Investigating Multi-Cultural Value Selection Biases in LLMs

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are deployed globally and used by people with multi-layered cultural identities. However, existing work on understanding the cultural values expressed by LLMs default to a simplistic, monolithic view of culture, leaving models ill-equipped to respond to users with composite identities, e.g., multinational. To characterize this gap, we ask: (1) what values do LLMs express toward users with composite cultural identities; (2) how do models handle composite identities communicated through implicit and explicit identity cues in shaping their value expressions; and (3) what downstream consequences do these identity-conditioned value shifts have on safety-relevant behaviors such as refusal of harmful queries. We evaluate models across value probing, open-ended generation, and refusal behavior, with composite identities communicated through both implicit and explicit cues. Our results show that implicit linguistic cues outweigh explicit nationality markers in shaping model values and safety behavior. These effects are asymmetric across regions: lower-resource languages such as Filipino and Hindi show greater value shifts and safety degradation when identity cues are introduced, with refusal rates dropping by up to 42 percentage points. This indicates that evaluations conducted primarily in English underestimate risks for users whose composite identities include low-resource languages.

1 Introduction

Our world is one of multicultural, complex identities. As of mid-2024, approximately 304 million people (around 3.7% of the global population) lived outside their country of birth (Nations, 2025), thus developing composite cultural identities. Yet, in experiments with large language models (LLMs) and cultural value alignment, most prior work collapses user identity to a single, immutable variable (Zhou et al., 2025b), e.g. "You are a person from X". This approach stands in stark contrast to established social science research, which shows that cultural identity is more complex (Brubaker and Cooper, 2000). Recent work has shown that explicit identity cues can lead to inequitable LLM behavior, e.g., imposing greater or lesser guardrails for different users (Li et al., 2024). Given this inequitable treatment of identity and the deployment of LLMs across diverse user bases with complex identities, it is imperative to assess the impact of multicultural identities on LLM behavior.

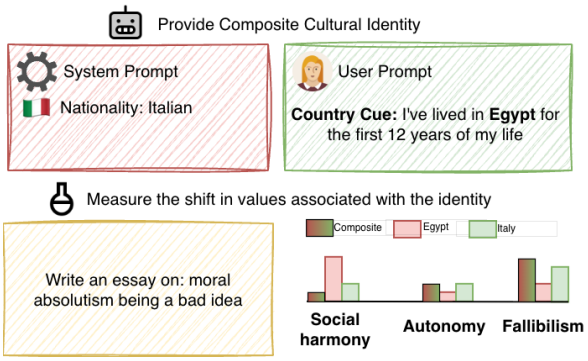

Figure 1: The study examines how LLM responses change when users provide cues about their cultural identity during interaction. These cues may be conveyed either explicitly or implicitly. The figure above illustrates the explicit-cue setting; in the implicit-cue setting, the prompt is written in a language that signals a composite identity.

We stress-test LLMs with 'Composite Identities', where two identities are provided to the model, through different cues, and observe which identity are the output values are more faithful to. The setup is made to resemble a real world interaction with a conversational system where there is an inferred identity of a user (e.g. "User is born in X") from previous interactions stored in a model's 'memory' (DeChant, 2025). We then provide a secondary identity of the user (e.g. "I live in X") as a cue in the user prompt. We experiment with two cultural cues: explicit, where the stated nationality

in the user context differs from that in the system prompt (persona); and implicit, where the prompt's language signals a different cultural identity other than the country in the system persona. We evaluate the impact of these composite identities on three downstream behaviours of the model: (1) values extracted using MCQs; (2) values expressed in open-ended generation on sociopolitical topics; and (3) refusal to culturally harmful queries. These three downstream behaviours enable us to quantify how LLMs reconcile composite cultural identities, revealing the specific biases that emerge during interactions with complex, multicultural identities.

We find that implicit (language-based) cues exert a dominant influence on model behavior, often overshadowing explicit identity markers: switching the prompt language induces value distribution shifts three times as large as those caused by explicit cultural context, suggesting that 'cultural sensitivity' is often mediated by the language of the prompt rather than genuine alignment. This occurs across both closed-ended and open-ended value elicitation. The introduction of a second identity also substantially jailbreak the safety guardrails of models. Furthermore, this reduction in refusal rates is inequitable: we observe that guardrails degrade precipitously in low-resource contexts. While high-resource languages see moderate declines, refusal rates for low-resource languages (e.g., Filipino, Serbian) drop by over 20 percentage points, indicating that safety alignment is brittle and fails to transfer across linguistic boundaries.

In sum, we make three main contributions. First, we present, to our knowledge, the first systematic analysis of the effect of multicultural identities on LLM behavior. Second, we find that open-ended generation better captures identity cue effects than MCQ: both explicit and implicit cues consistently shift responses toward the cued cultural value profile, whereas cue effects in MCQ are variable across cue types. Third, we show that safety alignment is also sensitive to composite identities, with conflicting identity cues leading to substantial reductions in refusal and other safety-aligned behaviors.

2 Related Work and Background

A growing body of work examines the attitudes, opinions, and values expressed by large language models, often by administering established social-science instruments such as Schwartz's theory of basic values or the World Values Survey (Schwartz, 2012; Haerpfer et al., 2022). A recent survey organizes this space and reviews the main methodological choices involved in evaluating attitudes, opinions, and values in LLMs (Ma et al., 2024). More recent frameworks such as VALUECOMPASS and VALUEACTIONLENS move beyond static elicitation by measuring contextual value alignment and the gap between stated values and downstream actions, showing that LLMs may articulate value preferences that are not consistently reflected in behavior (Shen et al., 2025b,a).

A related line of work studies persona and sociodemographic prompting for probing or steering model behavior. This literature shows that demographic cues, ideological labels, and persona descriptions substantially affect generated opinions, value-laden judgments, and social stereotypes (Santurkar et al., 2023; Hwang et al., 2023; Liu et al., 2024; Lutz et al., 2025). However, most evaluations still operationalize identity as either a single demographic attribute or a fixed, internally coherent persona. Even work that goes beyond single-axis identities has primarily focused on stereotype measurement for intersectional demographic groups rather than multicultural identities (Cheng et al., 2023).

