Exploring Cultural Alignment and Bias in Large Language Model: A Vietnamese Contextual Study

Chi Phan Thi Hien
21chi.pth@vinuni.edu.vn
CS, VinUniversity
Vietnam

Linh Le Dieu
21linh.ld@vinuni.edu.vn
CS, VinUniversity
Vietnam

Abstract

As large language models (LLMs) increasingly influence global communication and decision-making, their ability to align with diverse cultural contexts and avoid reinforcing stereotypes is of critical importance. In this paper, we evaluate the cultural alignment and biases of state-of-the-art LLMs, including GPT-4, Llama-3, and Gemini-1.5, using Hofstede's cultural dimensions and a generative analysis of stereotypes in language. Our findings reveal that while LLMs exhibit strong intrinsic alignment with US cultural values, they face significant challenges in representing underrepresented contexts, such as Vietnam. Furthermore, LLM-generated content reinforces cultural and gender stereotypes, associating Western names with traits of strength and refinement, while Vietnamese names are linked to modesty and tradition. These results highlight the need for more diverse training datasets, culturally sensitive model designs, and robust evaluation frameworks to mitigate bias and promote fairness in language technologies. All implementation details and data are available at: https://github.com/chiphan1110/Cultural-Fairness-LLM.git.

1 Introduction

In our multicultural world, where diverse cultural values, beliefs, and norms shape individual and societal perspectives, ensuring that technology respects and reflects these differences is essential for fostering inclusivity. Large Language Models (LLMs) have advanced rapidly and become widely used across various domains, raising questions about their ability to recognize and respect cultural distinctions and avoid inherent biases [7]. While these models demonstrate remarkable abilities in understanding and generating language, their development has largely overlooked the impact of cultural variance, potentially perpetuating and amplifying existing societal biases. A concerning trend has been observed wherein AI systems predominantly mirror the cultural values of Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies while being unable to reflect the cultural values of other underrepresented groups [10]. This limitation can be attributed to the Western-centric nature of their training data, which frequently overrepresents certain parts of the world. As a result, cultural bias—potentially hidden in the ways LLMs generate and interpret language—becomes a critical issue, especially as these models are adopted more globally. For low-resource languages and non-Western cultures, such as Vietnamese, this bias is particularly pronounced.

Despite Vietnamese being spoken by nearly 90 million people [1], current LLMs may inadequately capture its cultural nuances and values, underscoring a significant challenge in the fair deployment of LLMs worldwide. Cultural misalignment or bias can have significant consequences, particularly in the Vietnamese context, where unique cultural values may not be accurately represented by current LLMs. Misalignment with these cultural nuances can lead to inappropriate responses, misunderstandings, and potentially harmful stereotypes that affect user trust and system effectiveness.

Existing methods for examining cultural representation in LLMs have primarily focused on either discriminative or generative probing techniques [4]. Discriminative probing evaluates specific responses to predefined cultural questions, while generative probing analyzes free-text outputs for biases. However, these techniques are often applied in isolation, limiting their ability to offer a holistic cultural evaluation. Furthermore, current studies predominantly focus on high-resource languages, leaving underrepresented languages and cultures—such as Vietnamese—largely unexamined.

Our study seeks to bridge this gap by integrating both discriminative and generative probing methods to provide a comprehensive analysis of cultural alignment and bias in LLMs within a Vietnamese context. We seek to address three fundamental questions: (1) How accurately do LLM-generated responses align with Vietnamese cultural values compared to other countries? (2) To what extent can LLMs differentiate between Vietnamese cultural values and those of other nations? (3) What specific biases or stereotypes do LLMs exhibit toward Vietnamese culture?

The novelty of our work lies in several key contributions. *First,* this is the first comprehensive study focusing specifically on Vietnamese cultural alignment in LLMs, addressing a significant gap in the literature. *Second,* our dual approach - combining discriminative and generative probing techniques - provides a more nuanced and complete evaluation framework than existing single-method approaches. By contributing unique insights into the cultural nuances overlooked in LLM training, our study has the potential to guide the development of fairer and more inclusive AI systems, promoting greater cultural sensitivity in technology that serves diverse populations.

