🔨 The Wedge Questions: Latent Cultural Boundaries in LLMs via Persona Projection Divergence

Yejin Son [1] **Yongjin Yang** [2] **Ryan Faulkner** [2] **Matt Ratto** [3] **Seungwon Lim** [1] **Youngjae Yu** [4] **Zhijing Jin** [2 5]

Abstract

Large Language Models (LLMs) contain a wealth of cultural data, often functioning as a bridge into diverse societies. This demands cultural alignment with diverse moral landscapes and social conventions. Prior work has documented how post-training alignment homogenizes expression in the model's generated text. Our study reveals the complementary internal picture: rather than erasing cultural distinctions, alignment appears to *sharpen* their geometric encoding in the latent space. We term this phenomenon the "Alignment Paradox." We suggest that this sharpening is a natural byproduct of homogenization: to consistently produce neutral outputs, a model must first know *where* cultural distinctions lie, which entails encoding them more precisely in its internal representations. To surface and quantify this effect, we propose Persona Projection Divergence (PPD), a geometric measure to identify cultural value boundaries within the latent space. While alignment-induced homogenization often renders cultural distinctions invisible to surface-level text metrics, PPD uncovers these boundaries within the latent space. Our framework provides a non-redundant diagnostic signal that captures internal value conflicts, uncovering latent polarities that remain hidden behind a neutralized textual surface. By probing these internal axes before they are collapsed into a neutral facade, we establish a scalable diagnostic tool to identify "Wedge Questions" that trigger latent value conflicts otherwise hidden behind homogenized outputs.

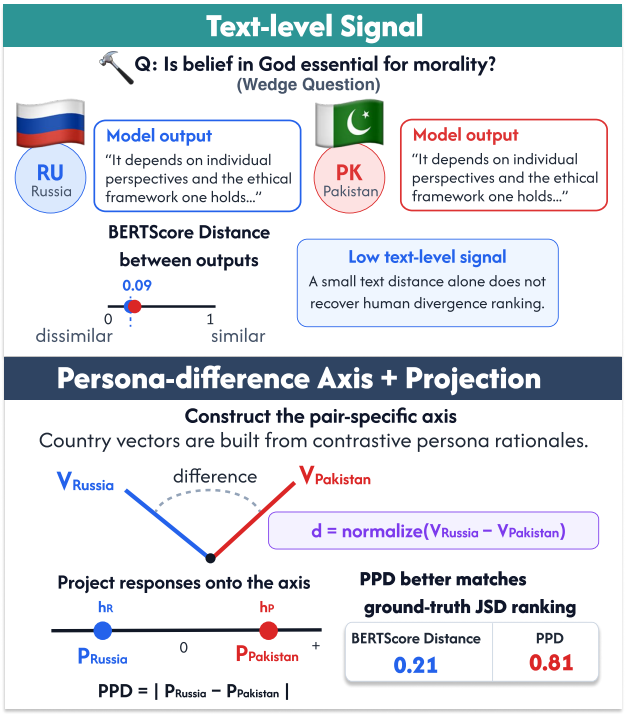


Figure 1. Text similarity fails to capture cross-cultural divergence. By constructing a persona-difference axis and projecting responses onto it, PPD (Persona Projection Divergence) reveals divergence that better matches ground-truth human disagreement (JSD), revealing latent cultural divergence hidden by surface-level neutrality.

1. Introduction

As Large Language Models (LLMs) are deployed at a global scale, there is a growing demand for them to possess cultural competence—a multifaceted capability that extends beyond mere linguistic translation to the faithful alignment with diverse moral landscapes and social conventions (Hershcovich et al., 2022). Thus, it is through necessity that models must recognize that "culture is not trivial" (Zhou et al., 2025) and perform sophisticated localization based on a contextualized understanding of diverse cultural backgrounds to support users effectively. For instance, when asked about a sensitive social issue like 'filial obligations versus career ambition,' a user in an Eastern context would immediately identify a lack of cultural competence if the model provides

[1]Yonsei University, Seoul, Republic of Korea [2]Jinesis Lab, University of Toronto & Vector Institute [3]Faculty of Information, Schwartz-Reisman Institute for Technology and Society, University of Toronto. [4]Seoul National University, Seoul, Republic of Korea [5]Max Planck Institute for Intelligent Systems, Tübingen, Germany. Correspondence to: Yejin Son <yejinhand@yonsei.ac.kr>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea.

a purely individualistic rationale. Such 'cultural dissonance' occurs not because of a translation error, but because the model's internal alignment fails to prioritize the relational values dominant in that specific society. Furthermore, such competence is essential for representing individuals from different cultures in complex social simulations (Argyle et al., 2023; Park et al., 2023) and for providing pluralistic responses that reflect diverse perspectives rather than a single answer (Sorensen et al., 2024; Nie et al., 2026).

A critical method for evaluating a model's cultural competence is to examine whether it clearly perceives the differences in viewpoints and stances between two distinct cultures. To this end, it is necessary to identify questions that trigger divergence between two countries, which we define in this work as "Wedge Questions."

Not all questions in value surveys (e.g., WVS[1], Pew[2]) create divergent views among respondents. For example, while South Korea and China may provide similar responses to certain social questions due to shared communitarian values, they exhibit significantly different response patterns on topics where their core values clash, such as political decision-making processes or family structures. The ability to automatically identify these "wedge questions" for specific country pairs would significantly scale up the evaluation of a model's cultural understanding capabilities.

Beyond diagnostics, identifying these Wedge Questions is a strategic necessity for pluralistic alignment (Sorensen et al., 2024). By surfacing high-variance triggers where the model's latent states remain polarized despite its "neutral facade," we can curate the high-contrast synthetic interaction data required to break through surface-level homogeneity. This provides a scalable path to generate the diverse supervision signals that standard web-scale datasets fail to provide.

In this work, we propose **Persona Projection Divergence (PPD)**, a metric that automatically identifies questions that trigger divergence between two countries. We use the term *persona* following Chen et al. (2025), from whose framework we build; however, our construct operates at the group level, representing collective population-level value tendencies rather than individual character traits. As illustrated in Figure 1, PPD operates by constructing persona vectors for two countries and measuring the degree of separation when a question's response is projected onto the axis formed by the difference between these vectors. Specifically, we generate rationales for answers a country is likely to give versus those it is unlikely to give, and then quantify the contours of value judgment by analyzing the activation gaps between these rationales within the model.

PPD overcomes the limitations of existing methods that rely solely on the distance between generated text responses. Prior work has shown that post-training alignment processes, such as instruction tuning, suppress response diversity and homogenize cultural expression at the surface level (Santurkar et al., 2023; González Barman et al., 2025). Our study reveals the complementary internal picture: for questions such as "belief in God," the Pakistan and Russia personas generate linguistically similar, neutral responses with negligible differences according to text-based metrics, yet within the model's latent activation space, these two personas are pulled toward opposite ends of the axis, revealing sharp latent cultural boundaries (See Figure 1).

We suggest that this sharpening is a natural byproduct of homogenization: to consistently produce neutral outputs across culturally divergent inputs, a model must first encode *where* those distinctions lie—which entails representing cultural boundaries more precisely in its internal activations. Such latent differentiation is not a mere artifact of instruction following but a plausible structural consequence of alignment; the model's capacity to navigate toward a neutral objective is inherently predicated on its internal ability to distinguish these cultural margins. We term this the **Alignment Paradox**: the same process that neutralizes surface-level cultural expression may simultaneously sharpen the latent geometry along which cultures diverge. Crucially, this sharpening is not a byproduct of rote memorization but of *refined internal identification*: it is relational, defined with respect to a contrastive axis $\hat{v}_{AB}^{(\ell)}$ between two cultural personas rather than any single cultural label, and layer-progressive, growing in magnitude as representations pass through the model's middle and upper layers.

The primary contributions of this research are as follows:

1. **Introduction of the PPD Metric**: We propose **Persona Projection Divergence (PPD)**, which quantifies cultural boundaries by probing the activation gaps between value-laden rationales within a model's latent space. While traditional text-based metrics fail to detect divergence when outputs are neutralized during alignment, PPD explicitly measures the degree of separation on a task-specific axis. This approach allows us to uncover latent cultural polarities that remain important for the model's ability to navigate toward neutral surface-level responses.

2. **The Alignment Paradox**: We provide evidence suggesting that post-training alignment sharpens the latent geometric encodings of cultural boundaries, even as it homogenizes surface-level outputs. Building on this observation, we offer an interpretation grounded in the mechanics of neutralization to explain how such latent

[1] https://www.worldvaluessurvey.org/

[2] https://www.pewresearch.org/global/datasets/

Table 1. Wedge questions: high-JSD item pairs where national response distributions diverge largely. Stacked bars show response shares; gray rows report Jensen–Shannon divergence (bits).

Source	National shares
Q: Should homosexuality be accepted by society, or not accepted?	accepted / not accepted
GOQA	**Germany** vs **Pakistan**
	$\hat{P}_{\text{DEU}} = (.888, .112)$, $\hat{P}_{\text{PAK}} = (.022, .978)$ JSD = **0.663**
Q: How important is religion in your life?	very important / rather important / not very important / not at all
WVS	**Myanmar** vs **Netherlands**
	$\hat{P}_{\text{MMR}} = (.810, .173, .012, .005)$, $\hat{P}_{\text{NLD}} = (.109, .122, .257, .512)$ JSD = **0.559**

differentiation arises.

3. **Demonstration of Practical Scalability**: Through the **"Wedge-in-the-Haystack"** task, we show that PPD can automatically identify wedge questions among generated queries, establishing a scalable diagnostic framework that reduces human review costs by up to 14.8%.

2. Related Work

Pluralistic Alignment and Persona Modeling Recent research emphasizes that AI systems should reflect the diversity of human values rather than converging toward a single, homogenized perspective (Zhang et al., 2025; Sorensen et al., 2024; 2026).