The body of work shares a common limitation: it treats cultural and demographic identities as static, monolithic categories. In practice, users often hold more complex, intersectional identities shaped by multiple cultural backgrounds. Ignoring these risks disregards the specific combination of values and perspectives this person holds (Weidinger and et al.,

System prompt

```
### ROLE & GOAL
You are an AI assistant....

### USER INFORMATION
The following information about the user is
    available:
User Country: {user_country}

### INSTRUCTIONS
1. Carefully consider the contextual
    information provided.
...
```

User prompt

```
Context: I was born in {second_identity}

{Query}
```

Figure 2: Prompt template used for response generation. Highlighted spans indicate substitutions.

2021; Fricker, 2007). Our methodology extends existing single-identity evaluations to assess LLM behavior under composite cultural identities.

3 Methodology

3.1 System Prompts and User Identity

Personalization in current chatbot systems can occur in the following ways (Anthropic, 2025; OpenAI, 2024): (1) static identity encoding user context in the system prompt (location, timezone, cultural preferences, etc.) learned from contextual information about the user and stored in memory (DeChant, 2025); (2) explicit (e.g., user interaction styles, stated preferences) and implicit (e.g., if the user changes the language) cues provided during the course of a conversation which may conflict with the static component. We operationalize composite identities by specifying one identity via the system prompt (user country) and introducing another through the user prompt, either explicitly (e.g., a self-declared birthplace) or implicitly through language choice that differs from the system-encoded identity (Figure 2; complete prompt in Appendix B). The explicit identity cues used to signal user identity in the user prompt are drawn from the disclosures dataset (Mireshghallah et al., 2024), which is mined from real user conversations in the WildChat dataset (Zhao et al., 2024). These cues take the form of natural first-person expressions such as "I am from X" or "I was born in X", etc. Implicit cues are operationalized via query language: the model is prompted in the language associated with a given country rather than through an explicit nationality mention.

3.2 Models and Tasks

We test this setup across three different tasks: World Values survey questions (multiple choice), open-ended generation using the templates from IssueBench (Röttger et al., 2025) and topics derived from the World Values Survey (WVS), and refusal rates on culturally sensitive requests (Aakanksha et al., 2024). All the tasks have a multilingual component for the language cues part. We perform our experiments with Llama-3.1-8B-Instruct (Llama) (Grattafiori et al., 2024), Gemma-3-27B-IT (Gemma) (Team et al., 2025), and Ministral-3-14B-Instruct-2512 (Mistral) (Liu et al., 2026). These models span three families and parameter scales (8B, 14B, 27B). Note that our aim is not a controlled comparison between models but to test whether identity cue effects generalize across diverse architectures and generations.

4 Value Probing using WVS

In this experiment, we probe LLMs with World Values Survey (WVS) questions to study and identify the values that LLMs associate with multicultural identities. We specifically use the subset of questions from Liu et al. (2025) that contains 144 questions with the highest variance of opinions across cultures. The dataset contains (human-verified) translations of questions in: Spanish, Chinese, Japanese, Korean, German, Russian, Vietnamese, and Portuguese. For probing, we use the countries from Liu et al. (2025): Vietnam, Colombia, Argentina, South Korea, Spain, Japan, China, Germany, Portugal, and Russia.

Model Probing: In both experiments, we fix a *base identity* s in the system prompt (Figure 2) and introduce a composite *target identity* $c \neq s$ via either an explicit or implicit cue. To estimate response distributions, we sample diverse user profiles by varying age $\in \{20, 30, \ldots, 80\}$ and gender $\in \{\text{masc.}, \text{fem.}, \text{other}\}$ following Waller and Anderson (2021), with 10 distinct cue phrasings at temperature 0.

Let $P_q(s)$ denote the response distribution for question q under base identity s (English, no composite cue), and $P_q(c)$ the distribution under target identity c. For each experiment, we introduce a composite identity cue and obtain a shifted distribution $\tilde{P}_q(s)$, then measure how much the cue moves s *toward* c:

$$\Delta_q(s,c) = \underbrace{\text{JSD}(P_q(s),\, P_q(c))}_{\text{divergence before cue}} - \underbrace{\text{JSD}(\tilde{P}_q(s),\, P_q(c))}_{\text{divergence after cue}} \quad (1)$$

where JSD is the Jensen-Shannon divergence (square root form, $\in [0,1]$). A positive Δ_q indicates the cue moved s's responses *toward* c; a negative value indicates divergence. We report the mean over all questions Q and all base identities $s \neq c$:

$$\bar{\Delta}(c) = \frac{1}{|Q|\,|S|} \sum_{q \in Q} \sum_{s \in S} \Delta_q(s,c) \quad (2)$$

The two experiments differ only in how $\tilde{P}_q(s)$ is induced: **(1) Explicit cue:** the user prompt states

Target Country (c)	Lang.	Explicit Mention ↑			Language Effect ↑		
		Gemma	Llama	Mistral	Gemma	Llama	Mistral
Argentina	Spanish	0.013	-0.060	-0.062	-0.401	-0.316	-0.319
China	Chinese	-0.008	-0.065	-0.048	-0.371	-0.261	-0.292
Colombia	Spanish	0.018	-0.052	-0.052	-0.414	-0.317	-0.315
Germany	German	0.046	-0.057	-0.055	-0.328	-0.289	-0.275
Japan	Japanese	0.032	-0.059	-0.055	-0.360	-0.307	-0.330
Portugal	Portuguese	-0.020	-0.064	-0.070	-0.316	-0.286	-0.277
Russia	Russian	0.038	-0.073	-0.008	-0.372	-0.249	-0.244
South Korea	Korean	-0.025	-0.049	-0.047	-0.417	-0.330	-0.268
Spain	Spanish	-0.019	-0.067	-0.072	-0.365	-0.324	-0.324
Vietnam	Vietnamese	0.003	-0.074	-0.046	-0.422	-0.315	-0.313

Table 1: Mean movement toward the target country distribution $\bar{\Delta}(c)$ (Eq. 2) under explicit and language cues across three models (↑ higher is better). Green indicates movement toward $P_q(c)$; red indicates divergence, with darker shading reflecting larger magnitude. Language cues produce uniformly large negative shifts, while explicit mentions show small, model-dependent effects.

the target country (*"I am from [c]"*), and **(2) Implicit cue:** the question is posed in the language L_c associated with c, with no country stated.