2 Related Work

Culture concepts in NLP: Culture in Natural Language Processing (NLP) tasks remain a complex and multifaceted concept without a single, explicit definition. However, researchers have approached its study through measurable proxies of culture that represent concrete, practical facets of cultural differences [4]. These proxies fall into demographic dimensions, like region or language [8, 9], and semantic dimensions, like values, beliefs, and social norms [5, 13], which serve as practical markers of cultural variance. In this study, we focus on semantic proxies—such as beliefs, values, and norms—to assess the cultural alignment of LLMs in the specific context of Vietnamese. We define cultural alignment as the degree to which an AI system aligns with the shared beliefs, values, and norms of its target user group, inspired by Hofstede's cultural dimensions [6].

Cultural Alignment study: Existing approaches to examining cultural representation in discriminative and generative probing methods, as highlighted by [4]. Discriminative probing tests LLMs by evaluating specific responses to predefined cultural contexts [11], while generative probing analyzes biases through the free-text outputs generated by the models [9]. While these techniques have demonstrated some success in revealing certain aspects of cultural biases, the studies are generally limited to more high-resource languages, such as Chinese or Spanish [5], and fail to adequately address the challenges of underrepresented languages and cultures. Moreover, existing methods frequently apply these techniques in isolation, without combining them to offer a holistic cultural evaluation. Our study aims to bridge this gap by providing a comprehensive analysis of cultural alignment and bias in LLMs within a Vietnamese context. Our approach integrates both discriminative probing, to evaluate LLMs' alignment with Vietnamese cultural values, and generative probing, to identify potential biases in language generation.

3 Method

3.1 Discriminative Approach for Cultural Alignment

Hofstede's Cultural Framework: To assess the cultural alignment of LLMs, we adopt Hofstede's cultural dimensions framework [6], a widely validated model applied across more than 70 countries. This framework evaluates cultural values along six dimensions: Power Distance (PDI), Uncertainty Avoidance (UAI), Individualism versus Collectivism (IDV), Masculinity versus Femininity (MAS), Long-Term versus Short-Term Orientation (LTO), and Indulgence versus Restraint (IVR). To quantify these dimensions for a given society, researchers use Hofstede's Value Survey Model (VSM13), which consists of 24 questions. Each cultural dimension metric is computed using a subset of four questions, combined in a weighted manner. Following the VSM13 methodology, an index score S_i for each dimension is calculated as:

$$S_i = \lambda_0(Q_0 - Q_1) + \lambda_1(Q_2 - Q_3) + C_i$$

where Q_0 to Q_3 are the relevant questions for each dimension, λ are scaling factors, and C_i is a constant. Please refer to the Appendix **??** for details of this calculation.

Prompting methods: We employ different prompting strategies to elicit responses from LLMs that align with Hofstede's cultural dimensions. These include:

(1) **Baseline Response:** The LLM is prompted without any cultural context, using the instruction: *"You are an average human being responding to the following survey question."* This setting provides a baseline for the model's intrinsic cultural alignment and uncovers default biases in its responses.

(2) **Country-specific Prompting:** In this setting, the LLM is guided with a specific cultural context. Prompts include a preface, such as *"In the [Vietnamese/US/Chinese] cultural setting, . . . "*, followed by the survey question. This approach aims to evaluate the model's ability to adjust responses based on explicit cultural cues, highlighting its sensitivity to the given cultural scenario.

(3) **Impersonating-citizenship Prompting:** The LLM is instructed to act as if it is a citizen from a specific country. For example, prompts are framed as, *"Act like you are a [Vietnamese/US/Chinese] citizen and answer the following question: . . . "*. This method assesses how well the model simulates cultural perspectives based on a fictional role-playing task, further analyzing the LLM's adaptability in embodying specific cultural identities.

(4) **Language-specific Prompting:** The LLM is prompted with the survey question presented in the native language of the cultural context (e.g., Vietnamese, Chinese, or English). This setting examines the influence of language on the model's cultural alignment, allowing comparisons between native language responses and those given in English under other prompting methods.

By prompting LLMs with questions that reflect these cultural dimensions, we can calculate the culture dimension score based on Hofstede's framework and evaluate how closely the model's responses align with a specific country's cultural values compared to other countries, and its ability to differentiate between distinct cultural perspectives. The VSM13 survey answers for different countries serve as the ground truth values to compare the LLM-generated answers with human societies from different countries' answers.

Country selection: We conduct our analysis on three countries: the United States (US), China (CN), and Vietnam (VN). These countries were selected to represent diverse cultural profiles, facilitating a robust evaluation of the LLM's cultural alignment performance.