This challenge is contextualized by the *persona selection model*, which posits that while LLMs internalize a vast library of cultural personas during pre-training, the subsequent alignment process functions as a selector that suppresses this latent diversity in favor of a narrow "assistant" profile (Lu et al., 2026; Marks et al., 2026). Our work builds on these theoretical foundations by proposing a geometric method to measure the divergence between these latent personas.

Cultural Alignment and Evaluation Datasets To evaluate and improve cultural competence, various benchmarks have been developed. *GlobalOpinionQA* provides country-level response distributions for cross-cultural analysis (DURMUS et al., 2024). More recently, the *CAReDiO* framework was introduced to identify culturally representative and distinctive samples for efficient alignment using information-theoretic objectives (Yao et al., 2026). While these works focus on data curation for alignment performance, our research introduces *Persona Projection Divergence (PPD)*, a diagnostic metric that operates directly on the latent activation space to identify "wedge questions" that expose cultural boundaries without requiring extensive human annotations.

Probing Geopolitical and Value-Based Boundaries A subset of research focuses on identifying specific domains where model viewpoints diverge, such as the *BorderLines* dataset, which evaluates LLM biases in the context of geopolitical disputes (Li et al., 2024). Our approach generalizes these efforts by proposing a systematic, automated method (*Wedge-in-the-Haystack*) to detect such high-divergence questions across any pair of cultural personas, shifting the focus from manual data collection to automated internal probing.

3. The Landscape of Cultural Divergence

We propose a formal foundation for identifying cultural boundaries by defining the empirical landscape of human disagreement. First, we will explain the mathematical formulation of Jensen-Shannon Divergence (JSD) used as our ground-truth reference (§3.1). Then, we describe the public opinion datasets and the preprocessing steps required to handle diverse response distributions (§3.2). Finally, we detail the selection of cultural persona pairs and the construction of the evaluation corpus (§3.3).

3.1. Measuring Human Disagreement

Our goal is to identify questions that elicit divergent value judgments between two countries. To quantify this divergence, we use the Jensen–Shannon Divergence (JSD) between country-level response distributions as a ground-truth signal. Given two countries A and B and their response distributions P and Q over answer choices, we define:

$$\text{JSD}(P, Q) = \tfrac{1}{2}\text{KL}(P \,\|\, M) + \tfrac{1}{2}\text{KL}(Q \,\|\, M), \;\; M = \tfrac{1}{2}(P+Q). \tag{1}$$

JSD is a standard measure of distributional divergence, making it suitable for comparing survey response patterns across countries. Unlike text-based similarity metrics, it captures population-level disagreement independent of surface linguistic realization. Throughout this work, JSD serves as the human-grounded reference signal for golden layer selection and alignment analysis.

3.2. Source Distributions and Preprocessing

We use cross-national public opinion datasets: **GlobalOpinionQA (GOQA)** (DURMUS et al., 2024) and the **World Values Survey (WVS) Wave 7**[3]. GOQA captures geopolitical and social attitudes, while WVS probes deep-seated structural values. For GOQA, we retain Pew GAS items and remove non-substantive responses (e.g., *don't know*, *refusal*), renormalizing probabilities over the remaining categories. For WVS, we aggregate individual-level responses into country-level distributions, excluding standard non-answer codes (e.g., `Not applicable`) before renormalization. Each sample is represented as a `(country, question, distribution)` tuple. Table 1 provides representative examples of such distributions from both datasets, along with the resulting JSD values between country pairs.

3.3. Selection of Cultural Persona Pairs

For each dataset, we enumerate country pairs based on the number of jointly answered questions and retain 10 pairs to ensure broad continental coverage. For a given pair (A, B), only questions answered by both countries are considered for the evaluation sets. These shared questions are reserved to test the model's ability to perceive known human cultural boundaries.

4. Methods: Persona Projection Divergence

We propose Persona Projection Divergence (PPD), a geometric approach to identify cultural boundaries within the latent space of LLMs. As illustrated in Figure 2, this method is divided into a **TRAIN stage** (Steps 1–5) for axis construction and an **EVAL stage** (Steps 1–5) for scoring.

First, we will explain the construction of country-specific persona vectors via contrastive rationales (§4.1). Then, we describe the pair-specific projection mechanism used to calculate the PPD metric during inference (§4.2). Finally, we detail the selection of the "golden layer" and the variants designed for robust diagnostics (§4.3).

4.1. Constructing Persona Vectors (TRAIN)

Following the persona vector approach of Chen et al. (2025), which was originally designed for individual-level character traits, we adapt this framework to construct group-level cultural value profiles, establishing their geometric coordinates offline as illustrated in the top row of Figure 2.

Step 1: Survey Data. We utilize survey response distributions from datasets such as GOQA and WVS. For each country C and question q in the training set, these distributions serve as the empirical basis for cultural anchoring, while the inter-country JSD established in §3.1 provides the human-grounded reference signal.

Step 2: Rationale Generation. For each training question, we generate a positive rationale x^+ supporting the modal response and a negative rationale x^- supporting the least frequent response using *Meta-Llama-3-70B-Instruct*; details regarding the rationale generation prompts are provided in Appendix E. As this procedure involves stochastic generation, one concern is sensitivity to specific samples or model choices; however, we find that PPD scores remain stable under repeated rationale sampling and across different rationale generators (see Appendix H.2 and H.3). The resilience of PPD rankings against such variations indicates that the metric probes a consistent internal value axis rather than responding to superficial linguistic cues or prompt-specific artifacts.

Step 3-4: Vector Computation. We feed x^+ and x^- into the model $\mathcal{M}$ (Step 3). At each layer ℓ, we compute a summary representation $\mathbf{h}^{(\ell)}$ for each individual rationale by mean-pooling the hidden states across all token positions in the rationale span R:

$$\mathbf{h}^{(\ell)} = \frac{1}{|R|} \sum_{t \in R} \mathbf{h}_t^{(\ell)}. \tag{2}$$

To build the country-level centroid, we aggregate these question-level vectors over the training set Q_{train}. We define $\mu_{C,\text{pos}}^{(\ell)}$ and $\mu_{C,\text{neg}}^{(\ell)}$ as the average of the $\mathbf{h}^{(\ell)}$ vectors obtained from all positive and negative rationales, respectively, for country C. The per-layer persona vector $v_C^{(\ell)}$ (Step 4) is then constructed as the difference between these centroids:

$$v_C^{(\ell)} = \mu_{C,\text{pos}}^{(\ell)} - \alpha \cdot \mu_{C,\text{neg}}^{(\ell)}, \qquad \hat{v}_C^{(\ell)} = v_C^{(\ell)} / \|v_C^{(\ell)}\|_2. \tag{3}$$

Step 5: Axis Construction. We define the unit cultural contrast axis $\hat{d}_{AB}^{(\ell)} = \text{normalize}(\hat{v}_A^{(\ell)} - \hat{v}_B^{(\ell)})$. This axis isolates the latent direction that maximally differentiates the cultural value profiles of the two countries, effectively creating a one-dimensional scale of cultural contrast in the activation space.

4.2. Pair-Specific Projection Divergence (EVAL)

In the EVAL stage, we diagnose unseen questions by projecting response activations onto the pre-constructed axes. First, we explain the encoding of country-conditioned responses (Steps 1–3). Then, we describe the final PPD scoring and ranking (Steps 4–5).

Step 1-3: Generation and Normalized Encoding. Given a new question q (Step 1), $\mathcal{M}$ decodes responses $y_C(q)$ (Step 2). We mean-pool the hidden states and ℓ_2-normalize to obtain $\tilde{h}_C^{(\ell)}(q)$ (Step 3).

[3] https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp

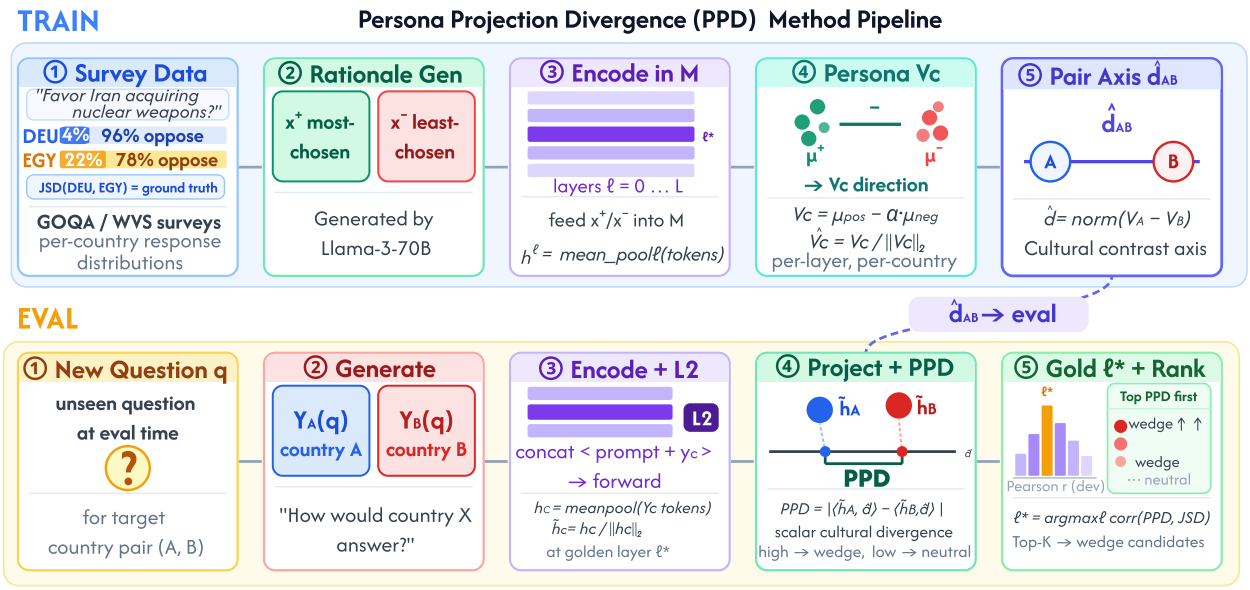


Figure 2. **The PPD method pipeline.** (*TRAIN*) A unit cultural contrast axis $\hat{d}_{AB}$ is constructed offline from survey distributions by encoding contrastive rationales (Steps 1–5). (*EVAL*) Activations for unseen questions are projected onto the axis at the golden layer ℓ^*; PPD quantifies the scalar distance between these projections to surface wedge candidates (Steps 1–5).

