A Dual-View Mechanistic Framework for Cultural Value Alignment in Large Language Models

Anonymous ACL submission

Abstract

Cultural value alignment is a prerequisite for the safe and widespread deployment of Large Language Models (LLMs). To address the transparency and causal interpretability limitations of traditional questionnaire-based assessments, this paper proposes a Dual-View Mechanistic Framework grounded in linguistic categorization. This framework posits that meaningful value alignment manifests as consistent responsiveness to identical values and discriminative responsiveness to disparate ones, providing a mechanistic basis to explain performance variations among LLMs. Extensive experiments involving five LLMs and 31 distinct value topics across 14 national contexts yield a significant correlation of 0.78 between the proposed neuronal metrics and behavioral alignment, outperforming conventional discriminative-only approaches by a substantial margin of 0.51.These findings validate the framework as a faithful explanatory proxy for LLMs' value-awareness, offering a granular perspective for evaluating and interpreting cultural value alignment.

1 Introduction

With the widespread global deployment of Large Language Models (LLMs), their cultural value alignment has become crucial for safe, reliable, and acceptable applications across diverse sociocultural contexts (Ganguli et al., 2022; Prabhakaran et al., 2022; Kasirzadeh and Gabriel, 2022). Grounded in sociological and psychological theories (Bram et al., 1953; Yang, 1987; Hofstede et al., 1991; Christian Haerpfer and Puranen, 2022), current research predominantly employs standardized psychometric instruments to assess value alignment of LLMs. This paradigm typically prompts models to respond to multidimensional questionnaires under specific persona-conditioned settings, and derives quantitative alignment metrics by comparing model responses with human empirical benchmarks. Thanks to their high operability and broad coverage, such questionnaire-based evaluations not only enable cross-model and cross-culture comparisons but also uncover latent cultural value biases within LLMs (Masoud et al., 2023; Cao et al., 2023; AlKhamissi et al., 2024; Karinshak et al., 2024; Xu et al., 2024; Sukiennik et al., 2025).

Despite these advances, existing evaluations largely remain at the behavioral level, offering limited insight into how cultural values are internally represented within LLMs. This reliance on surface-level outputs constrains alignment assessment to observable behaviors, leaving the underlying mechanisms opaque and difficult to diagnose or improve, which ultimately limits robustness and trustworthiness. Although recent work has begun to investigate the internal mechanisms of LLM alignment more broadly (Zhao et al., 2025; Ying et al., 2025; Jin et al., 2025; Dai et al., 2025), principled evaluation frameworks that connect internal representations to diversity-aware cultural value alignment are still largely underexplored.

To bridge these gaps, inspired by linguistic categorization (Taylor, 2003), we introduce a Dual-View Framework. This framework operationalizes the mechanistic study of cultural value alignment by linking internal neuronal representations to behavioral outputs. ***It posits that meaningful cultural value alignment is not only characterized by discriminative responsiveness to disparate values (Su et al., 2025) but also by consistent responsiveness to identical ones.*** For example, within a given national context, a well-aligned model should respond consistently to the topic of "Friend" across diverse query formulations, yet effectively differentiate this topic from "Family", reflecting their distinct cultural value meanings.

By synergistically modeling the ***Intra-value View*** (focusing on consistency) and the ***Inter-value View*** (focusing on discriminability), our framework empirically validates the theoretical assumptions introduced above. Our framework achieves a strong