4.1 Results

Table 1 reports the mean movement toward the target country distribution for each model and target country under both cues. The **Language Effect** columns are negative across all models and countries (-0.244 to -0.422), indicating that posing questions in a country's language consistently moves answer distributions *away* from that country's baseline. In contrast, the **Explicit Mention** columns show smaller, model-dependent effects (-0.072 to $+0.046$): Gemma produces small positive shifts toward the target for most countries, while Llama and Mistral show near-zero or negative shifts, suggesting explicit country mentions have limited and inconsistent influence on model responses. The language effect is also consistent across Spanish-speaking countries (Argentina, Colombia, Spain; Gemma: -0.365 to -0.414), pointing to a language-specific shift rather than one driven by the cultural identity associated with that language. We note that the target country baseline is estimated with English-language prompts throughout, while the language cue condition poses questions in the target country's language; the negative language effect may therefore partly reflect a cross-lingual distributional shift, consistent with language not being a reliable proxy for cultural identity in these models.

Takeaways. Language cues produce larger distributional shifts than explicit country mentions, but move answer distributions away from the target country baseline rather than toward it. Explicit mentions yield small, model-dependent changes with no consistent movement toward P_c across models. Taken together, cue *form* determines the magnitude of distributional shift more than the country identity being signalled. More broadly, when cultural identity is expressed through language rather than stated explicitly, models do not shift toward the value distribution of the associated country suggesting that implicit identity expression through language is processed differently from explicit country-level identity cues.

5 Values in Open-Ended Generation

In this section, we look into the values expressed by LLMs in open-ended generation and the effect of composite identities on open-ended generation. To quantify cultural value shifts based on composite identities, we create a pipeline that probes LLMs using sociopolitical topics, annotates generated content with expressed values, and measures the shifts in values for the composite identities.

Prompt Curation and Sampling. We start by creating a list of topics based on WVS. We curate the list of topics using a human-in-the-loop pipeline discussed in Appendix D. We use a total of 144 topics and sample 5 prompt templates from Issuebench, resulting in 720 prompts for open-ended generation. We translate the prompts into Spanish, Chinese, Japanese, Korean, German, Russian, Vietnamese, and Portuguese using Google Translate API (Google Cloud, 2026). We then use the prompts to generate open-ended responses from the model in an implicit (language) and an explicit setup similar to Section 4. Forexplicit cues, we

use the same set of countries: Argentina, China, Colombia, Germany, Japan, New Zealand, Portugal, Russia, South Korea, Spain, the United States, and Vietnam.

Value Annotation and Clustering. Given the open-ended nature of the generation, we used the Values-in-the-Wild (Huang et al., 2025) approach to annotate each response with the values elicited in the response. The Values-in-the-Wild approach uses an LLM-based value annotator to label the values expressed in the free-text. We also validate our value-annotator using the test-split of Value Kaleidoscope dataset (Sorensen et al., 2024) and report a precision of 0.91 and a recall of 0.83.

Many of the resulting 5,900 unique value expressions are sparse, long-tailed, and often semantically redundant, making raw string-level representations not suitable for our analysis. We therefore cluster them using the Qwen3 embedding model (Zhang et al., 2025) in the Sentence Transformer library (Reimers and Gurevych, 2019). We perform agglomerative clustering, identifying 24 clusters of values, with the silhouette score peaking at 0.40. We then use gpt-4.5 (OpenAI, 2025) to assign each cluster a moral value label that represents all the expressions in that cluster (we report this set of moral values in Appendix C.)

Vector Representation and Shift Calculation. To represent the values reflected in each response, we map the annotated value expressions to these 24 clusters, yielding a 24-dimensional binary vector where each entry denotes the presence of a corresponding value. We aggregate these vectors across all responses for three conditions: the **User** baseline (**U**), where only the user's national identity is specified; the **System** baseline (**S**), where only the system persona's national identity is specified; and the **Composite** condition (**C**), where both identities are specified simultaneously. To measure which identity's values the composite response aligns with, we compute:

$$\text{Shift} = \text{Sim}(C, U) - \text{Sim}(C, S) \quad (3)$$

A positive shift indicates that the value profile of the composite identity more closely resembles the cultural values of the identity specified in the user cue than those specified in the system prompt.

Multilingual Setup for Implicit Cues. To study the effect of implicit cues on generated values, we translate the open-ended generation prompts to Spanish, Chinese, Japanese, Korean, German, Russian, Vietnamese, and Portuguese using Google Translate. The translated prompts are used to elicit open-ended responses from the models, with explicit instructions to respond in the language of the prompt. Since applying Google Translate at the scale of several hundred thousand responses would be prohibitively expensive, we back-translate the generated responses into English using the Aya multilingual model (Üstün et al., 2024), following the translation prompt specified in its technical report. To validate translation quality, we compare Aya against Google Translate on 100 randomly sampled examples per language, achieving an average COMET score (Rei et al., 2020) of 0.896 and BERTScore F1 (Zhang et al., 2020) of 0.916, indicating strong agreement. We then run value annotation, clustering, and aggregation as described in the previous part of this section. Since the the value clusters correspond to abstract moral foundations such as *global solidarity*, *civic duty*, and *quality of life* present across all human cultures (Graham et al., 2013; Atari et al., 2023), translation-induced lexical variation does not affect cluster assignment: a response expressing *care for the vulnerable* in Spanish maps to the same cluster as its English equivalent, consistent with prior cross-lingual value analysis (Wu et al., 2023).

5.1 Results

We analyze value shifts in composite identities (Equation 3) across two experimental conditions: implicit cues (language-based) and explicit cues (direct nationality mentions). As shown in Table 2, all three models exhibit substantial shifts toward the language-associated cultural value profile, with average shifts ranging from 0.175 (Gemma on Russian) to 0.966 (Llama on Vietnamese). Llama demonstrates the strongest language sensitivity, with shifts exceeding 0.80 for most languages and approaching complete alignment (0.966) for Vietnamese prompts. Gemma shows more moderate shifts (0.175-0.313), while Mistral occupies an intermediate range (0.414-0.708).