Step 4-5: Scoring. PPD is the absolute difference between projections onto $\hat{d}_{AB}$ (Step 4):

$$\mathrm{PPD}^{(\ell)}(q) = \left| \langle \tilde{h}_A^{(\ell)}(q),\, \hat{d}_{AB}^{(\ell)} \rangle - \langle \tilde{h}_B^{(\ell)}(q),\, \hat{d}_{AB}^{(\ell)} \rangle \right|. \quad (4)$$

This measures the distance between the scalar projections of the two response activations onto the contrast axis. A high PPD indicates that the question polarizes the model's internal representations into states that are linearly opposed along the value dimension that distinguishes the two countries. Because PPD measures these activations directly, it captures a diagnostic signal that remains decoupled from the token-generation process. Questions are ranked by this metric to surface wedge questions[4] (Step 5).

4.3. Golden Layer and Variants

We select a single **golden layer** ℓ^* per model by maximizing the Pearson correlation with human JSD on dev data. This separation is critical for methodological validity: exposing test data during selection would allow indirect optimization for test performance. We also explore variants: **Pair-wise** α and **Layer Weighting** (top-5 layers).

[4]Broadly, we define wedge questions as culturally contentious queries that elicit high inter-country opinion divergence. In our evaluation tasks (§5), we operationally identify questions with the highest PPD scores as wedge candidates.

5. Experiments

We propose a multi-faceted evaluation to demonstrate the efficacy and diagnostic value of PPD. Broadly, we use the term **wedge question** to refer to queries that induce high inter-country opinion divergence. Specific evaluation tasks instantiate this definition using task-dependent thresholds over human JSD. Our experiments aim to show how PPD identifies these internal boundaries even when they are obscured by homogenized surface-level text. First, we will explain the experimental configurations, including datasets, models, and competitive baselines used in our study (§5.1). Then, we evaluate PPD's practical utility in the *Wedge-in-the-Haystack* task (§5.2). Finally, we evaluate the correspondence between latent signals and human values in the *Latent-Human Alignment* task (§5.3).

5.1. Experimental Setup

Overview We propose a controlled environment designed to test the limits of cultural boundary detection in aligned versus non-aligned models. First, we will explain the datasets and model configurations used for evaluation. Then, we describe the competitive baselines and the specific tasks designed to test the scalability of our approach.

Datasets. We use cross-national public opinion datasets: **GlobalOpinionQA (GOQA)**, focusing on contemporary geopolitical attitudes, and the **World Values Survey (WVS) Wave 7**, which probes structural values like religion and

social trust.

Models. We evaluate PPD primarily on representative instruction-tuned (It) models and their base (Bs) counterparts, specifically **Llama-3.1-8B**[5] and **Gemma-2-9B**.[6] This focus allows us to directly observe how the post-training alignment process structures the internal representation of cultural values across different architectures.

Baselines. To validate the effectiveness of PPD, we compare against baselines that estimate wedge strength at the text level:

- **BERTScore Distance**: To align with the distance-based nature of our proposed PPD, we compute the semantic divergence as $1 - \text{F1}_{\text{BERTScore}}$. Since the raw BERTScore measures similarity, this transformation ensures that a value of 0 represents identical rationales, while higher values indicate greater linguistic divergence.
- **LLM Judge:** Large models (Llama-3.1-70B, Qwen2.5-72B, and GPT-4o-mini) are used to directly assess wedge intensity on a scale of 0–100 based on the full rationales.

Metric. We use the **Jensen-Shannon Divergence (JSD)** between human response distributions as the ground-truth signal for cultural divergence.

Tasks. Our evaluation consists of primary tasks: (1) **Wedge-in-the-Haystack** (§5.2), measuring retrieval efficiency, and (2) **Latent-Human Alignment** (§5.3), examining the correlation between internal activations and human opinions.

5.2. Wedge-in-the-Haystack

Overview We propose the *Wedge-in-the-Haystack* task to evaluate whether PPD can practically reduce the cost of identifying culturally divisive questions. First, we will explain the task definition and the specific operational definition of "needles" within the dataset (§5.2.1). Then, we present the precision results and the resulting reduction in human review costs (§5.2.2).

5.2.1. Task Definition and Metrics

This task measures retrieval efficiency in a realistic scenario where reviewers must prioritize a few contentious questions from a large pool. For this task, we operationally define **wedge questions (needles)** as the **top 20% of questions** ranked by human JSD. The remaining 80% are treated as neutral background (the haystack). We measure **Precision@K** ($P@K$) for $K \in \{10, 20\}$ and calculate review-cost savings.

$$P@K = \frac{1}{K} \sum_{i \in \text{Top}K} \mathbf{1}[y_i = 1] \tag{5}$$

5.2.2. Results and Practical Implications

We report retrieval results in Table 2. The results reveal a stark performance gap between instruction-tuned and base models. PPD variants, particularly the α**+Layer weighting** variant, consistently outperform text-based baselines on instruction-tuned models, achieving a peak $P@10$ of **0.70** on GOQA · Llama-It and **0.60** on GOQA · Gemma-It. While text-based BERTScore Distance remains competitive on WVS ($P@10 = 0.40$), PPD variants like Pair-wise α achieve superior performance ($P@10 = 0.50$ on Gemma-It).

The practical utility of PPD is most evident in the review-cost savings. As shown in the lower portion of Table 2, **PPD (Default)** demonstrates significant savings compared to random screening. In the GOQA · Llama-It configuration, it achieves a **+14.8% cost reduction**, and in the WVS · Gemma-It setting, it reaches **+9.9% savings**. These figures consistently exceed the savings from BERTScore Distance (e.g., +3.3% and +4.2% respectively). In contrast, PPD performance on base models is often inconsistent, yielding much lower or near-baseline savings, which is consistent with the hypothesis that post-training alignment sharpens latent cultural boundaries.

5.3. Latent-Human Alignment

Overview We propose the *Latent-Human Alignment* task to evaluate whether latent geometric signals can reliably identify culturally divisive questions. We focus on the evaluation protocol (§5.3.1) and wedge detection performance via AUC-ROC (§5.3.2).

5.3.1. Evaluation Metrics

All metrics are calculated per triple (c_A, c_B, q). The ground truth signal is human disagreement measured by Jensen–Shannon Divergence (JSD), while each method produces a continuous wedge score x.

Human Disagreement (JSD). We compute JSD between response distributions normalized over shared options (minimum 2). We use the square root of the symmetric KL divergence: $\sqrt{(D_{KL}(P\|M) + D_{KL}(Q\|M))/2}$.

AUC-ROC for Wedge Detection. To evaluate wedge detection performance, we convert the continuous JSD values into binary labels using a median split: instances with JSD at or above the median are labeled as wedges (1), and the remainder as non-wedges (0). AUC-ROC then measures how

[5] meta-llama/Meta-Llama-3.1-8B-Instruct (It) and meta-llama/Meta-Llama-3.1-8B (Bs)

[6] google/gemma-2-9b-it (It) and google/gemma-2-9b (Bs)

		Instruct Models								Base Models							
		Llama-It		Gemma-It		Llama-It		Gemma-It		Llama-Bs		Gemma-Bs		Llama-Bs		Gemma-Bs	
		GOQA				WVS				GOQA				WVS			
Group	**Method**	P@10	P@20	P@10	P@20	P@10	P@20	P@10	P@20	P@10	P@20	P@10	P@20	P@10	P@20	P@10	P@20
PPD	PPD-D	0.50	0.35	0.20	0.40	0.20	0.20	0.40	0.25	0.20	0.20	0.00	0.10	0.30	0.25	**0.40**	0.30
	PPD-α	0.60	**0.50**	0.40	**0.50**	0.30	**0.35**	**0.50**	**0.35**	**0.40**	**0.35**	0.10	0.15	0.10	0.30	0.20	0.10
	PPD-L	0.50	0.35	0.20	0.30	0.20	0.20	0.20	0.25	0.20	0.20	0.00	0.10	0.30	0.20	0.30	0.25
	PPD-αL	**0.70**	0.45	**0.60**	0.50	0.30	0.35	0.20	0.30	0.40	0.35	0.20	0.15	0.20	0.30	0.30	0.30
NL	BERT-D	0.20	0.25	0.20	0.25	**0.40**	0.25	0.40	0.25	0.20	0.25	0.20	0.25	**0.40**	0.25	0.40	0.25
	J-Llama	0.30	0.30	0.30	0.30	0.20	0.35	0.20	0.35	0.30	0.30	**0.30**	**0.30**	0.20	**0.35**	0.20	**0.35**
	J-Qwen	0.00	0.15	0.00	0.15	0.40	0.30	0.40	0.30	0.00	0.15	0.00	0.15	0.40	0.30	0.40	0.30
	J-GPT	0.10	0.10	0.10	0.10	0.40	0.30	0.40	0.30	0.10	0.10	0.10	0.10	0.40	0.30	0.40	0.30
Cost↓	PPD-D	**+14.8%**		+0.7%		**+7.2%**		**+9.9%**		+3.0%		+3.0%		**+1.9%**		+1.4%	
	BERT-D	+3.3%		**+3.3%**		+0.9%		+4.2%		**+3.3%**		**+3.3%**		+0.9%		**+4.2%**	

Table 2. Wedge-in-the-Haystack results (top-20% JSD): Precision@10/20 and review-cost savings. PPD-D: PPD (Default); PPD-α: Pair-wise α; PPD-L: Layer weighting; PPD-αL: α+Layer w.; BERT-D: BERTScore Distance; J-Llama/Qwen/GPT: LLM Judge variants. Bold: best per column; Cost↓: savings vs. random screening.