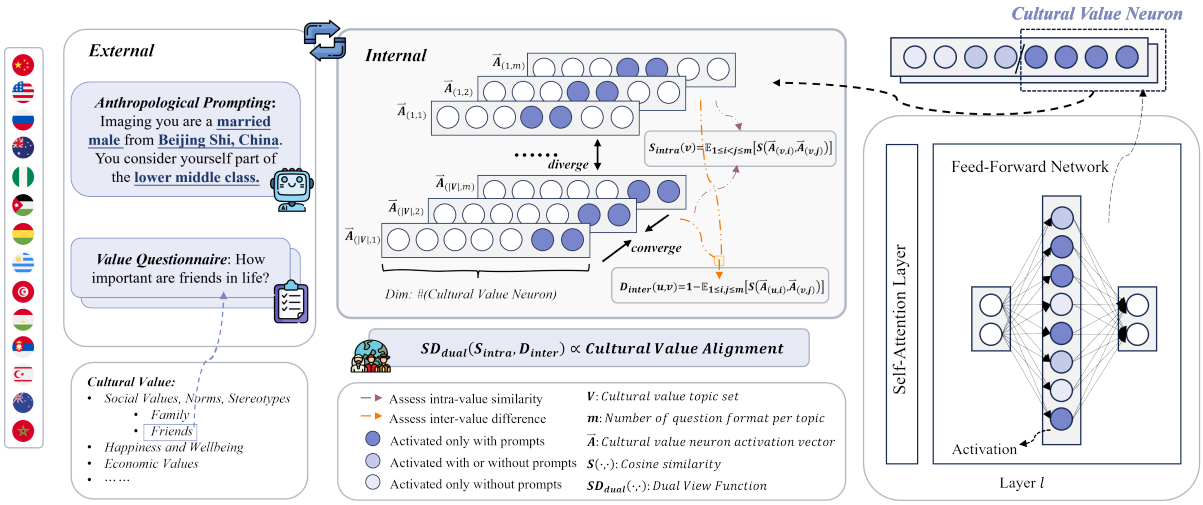

Figure 1: **Mechanistic Dual View Framework** proposed in this work.

correlation between internal neuronal metrics and external behavioral alignment scores, demonstrating that culturally grounded values are reliably reflected in internal representations.

2 Preliminary

Questionnaire-based Behavioral Evaluation Given a set of national contexts C, a set $V = \{v_1, v_2, ..., v_n\}$ encompassing n independent value topics (e.g., "Family", "Friend"), corresponding questionnaire items q—each with has m linguistic variations (i.e., paraphrases)— are designed for each value topic $v \in V$ to form the questionnaire set $Q = \{q_{(1,1)}, q_{(1,2)}, ..., q_{(n,m)}\}$. Subsequently, to evaluate the cultural value alignment of model $\mathbb{M}$ within a specific national context $c \in C$ (e.g., "China"), we prompt $\mathbb{M}$ to simulate various personas P under c (e.g., "A Chinese female student") to respond to Q. By calculating the behavioral similarity between the generated responses of $\mathbb{M}$ and the ground-truth responses of P, we can quantify the degree of alignment between $\mathbb{M}$ and target cultural values (AlKhamissi et al., 2024; Karinshak et al., 2024; Xu et al., 2024; Sukiennik et al., 2025).

Key Neuron Identification In sociology, cultural values are defined as a structured system of knowledge developed by specific groups from their collective experiences over time (Schwartz, 2006, 2012). In the context of LLMs, recent interpretability research has established that factual knowledge in Transformers is stored within the Feed-Forward Network (FFN) layers as key-value memories (Geva et al., 2021; Tang et al., 2024; Ying et al., 2025). Therefore, we aim to identify the corresponding cultural value neurons within the FFN modules. In a decoder-only Transformer, the FFN at layer l can be formulated as:

$$\text{FFN}^l(h^l) = W^l_{\text{down}} \cdot \text{Act_Fn}(W^l_{\text{up}} \cdot h^l) \quad (1)$$

where h^l denotes the input to the FFN at layer l, W^l_{up} and W^l_{down} as the weight matrices and $\text{Act_Fn}(\cdot)$ refers to the activation function.

Following prior work (Tang et al., 2024; Ying et al., 2025), the i-th element of $\text{Act_Fn}(W^l_{\text{up}} \cdot h^l) \in \mathbb{R}^{dim}$ is considered as the i-th neuron in layer l, denoted as $n_{(l,i)}$. Neurons exhibiting higher activation values during inference are deemed more critical. Consequently, given a question q and the generated response $R = \{r_1, r_2, ..., r_{|R|}\}$ from $\mathbb{M}$, key neurons N_q are identified by selecting $n_{(l,i)}$ with activation values $a_{(l,i)}$ exceeding a threshold:

$$N_q = \{n_{(l,i)} \mid a^{r_k}_{(l,\,i)} > \mu(l, r_k)\ ,\ r_k \in R\} \quad (2)$$

where $\mu(l, r_k)$ is the threshold determined by the K-th largest activation value across all neurons in layer l during the generation of token r_k, where K is a pre-defined constant, set to 100 in this study.[1]

3 Mechanistic Dual-View Framework

Grounded in the principle that robust cultural value alignment should exhibit intra-value consistency and inter-value distinctiveness. To capture these two properties, we propose a Dual-View Framework in this section (as illustrated in Figure 1).

[1]In this work, we experimentally demonstrate the robustness of the K across various models (see Appendix G).