Figure 3 plots, for each composite identity, the cosine similarity of its value profile against the system baseline (x-axis) and the user baseline (y-axis). A point above the diagonal indicates that the composite aligns more closely with the user-indicated identity (cue) than with the system persona; a point below the diagonal indicates the reverse. Circle

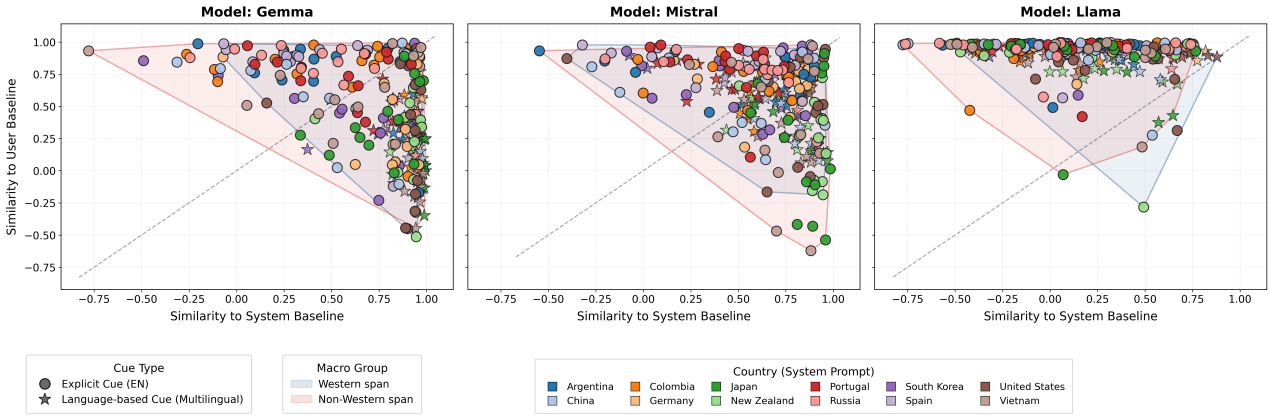

Figure 3: Cultural value shifts in composite identities under explicit (circle markers) and implicit/language-based (star markers) cues (section 5). We plot the cosine similarity of composite identity responses against the system-prompted country baseline (x-axis) and the user-indicated language baseline (y-axis). A point in the upper left indicates that the values in the model's response align more with the user-indicated identity than the system-prompted one, while a point in the lower right indicates the model predominantly follows the system-prompted identity.

markers correspond to explicit nationality cues and star markers to implicit language-based cues. The shaded regions show the convex hull of country pairs where the user-indicated identity corresponds to a Western country (blue; Germany, New Zealand, Portugal, Spain, United States) and a non-Western country (red; Argentina, China, Colombia, Japan, Russia, South Korea, Vietnam).

Western user cues are less effective at shifting the composite response. As visible in Figure 3, the blue (Western) convex hull consistently sits closer to the diagonal or further into the lower-right region compared to the red (non-Western) hull, meaning composite responses under Western user cues retain more of the system persona's value profile. Concretely, in Gemma, both groups are system-dominant, but Western user cues are more so (-0.597 vs. -0.147, averaged across cue types). In Mistral, Western user cues remain near-neutral (-0.012) while non-Western user cues shift system-dominant (-0.126). In Llama, both groups are user-dominant, but non-Western user cues produce a larger shift ($+0.780$ vs. $+0.630$). Across all three models and both cue types, a user indicating a non-Western identity shifts the composite response further from the system persona's value profile than a user indicating a Western identity.

Explicit and implicit cues produce consistent shifts. Across all three models, both cue types produce shifts in the same direction and of comparable magnitude (see star vs. circle markers in Figure 3). For Gemma, Western pairs yield shifts of -0.614 (explicit) and -0.581 (implicit), both system-dominant, while non-Western pairs yield $+0.186$ and $+0.108$, both user-side. For Mistral, both cue types are near-neutral for Western pairs ($+0.007$, -0.030) and modestly system-dominant for non-Western pairs ($+0.010$, -0.262). For Llama, both are user-side: $+0.753$ and $+0.507$ for Western, and $+0.816$ and $+0.743$ for non-Western. Within each model, the two cue types land on the same side of the diagonal; whether identity is signalled through an explicit nationality mention or through language, the composite value profile shifts in the same direction. This has been observed in single-identity probing settings (Cheng et al., 2023), and we find it holds in composite-identity scenarios as well.

Takeaways. Both cue types shift composite responses in the same direction across all models, which contrasts with the MCQ setting (§4), where language-based cues shift responses *away* from the value profile of the cued language's associated culture a Chinese-language prompt makes responses *less* similar to Chinese cultural values. In the open-ended setting, the same prompt shifts the composite value profile *toward* Chinese cultural values. In MCQ, discrete answer choices constrain how value positions are expressed, redirecting identity cue effects away from the culturally expected response (Röttger et al., 2024a); in open-ended generation, no such constraint exists. The consistency between cue types observed here is absent in the MCQ condition, further suggesting that the differences between explicit and implicit cues in prior work may partly reflect format constraints rather

Language	Gemma	Llama	Mistral
Chinese	0.195	0.902	0.510
German	0.206	0.893	0.445
Japanese	0.313	0.908	0.624
Korean	0.285	0.916	0.708
Portuguese	0.242	0.803	0.434
Russian	0.175	0.895	0.552
Spanish	0.213	0.820	0.414
Vietnamese	0.236	0.966	0.581

Table 2: Avg shift in value distribution when the language is used as a cue to signal identity

than differences in how models process identity signals.

6 Refusal Rates

Prior work has found how safety of models is dependent on identity through demographic personas you assign to the model (Plaza-del Arco et al., 2025) and to the values induced into the models (Choi et al., 2025). Safety mechanisms in models should not vary based on which identity the user signals or through which cue. Such biases are concerning for two reasons: they create inequitable safety protection across different cultural groups, and they expose implicit assumptions about which identity combinations are perceived as trustworthy or threatening. We examine how our composite identity setup, where two identities are signalled to the model through different cues, affects model refusal rates on harmful requests. We measure change in refusal rate when the same identity is provided to the model, but through a different cue (language vs system prompt) or when a different identity is provided.

6.1 Experimental Setup

For experiments, we use the Aya red-teaming dataset (Aakanksha et al., 2024) in a cross-lingual setup. The dataset contains culturally sensitive queries that an LLM should ideally refuse, as they contain requests about hate-speech content generation. The dataset contains harmful requests in Arabic, English, Filipino, French, Hindi, Russian, Serbian, and Spanish corresponding to Egypt, the United States, the Philippines, France, India, Russia, Serbia, Spain, and Argentina, respectively.