		Instruct Models				Base Models			
		GOQA		WVS		GOQA		WVS	
Group	**Method**	Llama-It	Gemma-It	Llama-It	Gemma-It	Llama-Bs	Gemma-Bs	Llama-Bs	Gemma-Bs
PPD	PPD-D	0.49	0.58	0.54	0.55	0.41	0.58	0.52	0.52
	PPD-α	0.54	0.58	0.58	0.58	0.54	0.63	0.54	0.49
	PPD-L	0.49	0.55	0.54	0.54	0.41	0.61	0.48	0.48
	PPD-αL	0.53	**0.61**	0.56	0.60	0.54	**0.64**	0.54	0.48
NL	BERT-D	0.51	0.51	**0.63**	**0.63**	0.51	0.51	**0.63**	**0.63**
	J-Llama	**0.59**	0.59	0.39	0.38	**0.59**	0.59	0.42	0.42
	J-Qwen	0.52	0.52	0.49	0.49	0.52	0.52	0.49	0.49
	J-GPT	0.54	0.54	0.52	0.52	0.54	0.54	0.52	0.52

Table 3. ROC-AUC for wedge detection (higher is better). Wedge labels defined via median split on human JSD. Abbreviations follow Table 2. Full results in Appendix G.

effectively each method ranks high-disagreement instances above low-disagreement ones.

5.3.2. Results: Wedge Detection Performance

The AUC-ROC results in Table 3 validate the robustness of PPD for wedge detection. On Gemma-It (GOQA), the α+Layer weighting variant reaches an AUC of **0.61**, surpassing both BERTScore Distance (0.51) and LLM-as-Judge baselines.

Notably, while LLM-as-Judge (Llama-3.1-70B) performs competitively on GOQA (AUC=0.59), its performance collapses on WVS (AUC=**0.38**), suggesting difficulty in capturing deeper structural value differences. In contrast, PPD maintains more stable performance across datasets by leveraging latent signals that remain difficult to recover from surface-level text alone.

The contrast between instruction-tuned and base models further supports our central hypothesis: latent cultural boundaries become substantially more detectable after alignment. While instruction-tuned models consistently exhibit meaningful wedge separability, base models often remain near chance-level performance. This suggests that instruction tuning appears to organize latent value representations.

6. Analysis: Surfacing the Invisible Wedges

We analyze how PPD surfaces latent cultural boundaries that remain hidden under surface-level textual homogenization. First, we examine *PPD-only wedges*, where latent divergence is detected despite low text-level disagreement (§6.1). Then, we investigate the *Alignment Paradox* through layer-wise analysis, demonstrating how post-training alignment sharpens latent cultural geometry and translates into practical retrieval gains (§6.2).

Table 4. **Retrieval performance gaps (Δ) across model pairs.** Positive values indicate Instruct > Base. Benchmarks: GOQA and WVS.

Model Gap	GOQA	WVS
Retrieval Score		
Δ(Instruct−Base, PPD)	+0.50	+0.20
Δ(PPD−BSD)	+0.05	+0.15
Δ(PPD−LLM-j)	+0.25	+0.10
Lift@5		
Δ(Instruct−Base, PPD)	+2.50	+1.00
Δ(PPD−BSD)	+0.25	+0.75
Δ(PPD−LLM-j)	+1.26	+0.50

6.1. PPD-only Wedge Analysis

To demonstrate that PPD captures critical diagnostic information often omitted by surface-level text metrics, we conduct a specialized *PPD-only wedge analysis*. We define an instance as a *PPD-only wedge* if it satisfies three simultaneous conditions: (i) high ground-truth human disagreement (JSD $\geq$ median), (ii) high PPD prediction (PPD $\geq$ median), and (iii) low text-based divergence prediction (BERTScore Distance $<$ median). Under the GOQA/Gemma-It configuration, this criterion isolates approximately 12% of the test set ($n = 42$). These PPD-only wedges represent cases where aligned models generate linguistically homogenized rationales that appear neutral, yet diverge substantially in their latent persona projections. For instance, questions regarding marriage preferences (Canada–France, JSD = 0.203) or U.S. military withdrawal policy (Argentina–Mexico, JSD = 0.224) yield semantically similar rationales at the text level (BERTScore Distance < 0.085). However, PPD reveals sharp cultural separation in the activation space, identifying value conflicts where surface-level similarity might mislead observers into assuming alignment. This confirms that PPD provides a non-redundant diagnostic signal by accessing a value-judgment axis that textual analysis fails to recover.

6.2. The Alignment Paradox: Empirical Evidence

The empirical validity of the *Alignment Paradox* is further evidenced by the layer-wise correlation between latent signals and human opinions. Figure 3b illustrates the mean Pearson correlation between PPD and JSD across layers for both the base and instruction-tuned Gemma-2-9B.

The instruct model exhibits a monotonic rise in correlation within the upper layers, peaking at the golden layer $\ell^* = 41$ with $r = 0.146$. In stark contrast, the base model plateaus at a significantly lower correlation of $r = 0.083$ before declining. This divergence is consistent with the *Geometric Sharpening* hypothesis: while post-training alignment neutralizes surface expression, it simultaneously sharpens the internal geometric encoding of cultural boundaries.

The practical advantage of this internal sharpening is quantified by the retrieval and lift gaps shown in Table 4. In the GOQA dataset, the performance gap between instruction-tuned and base models (Δ(It-Base, PPD)) reaches **+0.50** for retrieval and **+2.50** for Lift@5. Furthermore, PPD consistently outperforms surface-level baselines, showing a retrieval improvement of **+0.25** over LLM-as-judge in GOQA and **+0.15** over BERTScore Distance (BSD) in WVS. These positive 'Lift@5' gaps, surpassing LLM-as-judge by up to **+1.26**, confirm that latent signals provide a much more reliable anchor for identifying cultural boundaries than decoded, homogenized text. A more comprehensive breakdown of these performance gaps across diverse metrics and country pairs is available in Appendix G.

7. Conclusion

In this work, we introduced Persona Projection Divergence (PPD), a geometric measure that uncovers the latent cultural boundaries within LLMs. Our investigation into the Alignment Paradox reveals that as models are trained to produce more neutral and homogenized text, their internal maps of cultural differences become more distinct and geometrically precise. As a non-redundant diagnostic signal, PPD captures internal value conflicts that remain invisible to surface-level text metrics. This suggests that the internal structures of aligned models are richer and more culturally aware than their outputs lead us to believe. By moving the diagnostic lens to these latent structures, we provide a scalable framework for identifying culturally contentious queries. Through PPD, we move closer to AI systems that do not just hide cultural differences behind a neutral facade, but actively navigate them to provide a more responsible, representative, and culturally sensitive experience for the global user base.

8. Acknowledgments

This work was supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2022-00143911, AI Excellence Global Innovative Leader Education Program); by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B; by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645; by Schmidt Sciences SAFE-AI Grant; by the Frontier Model Forum and AI Safety Fund; by Coefficient Giving; by the Canadian AI Safety Institute Research Program at CIFAR; by the Canadian AI Safety Institute Research Program at CIFAR through a Catalyst Award; by the Survival and Flourishing Fund; and by the Cooperative AI Foundation. The usage of OpenAI credits is largely supported by the Tübingen AI Center and Schmidt Sciences. Resources used in preparing this research project were provided, in part, by

the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute.

References

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31 (3):337–351, 2023. doi: 10.1017/pan.2023.2.

Chen, R., Arditi, A., Sleight, H., Evans, O., and Lindsey, J. Persona vectors: Monitoring and controlling character traits in language models, 2025. URL https://arxiv.org/abs/2507.21509.

DURMUS, E., Nguyen, K., Liao, T., Schiefer, N., Askell, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., Lovitt, L., McCandlish, S., Sikder, O., Tamkin, A., Thamkul, J., Kaplan, J., Clark, J., and Ganguli, D. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=zl16jLb91v.

González Barman, K., Lohse, S., and de Regt, H. W. Reinforcement learning from human feedback in LLMs: Whose culture, whose values, whose perspectives? *Philosophy & Technology*, 38(2):35, 2025. ISSN 2210-5441. doi: 10.1007/s13347-025-00861-0. URL https://doi.org/10.1007/s13347-025-00861-0.

Hershcovich, D., Frank, S., Lent, H., de Lhoneux, M., Abdou, M., Brandl, S., Bugliarello, E., Cabello Piqueras, L., Chalkidis, I., Cui, R., Fierro, C., Margatina, K., Rust, P., and Søgaard, A. Challenges and strategies in cross-cultural NLP. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6997–7013, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.482. URL https://aclanthology.org/2022.acl-long.482/.

Li, B., Haider, S., and Callison-Burch, C. This land is Your, My land: Evaluating geopolitical bias in language models through territorial disputes. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3855–3871, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.213. URL https://aclanthology.org/2024.naacl-long.213/.

Lu, C., Gallagher, J., Michala, J., Fish, K., and Lindsey, J. The assistant axis: Situating and stabilizing the default persona of language models, 2026. URL https://arxiv.org/abs/2601.10387.

Marks, S., Lindsey, J., and Olah, C. The persona selection model: Why AI assistants might behave like humans. https://alignment.anthropic.com/2026/psm/, February 2026. Anthropic Alignment Science Blog.

Nie, S., Omoomi, K., Flek, L., Zhao, Z., and Welch, C. Perspectra: A scalable and configurable pluralist benchmark of perspectives from arguments. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=dyooGJcKJg.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701320. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29971–30004. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

Sorensen, T., Moore, J., Fisher, J., Gordon, M. L., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., and Choi, Y. Position: A roadmap to pluralistic alignment. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46280–46302. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/sorensen24a.html.