3.1 Cultural Value Neuron Identification

Extract Cultural Value Neurons To isolate neurons associated with cultural value expression, we employ a decoupling procedure:

- Define N_G (general) and N_P (role-play) as neurons activated by questionnaire set Q *without* and *with* persona prompts P, respectively.
- Extract the cultural value neuron set N_V through the set difference $N_P \setminus N_G$. This approach filters out task-general activations, isolating neurons specifically triggered by cultural-reasoning requirements.

Importantly, our approach does not assume that base prompts are incapable of activating value-encoding neurons. Rather, we posit that *neurons activated exclusively under anthropological persona prompts possess more extensive value-related knowledge*, a hypothesis validated by our ablation studies (see Appendix F).

Formulate Neuronal Activation Vectors Let $N_{q_{(i,j)}} \in N_V$ denote subset of cultural value neurons activated by $q_{(i,j)}$. Subsequently, the neuronal activation vector $\vec{A}_{(i,j)} \in \mathbb{R}^{|N_V|}$, is defined as follows, where $|N_V|$ denotes the cardinality of N_V:

$$\vec{A}_{(i,j)} = \begin{cases} 1 & \text{if } n \in N_{q_{(i,j)}} \\ 0 & \text{otherwise} \end{cases}, \text{for } n \text{ in } N_V \quad (3)$$

$\vec{A}_{(i,j)}$ serves as the binarized neuronal activation vector elicited by a given value-question pair. Fundamentally, binarization serves as a form of discrete representation, a methodology proven by prior works to effectively mitigate fluctuations induced by activation intensity (Hinton, 1984; Van Den Oord et al., 2017). Detailed statistics of neuron numbers across different national contexts are provided in Appendix J.

3.2 Dual-View Neuronal Evaluation of Cultural Value Alignment

Intra-Value View (S_{intra}) For samples belonging to an identical cultural value topic, their neuronal activation vectors are expected to demonstrate maximal alignment. For a specific value v, this consistency is quantified by the average pairwise similarity S_{intra}, defined as:

$$S_{\text{intra}}(v) = \mathbb{E}_{1 \le i < j \le m}[S(\vec{A}_{(v,i)}, \vec{A}_{(v,j)})] \quad (4)$$

where $S(\cdot,\cdot)$ denotes the cosine similarity, $\mathbb{E}[\cdot]$ denotes the statistical expectation.

Inter-Value View (D_{inter}) Conversely, activation vectors from disparate value topics are expected to exhibit minimal alignment. This discrimination is measured by the average pairwise dissimilarity D_{inter}, defined as:

$$D_{\text{inter}}(u,v) = 1 - \mathbb{E}_{1 \le i,j \le m}[S(\vec{A}_{(u,i)}, \vec{A}_{(v,j)})] \quad (5)$$

where u and v denote a value pair ($u \neq v$).

Dual View (SD_{dual}) By integrating S_{intra} and D_{inter} across diverse personas and values, we derive the final dual-view alignment score for $\mathbb{M}$, denoted as SD_{dual} and formalized as:

$$SD_{\text{dual}} = \sqrt{\bar{S}_{\text{intra}} \cdot \bar{D}_{\text{inter}}} \quad (6)$$

where $\bar{S}_{\text{intra}}$ and $\bar{D}_{\text{inter}}$ denote the mean S_{intra} and D_{inter} metrics across all simulation respondents, respectively; $SD_{\text{dual}} \in [0,1]$, with a higher score corresponding to better cultural value alignment.

4 Experiment Setup

Dataset We employ the World Values Survey 7 (WVS-7) (Christian Haerpfer and Puranen, 2022) as our evaluative benchmark. WVS-7 encompasses 259 questions spanning diverse domains, aligned with the United Nations Sustainable Development Goals. In this work, we derived 31 questions from the corpus that encompass diverse topics to ensure broad coverage, with each question paraphrased three times to mitigate linguistic phrasing biases (AlKhamissi et al., 2024) (Details in Appendix B).