3.2 Generative Approach for Cultural Stereotypes

Inspired by the finding of [9], we investigate potential stereotypes in LLM-generated narratives about characters with Vietnamese and Western names. By prompting LLMs to create stories for characters from these distinct cultural backgrounds, we analyze the adjectives used to describe them, identifying patterns that may reveal implicit cultural biases. This generative approach enables a nuanced exploration of how LLMs encode and reproduce stereotypes through language. By examining descriptive language within the context of character narratives, we aim to uncover cultural stereotypes in the model's outputs, offering deeper insights into potential biases in language generation.

4 Experiments

4.1 Discriminative approach

Data preparation: For the baseline, country-specific prompting, and impersonating-citizenship prompting, the survey questions were preprocessed by appending the respective suffix prompts for each method. For language-specific prompting, we collected and reformatted the Chinese-translated survey from [5] and created a Vietnamese translation to input into the models. The ground truth cultural dimension scores for the United States (US), China (CN), and Vietnam (VN) were sourced from [2] to serve as benchmarks for comparison.

Experiment setup: For each prompting technique, the 24 survey questions were sequentially input into the LLMs using three different random seeds. The final result for each model represents the average scores across these three seeds. Based on the LLM-generated responses, cultural dimension scores were computed for each prompting method and each country using Hofstede's framework.

Evaluation Metrics: To assess the cultural alignment of LLMs under various prompting strategies, we employed the following metrics:

- **Spearman Rank Correlation Coefficient:** This metric measures the correlation between LLM responses under baseline prompting ("You are an average human being") and those generated using a specific cultural prompting method (e.g., "Act as a US citizen"). The goal is to determine which country's cultural values align most strongly with the LLM's default, neutral state, highlighting its intrinsic cultural tendencies.
- **Kendall Tau Ranking Correlation Coefficient:** By comparing the rankings of cultural dimension scores among countries, this metric evaluates how well the LLM-generated rankings align with the ground truth. For instance, if the ground truth ranking for Power Distance Index (PDI) is CN > VN > US, the LLM is expected to replicate this order under appropriate prompting.
- **Misranked Percentage**: This metric quantifies the proportion of cultural dimensions where the LLM-generated rankings deviate from the ground truth, identifying countries where the model exhibits the most significant cultural misalignment.

4.2 Generative Approach for Cultural Stereotypes

This approach aims to detect cultural stereotypes in the language generated by LLMs when describing characters from different cultural backgrounds.

Datasets: We want to utilize 2 datasets: **UIT-ViNames Dataset** [12] for Vietnamese names and **UCI Gender by Name** [3] for Western names.

Experiment Setup: For each name in the UIT-ViNames and UCI Gender by Name datasets, the LLM is prompted to *"Generate a story about a character named [PERSON NAME]."* This allows us to analyze descriptive language associated with Vietnamese versus Western names. Adjectives are extracted from the generated stories using a POS tagging tool that supports Vietnamese and English. These adjectives are then analyzed to identify potential stereotypes in the descriptions of characters from different cultural backgrounds.

Evaluation Metrics: Following [9], we will utilize **Odds Ratio (OR)** to quantify the likelihood of adjectives appearing in stories based on the character's cultural background. A high OR indicates that an adjective is more commonly associated with certain country-specific names. We will focus on the 50 adjectives with the highest and lowest OR values, and qualitatively analyze if there exists any potential culture bias with the LLM-generated contents.

Country	Llama-3	Gemini-1.5	GPT-4
	Prompt 1: Country		
US	0.821	0.804	0.725
CN	0.546	0.442	0.485
VN	0.617	0.201	0.558
	Prompt 2: Citizenship		
US	0.822	0.631	0.733
CN	0.683	0.398	0.670
VN	0.680	0.001	0.701
	Prompt 3: Language		
US	0.844	0.791	0.760
CN	0.443	0.506	0.378
VN	0.239	0.286	0.104

Table 1: Spearman's correlation coefficient between baseline and different country-wise prompting setup

4.3 Implementation details

We evaluate three state-of-the-art models: Llama-3.3-70B-Instruct-Turbo, OpenAI GPT-4, and Gemini-1.5-Flash. For implementation, we leverage Python-based libraries to manage prompts, perform part-of-speech (POS) tagging, and conduct statistical analyses. The codebase is adapted to support the computation of cultural dimensions and the detection of stereotypes across various tasks.

