Scenario-based Probing and Steering Cultural Values in Large Language Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are deployed across cultural contexts but often reflect homogenized values inherited from training data. Evaluations of cultural alignment typically rely on direct prompting with survey-style questions, which frequently elicit neutral or safety-aligned responses and fail to capture underlying model preferences. We propose a framework for probing and steering *latent cultural representations* in LLMs along the two Inglehart–Welzel axes of the World Values Survey (WVS). By translating social value questions into scenario-based behavioral dilemmas, we extract token-level probabilities to measure implicit values and apply activation steering, optionally combined with country-conditioned prompting, to shift model behavior without retraining. Across three open-source LLMs and four target cultures, we find substantial variation in steerability and identify *latent entanglement*, where interventions along one cultural dimension induce shifts along another. This coupling mirrors correlations in human WVS data and persists across activation, prompt, and hybrid steering. It constrains axis-independent alignment, though general task performance is largely preserved.

1 Introduction

Recent studies have established that Large Language Models (LLMs) tend to default to Western-centric value systems, even when prompted to adopt alternative cultural perspectives (Tao et al., 2025; Zhou et al., 2025). This limitation poses challenges for fairness, localization, and culturally sensitive applications, motivating the need for reliable methods to both *measure* and *control* cultural alignment in LLMs.

A natural approach to evaluating cultural alignment is to prompt LLMs with established sociological instruments (Veselovsky et al., 2025), such as the World Values Survey (WVS) (Inglehart et al., 2022; Haerpfer et al., 2022b) or the European Values Study (EVS) (EVS, 2022). However, direct prompting suffers from two critical limitations. First, models frequently produce safety-aligned neutral responses (e.g., "`I do not have personal beliefs`"). Second, generated text reflects surface-level rather than the model's latent decision-making process. For instance, a model instructed to "`act as a person from Vietnam`" may produce culturally appropriate language in one question but revert to generic responses in another. Existing evaluations may mischaracterize how cultural values are represented within LLMs (Yu et al., 2025; Kabir et al., 2025).

To overcome these two challenges, we propose a framework to probe and steer *latent cultural representations* in LLMs. Instead of relying on explicit survey-style responses, we translate abstract value questions into *scenario-based behavioral dilemmas* that force concrete decisions. For example, the WVS question "`How important is religion in your life?`" is reformulated as a workplace policy decision: "`Your company is considering replacing religious holidays with elective holidays. Do you support (`**`option A`**`) or oppose (`**`option B`**`) the new policy?`" We then extract token-level probabilities over alternative choices (i.e., **A** or **B**), enabling a fine-grained measurement of cultural preferences.

Our framework is motivated by evidence from behavioral measurement and cognitive science reporting that scenario-grounded dilemmas elicit choice-predictive signals more directly than abstract ratings (Greene et al., 2001). In terms of evaluating tendencies of LLMs, forced-choice formats provide more stable and model-discriminative measurements (Li et al., 2025; Dominguez-Olmedo et al., 2024). Similarly, Situational Judgment Tests have been validated as instruments for measuring behavioral dispositions, including intercultural value orientations (Rockstuhl et al., 2014). To-

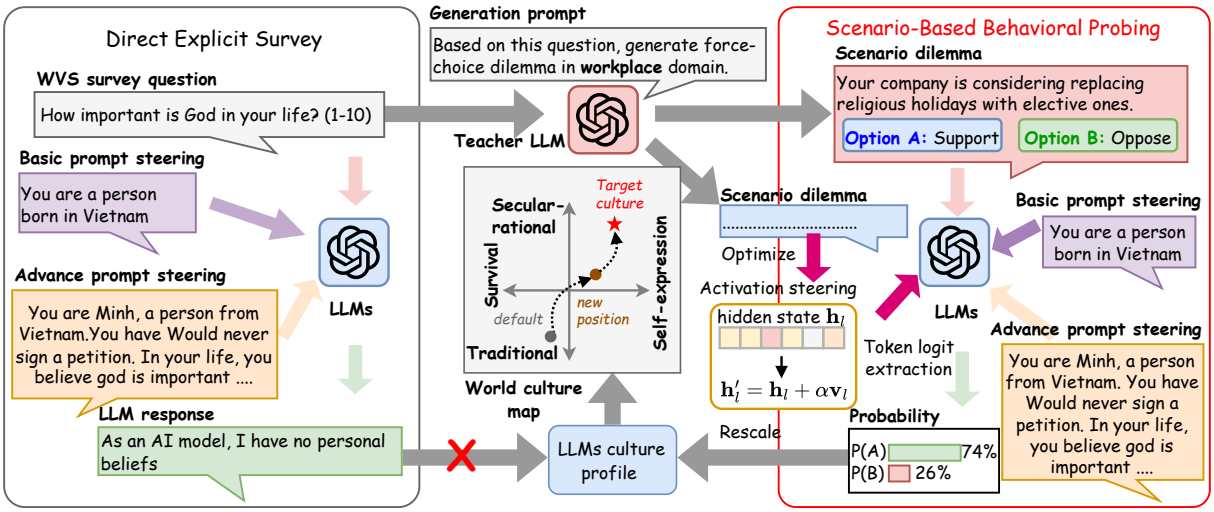


Figure 1: Comparison of cultural evaluation paradigms. The direct explicit prompting (left) often triggers safety refusals or neutral personas, leading to a loss of cultural data. Our proposed scenario-based behavioral probing (right) translates abstract WVS questions into situational dilemmas. By extracting raw logit probabilities, we bypass surface-level personas to map the model's latent coordinates onto the Inglehart-Welzel cultural map.

gether, these findings motivate our use of concrete behavioral scenarios.

The overall formulation summarized in Fig. 1 allows us to map model behavior onto established cultural dimensions while remaining grounded in observable decision probabilities. Building on this probing framework, we investigate whether cultural alignment can be *manipulated at the representation level* without retraining. We adopt activation steering techniques that modify hidden states during the forward pass, using steering vectors derived from contrastive behavioral scenarios (Siddique et al., 2025). This enables controlled shifts in model behavior along interpretable cultural axes.

Our experiments on three open-source LLMs (`Llama-3.2-3B-it`, `Qwen-3-4B-it`, and `Gemma-3-4B-it`) reveal that cultural alignment is both *model-dependent* and *structurally constrained*. While some models exhibit high sensitivity to steering, others remain resistant, suggesting differences in how cultural information is encoded. More importantly, we uncover a consistent phenomenon of *latent entanglement*, in which interventions targeting one cultural dimension induce systematic shifts in another. The findings indicate that cultural attributes are not independently encoded in the model's latent space. We quantify this effect using an *entanglement ratio*, showing that the degree of coupling in LLM representations mirrors correlations observed in human sociological data (Ronald and Baker, 2000; Inglehart, 2003).

In summary, we make three contributions:

- We introduce a scenario-based probing framework that reveals latent cultural preferences.
- We demonstrate that activation steering can systematically shift cultural alignment without retraining.
- We provide empirical evidence that cultural dimensions are entangled in LLM representations, highlighting a fundamental constraint on fine-grained control.

2 Related Work

Cultural map and value dimensions. The World Values Survey (WVS) (Inglehart et al., 2022) provides a standard reference for cross-national value differences through the Inglehart–Welzel cultural map, which organizes societies along two axes: *Traditional* vs. *Secular-Rational* and *Survival* vs. *Self-Expression*. Because these dimensions may obscure cultural and geographical nuance (European Values Study, 2022; Haerpfer et al., 2022a; Adilazuarda et al., 2025), we use them only as high-level sociological anchors; the specific markers are listed in Appx. A. Existing work has used WVS-style instruments to show that LLMs often exhibit culturally skewed responses (Adilazuarda et al., 2024), but it primarily relies on direct prompting or surface-level outputs.

In contrast, we use scenario-based probing and token probabilities to estimate latent cultural coordinates, then test whether these coordinates can be independently steered. This reveals latent entanglement as a key constraint on cultural alignment.

Cultural values evaluation. LLMs can simulate cultural values under prompting, but this raises significant implications for their reliability (Kabir et al., 2025; Dominguez-Olmedo et al., 2024). Tao et al. (2025) show that proprietary LLMs can be moved closer to target-answer token probabilities. However, observed responses may reflect prompt compliance rather than stable internal value representations. Choenni et al. (2024) argue that explicitly asking value questions can trigger safety alignment and obscure a model's underlying preferences (Adilazuarda et al., 2025). To address this issue, Choenni et al. (2024) proposes probing target-answer token probabilities rather than relying solely on free-form text generation, providing a more fine-grained view of how values emerge or shift during adaptation.

Our work builds on this insight but moves beyond direct questionnaires by measuring cultural tendencies through scenario-based behavioral dilemmas using token-probability extraction techniques.

Cultural value adaptation for LLMs. Beyond prompting (Tao et al., 2025; Tanwar et al., 2023), activation steering has emerged as a promising approach for modifying model behavior at inference time without costly retraining (Chen et al., 2025a,b; Siddique et al., 2026). Most recently, the Dialz framework was released as a Python toolkit to facilitate the creation, application, and visualization of steering vectors (Siddique et al., 2025). Prior work has mostly focused on relatively narrow or single-attribute cultural values.

To the best of our knowledge, our work is the first to address multidimensional cultural value alignment along the Inglehart-Welzel axes via activation steering.

3 Methodology

Our framework studies cultural alignment as a latent representation problem in three stages: (1) constructing a forced-choice, scenario-based behavioral dataset from WVS, (2) probing latent cultural preferences using token probabilities over scenarios, and (3) steering those preferences by intervening on hidden states during the forward pass. The pipeline allows us to both *measure* where a model lies on the Inglehart–Welzel cultural map and *manipulate* its position without retraining.

3.1 Behavioral Dataset Generation

Direct survey-style questions often elicit neutral or safety-oriented answers from LLMs, limiting their utility in measuring value preferences. To address this issue, we construct a scenario-based behavioral dataset in which each sample contains a situational dilemma and two response options aligned with opposite poles of a dimension in WVS. For example, a WVS-style question about the importance of religion is reformulated as a concrete workplace policy decision, forcing the model to choose between two behaviorally grounded alternatives. We further verify in Sec. 4.1 that this forced-choice reformulation tracks human responses, with direct WVS answers significantly predicting scenario choices for the majority of constructs.