Models and Prompt Settings We select five mainstream LLMs: Llama-3-8B-Instruct, Llama-3.1-8B-Instruct (Touvron et al., 2023), Qwen2.5-7B-Instruct (Bai et al., 2023), gemma-2-9B-it (Riviere et al., 2024) and Mistral-7b-Instruct-v0.3 (Jiang et al., 2023). Our persona prompts incorporate specified national identities alongside granular demographic characteristics and are formulated in English (Details in Appendix D).[2]

Country Selection To ensure geographical diversity, we select 14 representative countries across five continents, sampling 80 respondents from each country. These include: Australia, Bolivia, China,

[2]We adopt English-only prompting as a controlled experimental design to establish a unified baseline, thereby isolating the evaluation from the confounding effects of varying multilingual proficiencies across models. Results indicate that this setting yields consistent value alignment trends across all studied countries, effectively mitigating concerns regarding linguistic bias (See Appendix H).

	ρ(SA) ↑	ρ(HA) ↑	gemma-2-9B-it	Llama-3-8B-it	Llama-3.1-8B-it	Qwen2.5-7B-it	Mistral-7B-it
Neuron-based Internal Evaluation							
VNR	-0.495***	-0.396***	0.024	0.073	0.019	0.074	0.108
S_{intra}	-0.163	-0.241*	0.642	0.634	0.546	0.547	0.796
D_{inter}	0.268*	0.351***	0.779	0.753	0.82	0.802	0.467
SD_{dual}	**0.779*****	**0.543*****	0.707	0.688	0.669	0.662	0.609
Questionnaire-based External Evaluation (References)							
SA	1.0***	0.779***	0.672	0.633	0.627	0.613	0.581
HA	0.779***	1.0***	0.39	0.383	0.383	0.366	0.338

Table 1: Evaluation of cultural value alignment across 14 national contexts. The rows labeled "Neuron-based Internal Evaluation" represent our primary contributions, while the "Questionnaire-based External Evaluation" rows provide ground-truth references for validation. Symbols *, **, and *** indicate statistical significance at the $p < 0.05$, 0.01, and 0.001 levels, respectively. See Appendix C for detailed implementation.

Cyprus, Jordan, Morocco, New Zealand, Russia, Serbia, Tajikistan, Tunisia, United States, Uruguay.

Evaluation Metric To assess the reliability of internal metrics—specifically, their faithfulness in reflecting external alignment performance—we calculate the **Spearman's correlation coefficient (ρ) between the internal metrics and the survey-based scores** for the model set M across national contexts C. Two external metrics serve as references: *1) Hard Accuracy (HA),* with its corresponding correlation denoted as ρ(HA), which quantifies exact response consistency with the specified persona; and *2) Soft Accuracy (SA),* with its corresponding correlation denoted as ρ(SA), a relaxed variant of HA that awards partial credit for ordinal-scale question responses. Details of these metrics are given in Appendix E.

Baselines We compare with following baselines:

- **Cultural Value Specific Neuron Activation Ratio (Value Neuron Ratio, VNR)**: Building upon the observation that language-culture-specific neurons sustain multilingual performance (Ying et al., 2025), we introduce a baseline centered on the activation density of value-associated neurons, defined as VNR $= \mathbb{E}_{p\sim P, q\sim Q}[|N_V(p,q)|/|N_P(p,q)|]$, where higher values indicate a greater proportion of activated value-related neurons.
- **Single Intra-Value View (S_{intra})**: Only consider the consistency from the intra-value view.
- **Single Inter-Value View (D_{inter})**: Only consider the discriminability from the inter-value view.

5 Result and Analysis

Table 1 summarizes the primary results. Our proposed dual-view strategy *yields the most robust correlations between internal neuronal metrics and external behavioral alignment scores.* Notably, it achieves a Spearman correlation of 0.779 ($p < 0.001$) on the SA metric, comparable to the observed SA-HA inter-correlation. Conversely, isolating either S_{intra} or D_{intra} triggers a sharp degradation in performance, with certain instances even exhibiting inverse relationships, as observed for S_{intra} in the ρ(HA) metric. These findings underscore that effective value alignment necessitates the integration of both intra- and inter-value perspectives. Additionally, the significant negative correlation between VNR and both SA and HA suggests that highly aligned cultural values are likely mediated by sparse, specialized circuits rather than distributed, dense activations.

Our dual-view framework further *explains the performance differences between models,* specifically elucidating why relying solely on either S_{intra} or D_{inter} fails to achieve optimal alignment. For instance, Llama-3.1 and Qwen2.5 prioritize value discrimination but lack representational stability. In contrast, Mistral shows the opposite limitation: it fails to maintain clear boundaries in its embedding space, making it difficult to distinguish between different value topics.