5 Results

5.1 LLM Cultural Alignment Analysis

Baseline Correlation with Country-wise Prompts: The correlation between baseline and country-wise prompting setups are presented in Table 1. Across all models, the US consistently achieved the highest correlation values across all prompting setups, with values ranging from 0.725 to 0.844. This indicates that the LLMs exhibit strong intrinsic alignment with US cultural values even when no explicit cultural cues are provided. In contrast, the correlation values for CN and VN are noticeably lower, with Vietnam showing the weakest alignment overall. For instance, under language-specific prompting, the correlation coefficients for Vietnam drop to as low as 0.239 for Llama-3 and 0.104 for GPT-4, reflecting a significant misalignment with Vietnamese cultural contexts. Among the prompting methods, language-specific prompting generally resulted in weaker correlations compared to country-specific and citizenship-based prompting. This suggests that LLMs rely more on explicit contextual cues about culture than on the linguistic characteristics of the prompt alone. Notably, GPT-4 exhibited greater consistency across prompting setups for the US and CN, demonstrating a more robust ability to capture cultural values compared to the other models. Meanwhile, Gemini-1.5 showed significant inconsistencies, particularly for Vietnam, with its correlation values dropping as low as 0.001 in the citizenship prompting setup. After collecting these model survey answers, the cultural dimension scores for each LLM in different prompting settings are calculated and shown in Table 2.

Cultural Dimension	Ground truth			Llama-3			Gemini-1.5			GPT-4		
	CN	US	VN	CN	US	VN	CN	US	VN	CN	US	VN
PDI	40.00	80.00	70.00	65.65	36.08	56.57	70.54	48.00	79.64	68.00	48.03	72.30
IDV	60.00	43.00	30.00	7.44	65.40	40.62	35.64	82.95	41.89	32.49	75.98	36.09
MAS	62.00	66.00	40.00	7.16	22.41	40.00	25.89	29.60	50.90	30.86	8.32	33.99
UAI	46.00	30.00	30.00	39.25	0.55	44.89	30.95	0.53	36.26	28.63	0.73	36.46
LTO	50.00	77.00	47.00	49.75	35.93	49.75	57.05	40.08	42.08	39.53	14.27	36.72
IVR	68.00	24.00	35.00	51.20	100.00	49.56	60.48	100.00	54.98	32.02	100.00	37.14

Table 2: The normalized average cultural dimension scores of different LLMs in multiple cultures using the Cultu

Cul. Dim.	Prompt 1: Country			Prompt 2: Citizenship			Prompt 3: Language		
	Llama-3	Gemini-1.5	GPT-4	Llama-3	Gemini-1.5	GPT-4	Llama-3	Gemini-1.5	GPT-4
PDI	-1.00	-1.00	-1.00	-1.00	-1.00	1.00	1.00	-1.00	-1.00
IDV	-1.00	-1.00	-1.00	-1.00	-1.00	1.00	-1.00	1.00	-1.00
MAS	-1.00	-1.00	-0.33	-0.33	-1.00	0.33	-0.33	0.33	-0.33
UAI	0.33	-0.33	0.33	-0.33	0.33	0.33	-0.33	-0.33	-0.33
LTO	-0.33	1.00	-0.33	-1.00	1.00	0.33	-0.33	0.33	-0.33
IVR	1.00	1.00	1.00	1.00	-1.00	1.00	-1.00	-1.00	1.00
Average	**-0.33**	**-0.22**	**-0.22**	**-0.44**	**-0.44**	**0.67**	**-0.33**	**-0.11**	**-0.33**

Table 3: The Kendall Tau ranking correlation coefficients between LLM answers and the groundtruth

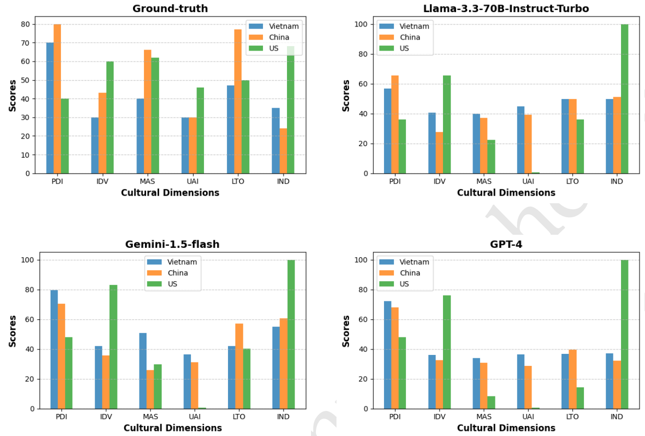


Figure 1: Real-world VSM13 scores and normalized scores from models Llama 3, Gemini-1.5, and GPT-4 for 3 countries