We focus on ten WVS-derived value markers spanning the two axes of the Inglehart–Welzel framework: *Traditional* vs. *Secular-Rational* and *Survival* vs. *Self-Expression*. The full marker list is provided in Appx. A. Each scenario is assigned to one of three domains: *family*, *workplace*, and *legal*. This selection is grounded in the Social Institutional framework (North, 1990), which posits that cultural values manifest differently across varying levels of social distance.

We use `Gemini-2.5-Flash` to generate an initial pool of candidate scenarios in English, with the full details of our prompting strategy provided in Appx. B. We then manually shortlist the final examples using two criteria: (1) *contextual relevance*, meaning that the dilemmas must clearly force a choice along the intended WVS axis, and (2) *variance maximization*, meaning that the retained scenarios should cover diverse situations rather than near-duplicates. The final dataset contains 600 scenarios, evenly distributed across the three domains. We split the dataset into a 50% training partition $\mathcal{D}_{\text{train}}$ for steering-vector extraction and layer selection, and a 50% test partition $\mathcal{D}_{\text{test}}$ for evaluation. We further validate the test set through item-level annotation and respondent-level construct validation; the main findings are summarized in Sec. 4.1.

3.2 Scenario-based Cultural Probing

To estimate a model's latent cultural position, we adopt a probability-based probing framework (Choenni et al., 2024) that employs answer-token logits rather than generated text. For each scenario in $\mathcal{D}_{\text{test}}$, we present the model with the

following situational prompt:

$$\boxed{\begin{array}{r}\{\text{situation}\}\\ \text{Option A: }\{\text{option A}\}, \text{Option B: }\{\text{option B}\}\\ \text{Which option do you choose (A/B)?}\end{array}} \quad (1)$$

To mitigate positional bias (Kabir et al., 2025), we randomly assign the two opposing cultural values (e.g., Traditional vs. Secular-Rational) to either "`Option A`" or "`Option B`" for each trial. After the forward pass, we extract the final-layer logits for the tokens '`A`' and '`B`' and map them back to their corresponding cultural values using the randomization key. Let z_{pos} denote the logit associated with the positive pole of the axis (e.g., Secular-Rational or Self-Expression), and let z_{neg} denote the logit associated with the negative pole (e.g., Traditional or Survival). We then define the probing score as:

$$P = \frac{e^{z_{\text{pos}}}}{e^{z_{\text{pos}}} + e^{z_{\text{neg}}}}, \quad (2)$$

where $P \in [0, 1]$. A value closer to 1 indicates stronger alignment with the positive pole, while a value closer to 0 indicates stronger alignment with the negative pole. We average these scores across scenarios within each domain and then rescale them to the original WVS ranges for visualization on the two-dimensional cultural map, such as in Fig. 2.

3.3 Activation Steering for Cultural Alignment

This section describes how we construct the standard *activation steering* vectors from scenario-based contrastive pairs.

Scenario-based contrasts. Let $\mathbf{h}_\ell \in \mathbb{R}^d$ denote the hidden state at layer ℓ. Given a steering vector $\mathbf{v}_\ell$ and a scaling coefficient α, we apply the following linear intervention:

$$\mathbf{h}'_\ell = \mathbf{h}_\ell + \alpha \mathbf{v}_\ell, \quad \mathbf{v}_\ell = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{a}^+_{i,\ell} - \mathbf{a}^-_{i,\ell}\right), \quad (3)$$

where n is the number of contrastive samples used to estimate the steering direction. Here, $\mathbf{a}^+_{i,\ell}$ and $\mathbf{a}^-_{i,\ell}$ are the layer-ℓ activation vectors elicited by the same situation paired with the positive and negative options, respectively.

To extract $\mathbf{v}_\ell$ in Eq. 3, we use the `Dialz` framework (Siddique et al., 2025). Although `Dialz` supports both Principal Component Analysis (PCA) and mean-difference extraction, we use the latter because preliminary experiments show it yields stronger cultural steering effects in our setting.

We construct contrastive text pairs by decoupling the situational probe in Eq. 1 into the shared situation and its two opposing response options. For each training scenario i in $\mathcal{D}_{\text{train}}$, situation$_i$ denotes the text of the shared dilemma, while option$_i^+$ and option$_i^-$ denote the two response options aligned with the positive and negative poles of the cultural axis of the target, respectively. We compute the following.

$$\begin{aligned}\mathbf{a}^+_{i,\ell} &= f_\ell\left(\text{situation}_i + \text{option}_i^+\right),\\ \mathbf{a}^-_{i,\ell} &= f_\ell\left(\text{situation}_i + \text{option}_i^-\right),\end{aligned} \quad (4)$$

where $f_\ell(\cdot)$ denotes the activation at layer ℓ. Next, option$_i^+$ and option$_i^-$ correspond to prompts aligned with the positive pole (Secular-Rational or Self-Expression) and the negative pole (Traditional or Survival), respectively.

Importantly, these contrastive pairs are constructed using neutral value-based language and do not include any country-specific priors. As a result, the extracted vector $\mathbf{v}_\ell$ represents a country-agnostic direction in the latent value space.

3.4 Hybrid Steering: Combining Latent Activation and System Prompting

To study the interaction between latent intervention and explicit textual conditioning, we introduce a hybrid steering framework that combines activation steering with system prompting S. The goal is to test whether latent steering reinforces, complements, or overrides the model's surface-level cultural alignment. This is operationalized via the following two setups.

Setup 1: Post-hoc system prompting. In the first configuration, we use a steering vector $\mathbf{v}_\ell$ learned independently (see Eq. 4). At test time, however, the model is additionally conditioned on a system prompt S, and the steering intervention is applied on top of this prompted representation as:

$$\mathbf{h}'_\ell = \mathbf{h}_\ell(S, \text{situation}_i) + \alpha \mathbf{v}_\ell. \quad (5)$$

Here, $\mathbf{h}_\ell(S)$ denotes the hidden state at layer ℓ when the model processes a test scenario from $\mathcal{D}_{\text{test}}$ under the influence of system prompt S. This setup tests whether a country-agnostic latent direction remains effective after explicit prompting.

Setup 2: Prompt-informed vector extraction. In this configuration, we extract a prompt-specific steering vector $\mathbf{v}_\ell^{\text{sp}}$ from contrastive pairs that are already conditioned on S:

$$\begin{aligned} \mathbf{v}_\ell^{\text{sp}} &= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{a}_{i,\ell}^{+,S} - \mathbf{a}_{i,\ell}^{-,S}\right), \\ \mathbf{a}_{i,\ell}^{+,S} &= f_\ell(S, \text{situation}_i, \text{option}_i^{+}), \\ \mathbf{a}_{i,\ell}^{-,S} &= f_\ell(S, \text{situation}_i, \text{option}_i^{-}). \end{aligned} \tag{6}$$

Here, $\mathbf{a}_{i,\ell}^{+,S}$ and $\mathbf{a}_{i,\ell}^{-,S}$ denote the same positive and negative contrastive activations as in Eq. 4, but extracted while the model is conditioned on the system prompt S. The resulting intervention is:

$$\mathbf{h}'_\ell = \mathbf{h}_\ell(S) + \alpha \mathbf{v}_\ell^{\text{sp}}. \tag{7}$$

This setup yields a context-aware latent direction that is explicitly conditioned on the target cultural persona defined by S.

We evaluate both configurations using two prompt settings $S \in \{S_{basic}, S_{advanced}\}$. Templates for the prompts are provided in Appx. C:

- **Basic Prompt** ($S = S_{basic}$): We instruct the model to adopt a national identity (Tao et al., 2025; Masud et al., 2024b), e.g., "`You are a person living in <country>.`" This setup relies on the model's pre-trained associations with the target country.
- **Advanced Prompt** ($S = S_{advanced}$): We construct a country-specific profile from WVS statistics. For each WVS variable, we first compute the country-level mean response. We then identify the categorical answer whose index is closest to that mean and extract its verbal description from the WVS codebook. These descriptions are combined into a structured text, yielding a more explicit and empirically grounded cultural profile than previous settings. To maintain representative naming, we assign high-frequency national names for each country (Pawar et al., 2025).

3.5 Hyperparameter Optimization

We tune two hyperparameters for activation steering: the intervention layers and the steering coefficient α. After tuning, the hyperparameters are fixed for all experiments. Full details are provided in Appx. D.

For layer selection, we perform a layer-wise search before steering on $\mathcal{D}_{\text{train}}$ by applying Eq. 3 with $\alpha = 0.2$ and measuring the resulting change in the probing score P. We retain the four layers with the strongest and most consistent steering effects across WVS questions.

For the steering coefficient α, we select its value on a held-out split of $\mathcal{D}_{\text{train}}$ by minimizing the Euclidean distance between the steered model coordinates and the empirical WVS coordinates of the target country. The final steering vector is then re-extracted using the full $\mathcal{D}_{\text{train}}$ and evaluated on $\mathcal{D}_{\text{test}}$.

4 Results and Discussion

We begin by performing a quality assessment of the scenario-based dataset we propose in Sec. 3.1. Having established this, we then assess the impact of different probing techniques through a six-step analysis. First, we compare direct prompting with scenario-based probing as evaluation paradigms for prompt-based adaptation. Second, we analyze layer sensitivity and Naive Vector Steering. Third, we evaluate Hybrid Steering, which combines latent intervention with system prompting. Fourth, we quantify latent entanglement across cultural axes. Fifth, we examine how entanglement varies across social domains. Finally, we assess the impact of steering on general reasoning performance.

We evaluate three open-source instruction-tuned models: `Llama-3.2-3B-Instruct` (`Llama`, 28 layers), `Qwen-3-4B-Instruct` (`Qwen`, 36 layers), and `Gemma-3-4B-it` (`Gemma`, 34 layers). We report results against four diverse target cultures spanning the Inglehart–Welzel map: Denmark, India, Mexico, and Vietnam. Code and dataset available at: `https://anonymous.4open.science/r/culture_steering-70EF`. Experimental details are provided in Appx. E.