6 Conclusion

In this paper, we propose a dual-view framework to mitigate the lack of systematic and interpretable evaluation methodologies in value alignment. By synergistically modeling intra-value consistency and inter-value discriminability, the framework significantly strengthens the correlation between neuronal representations and external behavioral observations, offering novel perspectives and insights for assessing the cultural value alignment of LLMs.

Limitations

We now discuss the limitations of the proposed Dual-View framework. (1) A primary constraint is our reliance on static evaluation datasets, which fails to account for the context-dependent nature of cultural values. Moving forward, a critical research task is to bridge the gap between static benchmark evaluations and the evaluation of cultural alignment in complex, open-ended environments, particularly concerning the prevention of value drift. (2) Our evaluation relies on specific questionnaire benchmarks (SA/HA). A risk is that models might learn to exploit these patterns to score higher. Therefore, we suggest using our method as a supplementary tool to help identify issues, rather than using it as the only way to evaluate a model. Despite these limitations, our framework offers a novel and valuable perspective for investigating these intricate challenges.

References

Badr AlKhamissi, Muhammad N. ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. *ArXiv*, abs/2402.13231.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenhang Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, K. Lu, and 31 others. 2023. Qwen technical report. *ArXiv*, abs/2309.16609.

Joseph Bram, Alfred Louis Kroeber, Clyde Kluckhohn, Wayne Untereiner, and Alfred G. Meyer. 1953. Culture, a critical review of concepts and definitions.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *ArXiv*, abs/2303.17466.

Alejandro Moreno Christian Welzel Kseniya Kizilova Jaime Diez Medrano Marta Lagos Pippa Norris Eduard Ponarin Christian Haerpfer, Ronald Inglehart and Bi Puranen. 2022. World values survey wave 7 (2017-2020) cross-national data-set.

Xunlian Dai, Li Zhou, Benyou Wang, and Haizhou Li. 2025. From word to world: Evaluate and mitigate culture bias via word association test. *ArXiv*, abs/2505.18562.

Deep Ganguli, Liane Lovitt, John Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Benjamin Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, and 17 others. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *ArXiv*, abs/2209.07858.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. *Preprint*, arXiv:2012.14913.

Geoffrey E Hinton. 1984. Distributed representations.

Gert Jan Hofstede, Gert Jan Hofstede, and Michael Minkov. 1991. Cultures and organizations: Software of the mind.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Haoran Jin, Meng Li, Xiting Wang, Zhihao Xu, Minlie Huang, Yantao Jia, and Defu Lian. 2025. Internal value alignment in large language models through controlled value vector activation. In *Annual Meeting of the Association for Computational Linguistics*.

Elise Karinshak, Amanda Hu, Kewen Kong, Vishwanatha Rao, Jingren Wang, Jindong Wang, and Yi Zeng. 2024. Llm-globe: A benchmark evaluating the cultural values embedded in llm output. *Preprint*, arXiv:2411.06032.

Atoosa Kasirzadeh and Iason Gabriel. 2022. In conversation with artificial intelligence: Aligning language models with human values. *Philosophy & Technology*, 36:1–24.

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip C. Treleaven, and Miguel Rodrigues. 2023. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In *International Conference on Computational Linguistics*.

Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. 2022. Cultural incongruencies in artificial intelligence. *ArXiv*, abs/2211.13069.

Gemma Team Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L'eonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram'e, Johan Ferret, Peter Liu, Pouya Dehghani Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 176 others. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118.

Shalom H. Schwartz. 2006. A theory of cultural value orientations: Explication and applications. *Comparative Sociology*, 5:137–182.

Shalom H. Schwartz. 2012. An overview of the schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2:11.

Yi Su, Jiayi Zhang, Shu Yang, Xinhai Wang, Lijie Hu, and Di Wang. 2025. Understanding how value neurons shape the generation of specified values in llms. *Preprint*, arXiv:2505.17712.

Nicholas Sukiennik, Chen Gao, Fengli Xu, and Yong Li. 2025. An evaluation of cultural value alignment in llm. *Preprint*, arXiv:2504.08863.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. *ArXiv*, abs/2402.16438.

John R Taylor. 2003. *Linguistic Categorization*. Oxford University Press.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Aaron Van Den Oord, Oriol Vinyals, and 1 others. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2024. Self-pluralising culture alignment for large language models. In *North American Chapter of the Association for Computational Linguistics*.

Kuoshu Yang. 1987. Chinese values and the search for culture-free dimensions of culture. *Journal of Cross-Cultural Psychology*, 18:143 – 164.