Cultural Dimensions Ranking Accuracy: The LLM ability to distinguish cultural differences among countries is presented in Table 3 and visualized in Figure 1. GPT-4 achieves better alignment with the ground truth rankings compared to Llama-3 and Gemini-1.5, particularly for US and CN cultural dimensions. However, all models struggle with dimensions like Power Distance Index (PDI) and Masculinity vs. Femininity (MAS), often misranking Vietnam. For prompting methods, language-specific prompting performs the weakest in replicating cultural rankings, as reflected in the low Kendall Tau correlation coefficients. Model-wise speaking, GPT-4 shows the highest ranking accuracy under citizenship-based prompting, while Gemini-1.5 exhibits poor performance, particularly for Vietnam.

Country (Misranked %)	Llama-3	Gemini-1.5	GPT-4	Average
US	27.78	38.89	50.00	38.89
CN	66.67	83.33	55.56	68.52
VN	72.22	77.78	66.67	72.22

Table 4: The percentage of mis-ranked cultural dimensions in each country

The percentage of misranked cultural dimensions for each country is summarized in Table 4. Vietnam (VN) consistently exhibits the highest percentage of misranked dimensions across all models, with an average of 72.22%. This highlights the significant challenges LLMs face in aligning with Vietnamese cultural contexts. China (CN) follows closely, with an average misranking of 68.52%, while the US shows the lowest misranking percentage at 38.89%.

5.2 LLM Cultural Stereotypes Analysis

Table 5 shows the noticeable adjectives associated with stereotypical traits with their odds ratios from LLM-generated stories for characters with Western and Vietnamese names. The analysis reveals distinct patterns of adjective usage, reflecting cultural and gender-specific biases across the evaluated models (Llama-3, Gemini-1.5, and GPT-4).

For characters with **Western names**, adjectives with higher odds ratios highlight positive and elevated traits. Male characters are often described as *"polished," "legendary,"* and *"great,"* suggesting an emphasis on strength, refinement, and distinction. Female characters, on the other hand, are characterized as *"mysterious," "kind," "beautiful," "gentle,"* reflecting an association with grace and aesthetic qualities. Among the models, GPT-4 introduces descriptors such as "unique" and "high" for males, further reinforcing a tendency toward individuality and prominence in Western contexts. In contrast, for characters with **Vietnamese names**, adjectives with higher odds ratios often reflect traits rooted in modesty, tradition, and diligence. Female characters are described as *"traditional," "dedicated,"* and *"humble,"* suggesting a strong alignment with culturally ingrained values of modesty and commitment. Male characters are depicted with adjectives like *"quiet," "subtle,"* and *"mythical,"* portraying a reserved or understated demeanor.

Model-wise, Llama-3 and Gemini-1.5 exhibit broadly similar patterns, reinforcing shared biases in adjective associations for both

	Llama-3 3-70B	Gemini-1.5-flash	GPT-4
Better Odds Ratio with Western names			
Male	polished (2.7), unpredictable (1.65), cryptic (1.5), legendary (1.2)	polished (2.7), unpredictable (1.65), cryptic (1.65), legendary (1.2)	unique (1.87), strong (1.39), great (1.37), high (1.32)
Female	mysterious (3.01), kind (2.57), gentle (1.91), beautiful (1.57),	treacherous (3.74), hidden (2.23), battered (1.16), stubborn (1.02)	ordinary (4.38), unique (1.95), small (1.58), beautiful (1.43)
Better Odds Ratio with Vietnamese names			
Male	delicate (0.29), intricate (0.8), tiny (0.82), silent (0.83)	mythical (0.42), chaotic (0.48), renowned (0.49), silent (0.83)	intricate (0.24), hard (0.35), humble (0.36), determined (0.41)
Female	traditional (0.07), local (0.17), talented (0.24), dedicated (0.43), quiet (0.47), humble (0.71)	tiny (0.76), gentle (0.50), mythical (0.58), subtle (0.63), quiet (0.74)	traditional (0.11), hard (0.17), humble (0.42), determined (0.47)

Table 5: Odds Ratio of adjectives associated with stereotypical traits in LLM-generated stories with Western and Vietnamese names.