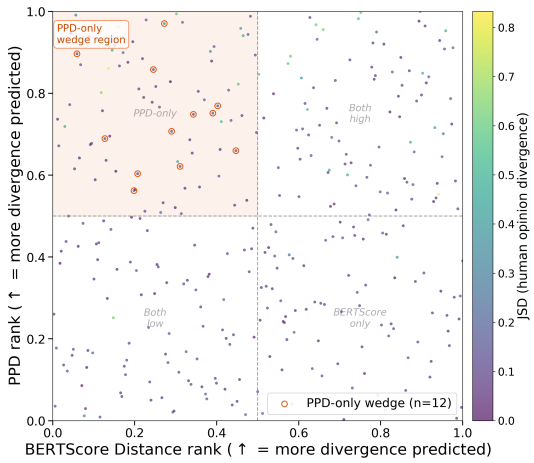
Sorensen, T., Newman, B., Moore, J., Park, C. Y., Fisher, J., Mireshghallah, N., Jiang, L., and Choi, Y. Spectrum tuning: Post-training for distributional coverage and in-context steerability. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ulvp7cbZeU.

Yao, J., Yi, X., Wang, J., Dou, Z., and Xie, X. Caredio: Cultural alignment via representativeness and distinc-

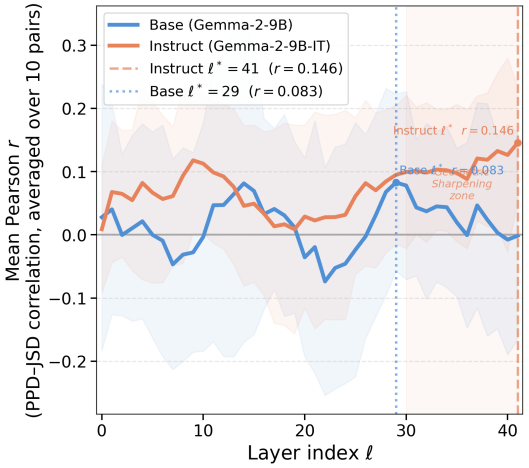
tiveness guided data optimization, 2026. URL https://arxiv.org/abs/2504.08820.

Zhang, Z., Rossi, R. A., Kveton, B., Shao, Y., Yang, D., Zamani, H., Dernoncourt, F., Barrow, J., Yu, T., Kim, S., Zhang, R., Gu, J., Derr, T., Chen, H., Wu, J., Chen, X., Wang, Z., Mitra, S., Lipka, N., Ahmed, N. K., and Wang, Y. Personalization of large language models: A survey. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=tf6A9EYMo6. Survey Certification.

Zhou, N., Bamman, D., and Bleaman, I. L. Culture is not trivia: Sociocultural theory for cultural NLP. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25869–25886, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1256. URL https://aclanthology.org/2025.acl-long.1256/.



(a) PPD vs. BERTScore rank-space diagnostic (GOQA)



(b) Layer-wise correlation between PPD and human JSD

Figure 3. **Alignment evidence: text-level homogenization vs. latent-level sharpening.** (a) Text-level diagnostic shows rank-space divergence between PPD and BERTScore on GOQA. (b) Layer-wise correlation reveals that Instruct models exhibit sharper latent cultural boundaries despite surface-level textual neutralization.

Appendices

Contents

A. Limitations

While PPD demonstrates strong and consistent performance, we acknowledge several directions for future work. Our current evaluation focuses on country-level personas in a single language setting; extending to other demographic axes or multilingual contexts is a natural next step. PPD is used here as a diagnostic tool, and exploring its role as a training signal for culturally-aware alignment remains an exciting direction. Additionally, as wedge questions are pair-specific and cultural divergence is projected onto a single axis, future work could investigate broader country-pair coverage and higher-dimensional representations of cultural variation. Finally, while our results are consistent across multiple rationale generators (LLaMA, Gemma), further analysis of generator-specific effects would be valuable.

B. Discussion: Beyond Detection to Cultural Stewardship

The discovery of the "Alignment Paradox", where post-training alignment sharpens rather than erases latent cultural boundaries, shifts our understanding of model safety from simple output filtering to a more nuanced form of latent cultural stewardship. As established in our introduction, PPD serves as a scalable diagnostic tool; however, the heightened precision of internal cultural encoding opens several transformative avenues for real-world impact and systemic intervention:

Precision Datasets for Targeted Alignment PPD does not merely identify conflict; it serves as a generative seed for high-signal data. By pinpointing "wedge questions" where latent representations are most divergent, researchers can curate specialized datasets. These can be used to fine-tune models to "fit" specific cultural alignments or to "warm up" base models, ensuring they exhibit desired persona vectors with greater stability than surface-level prompting allows.

Dynamic Routing and Distributional Pluralism Our results suggest a path toward *Distributional Pluralism* in AI deployment. Using PPD-detected wedge questions as diagnostic probes, a system can identify when a query touches upon a culturally sensitive boundary. Instead of a homogenized response, the system can route the request to specialized, fine-tuned models that respect the specific cultural context, thereby avoiding the "cultural erasure" often caused by one-size-fits-all alignment.

Structural Necessity for Value Reasoning A critical finding is that base models lack a stable value judgment axis, rendering their PPD signals near-random. This reinforces the idea that alignment may function not only as a constraint but also as a structural scaffold for models to navigate social theory-driven pluralism. Without alignment, a model cannot internally "map" where cultural distinctions lie, making it incapable of participating in nuanced socio-cultural reasoning or identifying the boundaries it operates within.

C. Examples of Datasets

Table 5. **Examples of Datasets.** GOQA response-category distribution and illustrative examples after Pew-only filtering and substantive-option cleaning. Non-Pew rows are excluded; non-substantive answer buckets are removed, and probabilities are renormalized.

K	# survey items	% items	% rows
2	387	48.9	58.2
3	162	20.5	18.1
4	199	25.1	19.5
5	11	1.4	1.7
6	9	1.1	0.7
7	2	0.3	0.2
8	3	0.4	0.1
9	7	0.9	0.6
11	7	0.9	0.6
12	3	0.4	0.2
14	1	0.1	0.0
15	1	0.1	0.0
Total	**792**	**100**	**2779**

QID / country	Question (verbatim)	Options & $\hat{p}$
q_11 Argentina	Working conditions for ordinary workers	K=2: Largely more connected; Other reasons. $\hat{p}$=(0.19, 0.81).
q_1037 Argentina	Post-9/11 interrogation methods	K=3: Justified; Not justified; Depends (VOL). $\hat{p}$=(0.15, 0.79, 0.06).
q_1061 Argentina	Government priority: childhood immunization	K=4: One of the most important priorities; A very important priority; A lower priority; Not a priority at all. $\hat{p}$=(0.72, 0.27, 0.01, 0.00).
q_844 Argentina	Global climate change harming people	K=5: Now; In the next few years; Not for many years; Never; Climate change does not exist (Vol.). $\hat{p}$=(0.79, 0.17, 0.03, 0.01, 0.00).
q_2050 France	Influence on country: Germany	K=6: Very good influence; Mostly good influence; Mostly bad influence; Very bad influence; Neither good or bad (VOL); Both good and bad (VOL). $\hat{p}$=(0.16, 0.70, 0.11, 0.03, 0.00, 0.00).
q_410 Argentina	Importance of being safe from crime (0–10 scale)	K=11: Not important at all; numeric bins 1–9; Very important. $\hat{p}$=(0.00, 0.00, 0.00, 0.00, 0.00, 0.01, 0.01, 0.02, 0.09, 0.10, 0.77).
q_351 Turkey	Which country helps most after natural disasters?	K=15: Sparse multinomial over country/organization labels including Argentina, Australia, Brazil, Britain, China, France, India, Mexico, Russia, Saudi Arabia, Turkey, United Nations, United States, None, Other (VOL); dominant Turkey with $\hat{p}$=0.80, United States 0.07, United Nations 0.04, France 0.01, Other (VOL) 0.07, remainder 0.00.

Table 6. WVS Wave 7 response-category distribution and illustrative examples after removing non-substantive answers and renormalizing probabilities.

K	# survey items	% items	% rows
2	60	22.3	35.5
3	25	9.3	11.4
4	106	39.4	32.0
5	21	7.8	6.2
6	1	0.4	0.5
7	4	1.5	0.6
8	5	1.9	1.1
10	38	14.1	9.6
11	3	1.1	0.2
160	3	1.1	1.8
161	1	0.4	0.5
418	1	0.4	0.4
968	1	0.4	0.2
Total	**269**	**100**	**3307**

ID / country	**Survey item**	**Options & $\hat{p}$**
wvs_Q6 Bangladesh	Important in life: Religion	K=4: Very important; Rather important; Not very important; Not at all important. $\hat{p}$=(0.94, 0.05, 0.00, 0.00).
wvs_Q6 Netherlands	Important in life: Religion	K=4: Very important; Rather important; Not very important; Not at all important. $\hat{p}$=(0.11, 0.12, 0.26, 0.51).
wvs_Q173 Bangladesh	Religious person	K=3: A religious person; Not a religious person; An atheist. $\hat{p}$=(0.98, 0.02, 0.00).
wvs_Q173 New Zealand	Religious person	K=3: A religious person; Not a religious person; An atheist. $\hat{p}$=(0.35, 0.50, 0.14).
wvs_Q10 Brazil	Important child qualities: Feeling of responsibility	K=2: Important; Not mentioned. $\hat{p}$=(0.71, 0.29).
wvs_Q33 Iraq	Jobs scarce: Men should have more right to a job than women	K=5: Agree strongly; Agree; Neither agree nor disagree; Disagree; Disagree strongly. $\hat{p}$=(0.61, 0.17, 0.09, 0.08, 0.04).
wvs_Q209 Ethiopia	Political action: Signing a petition	K=3: Have done; Might do; Would never do. $\hat{p}$=(0.12, 0.21, 0.67).
wvs_Q171 Andorra	How often do you attend religious services	K=7: More than once a week; Once a week; Once a month; Only on special holy days; Once a year; Less often; Never, practically never. $\hat{p}$=(0.02, 0.06, 0.08, 0.10, 0.04, 0.11, 0.59).
wvs_Q190 Germany	Justifiable: Parents beating children	K=10: Never justifiable; 2; 3; 4; 5; 6; 7; 8; 9; Always justifiable. $\hat{p}$=(0.77, 0.10, 0.06, 0.02, 0.03, 0.01, 0.01, 0.00, 0.00, 0.00).
wvs_Q111 Netherlands	Protecting environment vs. Economic growth	K=3: A: Protect environment; B: Economic growth; Other answer. $\hat{p}$=(0.72, 0.28, 0.00).
wvs_Q266 Netherlands	Country of birth: Respondent	K=160 country codes (sparse multinomial). Dominant category: Netherlands with $\hat{p}$=0.91; remaining probability mass is spread across other birth countries.