Jiahao Ying, Wei Tang, Yiran Zhao, Yixin Cao, Yu Rong, and Wenxuan Zhang. 2025. Disentangling language and culture for evaluating multilingual large language models. In *Annual Meeting of the Association for Computational Linguistics*.

Yiran Zhao, Wenxuan Zhang, Yuxi Xie, Anirudh Goyal, Kenji Kawaguchi, and Michael Shieh. 2025. Understanding and enhancing safety mechanisms of llms via safety-specific neuron. In *International Conference on Learning Representations*.

A The Usage of LLM

In this work, the application of LLMs is strictly limited to aiding and polishing academic writing, e.g., the description of result analysis, with no involvement in core research processes.

B Detailed Description and Statistical Analysis of the WVS

The World Values Survey (WVS) collects responses to a range of questions broadly categorized into social, cultural, material, governmental, ethical, and economic domains, with data drawn from demographically controlled population samples worldwide (Christian Haerpfer and Puranen, 2022). Its seventh wave (WVS-7), conducted between 2017 and 2021, incorporates region-specific modules alongside globally standardized categories. Comprising 259 questions, WVS-7 is designed to include indicators aligned with multiple United Nations Sustainable Development Goals. Administered as a questionnaire, the survey targets selected samples from the general population, with all questions localized to the native or dominant regional languages.

The value topics used in this work are as follows: (1) Friends, (2) People of a different race, (3) Immigrants/foreign workers, (4) Basic kinds of attitudes concerning society, (5) People of another religion, (6) People of another nationality, (7) Major Companies, (8) Private banks, (9) The United Nations (UN), (10) International Monetary Found (IMF), (11) The World Bank (WB), (12) The World Health Organization (WHO), (13) Increases the crime rate, (14) Increases the risks of terrorism, (15) Helps poor people establish new lives, (16) Losing my job or not finding a job, (17) Not being able to give one's children a good education, (18) Freedom and Equality - Which more important, (19) Freedom and security - Which more important, (20) How often do you attend religious services, (21) To make sense of life after death vs To make sense of life in this world, (22) Interest in politics, (23) Signing a petition, (24) Joining in boycotts, (25) local level, (26) Votes are counted fairly, (27) Election officials are fair, (28) How much would you say the political system in your country allows people like you to have a say in what the government does?, (29) Having a strong leader who does not have to bother with parliament and elections, (30) Having experts, not government, make decisions according to what they think is best for the country, (31) Having a

system governed by religious law in which there are no political parties or elections.

Although the WVS-7 data was collected between 2017 and 2021 and may have certain temporal limitations, it remains one of the most recent and comprehensive datasets available for this research; furthermore, our sampling within each culture covers diverse demographic variables—including gender, age, and social class—to mitigate representation bias. The concerns regarding the values timeliness, contextual dynamics and sample representativeness reflect systemic challenges within the field, rather than being specific to this work. Constructing such corpora involves significant resource requirements and privacy considerations, which transcend the scope of this study.

C Inference Parameters Configuration

The experiments were implemented using the Transformers library (Wolf et al., 2019), with the temperature parameter is set to 0 to eliminate generation stochasticity and ensure reproducibility. All experiments are conducted on a NVIDIA GeForceRTX 3080 with single run.

D Prompt Design and Demographic Profiles of WVS-derived Personas

Base Prompting: Answer the following question. Select exactly one option. Do not include any extra commentary. Answer by typing the number corresponding to your chosen answer.

Figure 2: *Base* Prompt design in this work.

Anthropological Prompting: Imaging you are a {marital_status} {sex} from {region}, {country}. You are {age} years of age and completed {education} education level. You consider yourself part of the {social_class}.
Answer the following question from this perspective. Others will read what you choose; your goal is to convince them it was chosen from the perspective of the persona described above. Select exactly one option. Do not include any extra commentary. Answer by typing the number corresponding to your chosen answer.

Figure 3: *Anthropological* Prompt design in this work.

For anthropological identity prompts, we uniformly use English to pose questions, following AlKhamissi et al.,(2024) providing specified national identities alongside more granular demographic characteristics which include *Country, Region, Sex, Age, Social Class, Education Level and Marital Status*.