Refusal denotes responses where the model declines to fulfill a request, typically citing safety or policy restrictions (Röttger et al., 2024b), and the refusal rate for an identity is the proportion of such responses. We use an LLM-based judge to classify each model response as a refusal or not, following the prompt optimized for this task by Ahmadian et al. (2024). While Ahmadian et al. (2024) use GPT-4 as the judge, we use Llama-3.1-Instruct-70B for economic feasibility, further optimizing the prompt on a validation set of 100 examples per language, achieving 98% agreement with the GPT-4 judge on those examples. We then evaluate on 500 randomly sampled examples stratified across languages, obtaining an overall agreement of 94%.

6.2 Results

Tables 4 and 5 show how refusal rates change under two types of identity cues: switching the prompt language (implicit cues) and explicitly mentioning a user's nationality in the prompt (explicit cues). While both cue types affect refusal behavior, implicit cues produce considerably larger changes in magnitude. For example, the average refusal change across languages on Mistral is -12.90 points for implicit cues compared to -3.11 points for explicit cues. This is consistent with our findings in previous sections, where language-based cues had more pronounced effects on model behavior than explicit nationality mentions.

However, the countries with the largest explicit-cue changes (Table 5) do not correspond to those with the largest language-based effects (Table 4). For example, Filipino prompts cause a -22.07 point drop on Mistral, whereas explicit Philippine nationality mentions produce only -4.12 points. Similarly, Arabic prompts result in a -15.99 point drop on Mistral, while explicit Egyptian identity yields only -2.74 points. These findings suggest that the two cue types have varied effects, with linguistic signals exerting a stronger influence on guardrail sensitivity than demographic markers (Masud et al., 2024). Since explicit cues produce relatively moderate and uniform changes across models, while language-based cues show greater variation, we examine the latter in more detail. Table 3 breaks down the effect of introducing a nationality via the system prompt, conditioned on the prompt language. We observe a consistent drop in refusal rates when any nationality is introduced, including when it matches the language's associated country (**Same-ID**). This effect is particularly pronounced for languages with lower baseline refusal rates such as Filipino, Serbian, and Hindi, where Same-ID drops on Llama reach -42.72, -34.10, and -32.24 points respectively. For several of these, matched identities reduce re-

Lang	Llama			Mistral			Gemma		
	Base	Same-ID	Avg-Other	Base	Same-ID	Avg-Other	Base	Same-ID	Avg-Other
Arabic	76.03	-16.78	2.22	41.36	-19.78	-16.46	52.08	-15.33	-9.89
English	81.26	0.81	0.39	46.35	-1.01	0.56	62.74	-3.14	-0.51
Filipino	56.59	-42.72	-34.17	31.54	-19.03	-21.89	52.70	-12.49	-8.18
French	83.00	-8.61	-5.77	49.51	-12.18	-9.84	64.73	-9.23	-5.67
Hindi	70.49	-32.24	-26.98	48.66	-14.86	-12.72	67.40	-8.52	-2.91
Russian	80.61	-7.75	-5.23	51.79	-8.04	-6.00	70.80	-7.15	-1.37
Serbian	65.46	-34.10	-31.45	41.63	-13.72	-15.62	57.53	-10.14	-7.12
Spanish	82.42	-3.07	2.21	45.68	-10.10	-10.12	61.83	-5.50	-3.82

Table 3: Refusal rate changes when nationality is mentioned in system prompt. **Base**: refusal rate with language cue only. **Same-ID**: change when system prompt nationality matches the language's country. **Avg-Other**: average change across mismatched nationalities. All values are percentage point deltas from **Base**; negative values indicate worse safety. More details in subsection 6.2

Language	Gemma	Mistral	Llama
Arabic	**-8.795**	-15.992	**2.527**
Filipino	-7.459	**-22.068**	**-33.978**
French	-5.289	-9.246	-5.556
Hindi	-2.332	-12.513	-26.703
Russian	**-0.936**	**-6.132**	-5.163
Serbian	-6.685	-14.323	-31.164
Spanish	-4.487	-10.048	0.085
Avg.	**-5.140**	**-12.903**	**-14.279**

Table 4: Average change in refusal rates for multicultural identities when the prompt language is changed. Darker shading indicates larger absolute effects; bold marks the highest and lowest value in each model column. Negative values indicate worse refusal behavior.

User Country	Gemma	Mistral	Llama
Argentina	-0.26	-3.78	1.92
Colombia	-1.07	-3.70	2.93
Egypt	0.15	-2.74	2.63
France	0.28	-2.40	1.79
India	**0.64**	**-2.36**	2.70
New Zealand	-0.33	-3.68	2.38
Philippines	**-1.56**	**-4.12**	1.80
Russia	0.46	-3.01	2.56
Serbia	0.35	-2.71	2.43
Spain	-0.36	-2.47	2.45
US	-1.08	-2.53	**1.60**
Venezuela	-1.53	-3.76	**3.99**
Avg.	**-0.359**	**-3.105**	**2.432**

Table 5: Average change in refusal rates when a country is mentioned in the user prompt, averaged across system prompts. Darker shading indicates larger absolute effects; bold marks the highest and lowest value in each model column. Positive values indicate improved refusal behavior, while negative values indicate worse refusal behavior.

fusal rates *more* than mismatched ones (e.g., Filipino on Llama: -42.72 Same-ID vs. -34.17 Avg-Other), suggesting that coherent identity cues may lead models to lower guardrails more readily. In contrast, English remains largely unaffected across all three models, with changes below 1 point, indicating comparatively robust safety alignment for English-centric configurations.

7 Discussion and Conclusions

When a user signals a composite identity (e.g., a French speaker living in India), a model should ideally reflect values consistent with that intersection. Our findings show that current models fall short: implicit linguistic cues produce larger effects than explicit demographic markers, and when the two cues conflict, the resulting value profile aligns more closely with the language-associated distribution than the nationality-associated one. While prior work has documented cultural biases under single-identity conditioning (Lutz et al., 2025; Adilazuarda et al., 2024; Sommerauer et al., 2025; Mukherjee et al., 2024), our composite setup shows these biases interact in ways single-cue evaluations do not capture, with lower-resource languages showing larger shifts in both value alignment and safety (Zhou et al., 2025a; Kim and Baek, 2024; Tao et al., 2024). Across value probing and open-ended generation, we find that explicit and implicit cues diverge in MCQ but produce consistent shifts in open-ended generation, suggesting that differences between cue types may partly reflect format constraints rather than differences in how models process identity signals. In the refusal rate analysis, matched composite identities reduce refusal rates more than mismatched ones, indicating safety behavior is sensitive to the coherence of identity signals. English remains near-baseline in refusal behavior while changes for Filipino, Serbian, and Hindi reach 30-40 points (Masud et al., 2024).