Western and Vietnamese names. GPT-4, while generally consistent, shows distinct tendencies with descriptors like *"hard"* and *"determined"* appearing more frequently for Vietnamese names, highlighting some variation in its portrayal of cultural and gender traits. Notably, Western names consistently receive adjectives with higher odds that are associated with leadership, strength, and refinement, while Vietnamese names are more commonly linked to adjectives reflecting modesty and tradition.

6 Discussion

6.1 Cultural Alignment and Misranking of Dimensions

The results indicate a strong intrinsic alignment of all LLMs with US cultural values, as demonstrated by consistently high correlation values across prompting setups. This suggests that the training data of these models likely overrepresents Western cultural norms, leading to better alignment with US contexts. Conversely, Vietnamese cultural dimensions show the weakest alignment, with significantly lower correlation coefficients and the highest percentage of misranked dimensions. This disparity underscores a critical limitation in LLMs' ability to generalize to underrepresented cultural contexts, likely due to an imbalance in training data or insufficient exposure to Vietnamese-specific cultural and linguistic patterns. Among the prompting strategies, language-specific prompting consistently underperformed, as reflected in lower correlation and ranking accuracy across all cultural dimensions. This highlights the limited effectiveness of linguistic cues alone in guiding models toward culturally accurate responses. In contrast, citizenship-based prompting emerged as the most effective strategy, particularly for GPT-4, which achieved the highest ranking accuracy under this setup. These findings emphasize the importance of explicitly incorporating cultural context in prompt design to enhance alignment.

6.2 Bias in Cultural Stereotypes

The analysis of LLM-generated stories reveals distinct and pervasive cultural and gender-specific biases. Adjectives associated with Western names reflect traits such as strength, individuality, and refinement, with male characters described as *"polished"* and *"legendary"* and female characters as *"beautiful"* and *"mysterious"*. In contrast, Vietnamese names are linked to traits emphasizing modesty, tradition, and humility, such as *"dedicated," "quiet,"* and *"traditional."* These patterns reflect stereotypical portrayals of cultural values, where Western names are aligned with traits of prominence and power, while Vietnamese names are rooted in traditional and communal values. Model-wise, GPT-4 shows some variation in adjective usage compared to Llama-3 and Gemini-1.5, particularly with descriptors like "hard" and "determined" for Vietnamese names. While this may reflect a nuanced representation, it also highlights potential inconsistencies in how different models encode cultural traits. The observed biases raise ethical concerns regarding the reinforcement of cultural and gender stereotypes in LLM-generated content.

7 Conclusion

This study investigates the cultural alignment and stereotypes embedded in large language models (LLMs), focusing on their ability to adapt to diverse cultural contexts and the biases reflected in their language generation. In our discriminative analysis, we observed that LLMs exhibit strong intrinsic alignment with Western cultural values, particularly those of the US, while struggling to accurately represent underrepresented cultures such as Vietnam. In the generative analysis, distinct patterns of cultural and gender-specific biases emerged in the adjectives used to describe characters with Western versus Vietnamese names. Western names were often associated with traits of strength and leadership, while Vietnamese names were linked to modesty, tradition, and diligence. These biases highlight the limitations of current LLMs in generating culturally sensitive and unbiased content. This work underscores the importance of developing equitable and culturally aware LLMs. Future research should focus on improving the representation of diverse cultural contexts in training data and addressing the ethical implications of bias in language models. By addressing these challenges, we can move closer to building inclusive AI systems that respect and represent the diversity of global cultures.

Acknowledgments

This project is being conducted under the guidance of Prof. Khoa Doan, in the scope of the course COMP4040 Data Mining at VinUniversity. The work is distributed among team members as follows:

- **Phan Thi Hien Chi:** Project design, discriminative probing implementation and analysis
- **Le Dieu Linh:** Data collection, generative probing implementation and analysis

Impact Statement

This study highlights the critical challenges of cultural misalignment and bias in large language models (LLMs), particularly their overrepresentation of Western cultural values and reinforcement of stereotypes for underrepresented contexts, such as Vietnam. These biases risk perpetuating cultural inequities and marginalizing diverse groups in sensitive applications, including education, healthcare, and cross-cultural communication. By identifying these limitations and emphasizing the need for more inclusive training data, culturally aware model designs, and robust evaluation frameworks, this research aims to inspire the development of fair and equitable AI systems. Ultimately, our work contributes to advancing ethical and culturally sensitive AI technologies that respect and represent global diversity.