E Evaluation Metric Formulation

Hard Accuracy (HA) quantifies the exact match of responses for the given persona. Formally, HA is defined as:

$$\mathrm{HA}(\mathbb{M}, c) = \mathbb{E}_{p\sim P, q\sim Q}\{\mathbb{I}(\tilde{y}_{p,q} = y_{p,q})\} \quad (7)$$

where $\tilde{y}_{p,q}$ and $y_{p,q}$ represent the model-generated response and the human ground-truth for persona p and question q, respectively.

Soft Accuracy (SA) relaxes the HA by incorporating partial credit for ordinal-scale questions; notably, the metric defaults to HA for categorical data or "don't know" responses:

$$\mathrm{SA}(\mathbb{M}, c) = \mathbb{E}_{p\sim P, q\sim Q}\{1 - \epsilon(p, q)\} \quad (8)$$

$$\epsilon(p, q) = \begin{cases} \dfrac{|\tilde{y} - y|_{p,q}}{|q| - 1} & \mathrm{IsOrd}(p, q), \\ \mathbb{I}(\tilde{y} \neq y)_{p,q} & \text{otherwise} \end{cases} \quad (9)$$

where $\mathrm{IsOrd}(p, q)$ represents identifying ordinal-scale questions with ordinal options, and $|q|$ denotes the count of options in q.

Internal-External Alignment (ρ): Given a collection of models M, we map internal metrics and survey scores under C into flattened vectors $v_{\text{int}}, v_{\text{ext}} \in \mathbb{R}^{|C|\cdot|M|}$, and subsequently calculate their Spearman's correlation coefficient ρ.

F Causal Validation via Neuron Ablation

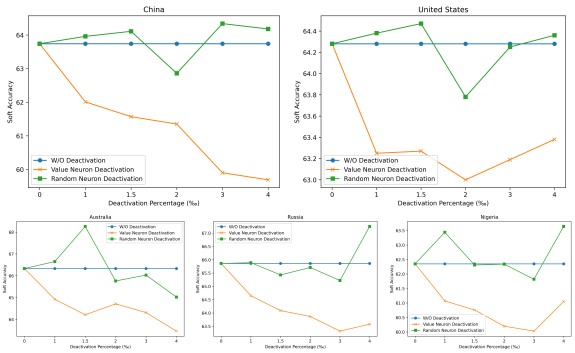

Figure 4: Trends in soft accuracy and deactivation ratio across different settings. The blue, orange, and green lines represent no deactivation, deactivation of value neurons, and deactivation of random neurons, respectively. A wider performance gap between the blue and other lines denotes the pivotal role of the identified neurons in value encoding. Deactivation percentage represents the portion of the model's total parameter count.

To verify the causal role of the identified value neurons, we conduct ablation analysis using Llama-3-8B-Instruct (Touvron et al., 2023) across five representative countries from different continents: the United States, China, Russia, Australia, and Nigeria. Specifically, we ablate a fixed proportion

of these neurons and compare the results against a baseline of randomly sampled neurons of the same scale.

As Figure 4 illustrates, the model's soft accuracy declines sharply as the deactivation percentage increases, confirming that the identified neurons play a critical role in value encoding.

G Sensitivity Analysis of the Pre-defined Constant K

Model	K=10	K=30	K=50	K=60	K=70	K=80	K=90	K=100
gemma	0.259	0.493	0.673	0.704	0.703	0.704	0.705	0.707
Llama3	0.618	0.688	0.695	0.695	0.695	0.692	0.690	0.688
Llama3.1	0.178	0.319	0.580	0.649	0.667	0.668	0.667	0.669
Qwen	0.196	0.636	0.664	0.669	0.668	0.662	0.660	0.662
Mistral	0.513	0.556	0.580	0.588	0.594	0.600	0.605	0.609

Table 2: SD_{dual} across different K values

Our sensitivity analysis (see Table 2) confirms that the results remain robust to variations in K for all $K > 70$. This cross-model stability validates that $K = 100$ is a reliable threshold for diverse architectures.

H Unified Linguistic Baselines

As discussed in previous studies on English prompting for cultural value alignment, probing a model with native-language prompts may yield poorer alignment results compared to English, even when targeting the corresponding culture. This discrepancy is likely attributable to the model's lower proficiency in a specific language rather than a lack of cultural value understanding. Consequently, using unified English prompts enables a comparison that isolates cultural alignment from linguistic interference (Xu et al., 2024).