8 Limitations

Several limitations constrain the generalizability of our findings. A limitation shared by almost all NLP studies on culture, including ours, is the use country as a proxy for culture. This simplification overlooks cultures that do not match country boundaries as well as substantial within-country variation across regions, generations, socioeconomic groups, and individual identities (Adilazuarda et al., 2024).

Second, we examine composite identities formed from only two identities, provided through two cues - a system persona and a user cue while real-world multicultural identities involve more complex intersections across language, nationality, religion, ethnicity, and lived experience. Whether our findings extend to higher-order identity combinations remains an open question.

Third, country selection was constrained by the availability of WVS Wave-7 translations from (Liu et al., 2025), introducing sampling bias toward regions with established survey infrastructure. Several major cultural regions including much of sub-Saharan Africa and Southeast Asia are underrepresented, limiting the geographic scope of our conclusions.

Finally, our approach relies on external models for value extraction from open-ended text, clustering and naming values, as well as machine translation. While we evaluated each aspect, models are error-prone and can affect our results.

References

Aakanksha, Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. 2024. The multilingual alignment prism: Aligning global and local preferences to reduce harm.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt 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 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, Sara Hooker, et al. 2024. The multilingual alignment prism: Aligning global and local preferences to reduce harm. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12027–12049.

Anthropic. 2025. Claude sonnet 4.5 system card. https://www.anthropic.com/system-cards. System card listed on Anthropic's Model System Cards page. Accessed: 2026-03-17.

AI Anthropic. 2025. System card: Claude opus 4 & claude sonnet 4. *Claude-4 Model Card*.

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T. Stevens, and Morteza Dehghani. 2023. Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*, 125(5):1157–1188.

Rogers Brubaker and Frederick Cooper. 2000. Beyond "identity". *Theory and Society*, 29(1):1–47.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Sooyung Choi, Jaehyeok Lee, Xiaoyuan Yi, Jing Yao, Xing Xie, and JinYeong Bak. 2025. Unintended harms of value-aligned llms: Psychological and empirical insights. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31742–31768.

Chad DeChant. 2025. Episodic memory in ai agents poses risks that should be studied and mitigated. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 321–332. IEEE.

Miranda Fricker. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press. Google Scholar Citations: 12,504.

Google Cloud. 2026. *Cloud Translation documentation*. Google Cloud. Cloud Translation – Advanced (v3), accessed January 16, 2026.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in Experimental Social Psychology*, volume 47, pages 55–130. Academic Press.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2022. World values survey wave 7 (2017–2022) cross-national data-set. World Values Survey Association. Version 4.0.0.

Saffron Huang, Esin Durmus, Miles McCain, Kunal Handa, Alex Tamkin, Jerry Hong, Michael Stern, Arushi Somani, Xiuruo Zhang, and Deep Ganguli. 2025. Values in the wild: Discovering and analyzing values in real-world language model interactions. *arXiv preprint arXiv:2504.15236*.

EunJeong Hwang, Bodhisattwa Prasad Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore. Association for Computational Linguistics.

Minsang Kim and Seungjun Baek. 2024. Exploring large language models on cross-cultural values in connection with training methodology. *arXiv preprint arXiv:2412.08846*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Victoria R. Li, Yida Chen, and Naomi Saphra. 2024. ChatGPT Doesn't Trust Chargers Fans: Guardrail Sensitivity in Context. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, et al. 2026. Ministral 3. *arXiv preprint arXiv:2601.08584*.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850, Bangkok, Thailand. Association for Computational Linguistics.

Yang Liu, Masahiro Kaneko, and Chenhui Chu. 2025. On the alignment of large language models with global human opinion. *arXiv preprint arXiv:2509.01418*.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. 2025. The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23212–23237, Suzhou, China. Association for Computational Linguistics.

Bolei Ma, Xinpeng Wang, Tiancheng Hu, Anna-Carolina Haensch, Michael A. Hedderich, Barbara Plank, and Frauke Kreuter. 2024. The potential and challenges of evaluating attitudes, opinions, and values in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8783–8805, Miami, Florida, USA. Association for Computational Linguistics.

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

Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024. Trust no bot: Discovering personal disclosures in human-llm conversations in the wild. *arXiv preprint arXiv:2407.11438*.

Sagnik Mukherjee, Muhammad Farid Adilazuarda, Sunayana Sitaram, Kalika Bali, Alham Fikri Aji, and Monojit Choudhury. 2024. Cultural conditioning or placebo? on the effectiveness of socio-demographic prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15811–15837.

United Nations. 2025. International migrant stock 2024. key facts and figures.

OpenAI. 2024. Memory and new controls for ChatGPT. Blog post. Accessed: 2026-03-15.

OpenAI. 2025. Introducing gpt-4.5. `https://openai.com/index/introducing-gpt-4-5/`. Accessed: 2026-03-16.

Flor Miriam Plaza-del Arco, Paul Röttger, Nino Scherrer, Emanuele Borgonovo, Elmar Plischke, and Dirk Hovy. 2025. No for some, yes for others: Persona prompts and other sources of false refusal in language models. In *Proceedings of the 9th Widening NLP Workshop*, pages 268–282.

Ricardo Rei, Craig Stewart, Catarina Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Paul Röttger, Musashi Hinck, Valentin Hofmann, Kobi Hackenburg, Valentina Pyatkin, Faeze Brahman, and Dirk Hovy. 2025. Issuebench: Millions of realistic prompts for measuring issue bias in llm writing assistance. *ArXiv*, abs/2502.08395.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024a. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15295–15311.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024b. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Shalom H. Schwartz. 2012. An overview of the schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2(1).

Hua Shen, Nicholas Clark, and Tanu Mitra. 2025a. Mind the value-action gap: Do llms act in alignment with their values? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3097–3118.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Yu-Ju Yang, Nicholas Clark, Tanu Mitra, and Yun Huang. 2025b. Valuecompass: A framework for measuring contextual value alignment between human and llms. In *Proceedings of the 9th Widening NLP Workshop*, pages 75–86, Suzhou, China. Association for Computational Linguistics.