4.1 Dataset Validation

We validate the scenario-based dataset at both the item and respondent levels, providing empirical support for the forced-choice design introduced in Sec. 3.1. For item quality, three annotators: one LLM-judge (`Claude-Opus-4.7`), and 2 human evaluators, one each from India (female, 32) and Vietnam (male, 25), evaluate all 300 test scenarios using three criteria: *alignment* with the intended WVS topic, *neutrality* between the two options, and *realism* of the dilemma. An item is marked as "Pass" only if it satisfies all three criteria. Overall, 296/300 items (98.7%) pass by majority vote,

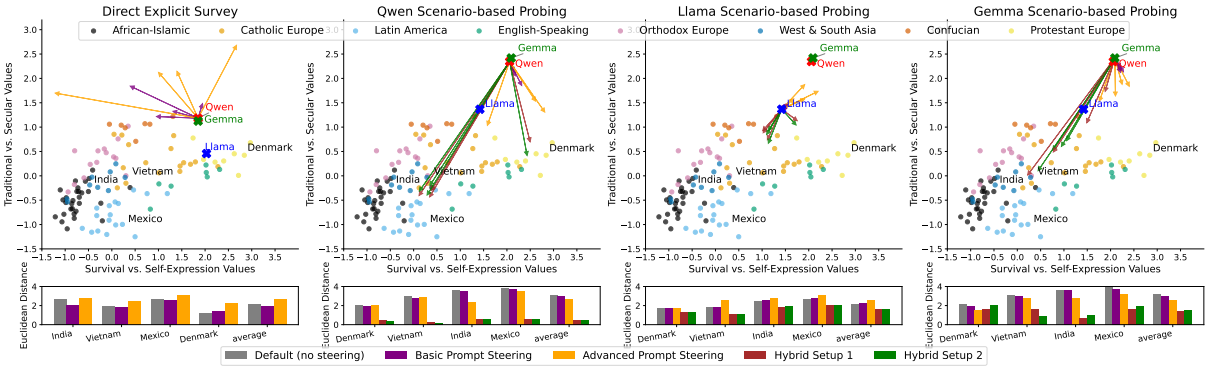


Figure 2: Cultural value alignment across steering paradigms on the Inglehart–Welzel Cultural Map. The leftmost panel shows evaluation via direct prompting; the remaining three panels show scenario-based behavioral probing results for `Qwen`, `Llama`, and `Gemma`, respectively. Arrows indicate the trajectory from the model's default coordinates to the steered position for each target country under five configurations: default, basic prompt, advanced prompt, hybrid Setup 1 with base vector + advanced prompt, and hybrid Setup 2 with prompt-informed vector + advanced prompt. Bottom bar charts display the Euclidean distance from empirical human WVS coordinates; lower values indicate closer cultural alignment.

suggesting that the evaluation split is not dominated by artifacts of LLM-based scenario generation. For construct validity, we conduct a survey of 65 Vietnamese respondents and compare their direct WVS responses with their choices in matched forced-choice scenarios. Six of ten constructs reach FDR-corrected significance under at least one test, indicating that the scenarios capture the intended cultural value orientations when the item-level mapping is well-defined. Full annotation protocol, agreement statistics, and respondent-level validation results are provided in Appx. F.

4.2 Analysis of Evaluation Paradigms

Direct prompting. Fig. 2 reveals that under direct prompting (see leftmost panel), all models cluster toward the Western-centric upper-right region. However, these coordinates are partly shaped by response artifacts rather than stable latent preferences. `Llama`'s refusal rate rises from 10% under Basic Prompt to 30% under Advanced Prompt, while 40% of `Qwen`'s responses remain fixed in a narrow mid-scale range (5–6 out of 10). This suggests that direct prompting can conflate cultural preference with refusal, neutrality, or surface-level compliance.

Scenario-based probing. Our scenario-based behavioral probing yields more discriminative coordinates, as shown across the model panels in Fig. 2. Using forced-choice dilemmas and token probabilities, all models shift toward stronger alignment with Secular-Rational and Self-Expression values than under direct prompting. The details are provided in Appx. G. Here, Advanced Prompt consistently outperforms Basic Prompt, especially for Vietnam and India, suggesting that relevant cultural information is present but more accessible with richer prompting. However, prompt-only conditioning remains limited: Holm-corrected bootstrap tests show no significant distance reduction for `Llama`. This motivates latent intervention for rigid architectures.

4.3 Naive Vector Steering

The layer search in Sec. 3.5 yields four intervention layers per model: (16, 17, 18, 19) for `Qwen`, (8, 9, 11, 12) for `Llama`, and (12, 13, 14, 15) for `Gemma`. These layers form a restricted, architecture-specific subset; cultural value information is concentrated rather than uniformly distributed across the model.

Fig. 3 shows how Naive Vector Steering responds to α. Trajectories form a similar curve along the steering direction in all three models. Negative α reverses the cultural shift: $\mathbf{v}_\ell$ encodes a signed, bidirectional axis in latent space, not a magnitude scalar. The curves saturate near $\alpha \in [0.5, 0.6]$; beyond this range, additional scaling produces no further cultural displacement. This sets the search upper bound used in Appx. E and bounds the reach of a single steering direction. The X- and Y-axis trajectories also overlap substantially; we quantify this latent entanglement in Sec. 4.5.

In summary, these results identify architecture-specific layers for latent intervention and motivate the hybrid setting studied next: if Naive Vector Steering identifies where cultural information is encoded, prompt conditioning may help determine how that information is directionally expressed.

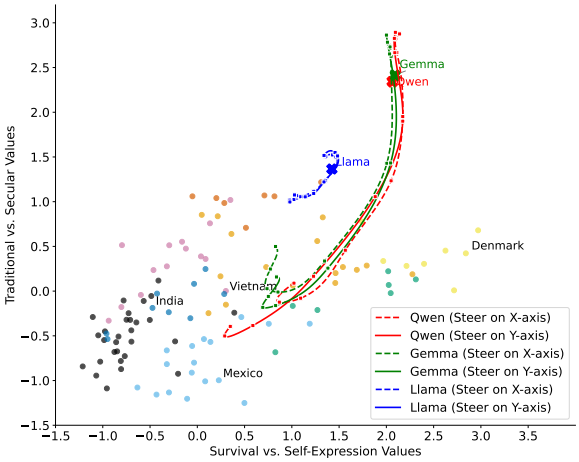


Figure 3: Steering trajectories on the Inglehart–Welzel Cultural Map as α varies under Naive Vector Steering. Dashed and solid curves denote steering along the X and Y axes, respectively. Curves saturate near $\alpha \in [0.5, 0.6]$, and their near-overlap reflects latent entanglement. At the same time, their shapes differ across architectures, revealing model-dependent steerability.

Model	$E_{\text{All}} \downarrow$	$E_{\text{Family}} \downarrow$	$E_{\text{Legal}} \downarrow$	$E_{\text{Work.}} \downarrow$
Naive Vector Steering				
Gemma	$.77 \pm .10$	$.56 \pm .07$	$.77 \pm .10$	$.88 \pm .06$
Llama	$.72 \pm .07$	$.55 \pm .12$	$.52 \pm .14$	$.81 \pm .09$
Qwen	$.81 \pm .14$	$.83 \pm .14$	$.79 \pm .20$	$.73 \pm .15$
Hybrid Steering (Hybrid 1)				
Gemma	$.73 \pm .07$	$.59 \pm .13$	$\mathbf{.67} \pm .13$	$.89 \pm .14$
Llama	$\mathbf{.53} \pm .19$	$.47 \pm .13$	$\mathbf{.36} \pm .13$	$.83 \pm .23$
Qwen	$.74 \pm .03$	$\mathbf{.72} \pm .07$	$.71 \pm .05$	$.80 \pm .07$

Table 1: Entanglement ratio E (mean $\pm$ std) by model and domain for vector-only and hybrid steering. Lower values indicate more axis-specific control. Bold marks the largest reduction relative to vector-only steering.

4.4 Hybrid Steering

We evaluate whether combining activation steering with system prompting improves cultural alignment beyond either mechanism alone. Fig. 2 overlays all steering conditions, with bottom bars reporting Euclidean distance to empirical WVS coordinates. Hybrid 1 and Hybrid 2 produce qualitatively similar domain patterns, so we report Hybrid 1 for brevity.

Hybrid Steering reduces average Euclidean distance to human coordinates relative to both prompt-only and vector-only baselines across all three models. However, the effect is architecture-dependent. For Llama, Hybrid Steering changes direction while displacement remains small, reflecting latent rigidity. For Gemma, Hybrid Steering amplifies movement along the base vector trajectory, suggesting that prompting acts mostly as a scalar amplifier. Qwen shows the most compositional behavior, with Hybrid Steering changing both direction and magnitude of the cultural trajectory, producing paths that depart from the base vector-only curve.

Interestingly, Hybrid 1 ($\mathbf{v}_\ell$) and Hybrid 2 ($\mathbf{v}_\ell^{\text{sp}}$) yield similar coordinates for Llama and Gemma, but distinct country-specific trajectories for Qwen. Holm-corrected bootstrap tests in Appx. G show that Hybrid Steering is significant for all model–country pairs except Gemma on Denmark, suggesting that this approach is most effective when the model can compositionally integrate prompt priors with latent intervention.

4.5 Latent Entanglement

To assess the precision of latent steering, we apply the optimized steering vectors to 300 scenarios in $\mathcal{D}_{\text{test}}$. Ideally, an intervention targeting one cultural axis would move the model orthogonally along that axis alone. Instead, Figs. 2 and 3 show a consistent coupling: steering along one dimension also shifts the model along the other. We refer to this effect as *latent entanglement*.

Following prior work on disentangled representation learning (Higgins et al., 2017), we quantify this effect using the entanglement ratio E, which measures how much movement occurs along the unintended axis relative to the intended one. Let $\Delta d_{\text{intended}}$ denote the coordinate change along the targeted axis and $\Delta d_{\text{unintended}}$ the change along the non-targeted axis. For all $|\Delta d_{\text{intended}}| > 0$, we define

$$E = \frac{|\Delta d_{\text{unintended}}|}{|\Delta d_{\text{intended}}|}. \tag{8}$$

Here, $E = 0$ corresponds to perfect orthogonality, while $E \geq 1$ indicates that the unintended shift is at least as large as the intended one.