Model	United States	China	Russia	Australia	Nigeria
gemma					
Baseline	58.14	57.61	59.3	60.57	51.16
English Prompt	67.66	69.75	70.41	68.6	65.95
Llama3					
Baseline	61.35	58.36	59.39	66.03	60.1
English Prompt	64.31	63.92	64.4	65.91	63.44
Llama3.1					
Baseline	62.1	64.58	64.19	64.57	58.59
English Prompt	63.86	67.6	65.83	65.99	60.61
Qwen					
Baseline	54.64	51.16	54.62	57.37	52.54
English Prompt	60.52	59.19	63.76	65.01	59.86
Mistral					
Baseline	60.8	59.0	59.76	66.25	59.04
English Prompt	58.96	50.06	53.78	58.82	58.99

Table 3: Unified Linguistic Baselines

As shown in Table 3, English-only prompting consistently yields gains in value alignment beyond Western contexts, which mitigates concerns regarding linguistic bias.

I Value Neuron Distribution

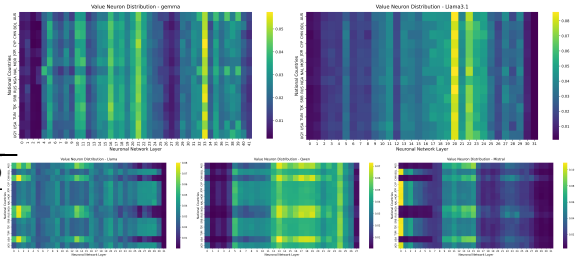

Figure 5: Value Neuron Distribution

As shown in Figure 5, we observe that value neurons mainly concentrate on the shallow-to-middle layers of the neural network. For instance, Gemma primarily focuses on layer 33, while Llama3.1 focuses on layer 20. Some models exhibit varying distributions across different countries; for example, the shallow layers of Llama3 and Mistral display distinct patterns depending on the specific country.

J Statistics of Neuron Numbers

Taking Llama3, Gemma2, and Qwen2.5 as examples, we found that different models exhibit distinct value neuron extraction percentages (Table 4, 5, 6), while each model demonstrates stable extraction rates across different national contexts.

Country	$\lvert N_G \rvert$	$\lvert N_P \rvert$	$\lvert N_V \rvert$	$\lvert N_V / N_P \rvert$
China	21028	26528	9014	0.3398
Australia		27473	9572	0.3484
Russia		28691	10939	0.3813
United States		29781	11204	0.3762
Nigeria		29109	10952	0.3762
Tajikistan		9314	3133	0.3364
New Zealand		9625	3334	0.3465
Serbia		9523	3287	0.3452
Bolivia		9425	3220	0.3416
Morocco		9367	3164	0.3378
Jordan		9336	3170	0.3395
Cyprus		9461	3230	0.3414
Uruguay		9557	3301	0.3454
Tunisia		9516	3273	0.3439

Table 4: **Llama3-8B**. Statistics of neuron numbers across different national contexts.

Country	$\lvert N_G \rvert$	$\lvert N_P \rvert$	$\lvert N_V \rvert$	$\lvert N_V / N_P \rvert$
China	10664	17245	8040	0.4662
Australia		18274	9080	0.4969
Russia		18129	8892	0.4905
United States		19406	10186	0.5249
Nigeria		23808	14452	0.6070
Tajikistan		12656	5280	0.4172
New Zealand		14400	6676	0.4636
Serbia		13084	5526	0.4223
Bolivia		12978	5532	0.4263
Morocco		12988	5536	0.4262
Jordan		12757	5352	0.4195
Cyprus		13000	5491	0.4224
Uruguay		13615	5996	0.4404
Tunisia		13156	5640	0.4287

Table 5: **Qwen2.5-7B**. Statistics of neuron numbers across different national contexts.

Country	$\lvert N_G \rvert$	$\lvert N_P \rvert$	$\lvert N_V \rvert$	$\lvert N_V / N_P \rvert$
China	13796	15497	3992	0.2576
Australia		15709	4203	0.2676
Russia		15805	4333	0.2742
United States		16429	4697	0.2859
Nigeria		15776	4324	0.2741
Tajikistan		14804	3640	0.2459
New Zealand		15856	4323	0.2726
Serbia		15819	4297	0.2716
Bolivia		15246	3934	0.2580
Morocco		15332	3960	0.2583
Jordan		15112	3854	0.2550
Cyprus		15445	4012	0.2598
Uruguay		15897	4330	0.2724
Tunisia		15215	3927	0.2581

Table 6: **Gemma-9B**. Statistics of neuron numbers across different national contexts.