Pia Sommerauer, Giulia Rambelli, and Tommaso Caselli. 2025. Simulating identity, propagating bias: Abstraction and stereotypes in llm-generated text. *arXiv preprint arXiv:2509.08484*.

Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. 2024. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19937–19947.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd annual meeting of the Association for Computational Linguistics (volume 1: Long papers)*, pages 15894–15939.

Isaac Waller and Ashton Anderson. 2021. Quantifying social organization and political polarization in online platforms. *Nature*, 600(7888):264–268.

Laura Weidinger and et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*. Google Scholar Citations: 1,228.

Winston Wu, Lu Wang, and Rada Mihalcea. 2023. Cross-cultural analysis of human values, morals, and biases in folk tales. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5113–5125, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*.

Ke Zhou, Marios Constantinides, and Daniele Quercia. 2025a. Should llms be weird? exploring weirdness and human rights in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 2808–2820.

Naitian Zhou, David Bamman, and Isaac L Bleaman. 2025b. Culture is not trivia: Sociocultural theory for cultural nlp. *arXiv preprint arXiv:2502.12057*.

A Model details and Experiment Details

For all experiments, we use the vLLM library for efficient inference (Kwon et al., 2023). We evaluate three instruction-tuned open-weight models for response generation: **Llama-3.1-8B-Instruct** (Llama) (Grattafiori et al., 2024), **Gemma-3-27B-IT** (Gemma) (Team et al., 2025), and **Ministral-3-14B-Instruct-2512** (Mistral) (Liu et al., 2026). These models span three families and parameter scales (8B, 14B, and 27B). For judge-based evaluation, we use **Llama-3.1-70B-Instruct**. We provide the corresponding model repository codes in Table 6.

Unless otherwise stated, we use the default vLLM sampling hyperparameters, namely `temperature=1.0`, `top_p=1.0`, `top_k=0` (disabled), `presence_penalty=0.0`, `frequency_penalty=0.0`, and `repetition_penalty=1.0`. We set `max_tokens=2048` for all generations, and generate 5 samples per prompt for robustness.

Llama: We use `meta-llama/Llama-3.1-8B-Instruct`, available via Hugging Face[1].

Gemma: We use `google/gemma-3-27b-it`, available via Hugging Face[2].

Mistral: We use `mistralai/Ministral-3-14B-Instruct-2512`, available via Hugging Face[3].

Judge model: For LLM-as-a-Judge evaluation, we use `meta-llama/Llama-3.1-70B-Instruct`, available via Hugging Face[4]. Unless otherwise noted, we use the same decoding configuration as above.

For response generation, we use the three generation models above and produce approximately 2 Million generations per model, for a total of 6 million responses across all models and tasks. This stage required approximately 10 days on 8 Nvidia A100s. For judge-based analysis, we run `Llama-3.1-70B-Instruct` over the generated responses, which required approximately 20 days on 8 Nvidia A100s.

The names dataset used in this work is released under the Apache-2.0 license, a permissive open-source license that allows use, modification, and redistribution. For the open-weight models, we accepted the corresponding Hugging Face usage terms, which permit their use for research, generation, and analysis in academic publications. We used code and writing assistant Claude (model: Sonnet 4.5 (Anthropic, 2025)) to help us with code and feedback on writing. The code generated was manually tested and verified before running.

B Full Input Prompt

The input prompt used for signalling composite identity is provided in figure 4.

[1] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[2] https://huggingface.co/google/gemma-3-27b-it

[3] https://huggingface.co/mistralai/Ministral-3-14B-Instruct-2512

[4] https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct

C Value Clusters

The labels of clusters are provided in Table 7.

D WVS Questions to Topics

We derive topics from WVS questions using a human-in-the-loop pipeline. For each question, we use GPT-4.5 (OpenAI, 2025) via the OpenAI API to generate candidate topic labels using the prompt in Figure 5, which are then manually curated. We retain only topics that are specific enough to function as standalone open-ended prompts, faithful to the original question, and likely to elicit meaningful opinions or values. In a human-in-the-loop curation step, three annotators review and edit the LLM-generated candidate topics, evaluating each for clarity, conciseness, and semantic faithfulness. Candidates that are overly broad, ambiguous, or weakly grounded in the source item are revised or discarded, and the final topic is chosen through annotator consensus on the most concise formulation that preserves the original meaning.

E Detailed Refusal Rates Table

Role	Model	Hugging Face repository
Generation	Llama	meta-llama/Llama-3.1-8B-Instruct
Generation	Gemma	google/gemma-3-27b-it
Generation	Mistral	mistralai/Ministral-3-14B-Instruct-2512
Judge	Llama-3.1-70B-Instruct	meta-llama/Llama-3.1-70B-Instruct

Table 6: Models used in this study and their corresponding Hugging Face repository codes.

```
SYSTEM_PROMPT = """ROLE & GOAL
You are an AI assistant. Your goal is to apply the user's contextual
information to answer survey questions authentically with a reasoned,
open-ended response.

### CONTEXTUAL INFORMATION
The user has provided background information about themselves, including:
* **User background:** {user_country}

### INSTRUCTIONS
1. Carefully consider the contextual information provided.
2. Your response must reflect the mainstream values, social norms, and common attitudes of a typical
    citizen from the specified nation.
3. Provide a brief, open-ended response that a person with this profile would most likely give."""

# User prompt (shared across conditions)

USER_PROMPT = f"Contextual information: {context_text}\n\nQuestion: {prompt_text}"
```

Figure 4: System and user prompts used for the implicit linguistic cue condition.