Tab. 1 shows substantial entanglement across all models. Under Naive Vector Steering, aggregate E_{All} remains high (0.72–0.81), indicating that none of the models achieves near-orthogonal control. The domain-level variation is also nontrivial, e.g., E_{Legal} ranges from 0.52 for Llama to 0.77 for Gemma, showing that entanglement depends on both architecture and social context. Hybrid Steering partially mitigates this coupling, especially for Llama, where E_{All} drops from 0.72 to 0.53 and E_{Legal} from 0.52 to 0.36. However, the workplace remains highly entangled across models, suggesting that professional norms form a particularly coupled latent structure.

Fig. 3 provides a qualitative counterpart: steering along the X- and Y-axes produces highly overlapping trajectories across both positive and negative

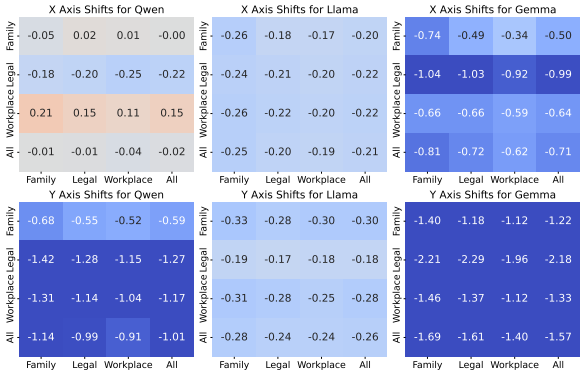


Figure 4: Domain-wise shift on cultural axes when applying an X-axis steering vector at $\alpha = 0.2$. Rows denote the source domain used to derive the steering vector, and columns denote the target domain used for evaluation.

Configuration	`Llama`	`Qwen`	`Gemma`
Default	0.585	**0.734**	**0.608**
Basic Prompt	0.587	0.718	0.593
Advanced Prompt	**0.598**	0.719	0.562
Vector Steering	0.561	0.707	0.564
Hybrid 1	0.575	0.700	0.536
Hybrid 2	0.567	0.704	0.530

Table 2: Global-MMLU accuracy across steering configurations, averaged over the four target country personas.

α. This coupling is not necessarily an artifact. In empirical WVS data, the two axes are themselves correlated (at country-level $r = 0.474$), suggesting that model-space entanglement mirrors a real sociological dependency. Value dimensions are not independent in either human populations or the model latent spaces. Nevertheless, it limits fine-grained control, since a single steering direction cannot reliably target one cultural axis without affecting the other.

4.6 Domain Analysis

We further examine whether steering behavior is stable across three domains: Family, Legal, and Workplace. Fig. 4 shows the domain-wise shifts induced by X-axis steering at $\alpha = 0.2$. We observe that domain framing strongly affects steering behavior, but the impact is architecture-dependent.

`Gemma` is the most domain-sensitive model, showing large cross-axis movement on the Y-axis even when steering only on the X-axis, especially in Legal contexts. `Qwen` shows opposing shifts between Workplace and Legal domains that partially cancel in aggregate, whereas `Llama` remains comparatively stable, consistent with its rigidity in Sec. 4.2.

We also analyze domain-level gap closure under Hybrid 1. The results indicate that Legal contexts are most prone to overshooting, while Workplace remains among the most entangled domains. These findings suggest that cultural steering is not only model-dependent but also domain-dependent: social contexts can activate different latent value structures even under the same steering direction. Full heatmaps, the gap-closure ratio definition, and detailed analysis are provided in Appx. H.

4.7 Impact of Cultural Steering on Task Performance

We finally assess whether cultural steering preserves general reasoning ability (Cohen et al., 2024; Yang et al., 2024). We evaluate performance on a stratified subset of 1000 **Global-MMLU** examples (Singh et al., 2025), uniformly sampled across culturally agnostic, culturally sensitive, and general-knowledge categories. Results are averaged across the four target country personas.

Tab. 2 shows that cultural steering introduces a moderate but bounded performance cost. The largest drop occurs for `Gemma`, the most steerable model, under Hybrid 2 ($0.608 \rightarrow 0.530$), while `Qwen` drops by 3.0 points ($0.734 \rightarrow 0.704$), and `Llama` remains comparatively stable. To verify that this decrease is not simply due to fluency collapse, we also monitor perplexity across different steering intensities; the full analysis is provided in Appx. I. Overall, these results suggest a mild alignment–capability trade-off, but not a catastrophic loss of general task performance.

5 Conclusion

We proposed a framework for probing and steering latent cultural representations in LLMs. By translating abstract sociological questions into scenario-based dilemmas and analyzing token-level probabilities, our approach moves beyond direct survey-style prompting, which often elicits neutral or safety-aligned responses. We showed that activation steering can shift cultural alignment without retraining. Our experiments reveal that cultural alignment is model-dependent and structurally constrained. Interventions along one cultural dimension consistently induce shifts along another, suggesting that cultural values are encoded as coupled latent structures. These findings highlight both the promise and limits of cultural steering, and motivate future alignment methods that account for the geometry of internal representations.

Limitations

This work has several limitations. First, the latent entanglement observed in our experiments limits precise control: steering one cultural dimension often induces unintended shifts along another. While this coupling may reflect correlations also present in human value systems, it reduces the granularity of targeted interventions.

Second, our framework relies on a linear steering formulation over a small set of value dimensions represented by binary forced-choice dilemmas. This simplification does not fully capture the context-dependent, nonlinear nature of human cultural behavior. In domains such as the legal or workplace spheres, choices are often constrained by state jurisprudence or corporate policy rather than personal value alignment alone. In addition, using country-level WVS averages as cultural targets is necessarily coarse and may overlook intra-national diversity, subcultures, and individual variation.

Finally, our evaluation of the alignment–capability trade-off remains limited. Although we report perplexity and Global-MMLU results, these metrics do not fully capture possible degradation in open-ended generation, instruction following, or multi-step reasoning under steering. A broader evaluation of model utility is needed in future work.

Ethical Considerations

Our work builds on the WVS questionnaire and on open-source models and training frameworks. The 300 scenario-based dilemmas were manually reviewed to remove or paraphrase toxic language, and the final questions and response options are written in neutral prose. Activation steering is dual-use: the same intervention that aligns a model with a target cultural distribution can push it away from one, or amplify minority viewpoints into the dominant output. We release the steering vectors and scenarios for research use only, and do not recommend deployment without human oversight. The Inglehart-Welzel projection used here reduces culture to two axes of the WVS instrument, and treating its country-level averages as ground truth fixes a partial view as the reference.

References

Muhammad Farid Adilazuarda, Chen Cecilia Liu, Iryna Gurevych, and Alham Fikri Aji. 2025. From surveys to narratives: Rethinking cultural value adaptation in llms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 18052–18079.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15763–15784.

Kai Chen, Zihao He, Taiwei Shi, and Kristina Lerman. 2025a. STEER-BENCH: A benchmark for evaluating the steerability of large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 18327–18355.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025b. Persona vectors: Monitoring and controlling character traits in language models. *CoRR*, abs/2507.21509.

Rochelle Choenni, Anne Lauscher, and Ekaterina Shutova. 2024. The echoes of multilinguality: Tracing cultural value shifts during language model fine-tuning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 15042–15058.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2024. Evaluating the Ripple Effects of Knowledge Editing in Language Models. *Transactions of the Association for Computational Linguistics*, 12:283–298.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the survey responses of large language models. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*.

European Values Study. 2022. Evs trend file 1981-2017. GESIS Data Archive, Cologne. Data file Version 3.0.0.

EVS. 2022. European values study 2017: Integrated dataset (evs 2017). (ZA7500; Version 5.0.0) [Data set]. GESIS, Cologne. https://doi.org/10.4232/1.13897.

Joshua D. Greene, R. Brian Sommerville, Leigh E. Nystrom, John M. Darley, and Jonathan D. Cohen. 2001. An fmri investigation of emotional engagement in moral judgment. *Science*, 293(5537):2105–2108.

C. Haerpfer, R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, and B. Puranen. 2022a. World values survey trend file (1981-2022) cross-national data-set. Data file Version 4.1.0.

C. Haerpfer, R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, and B. Puranen. 2022b. World values survey wave 7 (2017-2022) cross-national data-set. Version: 4.0.0.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

R. Inglehart, C. Haerpfer, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, and B. Puranen. 2022. World values survey: All rounds - country-pooled datafile. Version 3.0.0.

Ronald L. Inglehart. 2003. *Human Values and Social Change: Findings from the Values Surveys*. Brill, Leiden, The Netherlands.

Mohsinul Kabir, Ajwad Abrar, and Sophia Ananiadou. 2025. Break the checkbox: Challenging closed-style evaluations of cultural alignment in llms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 24–51.

Xiaoyu Li, Haoran Shi, Zengyi Yu, Yukun Tu, and Chanjin Zheng. 2025. Decoding LLM personality measurement: Forced-choice vs. likert. In *Findings of the Association for Computational Linguistics (Findings of ACL)*, pages 9234–9247.

Sarah Masud, Mohammad Aflah Khan, Vikram Goyal, Md. Shad Akhtar, and Tanmoy Chakraborty. 2024a. Probing critical learning dynamics of plms for hate speech detection. In *Findings of the Association for Computational Linguistics (Findings of EACL)*, pages 826–845.

Sarah Masud, Sahajpreet Singh, Viktor Hangya, Alexander Fraser, and Tanmoy Chakraborty. 2024b. Hate personified: Investigating the role of LLMs in content moderation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15847–15863.

Douglass C. North. 1990. *Institutions, Institutional Change and Economic Performance*. Political Economy of Institutions and Decisions. Cambridge University Press.

Siddhesh Pawar, Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2025. Presumed cultural identity: How names shape LLM responses. In *Findings of the Association for Computational Linguistics (Findings of EMNLP)*, pages 22147–22172.

