbf{P}_{\mathcal{J}}\mathbf{h}\|_2^2$ (Equation 3) is a *model* of LLM behavior, not a literal description of internal mechanics. Actual LLMs employ multi-layer attention mechanisms, non-linear activations, and stochastic sampling, none of which are strictly linear projections. Our projection operator framework captures the *effective behavior* (low-rank score variance, orthogonal subspaces) without claiming LLMs literally perform SVD or matrix projection during inference. The utility of the framework lies in its predictive and explanatory power, not mechanistic fidelity. Future work integrating attention analysis, probing classifiers, and mechanistic interpretability (**?**) could validate or refine the projection metaphor.

Semantic Decomposition Assumptions. Equation (1) posits orthogonal decomposition $\mathbf{h} = \mathbf{h}_{\text{style}} + \mathbf{h}_{\text{fact}} + \mathbf{h}_{\text{cult}}$ with $\langle \mathbf{h}_i, \mathbf{h}_j \rangle = 0$. In reality, these components likely interact: cultural appropriateness affects perceived clarity, factual content influences helpfulness judgments, style impacts trust in factual claims. Our operationalization via synthetic linguistic features (fluency metrics, entity density, cultural markers) is a proxy, not ground truth. The observed correlation patterns (style features $r < 0.2$ with residuals, cultural features $r > 0.58$) support the decomposition empirically, but $\mathbf{h}_{\text{style}}$, $\mathbf{h}_{\text{fact}}$, $\mathbf{h}_{\text{cult}}$ are latent constructs, not directly measurable.

Null-Space Fraction Interpretation. The null-space fraction $\alpha_{\text{null}} = 1 - \text{Var}_{\mathcal{J}}/\text{Tr}(_H)$ quantifies variance loss but does not distinguish between: (1) **perceptual blindness** (LLM cannot detect the dimension), versus (2) **weighting choice** (LLM detects but assigns zero weight). Our theory emphasizes (1), geometric orthogonality, but cannot fully rule out that LLMs perceive cultural dimensions yet choose (via RLHF or safety training) to ignore them. Mechanistic studies analyzing internal representations could disambiguate these scenarios.

C.6 Generalizability and External Validity

Beyond Text Evaluation. Our findings address text-based evaluation of text outputs. Multimodal LLMs evaluating image+text (medical imaging reports), video+text (telehealth consultations), or speech+text (voice-based health queries) may exhibit different geometric properties. Cultural context in multimodal settings involves visual symbolism, gesture norms, and audio prosody, dimensions not captured by our text-only framework.

Interactive vs. Static Evaluation. We analyze single-turn evaluation (judge scores a fixed response). Real-world deployment often involves multi-turn dialogue where evaluators assess conversational quality, empathy evolution, and context accumulation. Geometric mode collapse in interactive settings may manifest differently, sequential projections could compound orthogonality or, alternatively, multi-turn context could enable partial recovery of cultural signal.

Human-in-the-Loop Mitigation. Our work assumes fully automated LLM-as-Judge. In practice, human-AI collaboration (LLM pre-screens, human reviews edge cases) may mitigate some failure modes. If humans predominantly review cases where LLM confidence is low, and low confidence correlates with high cultural embedding, then the 85.9% null-space might be partially recoverable through hybrid workflows. However, this assumes LLMs can reliably detect their own cultural blindness, an empirical question we do not address.

C.7 Reproducibility and Data Access

Proprietary Model Constraints. GPT-4o is a proprietary model accessed via API, with limited transparency about training data, architectural details, or RLHF procedures. We cannot guarantee exact reproducibility as OpenAI may update model weights. Similarly, Sarvam-2B,

while marketed for Indic languages, provides limited documentation. Future replication studies may encounter version drift or API changes affecting results.

Sensitive Data and Privacy. Our health queries, though de-identified and sourced through CSO partnerships with informed consent, involve potentially sensitive topics (reproductive health, mental health, stigmatized conditions). We cannot publicly release the full dataset without additional anonymization, limiting independent verification of our findings. We plan to release aggregated annotations, prompt templates, and evaluation scripts, but raw responses remain confidential.

Language-Specific Resources. Reproducing our Malayalam analysis requires access to Malayalam-fluent medical professionals, a scarce resource even within India. Other low-resource language studies face similar constraints, limiting the pace at which our findings can be validated across diverse linguistic contexts. Crowdsourcing non-expert annotations for health content raises ethical concerns, creating a methodological tension between scale and expertise.

C.8 Implications and Scope of Claims

Despite these limitations, our core claims remain robust within the studied scope:

1. **Geometric mode collapse is real**: LLM judges compress variance and operate in low-rank subspaces across multiple models, languages, and criteria.
2. **Orthogonality is systematic**: $> 79°$ angles between LLM and human subspaces are too large to explain via calibration failure, they reflect architectural constraints.
3. **Resource stratification validates theory**: Malayalam's 85.9% null-space exceeding Hindi's 79.7% confirms that corpus size drives orthogonality, not model capability.
4. **Consensus does not imply validity**: Inter-LLM agreement (48° angles) is smaller than LLM-human misalignment (76-83°), proving LLMs agree *with each other* while failing to align *with humans*.

Future work should extend to additional languages (especially tonal, logographic, and polysynthetic languages), domains (legal, educational, creative), models (open-source, encoder-decoder, retrieval-augmented), and evaluation paradigms (multi-turn, multimodal, adversarial). Each extension will test whether our geometric framework generalizes or requires refinement. The *existence* of mode collapse is established; the *universality* remains an open question.

D Ethical Considerations

Our ethical considerations follow the framework proposed by Bender and Friedman (2018), with attention to institutional oversight, data provenance, and annotator welfare.

Institutional Process and Oversight The community-centered data collection was conducted in collaboration with Civil Society Organizations (CSOs) and local data workers. All participants were briefed about the purpose of the study, and consent was obtained prior to participation.[1]

Data Provenance and Quality Assurance The dataset was created through participatory design with CSOs to reflect real-world, multilingual health queries. This ensured that the evaluation captured culturally grounded and socially consequential dilemmas. To minimize potential harms and ensure quality, all responses were reviewed by both human evaluators and automated tools. No sensitive personal information was included in the dataset.

Annotator Demographics Fifteen community data workers contributed to the creation of queries, representing diverse age, gender, and educational backgrounds (Table 5). Their linguistic expertise and lived experience were crucial for ensuring that the data reflected genuine community priorities. The CSOs that facilitated collaboration and deployment contexts are summarized in Table 6.

[1]We do not disclose the names or organizational details of the CSOs or data workers to preserve anonymity.

Number of participants	15
Age range	19–36
Gender	Male: 4 ; Female: 11
Education	High-school: 1 ; Undergraduate: 10 ; Graduate: 4

Table 5: Demographic details of community data workers.

CSO ID	Chatbot deployment	Languages supported
CSO-1	Yes	Hindi, Marathi, Telugu
CSO-2	Yes	Kannada, English
CSO-3	Yes	Hindi, English
CSO-4	No	Malayalam
CSO-5	Yes	Hindi, Kannada, English

Table 6: Participating CSOs. Details of organizations remain anonymized for confidentiality.

Use of LLMs in 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.