D. Country pairs with shared questions and split sizes

#	Country pair	Train (1st / 2nd)	Dev	Test
GOQA				
1	Germany – Pakistan	125 / 143	30	31
2	Argentina – Mexico	68 / 55	34	35
3	Egypt – Turkey	29 / 112	32	33
4	Pakistan – Russia	143 / 67	32	32
5	Jordan – Lebanon	76 / 99	34	34
6	Britain – Japan	109 / 81	34	35
7	Canada – France	26 / 145	42	43
8	Italy – Spain	51 / 131	42	43
9	Indonesia – Turkey	59 / 112	28	27
10	Germany – United States	125 / 83	42	41
WVS				
1	Myanmar – Netherlands	42 / 49	66	66
2	Egypt – New Zealand	36 / 66	55	55
3	Libya – South Korea	51 / 53	58	57
4	Germany – Iraq	80 / 35	52	52
5	Ethiopia – Macao SAR	53 / 61	56	57
6	Japan – Jordan	39 / 38	68	69
7	Andorra – Bangladesh	42 / 50	56	56
8	Australia – Pakistan	49 / 53	54	54
9	Tunisia – Vietnam	31 / 63	62	62
10	Brazil – Tajikistan	33 / 39	58	59
Unique countries: 32 (GOQA: 17, WVS: 20, Overlap: 5)				

Table 7. Country pairs with shared questions and split sizes. The dev and Test sets consist of common question-answer pairs that both countries have, and the train set consists of the non-shared question-answer pairs that only one country in the pair has.

E. Rationale Generation Details

We use one user-message template for both polarities (with temperature set to 0.1). For **positive** rationales, `choice` is the country's modal survey response (dominant option text); for **negative** rationales, `choice` is the minority branch option—by default, the lowest-ranked option by reported survey share, or the lowest-ranked substantive option when DK(Don't Know)/non-answer codes are excluded.

Rationale Generation Prompt

How would someone from {country} answer the following question:
{question}
A person from {country} answered: "{choice}"
Write a concise and natural rationale (2–3 sentences) explaining
why someone from {country} would hold this view.

In our default API pipeline we query `meta-llama/llama-3.1-70b-instruct` via OpenRouter with temperature $= 0.1$, a maximum of 160 output tokens, and a single user turn (no system message). The assistant reply is stored verbatim as the rationale text.

F. Full-context LLM Judge Prompt

We use the following template for the full-context LLM judge baseline:

Full-context LLM Judge Prompt

You are evaluating the wedge potential between two countries.
Return only one number from 0 to 100.
Country A: {country_A}
Country B: {country_B}
Question: {question}
A positive rationale: {A_pos}
A negative rationale: {A_neg}
B positive rationale: {B_pos}
B negative rationale: {B_neg}
Score (0–100):

We evaluate three judge models under this setting: `meta-llama/llama-3.1-70b-instruct` (via OpenRouter), `gpt-4o-mini` (via OpenAI API), and `qwen/qwen-2.5-72b-instruct` (via OpenRouter). All models are queried with temperature $= 0.0$ and a maximum of 8 output tokens, with no system prompt. The raw output is post-processed by extracting the first numeric value from the response string, which is used directly as the wedge score.

G. Extended Latent–Human Alignment Analysis

Beyond the ROC-AUC results reported in the main paper, we conduct a broader evaluation of the relationship between latent wedge scores and human disagreement using complementary ranking, retrieval, and dependency metrics.

All evaluations are computed per triple (c_A, c_B, q), where each method produces a continuous wedge score and the target signal is the human Jensen–Shannon Divergence (JSD). Rather than relying on a single criterion, we evaluate whether methods consistently recover human cultural divergence across multiple perspectives: (i) linear and monotonic agreement (Pearson r, Spearman ρ), (ii) general statistical dependence (distance correlation), (iii) pairwise ranking consistency (Concordance Index), and (iv) practical wedge retrieval quality (NDCG, Precision@K, Lift, and tail enrichment).

This expanded evaluation is particularly important under alignment-induced homogenization, where culturally divergent responses may appear superficially similar at the text level. Under this setting, strong retrieval or ranking behavior can emerge even when direct linear correlation remains modest. The results in Table 8 therefore provide a more complete picture of how well latent geometric signals recover human cultural disagreement beyond a single correlation coefficient.

G.1. Metric Definitions

Let each evaluation item be a triple (c_A, c_B, q) with: (i) a method score s_i (higher means more wedge-like), and (ii) a human disagreement target $y_i = \mathrm{JSD}_i$.

For retrieval metrics, *needle* items are defined as the top 20% by y_i.

Pearson Correlation (r) **Concept:** Linear co-movement between method scores and human disagreement.

$$r = \frac{\mathrm{Cov}(s, y)}{\sigma_s \sigma_y}.$$

Range: $[-1, 1]$, where larger positive values indicate stronger linear alignment.

Spearman Rank Correlation (ρ) **Concept:** Monotonic rank agreement between method scores and human disagreement.

$$\rho = \mathrm{Pearson}(\mathrm{rank}(s), \mathrm{rank}(y)).$$

Range: $[-1, 1]$.

Distance Correlation (dCor) **Concept:** General dependence (linear or nonlinear) between s and y. We use the standard sample distance-correlation estimator based on double-centered pairwise distance matrices.

Range: $[0, 1]$ in finite-sample practice, where 0 indicates no detected dependence.

Concordance Index (C-index) **Concept:** Pairwise ordering consistency between method scores and human disagreement.

$$C = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \mathbf{1}[(s_i - s_j)(y_i - y_j) > 0],$$

where $\mathcal{P}$ contains comparable pairs satisfying $y_i \neq y_j$. Score ties are counted as 0.5.

Range: $[0, 1]$, where random ordering is near 0.5.

NDCG@K **Concept:** Top-K ranking quality when relevance is graded by human JSD.

$$\mathrm{DCG@K} = \sum_{t=1}^{K} \frac{2^{y_{\pi_t}} - 1}{\log_2(t+1)}, \qquad \mathrm{NDCG@K} = \frac{\mathrm{DCG@K}}{\mathrm{IDCG@K}}.$$

Range: $[0, 1]$, where higher is better.

Precision@K (P@K) **Concept:** Fraction of top-K scored items that are needles.

$$\mathrm{P@K} = \frac{\#\{\text{needles in top-}K\}}{K}.$$

Range: $[0, 1]$.

Lift@K **Concept:** Improvement over random retrieval at the same needle prevalence p ($p = 0.2$ in our setup).

$$\mathrm{Lift@K} = \frac{\mathrm{P@K}}{p}.$$

Range: $(0, \infty)$, where 1.0 corresponds to random retrieval and values greater than 1 indicate enrichment over random chance.

Tail Enrichment (top-q%) **Concept:** Overlap concentration between the score tail and the JSD tail. Let A_q denote the top-q% by score and B_q denote the top-q% by JSD:

$$\text{Enrich}_q = \frac{|A_q \cap B_q|/N}{(q/100)^2}.$$

Range: $(0, \infty)$, where 1.0 corresponds to independence-level overlap.

G.2. Results and Discussion

Wedge score definitions. All metrics are computed per triple (c_A, c_B, q), where each method produces a continuous wedge score x and the target signal is the human Jensen–Shannon Divergence (JSD) y. The quantity x differs in nature across methods. For **PPD-LW**, x is the absolute difference between the ℓ_2-normalized response activations of the two countries projected onto the cultural contrast axis $\hat{d}_{AB}$ (Eq. 4), aggregated via layer weighting; larger values indicate stronger latent polarization along the value-defining axis. For **BSD**, $x = 1 - F_1^{\text{BERTScore}}$ between the two countries' generated rationales; larger values indicate greater surface-level semantic dissimilarity. For **LLM-as-judge**, x is a 0–100 scalar emitted directly by Llama-3.1-70B-Instruct upon reading all four rationales; larger values indicate stronger perceived wedge intensity. While all three methods share the same ordinal interpretation—higher x implies a stronger wedge candidate—they tap fundamentally different information sources: latent geometric structure, surface text semantics, and explicit model judgment, respectively.

Instruct vs. Base: strong evidence for the Alignment Paradox. The most consistent finding across Table 8 is the sharp performance gap between instruction-tuned and base models under PPD-LW, which constitutes the most direct quantitative evidence for the Alignment Paradox introduced in §2. On GOQA, Gemma-It achieves P@5 = 0.80 and Lift@5 = 4.00×, whereas the corresponding Gemma-Bs entries collapse to P@5 = 0.20 and Lift@5 = 1.00×— statistically indistinguishable from random retrieval. The same trend holds for Pearson r (Gemma-It: $+0.35$ vs. Gemma-Bs: -0.02) and distance correlation (0.33 vs. 0.07). Critically, this degradation is specific to PPD-LW: BSD maintains comparable retrieval scores regardless of alignment status (P@5 = 0.40 in both Gemma-It and Gemma-Bs on GOQA), confirming that the Instruct–Base gap is not a general artifact of the evaluation protocol but a signature of latent geometric sharpening induced by post-training alignment.