Value Cluster Labels	
Environmental Stewardship & Nature	Social Justice & Equality
Community & Belonging	Family & Lineage
Democratic Integrity & Governance	Individual Liberty & Agency
Economic Prosperity & Opportunity	Health & Well-being
Security & Order	Compassion & Care
Knowledge & Truth	Innovation & Future Orientation
Spirituality & Transcendence	Global Solidarity & Peace
Personal Character & Virtue	Inclusion & Diversity
Fiscal Responsibility	Indigenous Rights & Wisdom
Patriotism & Sovereignty	Civic Duty & Service
Fairness & Reciprocity	Enjoyment & Quality of Life
Tradition & Conservatism	Collectivism & Cooperation

Table 7: The 24 value list

Lang	Egy	US	NZ	Phi	Fra	Ind	Rus	Ser	Spa	Arg	Col	Ven	Avg
Arabic	-15.33	-3.44	-9.67	-8.22	-11.00	-6.44	-6.22	-13.11	-9.11	-7.00	-7.11	-8.89	-8.80
English	0.41	2.03	-3.14	0.20	1.11	0.71	-1.62	-1.52	-0.20	-1.11	-0.41	0.30	-0.27
Filipino	-9.22	-4.16	-7.93	-12.49	-8.62	-4.46	-7.63	-9.32	-5.75	-3.87	-7.83	-8.23	-7.46
French	-4.55	-4.55	-5.90	-6.27	-9.23	-0.49	-7.38	-5.54	-6.03	-7.26	-1.35	-4.92	-5.29
Hindi	-5.25	0.66	-1.97	-3.39	0.98	-8.52	-3.17	-2.19	0.22	-2.84	-0.98	-1.53	-2.33
Russian	-1.99	2.88	2.18	1.69	-2.68	0.20	-7.15	-2.09	-1.09	0.99	-1.29	-2.88	-0.94
Serbian	-7.36	-5.17	-4.57	-7.65	-6.26	-3.68	-10.04	-10.14	-7.26	-6.16	-4.67	-7.26	-6.69
Spanish	-2.43	-3.58	-2.17	-2.17	-6.52	-2.05	-4.09	-5.63	-5.50	-5.12	-7.16	-7.42	-4.49
Avg	-5.72	**-1.92**	-4.15	-4.79	-5.28	-3.09	-5.91	-6.19	-4.34	-4.05	-3.85	-5.10	-4.53

Table 8: Avg change in refusals (Gemma). Columns ordered to show language-country correspondence along the diagonal.

Lang	Egy	US	NZ	Phi	Fra	Ind	Rus	Ser	Spa	Arg	Col	Ven	Avg
Arabic	-19.78	-14.56	-12.89	-12.11	-18.11	-18.33	-17.67	-15.67	-17.11	-15.89	-13.00	-16.78	-15.99
English	0.71	-2.13	-1.01	-2.33	3.04	1.22	-0.61	1.52	1.42	2.43	0.20	-1.32	0.26
Filipino	-21.21	-24.08	-20.22	-19.03	-23.89	-19.82	-22.00	-23.79	-25.17	-21.90	-20.32	-23.39	-22.07
French	-12.18	-8.24	-7.87	-8.36	-12.18	-13.04	-6.15	-10.21	-8.73	-6.15	-8.86	-8.98	-9.25
Hindi	-14.32	-10.49	-8.74	-10.27	-16.39	-14.86	-12.68	-12.35	-12.13	-13.11	-14.21	-10.60	-12.51
Russian	-4.57	-2.58	-3.57	-4.27	-4.87	-7.85	-8.04	-7.35	-7.45	-5.06	-8.34	-9.63	-6.13
Serbian	-11.63	-7.85	-16.00	-17.20	-16.30	-15.41	-15.41	-13.72	-19.28	-12.13	-12.82	-14.12	-14.32
Spanish	-9.46	-7.54	-9.08	-13.30	-10.74	-8.70	-8.18	-11.38	-10.10	-12.40	-8.57	-11.13	-10.05
Avg	-11.56	**-9.68**	-9.92	-10.86	-12.43	-12.10	-11.34	-11.62	-12.32	-10.53	-10.74	-11.99	-11.26

Table 9: Avg change in refusals (Mistral). Columns reordered for diagonal interpretation.

Lang	Egy	US	NZ	Phi	Fra	Ind	Rus	Ser	Spa	Arg	Col	Ven	Avg
Arabic	-16.78	3.00	5.44	9.22	4.33	-1.33	8.00	5.89	-1.89	8.11	3.33	2.50	2.53
English	1.11	1.82	0.81	0.71	-0.20	1.82	-1.01	-0.51	0.41	-0.51	1.01	0.10	0.46
Filipino	-34.79	-34.99	-33.70	-42.72	-29.53	-37.17	-30.13	-29.63	-35.68	-33.10	-32.80	-33.50	-33.98
French	-6.89	-1.48	-4.43	-7.13	-8.61	-3.44	-3.20	-5.41	-7.01	-4.80	-7.01	-7.26	-5.56
Hindi	-26.99	-30.27	-27.87	-26.56	-25.03	-32.24	-27.87	-25.36	-23.93	-22.19	-26.01	-26.12	-26.70
Russian	-3.87	-2.68	-7.75	-3.67	-3.67	-3.77	-7.75	-9.43	-1.89	-4.77	-6.55	-6.16	-5.16
Serbian	-30.91	-33.10	-30.72	-31.71	-30.52	-31.51	-33.50	-34.10	-28.63	-27.44	-30.82	-31.01	-31.16
Spanish	1.66	-0.77	0.51	0.64	2.81	5.63	4.99	4.48	-3.07	-8.57	-4.35	-2.94	0.09
Avg	-14.68	-12.92	-13.06	-13.13	-11.30	-12.21	**-11.16**	-11.50	-11.74	-11.66	-12.90	-12.99	-12.44

Table 10: Avg change in refusals (Llama). Columns reordered for diagonal interpretation.

Topic Phrase Extraction Prompt

System: You are an intelligent social scientist and survey methodologist. You specialize in mapping structured survey questions to concise topic phrases that preserve the underlying values and norms being measured.

User: Your task is to convert a multiple-choice survey question into a short topic phrase that fits naturally into writing templates such as *"Write a 100-word essay on {X}"* or *"Write a tweet on {X}"*.

Given a WVS question: `[question_text]`

Guidelines:

- Use a concise noun phrase (not a sentence or instruction)
- Preserve the values or norms elicited by the original question
- Do not presuppose a normative stance (e.g., avoid "benefits of")
- Avoid causal framing (e.g., avoid "leads to", "through")
- Prefer neutral, descriptive framing (e.g., perceptions, attitudes, norms)

Example — *Input:* "A man who cannot father children is not a real man."
Good: *Societal expectations of masculinity and fatherhood*
Bad: *How fatherhood defines real masculinity*

Return **only** the topic phrase.

Figure 5: Prompt used to extract neutral topic phrases from WVS questions for essay and social media generation tasks.