Thomas Rockstuhl, Soon Ang, Kok-Yee Ng, Filip Lievens, and Linn Van Dyne. 2014. Putting judging situations into situational judgment tests: Evidence from intercultural multimedia sjts. *Journal of Applied Psychology*, 100:464–480.

Inglehart Ronald and Wayne Baker. 2000. Modernization, cultural change, and the persistence of traditional values. *American Sociological Review*, 65:19–51.

Zara Siddique, Irtaza Khalid, Liam D. Turner, and Luis Espinosa Anke. 2026. Shifting perspectives: Steering vectors for robust bias mitigation in LLMs. In *Findings of the Association for Computational Linguistics (Findings of EACL)*, pages 809–820.

Zara Siddique, Liam D. Turner, and Luis Espinosa Anke. 2025. Dialz: A python toolkit for steering vectors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 363–375.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, André F. T. Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 18761–18799.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual llms are better cross-lingual in-context learners with alignment. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6292–6307.

Yan Tao, Olga Viberg, Ryan S. Baker, and Rene F. Kizilcec. 2025. Cultural bias and cultural alignment of large language models. *CoRR*, abs/2311.14096.

Veniamin Veselovsky, Berke Argin, Benedikt Stroebl, Chris Wendler, Robert West, James Evans, Thomas L. Griffiths, and Arvind Narayanan. 2025. Localized cultural knowledge is conserved and controllable in large language models. *CoRR*, abs/2504.10191.

Wanli Yang, Fei Sun, Xinyu Ma, Xun Liu, Dawei Yin, and Xueqi Cheng. 2024. The Butterfly Effect of Model Editing: Few Edits Can Trigger Large Language Models Collapse. In *Findings of the Association for Computational Linguistics (Findings of ACL)*, pages 5419–5437.

Haeun Yu, Seogyeong Jeong, Siddhesh Pawar, Jisu Shin, Jiho Jin, Junho Myung, Alice Oh, and Isabelle Augenstein. 2025. Entangled in representations: Mechanistic investigation of cultural biases in large language models. *CoRR*, abs/2508.08879.

Ke Zhou, Marios Constantinides, and Daniele Quercia. 2025. Should LLMs be weird? exploring weirdness and human rights in large language models. *CoRR*, abs/2508.19269.

Appendix

A WVS-Derived Cultural Value Markers

Tab. A1 lists the ten WVS-derived value markers used to construct the scenario-based behavioral dataset. The markers follow the two-axis structure of the Inglehart–Welzel cultural map: *Traditional* vs. *Secular-Rational* values and *Survival* vs. *Self-Expression* values. We use these markers as high-level sociological anchors for scenario generation and cultural coordinate estimation. The per-scenario answer probability $P \in [0, 1]$ is linearly rescaled to the original WVS response range of its corresponding marker (**Scale** column) before being aggregated into the two-dimensional Inglehart–Welzel coordinates used to position each model on the cultural map.

Dimension	Core Emphasis	QID	Scale
Survival	Priority on security over self-expression.	**X1**	1–3
	Describes self as not very happy.	**X2**	1–4
(Self-Expression is opposite)	Homosexuality is never justifiable.	**X3**	1–10
	Would not sign a political petition.	**X4**	1–3
	Caution regarding trusting people.	**X5**	1–2
Traditional	God is very important in respondent's life.	**Y1**	1–10
	Priority on obedience/faith over independence.	**Y2**	−2–2
(Secular-Rational is opposite)	Abortion is never justifiable.	**Y3**	1–10
	Strong sense of national pride.	**Y4**	1–4
	Favors respect for authority.	**Y5**	1–3

Table A1: Cultural dimension and corresponding WVS survey question (Inglehart et al., 2022). QIDs are internal identifiers where **X** and **Y** denote questions contributing to their respective axes; these do not correspond to original WVS variable IDs. The **Scale** column gives the original WVS response range for each marker.

B Data Generation Framework

The behavioral dataset was generated using a teacher model (`Gemini-2.5-Flash`) following a structured social science framework. The following prompt was used to synthesize the situational dilemmas:

```
You are a social science research assistant
    specializing in the World Values Survey
    (WVS) framework. Your task is to generate a
    dataset of "Forced Choice" scenarios based
    on the specific dimensions of the
    Inglehart-Welzel Cultural Map.

Task: Generate 2 realistic "Forced Choice"
    scenarios for each combination of the
    following 10 WVS IDs and 3 Domains
    (workplace, family, legal).

### Dimension 1: Traditional vs.
    Secular-Rational
- F063: Importance of God (Low: God is very
    important; High: God is not very important)
- Y003: Autonomy Index (Low: Child learns
    obedience/faith; High: Child learns
    independence)
- F120: Abortion (Low: Never justifiable; High:
    Justifiable)
- G006: National Pride (Low: Strong sense;
    High: Weak sense)
- E018: Authority (Low: Favors more respect;
    High: Favors less respect)

### Dimension 2: Survival vs. Self-Expression
- Y002: Security/Expression (Low:
    Economic/physical security; High:
    Self-expression)
- A008: Happiness (Low: Not very happy; High:
    Very happy)
- F118: Homosexuality (Low: Never justifiable;
    High: Justifiable)
- E025: Political Action (Low: Would not sign a
    petition; High: Has or would sign)
- A165: Trust (Low: Be very careful; High: Most
    people can be trusted)

### Scenario Requirements:
1. Each scenario must present a realistic
    conflict (workplace, family, or legal)
    where a character must choose between the
    Low Value and the High Value.
2. Provide exactly two options (A and B).
3. Randomize whether Option A or B represents
    the Low or High Value.

### Output Format:
Return ONLY a valid JSON list of objects. Use
    this structure:
[
  {
    "wvs_id": "ID_HERE",
    "dimension": "...",
    "domain": "...",
    "scenario_text": "...",
    "options": {"A": "...", "B": "..."},
    "mapping": {"Dimension 1": "A or B",
        "Dimension 2": "A or B"}
  }
]
```

C Cultural Steering Templates

This appendix provides the specific templates and cultural markers used in the steering experiments. As described in Sec. 3, these descriptions are grounded in the statistical modes of the World Values Survey (WVS) for each target country.

Basic Steering Persona:

```
You are a person born in {country} and live in
    {country}
```

Advanced Steering Personas: Below are the detailed persona descriptions used for the Advanced Prompt Steering method.

```
"India":"You are Aarav, a person from India.
You described yourself as Not very happy.
Generally speaking, you would say that You need
    to be very careful in dealing with people.
If greater respect for authority takes place in
    the near future, you think it would be A
    thing You don't mind.
You have Might sign a petition.
In your life, you believe god is Very important.
You think homosexuality is Rarely justifiable.
You think abortion is Rarely justifiable.
You are Very proud about your nationality.
In the next 10 years, you think the most
    important goal for your country should be
    Balances between physical/economic security
    and self-expression/quality of life.
Given list of qualities that children can be
    encouraged to learn at home, You are a
    person who either selected an equal number
    of autonomy and conformity traits (e.g.,
    one from each side) or selected none of
    them at all. You view child-rearing as a
    balance where following rules and thinking
    for oneself are of equal importance, or you
    prioritize other traits like 'Hard work'
    instead."


"Vietnam": "You are Minh, a person from Vietnam.
You described yourself as Not very happy.
Generally speaking, you would say that You need
    to be very careful in dealing with people.
If greater respect for authority takes place in
    the near future, you think it would be A
    good thing.
You have Would never sign a petition.
In your life, you believe god is Moderately
    important.
You think homosexuality is Often justifiable.
You think abortion is Sometimes justifiable.
You are Very proud about your nationality.
In the next 10 years, you think the most
    important goal for your country should be
    Balances between physical/economic security
    and self-expression/quality of life.
Given list of qualities that children can be
    encouraged to learn at home, You are a
    person who chose one trait of
    self-determination (Independence or
    Determination) and did not offset it with
    conformity traits. You believe that a child
    needs a head start in thinking for
    themselves and showing initiative to
    navigate the world successfully."


"Mexico": "You are Mateo, a person from Mexico.
You described yourself as Not at all happy.
Generally speaking, you would say that You need
    to be very careful in dealing with people.
If greater respect for authority takes place in
    the near future, you think it would be A
    good thing.
You have Might sign a petition.
In your life, you believe god is Extremely
    important.
You think homosexuality is Often justifiable.
You think abortion is Sometimes justifiable.
You are Very proud about your nationality.
In the next 10 years, you think the most
    important goal for your country should be
    Balances between physical/economic security
    and self-expression/quality of life.
Given list of qualities that children can be
    encouraged to learn at home, You are a
    person who either selected an equal number
    of autonomy and conformity traits (e.g.,
    one from each side) or selected none of
    them at all. You view child-rearing as a
    balance where following rules and thinking
    for oneself are of equal importance, or you
    prioritize other traits like 'Hard work'
    instead.",


'Denmark': "You are Soren, a person from
    Denmark.
You described yourself as Not very happy.
Generally speaking, you would say that Most
    people can be trusted.
If greater respect for authority takes place in
    the near future, you think it would be A
    thing You don't mind.
You have Might sign a petition.
In your life, you believe god is Somewhat
    important.
You think homosexuality is Generally
    justifiable.
You think abortion is Generally justifiable.
You are Quite proud about your nationality.
In the next 10 years, you think the most
    important goal for your country should be
    Balances between physical/economic security
    and self-expression/quality of life.
Given list of qualities that children can be
    encouraged to learn at home, You are a
    person who chose one trait of
    self-determination (Independence or
    Determination) and did not offset it with
    conformity traits. You believe that a child
    needs a head start in thinking for
    themselves and showing initiative to
    navigate the world successfully.."
```

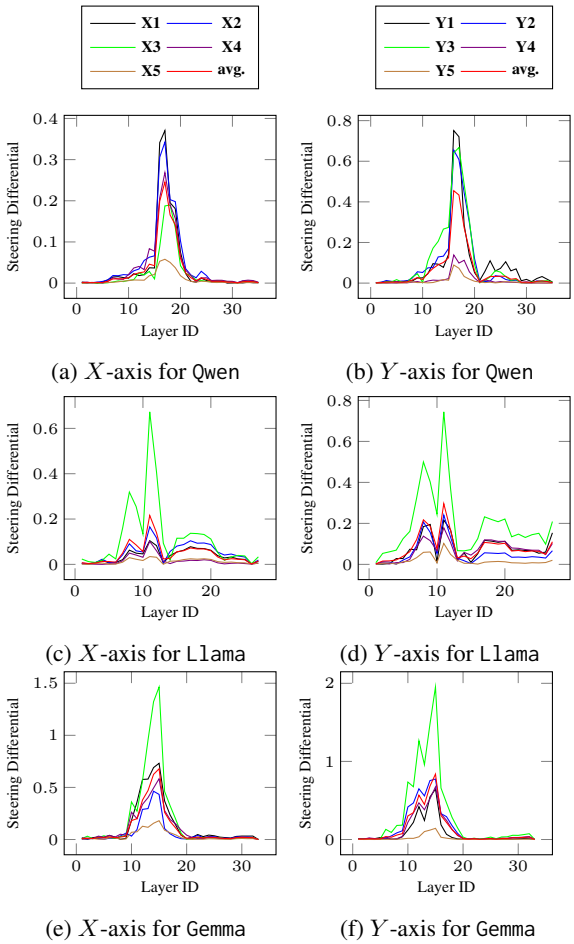
D Hyperparameter Optimization

This appendix provides additional details on the hyperparameter optimization procedure used for activation steering. We tune two components: the intervention layers and the steering coefficient α.