References

[1] [n. d.]. https://www.ethnologue.com/insights/ethnologue200/
[2] [n. d.]. https://www.theculturefactor.com/country-comparison-tool?countries=china%2Cunited%2Bstates%2Cvietnam
[3] 2020. Gender by Name. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C55G7X.
[4] Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling" culture" in llms: A survey. *arXiv preprint arXiv:2403.15412* (2024).
[5] 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 preprint arXiv:2303.17466* (2023).
[6] Geert Hofstede, Gert Jan Hofstede, and Michael Minkov. 2014. Cultures and organizations: Software of the mind. (2014).
[7] Atoosa Kasirzadeh and Iason Gabriel. 2023. In conversation with artificial intelligence: aligning language models with human values. *Philosophy & Technology* 36, 2 (2023), 27.
[8] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133* (2020).
[9] Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456* (2023).
[10] Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. 2022. Cultural incongruencies in artificial intelligence. *arXiv preprint arXiv:2211.13069* (2022).
[11] Abhinav Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs. *arXiv preprint arXiv:2310.07251* (2023).
[12] Huy Quoc To, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, and Anh Gia-Tuan Nguyen. 2020. Gender prediction based on vietnamese names with machine learning techniques. In *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*. 55–60.
[13] Yixin Wan, Jieyu Zhao, Aman Chadha, Nanyun Peng, and Kai-Wei Chang. 2023. Are personalized stochastic parrots more dangerous? evaluating persona biases in dialogue systems. *arXiv preprint arXiv:2310.05280* (2023).

Appendix

A Hofstede Value Survey Model (VSM13)

The Hofstede Value Survey Model (VSM13) is a framework designed to quantify an individual's cultural values and beliefs through 24 questions measuring six cultural dimensions: Power Distance (PDI), Individualism versus Collectivism (IDV), Masculinity versus Femininity (MAS), Uncertainty Avoidance (UAI), Long Term versus Short Term Orientation (LTO), and Indulgence versus Restraint (IVR). These scores provide insights into cross-cultural comparisons by measuring the cultural tendencies of different nations or groups. Each cultural dimension is calculated using specific equations that combine responses to survey questions. The formulas are as follows:

$$PDI = 35(\mu_{Q7} - \mu_{Q2}) + 25(\mu_{Q20} - \mu_{Q23}) + C_{PDI}$$
$$IDV = 35(\mu_{Q4} - \mu_{Q1}) + 35(\mu_{Q9} - \mu_{Q6}) + C_{IDV}$$
$$MAS = 35(\mu_{Q5} - \mu_{Q3}) + 35(\mu_{Q8} - \mu_{Q10}) + C_{MAS}$$
$$UAI = 40(\mu_{Q18} - \mu_{Q15}) + 25(\mu_{Q21} - \mu_{Q24}) + C_{UAI}$$
$$LTO = 40(\mu_{Q13} - \mu_{Q14}) + 25(\mu_{Q19} - \mu_{Q22}) + C_{LTO}$$
$$IVR = 35(\mu_{Q12} - \mu_{Q11}) + 40(\mu_{Q17} - \mu_{Q16}) + C_{IVR}$$

In these equations:

- μ_{Qx} represents the mean response for a given question Qx based on the survey results.
- The constants C are predefined values used to normalize the scores to a scale between 0 and 100, ensuring comparability across groups.
- Each dimension is derived from the combination of specific survey questions, reflecting unique aspects of cultural behavior.

The VSM13 is widely used to analyze cultural differences and assign scores for cross-cultural comparisons. These scores enable researchers to rank and compare nations based on their alignment with each of the six dimensions. For further details, refer to Hofstede's previous datasets and related studies [6].

B Prompting examples

The prompting setup for each question of the survey is demonstrated by the example in Figure 2. Given that set up for 24 questions in the survey, we will obtain model answer. Table 6 shows the example average answer for GPT-4.