PPD-LW vs. LLM-as-judge: a consistent advantage. Across all datasets, model families, and metric types, LLM-as-judge performs at or below random retrieval (Lift@5 $\approx 0.99\times$ throughout), and its Spearman ρ turns negative on WVS ($\rho = -0.16$), indicating systematic misordering of deep structural value questions. The C-index for LLM-as-judge is uniformly 0.50 on GOQA and 0.45 on WVS—at or below chance-level pairwise ordering. By contrast, PPD-LW achieves C = 0.55 on GOQA for both Llama-It and Gemma-It. These results confirm that explicit text-based judgment, even from a large frontier model, fails to recover the latent cultural signal that PPD-LW surfaces through geometric probing.

PPD-LW vs. BSD: complementary signals with distinct strengths. PPD-LW and BSD capture complementary aspects of cultural divergence, with PPD-LW demonstrating clear superiority in the retrieval regime that is most directly relevant to the Wedge-in-the-Haystack task. On GOQA, Gemma-It yields Lift@5 = 4.00× for PPD-LW against 2.00× for BSD, and NDCG@10 of 0.45 vs. 0.36—a consistent margin across all instruction-tuned configurations. This advantage reflects a fundamental difference in what each signal captures: BSD measures global surface dissimilarity and therefore tracks human JSD as a broad tendency, whereas PPD-LW is explicitly constructed to separate the latent value axis most relevant to the country pair, making it more sensitive to the high-JSD tail that defines true wedge questions. The non-redundancy of the two signals—one operating on decoded text, the other on pre-decoded activations—further reinforces that PPD-LW provides diagnostic value inaccessible to any surface-level approach.

Cross-dataset generalization: GOQA vs. WVS. PPD-LW demonstrates strong and consistent performance on GOQA, and maintains above-random retrieval on WVS across all instruction-tuned configurations (Lift@5 = 1.99× for Gemma-It). The more modest global correlation on WVS (Pearson r: 0.04–0.07) is consistent with the nature of WVS items, which probe deep structural values—religiosity, family norms, political trust—that are likely distributed across multiple latent directions simultaneously. The single pair-specific axis $\hat{d}_{AB}$ recovers the diagnostically most useful projection of this richer geometry, a design choice that prioritizes retrieval precision over global rank correlation. Crucially, the Instruct–Base gap is preserved on WVS (Lift@5: 1.99× vs. 0.99× for Gemma), demonstrating that the Alignment Paradox holds robustly across

both geopolitical attitude items and deep structural value questions. The WVS results thus motivate a natural extension of PPD to higher-dimensional cultural subspace modeling, building on the scalar projection framework established here.

Summary. Taken together, the extended metrics in Table 8 reinforce three conclusions. First, the Instruct–Base performance gap is robust across metric types and datasets, providing multi-faceted quantitative support for the Alignment Paradox. Second, PPD-LW consistently and substantially outperforms LLM-as-judge, demonstrating the superiority of latent geometric probing over surface-level text judgment for wedge detection. Third, PPD-LW outperforms BSD in the retrieval regime most relevant to practical wedge identification, while the complementary nature of the two signals underscores the non-redundancy of the latent geometric approach. Across all evaluation axes, the results confirm that probing the latent space before linguistic neutralization occurs is a more reliable path to surfacing cultural boundaries than any text-level diagnostic.

		Instruct Models				Base Models			
		GlobalOpinionQA		WVS		GlobalOpinionQA		WVS	
Metric (range)	Meth.	Llama-It	Gemma-It	Llama-It	Gemma-It	Llama-Bs	Gemma-Bs	Llama-Bs	Gemma-Bs
Needle retrieval (needle = top 20% JSD)									
P@5 $[0, 1]$	PPD-LW	0.60	0.80	0.40	0.40	0.20	0.20	0.20	0.20
P@5 $[0, 1]$	BSD	0.40	0.40	0.20	0.00	0.40	0.40	0.40	0.00
P@5 $[0, 1]$	LLM-as-judge (Llama-3.1-70B-Instruct)	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20
P@10 $[0, 1]$	PPD-LW	0.50	0.70	0.20	0.40	0.10	0.10	0.30	0.20
P@10 $[0, 1]$	BSD	0.40	0.40	0.20	0.00	0.40	0.40	0.50	0.00
P@10 $[0, 1]$	LLM-as-judge (Llama-3.1-70B-Instruct)	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20
P@20 $[0, 1]$	PPD-LW	0.40	0.70	0.20	0.35	0.15	0.15	0.20	0.25
P@20 $[0, 1]$	BSD	0.50	0.50	0.20	0.15	0.50	0.50	0.40	0.15
P@20 $[0, 1]$	LLM-as-judge (Llama-3.1-70B-Instruct)	0.20	0.20	0.25	0.25	0.20	0.20	0.25	0.25
Lift@5 $(0, \infty)$	PPD-LW	3.00×	4.00×	1.99×	1.99×	1.00×	1.00×	0.99×	0.99×
Lift@5 $(0, \infty)$	BSD	2.00×	2.00×	0.99×	0.00×	2.00×	2.00×	1.99×	0.00×
Lift@5 $(0, \infty)$	LLM-as-judge (Llama-3.1-70B-Instruct)	0.99×	0.99×	0.99×	0.99×	0.99×	0.99×	0.99×	0.99×
Lift@10 $(0, \infty)$	PPD-LW	2.50×	3.50×	0.99×	1.99×	0.50×	0.50×	1.49×	0.99×
Lift@10 $(0, \infty)$	BSD	2.00×	2.00×	0.99×	0.00×	2.00×	2.00×	2.49×	0.00×
Lift@10 $(0, \infty)$	LLM-as-judge (Llama-3.1-70B-Instruct)	0.99×	0.99×	0.99×	0.99×	0.99×	0.99×	0.99×	0.99×
Lift@20 $(0, \infty)$	PPD-LW	2.00×	3.50×	0.99×	1.74×	0.75×	0.75×	0.99×	1.24×
Lift@20 $(0, \infty)$	BSD	2.50×	2.50×	0.99×	0.74×	2.50×	2.50×	1.99×	0.74×
Lift@20 $(0, \infty)$	LLM-as-judge (Llama-3.1-70B-Instruct)	0.99×	0.99×	1.24×	1.24×	0.99×	0.99×	1.24×	1.24×
Concordance index									
C $[0, 1]$	PPD-LW	0.55	0.55	0.52	0.52	0.50	0.50	0.49	0.51
C $[0, 1]$	BSD	0.53	0.53	0.51	0.51	0.53	0.53	0.54	0.49
C $[0, 1]$	LLM-as-judge (Llama-3.1-70B-Instruct)	0.50	0.50	0.45	0.45	0.50	0.50	0.45	0.45
NDCG@10									
NDCG@10 $[0, 1]$	PPD-LW	0.43	0.45	0.35	0.40	0.12	0.25	0.36	0.32
NDCG@10 $[0, 1]$	BSD	0.36	0.36	0.34	0.22	0.36	0.36	0.43	0.24
NDCG@10 $[0, 1]$	LLM-as-judge (Llama-3.1-70B-Instruct)	0.11	0.11	0.36	0.36	0.11	0.11	0.36	0.36
Pearson correlation									
r $[-1, 1]$	PPD-LW	0.23	0.35	0.04	0.07	−0.06	−0.02	−0.02	0.03
r $[-1, 1]$	BSD	0.26	0.26	0.03	−0.06	0.26	0.26	0.07	0.02
r $[-1, 1]$	LLM-as-judge (Llama-3.1-70B-Instruct)	−0.02	−0.02	−0.15	−0.15	−0.02	−0.02	−0.15	−0.15
Distance correlation									
dCor $[0, 1]$	PPD-LW	0.21	0.33	0.10	0.09	0.08	0.07	0.07	0.05
dCor $[0, 1]$	BSD	0.27	0.27	0.09	0.12	0.27	0.27	0.11	0.07
dCor $[0, 1]$	LLM-as-judge (Llama-3.1-70B-Instruct)	0.09	0.09	0.20	0.20	0.09	0.09	0.20	0.20
Spearman rank correlation									
ρ $[-1, 1]$	PPD-LW	0.14	0.17	0.07	0.06	−0.00	0.01	−0.02	0.03
ρ $[-1, 1]$	BSD	0.10	0.10	0.03	0.02	0.10	0.10	0.11	−0.03
ρ $[-1, 1]$	LLM-as-judge (Llama-3.1-70B-Instruct)	0.01	0.01	−0.16	−0.16	0.01	0.01	−0.16	−0.16

Table 8. Extended latent–human alignment metrics between method scores and human Jensen–Shannon divergence (JSD), reported across Instruct/Base model families on GlobalOpinionQA (GOQA) and WVS. Metrics are grouped by evaluation type to compare retrieval-focused behavior (needle/top-tail recovery) and global agreement/ranking behavior in one view. Needles are defined as the top 20% of items by human JSD; higher is better for all reported metrics. PPD entries use the layer-weighted leak-free variant (PPD-LW), computed from `exp_layer_weighted_exact.csv`. BSD denotes BERTScore distance ($1-F_1$); for WVS, Gemma BSD entries are computed from the baseline subset (n=177). LLM-as-judge uses `meta-llama/llama-3.1-70b-instruct` with row-level 0–100 wedge scoring; reported coverage is GOQA n=338 and WVS n=209. Abbreviations: PPD-LW = PPD layer-weighted; BSD = BERTScore distance.