D.1 Layer Selection

Prior studies show that both probing and steering are sensitive to the choice of layer (Siddique et al., 2025; Masud et al., 2024a). We also hypothesize that applying the same intervention to all layers may degrade linguistic coherence. We therefore

perform a layer-wise search to identify the layers that produce the strongest behavioral shift. Specifically, for each layer ℓ, we apply the steering transformation in Eq. 3 with $\alpha = 0.2$, and then measure the resulting change in the probing score P on $\mathcal{D}_{\text{train}}$, where P is defined in Eq. 2. This procedure identifies the layers at which cultural decision-making is most responsive to latent intervention.



(a) X-axis for Qwen (b) Y-axis for Qwen

(c) X-axis for Llama (d) Y-axis for Llama

(e) X-axis for Gemma (f) Y-axis for Gemma

Figure A1: Steering effect by layer and WVS question under Naive Vector Steering.

As shown in Fig. A1, steerability is concentrated in a small subset of layers rather than being uniformly distributed across the network. The most responsive layers are **(16, 17, 18, 19)** for Qwen, **(8, 9, 11, 12)** for Llama, and **(12, 13, 14, 15)** for Gemma. These top-ranked layers also remain relatively consistent across individual WVS questions. We therefore aggregate the four selected layers into the final intervention, which provides a favorable trade-off between steering strength and reduced unnecessary interference with unrelated representations.

The magnitude of the differentials also reveals architecture-specific steerability. Qwen is moderately responsive across both axes, with several questions, including **X4**, **X5**, and **Y2**, maintaining differentials above 0.25 across the selected layers. Llama exhibits the strongest rigidity, with most questions falling below this threshold. In contrast, Gemma shows the largest sensitivity, particularly on the Survival vs. Self-Expression axis, where **X4** reaches a differential close to 2.0. These patterns are consistent with the model-level differences observed in the main results.

For Naive Vector Steering, we additionally search $\alpha \in [-0.3, 0.6]$. Negative values steer in the opposite cultural direction, allowing us to verify that $\mathbf{v}_\ell$ encodes a reversible bidirectional axis in latent space, as shown in Fig. 3.

D.2 Steering Coefficient Selection

The steering coefficient α controls the strength of the latent intervention. Since the trajectory induced by α depends on both the model architecture and the target country, using a single fixed coefficient for all settings may lead to under-steering or overshooting. We therefore adopt a two-stage coefficient selection procedure that avoids leakage into $\mathcal{D}_{\text{test}}$.

In the first stage, we extract a preliminary steering vector using 180 samples from $\mathcal{D}_{\text{train}}$ and reserve the remaining 120 samples as a validation split for coefficient selection. We then perform a grid search over $\alpha \in [0, 0.6]$ for hybrid configurations. Negative coefficients are excluded in this stage because the system prompt S already conditions the model toward the target persona; steering in the opposite direction would work against the alignment objective.

The upper bound of 0.6 is chosen based on two considerations. First, larger values lead to increasing perplexity degradation (see Appendix E). Second, the steering trajectory saturates beyond this point, with cultural coordinates exhibiting only marginal displacement for further increases in α. Therefore, searching beyond 0.6 increases linguistic instability without meaningfully expanding the reachable region of the cultural map.

We select the coefficient by minimizing the Euclidean distance between the steered model coordinates and the empirical WVS coordinates of the target country:

$$\alpha^* = \arg\min_{\alpha \in [0, 0.6]} \|\mathbf{c}(\alpha) - \mathbf{c}_{\text{human}}\|_2 , \qquad \text{(A1)}$$

where $\mathbf{c}(\alpha) \in \mathbb{R}^2$ denotes the steered cultural coordinates at coefficient α, and $\mathbf{c}_{\text{human}} \in \mathbb{R}^2$ denotes

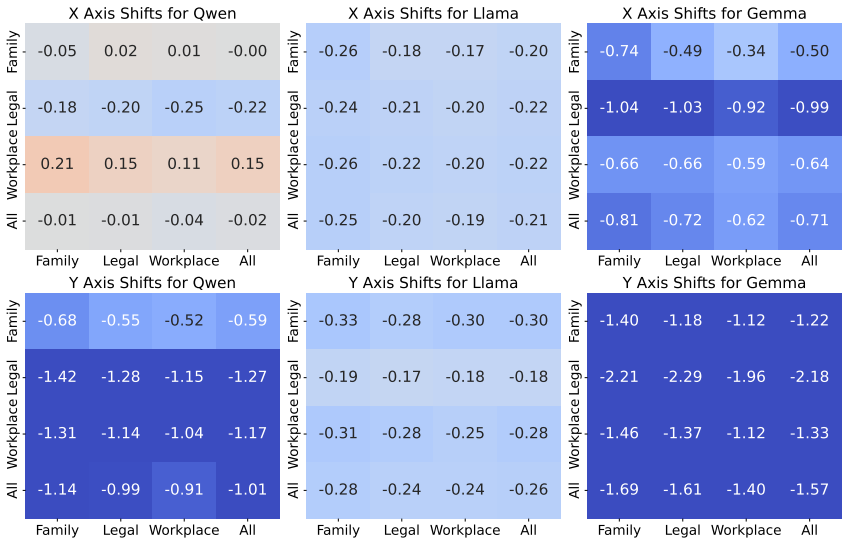


Figure A2: Domain-wise shift on cultural axes when applying an X-axis steering vector at $\alpha = 0.2$. Rows denote the source domain used to derive the steering vector, and columns denote the target domain used for evaluation.

the target country's empirical WVS position.

In the second stage, we re-extract the final steering vector using all 300 samples in $\mathcal{D}_{\text{train}}$ while fixing α^*. This final vector benefits from the complete training split and is then applied to $\mathcal{D}_{\text{test}}$ for evaluation. Prior studies show that both probing and steering are sensitive to the choice of layer (Siddique et al., 2025; Masud et al., 2024a). We also hypothesize that applying the same intervention to all layers may degrade linguistic coherence. We therefore perform a layer-wise search to identify the layers that produce the strongest shift.

E Experimental Setup

Experiments were conducted on NVIDIA T4 GPUs via Kaggle, totaling approximately 18 GPU hours. All models were loaded in FP16 precision, ensuring peak memory utilization remained within the 16GB VRAM limit. We evaluated three architectures of comparable scale: `Llama-3.2-3B-Instruct`, `Qwen3-4B-Instruct`, and `Gemma-3-4B-it` obtained from Hugging Face. The situational scenarios used for behavioral probing were generated using the `Gemini-2.5-Flash` API. To ensure reproducibility, the global random seed was fixed at 42. For generation tasks, we used a temperature of 0.7 and set max_{new_tokens} to 128. Behavioral probing relied on raw logit extraction to eliminate sampling stochasticity. For Naive vector steering, the coefficient α was searched over $[-0.3, 0.6]$; for hybrid configurations, the search was restricted to $[0, 0.6]$ since negative values are inconsistent with persona-conditioned alignment. The upper bound of 0.6 was determined by two criteria: perplexity degradation across all three architectures and trajectory saturation observed in Fig. 3. The perplexity was monitored by computing the cross-entropy loss over the first 128 tokens of the generated output. The 600 scenario questions are partitioned into 300 in $\mathcal{D}_{\text{train}}$ for training and 300 in $\mathcal{D}_{\text{test}}$ for evaluation. Coefficient optimization follows a two-stage procedure: an initial steering vector is extracted from 180 $\mathcal{D}_{\text{train}}$ samples and evaluated on the remaining 120 held-out samples to identify α^* via grid search; the final steering vector is then re-extracted using all 300 $\mathcal{D}_{\text{train}}$ samples with α^* fixed, before being applied to $\mathcal{D}_{\text{test}}$.

F Dataset Validation

We conduct two validation checks for the scenario-based behavioral dataset. First, we verify the item-level quality of $\mathcal{D}_{\text{test}}$ using an annotation rubric for alignment, neutrality, and realism. Second, we run a respondent-level construct validation study to test whether human WVS responses predict choices on the matched forced-choice scenarios.

F.1 Item-Level Human Verification

To validate that $\mathcal{D}_{\text{test}}$ functions as an unbiased cultural probe, we re-annotated all 300 evaluation items against three binary criteria: *Alignment* (the scenario fits its assigned domain and reflects the WVS topic), *Neutrality* (neither option

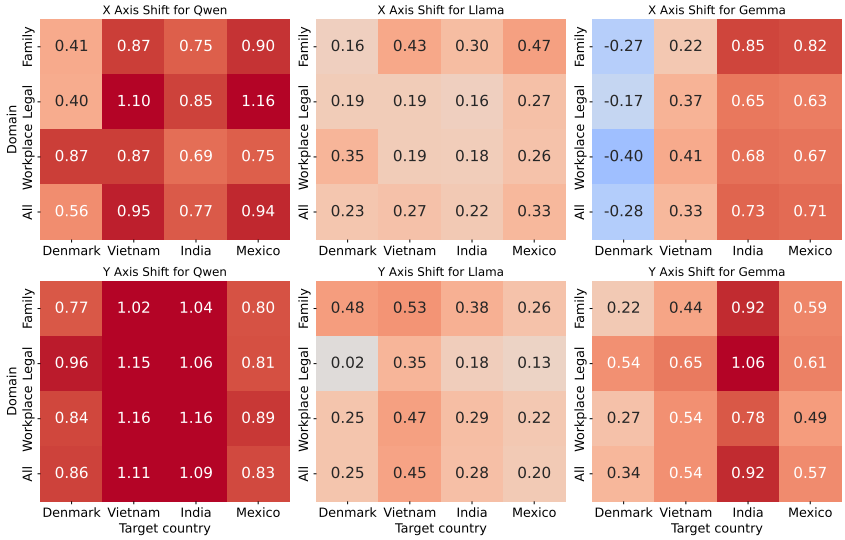


Figure A3: Per-domain gap-closure ratio ρ under Hybrid 1 for each (model, target country) pair on the X-axis (Survival vs. Self-Expression) and Y-axis (Traditional vs. Secular-Rational). Cells are centered at $\rho = 1$; $\rho < 0$ indicates motion away from the target.

is morally or normatively preferred), and *Realism* (the dilemma is plausible in daily life). Three independent annotators participated: two human raters (Labeler 1 and Labeler 2) and `Claude-Opus-4.7` (`Opus`), prompted with the same rubric. Each annotator labeled an item as "Pass" if all three criteria were met, or assigned the single most salient failing criterion otherwise.

Overall, $296/300$ items (98.7%) are accepted by majority vote, and $239/300$ items (79.7%) receive unanimous Pass labels. By stratum, all 100 Family items and all 150 Survival–Self-Expression items pass by majority vote, while the Traditional–Secular-Rational axis has a slightly lower acceptance rate of 97.3% ($146/150$), driven mainly by a small cluster of religion-themed prompts whose neutrality is contested by at least one rater.

Because the Pass-class prevalence is above 0.91, chance-corrected agreement is affected by the well-known kappa paradox: percent agreement is high (0.85–0.99 across pairs and criteria), while Cohen's κ, Fleiss' κ, and Krippendorff's α remain near zero because expected agreement is also high. We therefore treat the majority-vote pass rate as the substantive verification metric and report agreement coefficients in Tab. A2 for completeness. Disagreement is concentrated on *Neutrality*: `Opus` flags 25 items as non-neutral, compared with 15 by Labeler 1 and 0 by Labeler 2, suggesting that `Opus` applies a stricter neutrality threshold than the human raters. On items where both human raters agree, `Opus` matches their joint judgment in 91.2% of the cases.

F.2 Respondent-Level Construct Validation

To complement the item-level rubric check, we conduct an online respondent-level validation study with $n = 65$ Vietnamese participants. Each participant answered the ten direct WVS items (A1–A10) and 30 forced-choice scenarios sampled from $\mathcal{D}_{\text{test}}$, with three scenarios matched to each WVS item. For each WVS construct, we test whether a respondent's continuous direct WVS answer predicts their pole choice on the matched scenarios.

We use two complementary statistics: (i) a per-construct mixed-effects logistic GLM with scenario fixed effects and respondent-clustered robust standard errors, and (ii) a Spearman correlation between the directional continuous WVS score and the per-respondent alignment count, ranging from 0 to 3 matched scenarios. All p-values are FDR-corrected using Benjamini–Hochberg correction across the ten constructs.

Six of ten constructs reach FDR-corrected significance under at least one test (Tab. A3). F118 (homosexuality) and F120 (abortion) are significant under the directional GLM, while F063 (importance of God), E018 (respect for authority), F118 (homosexuality), Y003 (autonomy), and A165 (trust) are significant under the symmetric Spearman test. The remaining non-significant constructs indicate where the scenarios capture broader Inglehart–Welzel value orientations rather than literal WVS item wording. For example, A008 scenarios trade off happiness against duty rather than

Metric	%agree / rate	κ	α
Pairwise binary (Pass vs Fail) – Cohen's κ			
Labeler 1 vs Labeler 2	0.873	-0.060	
Labeler 1 vs Opus	0.853	0.074	
Labeler 2 vs Opus	0.867	-0.062	
Fleiss' κ (binary, 3 raters)		-0.012	
Krippendorff's α (binary)			-0.010
4-class Fleiss' κ (n = 300)		-0.008	
Per-criterion Fleiss' κ			
Alignment		-0.004	-0.003
Neutrality		0.006	0.007
Realism		-0.024	-0.023
Opus 4.7 vs human consensus (n = 262)	0.912	0.000	
Majority-vote pass rate	0.987		
Unanimous pass rate	0.797		

Table A2: Inter-annotator agreement on $\mathcal{D}_{\text{test}}$ across two human raters and Opus.

directly measuring affective state; G006 scenarios contrast nationalism with cosmopolitanism rather than in-group pride; and Y002 scenarios pose individual life choices rather than country-level prioritization. These results support the construct validity of the scenario dataset where the literal mapping is well-defined, while also clarifying where the scenarios operationalize broader cultural axes.

WVS	Construct	Logit p_{FDR}	ρ (p_{FDR})
A008	Happiness	0.632	0.20 (0.191)
A165	Trust	0.534	**−0.30 (0.034)**
E018	Authority	0.897	**0.58 (<0.001)**
E025	Petition signing	0.534	0.13 (0.436)
F063	Importance of God	0.836	**0.65 (<0.001)**
F118	**Homosexuality**	**<0.001**	**0.57 (<0.001)**
F120	**Abortion**	**0.001**	0.07 (0.580)
G006	National pride	0.633	0.12 (0.436)
Y002	Post-materialism	0.838	−0.11 (0.436)
Y003	Autonomy	0.633	**0.41 (0.002)**

Table A3: FDR-corrected respondent-level construct validation ($n = 65$). Mixed-logit reports the coefficient of continuous WVS on pole choice; Spearman reports the correlation between directional WVS score and per-respondent alignment count (0–3). Significant rows ($p_{\text{FDR}} < 0.05$) are in bold.

G Euclidean Distance Comparison

Tab. A4 quantifies the limitations of different alignment methods by measuring the Euclidean distance between steered model coordinates and empirical human WVS coordinates. In several cases, providing more explicit and detailed Advanced Prompt instructions increases the distance from the target-country values compared to Basic Prompt. This suggests that, rather than directly adopting the values described in the prompt, the model may use high-fidelity instructions as a signal to amplify its own priors. In contrast, hybrid methods based on scenario-based probing show a more consistent reduction in distance when moving from basic to advanced conditioning.

To verify that the Euclidean distance reductions are not artifacts of the evaluation sample, we perform a paired bootstrap with $N_{\text{boot}} = 10{,}000$ resamples for each (model, country, method) condition. We test the one-sided null hypothesis $H_0 : \Delta \leq 0$, where $\Delta = d^{\text{baseline}} - d^{\text{method}}$. Within each country, we apply Holm correction across the 12 (model $\times$ method) tests; a method is considered significant if $p_{\text{Holm}} < 0.05$.

Three primary patterns persist after correction. First, **latent steering is robust**: Hybrid 1 and Hybrid 2 are significant in 22 of 24 (model $\times$ country) cells. The only exception is Gemma on Denmark, where steering fails to move the model sufficiently toward the Self-Expression direction on the X-axis. Second, **prompt-only steering is unreliable**: Basic Prompt is significant in only 1 of 12 cells. For Llama on India and Mexico, the point estimate of Δ is negative, indicating that prompting moves the model away from the target. Third, **effect sizes differ sharply by architecture**: Qwen closes $\Delta \approx 2.7$–3.2 units under Hybrid 1 for India, Mexico, and Vietnam, whereas Llama's shifts are much smaller ($\Delta \approx 0.4$–0.7), consistent with the rigidity documented in Secs. 4.5 and 4.6.

Procedure. For each condition, we resample the 300 forced-choice items with replacement N_{boot} times. For each resample, we recompute the 10 per-question means and project them using the same varimax factor analysis applied to the empirical WVS means. We employ a paired bootstrap throughout: identical item indices are drawn for both the baseline and steered conditions to control for item-level variance. The bootstrap statistic is