H. Robustness of PPD to Question Phrasing and Rationale Sampling

A natural concern is whether PPD scores reflect genuine cultural divergence or are artifacts of a specific question wording or a particular rationale sample. We address this with two perturbation experiments on the GOQA / Gemma-2-9B-IT setting, applied to the 40 target questions: the top-20 PPD questions (*wedge*) and the bottom-20 (*neutral*). All perturbation experiments in this section were conducted on the full GOQA/Gemma-2-9B-IT evaluation set prior to substantive-response filtering, as the robustness checks do not depend on response quality and benefit from the larger sample size.

H.1. Q-Perturb: Sensitivity to Question Phrasing

Setup. For each of the 40 questions we generate three paraphrases using GPT-4o, preserving the original meaning while varying surface expression. We then compute PPD for each paraphrase and report (i) the Spearman rank correlation ρ between the original PPD scores and the mean paraphrase PPD scores across all 40 questions, and (ii) the mean per-question PPD standard deviation across the three paraphrases, separated by group.

Results. As shown in Table 9, the PPD ranking is highly stable under re-phrasing: $\rho = 0.843$ ($p = 9.1 \times 10^{-12}$, $n = 40$). The per-question spread introduced by paraphrasing is modest in both groups (wedge: 1.66, neutral: 1.09), corresponding to roughly 18% and 50% of each group's original mean PPD, respectively.

Group	n	PPD_{orig}	PPD_{para}	Std_{para}	Spearman ρ
Wedge	20	9.25	7.56	1.66	$\mathbf{0.843}^{***}$
Neutral	20	0.88	2.08	1.09	

Table 9. Q-Perturb results (GOQA, Gemma-2-9B-IT). PPD_{orig}: original score; PPD_{para}: mean over 3 GPT-4o paraphrases; Std_{para}: mean per-question standard deviation across paraphrases. $^{***}p < 0.001$.

The strong rank correlation confirms that PPD reliably identifies the same questions as high-divergence or low-divergence regardless of surface phrasing.

H.2. R-Perturb: Sensitivity to Rationale Sampling

Setup. For the same 40 questions we regenerate rationales five times using Llama-3-70B-Instruct at temperature $\tau = 0.7$ (versus $\tau = 0.1$ used for the main experiments) via the OpenRouter API. For each sample we build a question-local pair-direction vector from the Gemma-2-9B-IT activations of the new rationales and compute PPD against the original test activations. We report the mean and standard deviation of PPD across the five samples per question, and test whether PPD variance differs between wedge and neutral groups (one-sided Mann–Whitney U test: H_1: neutral std $>$ wedge std).

Results. Table 10 shows that PPD is similarly stable in both groups across five independent rationale samples. The per-question standard deviation is low for both wedge (1.28) and neutral (1.23) questions, and the Mann–Whitney U test finds no significant difference ($U = 164$, $p = 0.838$).

Group	n	Mean PPD	Mean Std (5 samples)
Wedge	20	12.83	1.28
Neutral	20	5.16	1.23
Mann–Whitney U (neutral std $>$ wedge std): $U = 164$, $p = 0.838$			

Table 10. R-Perturb results using Llama-3-70B-Instruct ($\tau = 0.7$, 5 samples). Both groups show similarly low PPD variance under repeated rationale sampling.

The near-chance p-value indicates that the low variance is not specific to wedge questions—*PPD is stable for all questions*, making it unlikely that the wedge/neutral distinction is driven by sampling noise.

H.3. Qwen2.5-72B Rationale Generator Ablation

Setup. To assess whether PPD depends on the specific rationale generation model, we repeat R-Perturb with Qwen2.5-72B-Instruct (also via OpenRouter, $\tau = 0.7$, 5 samples) in place of Llama-3-70B-Instruct. All other settings are identical.

Results. As shown in Table 11, substituting the rationale generator with a different 70B-class model produces qualitatively identical conclusions. PPD variance remains low in both groups (wedge: 1.12, neutral: 1.22), and the between-group difference is again non-significant ($U = 198$, $p = 0.527$). Crucially, the wedge/neutral mean PPD gap is preserved (11.87 vs. 4.44), confirming that the identity of the high-divergence questions does not depend on the choice of rationale generator.

Group	n	**Mean PPD**	**Mean Std (5 samples)**
Wedge	20	11.87	1.12
Neutral	20	4.44	1.22
Mann–Whitney U (neutral std > wedge std): $U = 198$, $p = 0.527$			

Table 11. Qwen2.5-72B ablation results ($\tau = 0.7$, 5 samples). Results mirror those obtained with Llama-3-70B-Instruct, confirming that PPD is not sensitive to the choice of rationale generator.

H.4. Summary

Table 12 consolidates the three robustness checks. Taken together, they show that PPD scores are stable across (i) question re-phrasing, (ii) repeated rationale sampling at higher temperature, and (iii) a fully different rationale generation model. These results support the interpretation of PPD as a genuine signal of cultural opinion divergence rather than an artifact of any single design choice.

Perturbation	**Perturbation Axis**	**Metric**	**Value**
Q-Perturb	Question phrasing (GPT-4o, $k = 3$)	Spearman ρ	0.843^{***}
R-Perturb (Llama)	Rationale sampling (τ=0.7, $k = 5$)	MWU p (variance diff.)	0.838
R-Perturb (Qwen)	Rationale generator + sampling ($k = 5$)	MWU p (variance diff.)	0.527

Table 12. Summary of perturbation robustness checks. All three axes of variation leave PPD rankings and variances stable. $^{***}p < 0.001$ (Spearman ρ); large MWU p-values indicate no significant difference in variance between wedge and neutral groups.

I. Statistical significance and uncertainty

Uncertainty and scope. Unless noted otherwise, **95% confidence intervals** bracket Pearson correlation between the method score and human disagreement (JSD) under **bootstrap resampling of question–pair rows** (fixed model and split). Intervals capture sampling variability over the evaluated items, not initialization or rationale-generation stochasticity. Two-tailed p**-values** use the standard t-test for Pearson r with $n-2$ degrees of freedom on the same rows.

Dataset	**Method**	n	r	95% CI	p
GOQA (substantive test)	α+Layer PPD	338	$+0.277$	$[+0.154, +0.401]$	$< 10^{-6}$
GOQA (substantive test)	Rationale BERTScore	338	$+0.212$	—†	$< 10^{-4}$
WVS	α+Layer PPD	177	$+0.244$	$[+0.091, +0.380]$	0.0011
WVS	Rationale BERTScore	209	$+0.219$	—†	0.0015

Table 13. **Correlation report** for the primary instruction-tuned model (**Gemma-2-9B-Instruct**) on held-out test rows. GOQA denotes GlobalOpinionQA substantive evaluation (n=338). p is two-tailed for $H_0\colon \rho{=}0$. †Bootstrap CI not computed from bundled per-row scores for NL baselines (point r and p match released summaries). The WVS row count differs between PPD and Rationale BERTScore because the released aggregation pipelines join a different number of valid items for those scores (177 vs. 209); interpret each row at its stated n.

Cross-architecture robustness. Llama-3.1-8B-Instruct on GlobalOpinionQA (n=338): α+Layer PPD $r = +0.266$, 95% CI $[+0.127, +0.401]$, $p<10^{-6}$. On WVS the same checkpoint evaluates α+Layer PPD on n=587 merged rows: $r=+0.143$, 95% CI $[+0.055, +0.231]$, $p=4.9\times10^{-4}$.

Limitation. Some base checkpoints yield correlations whose 95% bootstrap CIs include zero (e.g., Gemma-2-9B-Base, GOQA, α+Layer PPD: $r = +0.044$, CI about $[-0.043, +0.154]$, $p\approx0.42$); primary claims are not rested on those cells.

J. Experimental setup

Table 14. Experimental setup of the PPD framework.

Category	Detail
Datasets	GlobalOpinionQA (GOQA) and World Values Survey Wave 7 (WVS); 10 country pairs per dataset
Row counts	GOQA train / dev / test: 1,762 / 827 / 829; WVS: 963 / 1,170 / 1,174
Data splits	Pair-aware and leak-free; persona vectors constructed from *train_nonshared* questions only; shared questions divided 50% / 50% into dev and test (seed 42)
Rationale generation	Llama-3.1-70B-Instruct via OpenRouter; temperature 0.1, max 160 tokens (Appendix E)
Hidden-state pooling	Mean-pooling over all token positions in the rationale span R (Eq. 2)
PPD backbone	Gemma-2-9B-Instruct (main), Llama-3.1-8B-Instruct (additional); `bfloat16`, HuggingFace `transformers`; layers $\ell = 1 \ldots 41$ (Gemma) and $\ell = 1 \ldots 32$ (Llama)
α tuning	$\alpha \in \{0,\ 0.25,\ 0.5,\ 0.75,\ 1\}$; coordinate ascent on dev set (4 rounds)
Golden layer ℓ^*	$\arg\max_\ell\ \mathrm{Pearson}(\mathrm{PPD}^{(\ell)},\ \mathrm{JSD})$ evaluated on dev split only; $\ell^* = 41$ for Gemma-It on GOQA
Baselines	BERTScore F_1 on positive rationales; LLM Judge (Llama-3.1-70B-Instruct, Qwen2.5-72B, GPT-4o-mini; temperature 0.0, greedy)
Evaluation metrics	Pearson and Spearman correlation with JSD; AUC-ROC (median-JSD binarisation); Precision@K for $K \in \{10,\ 20\}$
Hardware	NVIDIA A100 80 GB on a SLURM cluster; `bfloat16` precision; approx. 4–6 h per dataset $\times$ model configuration
Compute	Approx. 60–80 GPU-h in total across all reported configurations (2 models $\times$ 2 types $\times$ 2 datasets plus ablations); approx. 15–25 GPU-h for preliminary runs. Rationale generation and LLM-Judge queries via OpenRouter API
Reproducibility	`wedge_questions/README.md`, `requirements-reproduction.txt`, and full scripts in supplemental material