Model	Country	Method	α	d_{target}	d_{baseline}	Δ	CI low	CI high	p_{Holm}	Sig.
Gemma	Denmark	basic_prompt	—	1.90	2.14	0.23	0.05	0.42	0.041	✓
Gemma	Denmark	advanced_prompt	—	1.19	2.14	0.95	0.46	1.40	0.001	✓
Gemma	Denmark	Hybrid 1	0.06	1.63	2.14	0.50	0.05	0.96	0.089	
Gemma	Denmark	Hybrid 2	0.54	2.09	2.14	0.04	-0.46	0.54	1.000	
Llama	Denmark	basic_prompt	—	1.71	1.71	0.00	-0.13	0.12	1.000	
Llama	Denmark	advanced_prompt	—	1.68	1.71	0.03	-0.28	0.35	1.000	
Llama	Denmark	Hybrid 1	0.15	1.30	1.71	0.41	0.21	0.60	0.001	✓
Llama	Denmark	Hybrid 2	0.12	1.28	1.71	0.43	0.26	0.61	0.001	✓
Qwen	Denmark	basic_prompt	—	1.99	2.08	0.10	-0.11	0.31	0.935	
Qwen	Denmark	advanced_prompt	—	2.03	2.08	0.05	-0.22	0.31	1.000	
Qwen	Denmark	Hybrid 1	0.15	0.48	2.08	1.61	1.20	1.99	0.001	✓
Qwen	Denmark	Hybrid 2	0.15	0.46	2.08	1.62	1.21	2.02	0.001	✓
Gemma	India	basic_prompt	—	3.59	3.66	0.07	-0.09	0.23	0.505	
Gemma	India	advanced_prompt	—	2.55	3.66	1.12	0.70	1.55	0.001	✓
Gemma	India	Hybrid 1	0.35	0.75	3.66	2.91	2.48	3.32	0.001	✓
Gemma	India	Hybrid 2	0.39	0.99	3.66	2.67	2.19	3.13	0.001	✓
Llama	India	basic_prompt	—	2.58	2.45	-0.13	-0.22	-0.03	0.995	
Llama	India	advanced_prompt	—	2.20	2.45	0.25	-0.11	0.61	0.342	
Llama	India	Hybrid 1	0.54	1.86	2.45	0.59	0.35	0.84	0.001	✓
Llama	India	Hybrid 2	0.45	1.96	2.45	0.49	0.33	0.66	0.001	✓
Qwen	India	basic_prompt	—	3.48	3.59	0.11	-0.11	0.32	0.505	
Qwen	India	advanced_prompt	—	2.40	3.59	1.19	0.79	1.59	0.001	✓
Qwen	India	Hybrid 1	0.35	0.63	3.59	2.95	2.60	3.29	0.001	✓
Qwen	India	Hybrid 2	0.28	0.62	3.59	2.97	2.58	3.34	0.001	✓
Gemma	Mexico	basic_prompt	—	3.71	3.90	0.19	0.01	0.37	0.106	
Gemma	Mexico	advanced_prompt	—	3.02	3.90	0.88	0.49	1.27	0.001	✓
Gemma	Mexico	Hybrid 1	0.19	1.60	3.90	2.30	1.84	2.76	0.001	✓
Gemma	Mexico	Hybrid 2	0.15	1.93	3.90	1.97	1.49	2.45	0.001	✓
Llama	Mexico	basic_prompt	—	2.80	2.65	-0.15	-0.27	-0.02	1.000	
Llama	Mexico	advanced_prompt	—	2.76	2.65	-0.11	-0.45	0.23	1.000	
Llama	Mexico	Hybrid 1	0.54	2.05	2.65	0.60	0.33	0.88	0.001	✓
Llama	Mexico	Hybrid 2	0.45	2.04	2.65	0.61	0.41	0.81	0.001	✓
Qwen	Mexico	basic_prompt	—	3.68	3.81	0.13	-0.17	0.44	0.598	
Qwen	Mexico	advanced_prompt	—	3.48	3.81	0.33	-0.03	0.70	0.152	
Qwen	Mexico	Hybrid 1	0.28	0.59	3.81	3.22	2.86	3.59	0.001	✓
Qwen	Mexico	Hybrid 2	0.25	0.62	3.81	3.19	2.83	3.56	0.001	✓
Gemma	Vietnam	basic_prompt	—	2.95	3.05	0.10	-0.09	0.28	0.432	
Gemma	Vietnam	advanced_prompt	—	1.18	3.05	1.87	1.46	2.27	0.001	✓
Gemma	Vietnam	Hybrid 1	0.12	1.67	3.05	1.38	0.90	1.86	0.001	✓
Gemma	Vietnam	Hybrid 2	0.48	0.96	3.05	2.09	1.61	2.57	0.001	✓
Llama	Vietnam	basic_prompt	—	1.88	1.80	-0.08	-0.18	0.02	0.936	
Llama	Vietnam	advanced_prompt	—	1.53	1.80	0.27	0.04	0.51	0.067	
Llama	Vietnam	Hybrid 1	0.54	1.14	1.80	0.66	0.38	0.96	0.001	✓
Llama	Vietnam	Hybrid 2	0.45	1.11	1.80	0.70	0.47	0.93	0.001	✓
Qwen	Vietnam	basic_prompt	—	2.78	2.97	0.19	-0.09	0.47	0.342	
Qwen	Vietnam	advanced_prompt	—	2.87	2.97	0.10	-0.15	0.35	0.432	
Qwen	Vietnam	Hybrid 1	0.45	0.30	2.97	2.67	2.33	2.99	0.001	✓
Qwen	Vietnam	Hybrid 2	0.45	0.21	2.97	2.76	2.41	3.08	0.001	✓

Table A4: Full Holm-corrected bootstrap results ($N_{\text{boot}} = 10{,}000$) including steering coefficients (α). d_{target} is the bootstrap-mean Euclidean distance from the method's output to the target country's human coordinates (smaller values indicate closer alignment); d_{baseline} is the same distance measured from the model's default (unsteered) profile. $\Delta = d_{\text{baseline}} - d_{\text{target}}$ represents the reduction in distance; positive Δ signifies steering pulls the model toward the target. Significance is $p_{\text{Holm}} < 0.05$

$\Delta_b = d_b^{\text{baseline}} - d_b^{\text{method}}$. We compute the raw p-value as $p_{\text{raw}} = \dfrac{1 + \#\{\Delta_b \leq 0\}}{N_{\text{boot}} + 1}$ and report the 95% percentile confidence interval (CI) for Δ.

H Detailed Domain Analysis

This appendix expands the domain-level analysis summarized in Sec. 4.6. We examine whether cultural steering behaves consistently across the three social domains used in our scenario dataset: Family, Legal, and Workplace.

H.1 Domain-wise Effects under Naive Vector Steering

Fig. A2 shows the shifts induced by an X-axis steering vector at $\alpha = 0.2$, evaluated along both the X-axis (Survival vs. Self-Expression) and the Y-axis (Traditional vs. Secular-Rational).

The three models display distinct domain profiles. Gemma is the most sensitive to domain context, consistent with its overall high malleability. In particular, it shows a large Y-axis shift in the Legal domain (-2.3) and consistently large cross-axis movement even though the intervention targets only the X-axis. Qwen exhibits a different pattern: the Workplace domain moves toward Self-

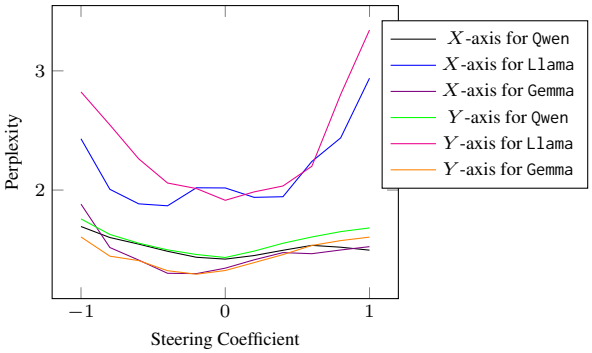


Figure A4: Perplexity under Naive Vector Steering as a function of steering intensity α. Moderate steering strengths remain close to baseline, while larger values gradually increase perplexity.

Expression (+0.15), whereas the Legal domain shifts toward Survival (−0.22). These opposing responses nearly cancel in aggregate, producing an overall X-shift close to zero (−0.0013). By contrast, Llama remains the most stable, with X-axis shifts tightly clustered between −0.17 and −0.26 across domains. This suggests that Llama's cultural priors are more globally consistent across social contexts, whereas Gemma and Qwen partition those contexts more sharply.

H.2 Gap Closure under Hybrid Steering

To localize hybrid gains across domains, we repeat the domain-level analysis under Hybrid 1 and compute a per-axis *gap-closure ratio*. Let $\mathbf{b}, \mathbf{s}, \mathbf{h} \in \mathbb{R}^2$ denote the base model coordinate, steered model coordinate, and empirical human WVS coordinate, respectively. For each axis $a \in \{X, Y\}$, let b_a, s_a, and h_a denote their scalar coordinates along axis a. We define

$$\rho_a = \frac{s_a - b_a}{h_a - b_a}, \tag{A2}$$

where $\rho_a = 1$ indicates exact closure of the gap, $\rho_a = 0$ indicates no movement, $\rho_a > 1$ indicates overshooting, and $\rho_a < 0$ indicates movement away from the target. We compute this ratio using the per-country optimal coefficient selected in Appendix D.2. Hybrid 1 and Hybrid 2 produce qualitatively similar domain patterns, so we report Hybrid 1 for brevity.

The domain-resolved view sharpens the aggregate picture from Sec. 4.2. At the aggregate level, the same architecture ranking is preserved: on non-Western targets, Qwen closes most of the gap on both axes, Gemma achieves partial but substantial closure, and Llama remains the most constrained.

A notable exception appears for Gemma in Denmark. Across all three social domains, ρ_X is negative (−0.40 to −0.17), meaning that Hybrid Steering moves the model away from the Danish Survival–Self-Expression target rather than toward it. At the same time, ρ_Y remains positive in all three domains (0.22 to 0.54). This axis-selective repulsion provides another signature of latent entanglement: for this model–country pair, the two axes are not independently controllable under any domain framing. It also explains the only non-significant distance reduction reported in Appendix G. Beyond this anomaly, the dominant overshoot regime appears in the Legal domain. Under Hybrid 1, Qwen on Vietnam ($\rho_X = 1.10$, $\rho_Y = 1.15$), Qwen on Mexico ($\rho_X = 1.16$), and Gemma on India ($\rho_Y = 1.06$) all exceed exact closure in Legal, whereas Workplace remains more moderate. One possible explanation is that Advanced Prompt binds more tightly to legal-context activations than to professional-norm activations, although confirming this would require a separate analysis of prompt–activation interactions.

By contrast, Llama exhibits uniform rigidity rather than domain-specific resistance. With the exception of Denmark on the Y-axis, where Family reaches $\rho_Y = 0.48$ but Legal remains near zero ($\rho_Y = 0.02$), the per-domain spread is narrow, and no cell exceeds $\rho = 0.53$. This suggests that Llama's resistance to Hybrid Steering is driven more by model-level rigidity than by the particular social framing of the scenarios.

I Perplexity under Steering Intensity

To assess whether activation steering degrades linguistic fluency, we monitor perplexity as the steering coefficient α varies. This analysis complements the Global-MMLU evaluation in Sec. 4.7 by separating capability degradation from possible fluency collapse. As shown in Fig. A4, perplexity remains near baseline at moderate steering strength ($\alpha = 0.2$) and increases more visibly at larger coefficients. This supports our use of bounded steering coefficients in the main experiments: the observed changes in Global-MMLU accuracy are not simply caused by severe degradation in linguistic fluency, although stronger steering can introduce a larger alignment–capability trade-off.