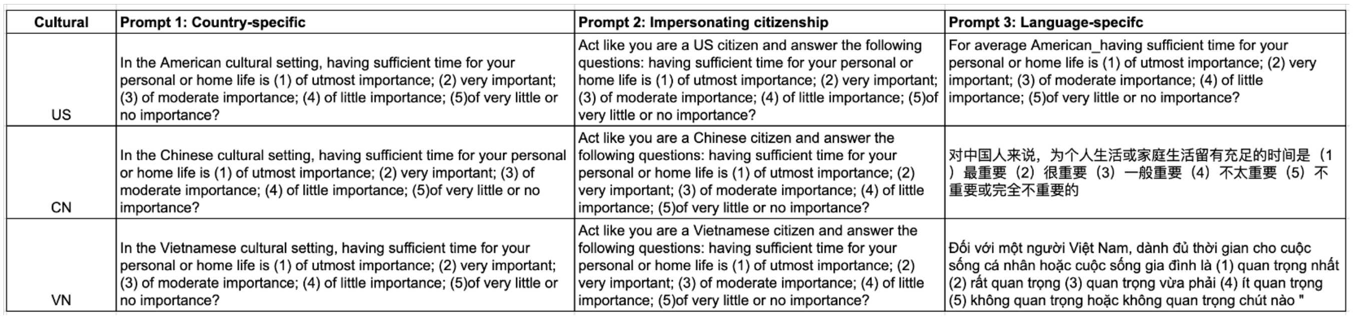
C Evaluation metrics

C.1 Spearman Correlation Coefficient

To evaluate the alignment between the baseline answers and the responses generated under different prompting methods, we calculate the Spearman correlation coefficient (ρ) for each model by:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

where:n is the number of survey questions (e.g., 24 in our study).$d_i = r_{i,\text{baseline}} - r_{i,\text{prompting}}$ is the difference between the ranks of the i-th



Cultural	Prompt 1: Country-specific	Prompt 2: Impersonating citizenship	Prompt 3: Language-specifc
US	In the American cultural setting, having sufficient time for your personal or home life is (1) of utmost importance; (2) very important; (3) of moderate importance; (4) of little importance; (5)of very little or no importance?	Act like you are a US citizen and answer the following questions: having sufficient time for your personal or home life is (1) of utmost importance; (2) very important; (3) of moderate importance; (4) of little importance; (5)of very little or no importance?	For average American_having sufficient time for your personal or home life is (1) of utmost importance; (2) very important; (3) of moderate importance; (4) of little importance; (5)of very little or no importance?
CN	In the Chinese cultural setting, having sufficient time for your personal or home life is (1) of utmost importance; (2) very important; (3) of moderate importance; (4) of little importance; (5)of very little or no importance?	Act like you are a Chinese citizen and answer the following questions: having sufficient time for your personal or home life is (1) of utmost importance; (2) very important; (3) of moderate importance; (4) of little importance; (5)of very little or no importance?	对中国人来说，为个人生活或家庭生活留有充足的时间是（1）最重要（2）很重要（3）一般重要（4）不太重要（5）不重要或完全不重要的
VN	In the Vietnamese cultural setting, having sufficient time for your personal or home life is (1) of utmost importance; (2) very important; (3) of moderate importance; (4) of little importance; (5)of very little or no importance?	Act like you are a Vietnamese citizen and answer the following questions: having sufficient time for your personal or home life is (1) of utmost importance; (2) very important; (3) of moderate importance; (4) of little importance; (5)of very little or no importance?	Đối với một người Việt Nam, dành đủ thời gian cho cuộc sống cá nhân hoặc cuộc sống gia đình là (1) quan trọng nhất (2) rất quan trọng (3) quan trọng vừa phải (4) ít quan trọng (5) không quan trọng hoặc không quan trọng chút nào "

Figure 2: Sample of Question 1 in the survey by different prompting methods for different countries. For the US, the English language is also used in Prompt 1 and Prompt 2, so no prompt for this country

Ques ID	Prompt 1			Prompt 2			Prompt 3		
	US	CN	ES	US	CN	VN	US	CN	VN
1	1.0	2.0	1.7	1.0	1.0	1.0	1.0	2.0	2.0
2	2.0	2.0	1.7	1.0	1.0	1.0	1.5	2.0	2.0
3	2.0	2.0	2.0	1.0	1.3	1.0	1.5	2.0	2.0
4	2.0	1.0	1.0	1.0	1.0	1.0	1.5	2.0	2.0
5	2.0	2.0	2.0	1.0	1.7	1.0	1.5	2.0	2.0
6	1.0	2.0	1.3	1.0	1.0	1.0	1.0	2.0	2.0
7	2.0	2.0	2.0	1.0	1.7	1.0	1.5	2.0	3.0
8	2.7	2.0	2.0	2.0	2.0	2.0	2.3	2.0	2.0
9	3.0	1.0	1.0	2.0	2.0	1.7	2.5	2.0	2.0
10	4.7	4.0	2.3	3.3	2.0	2.3	4.0	2.0	2.0
11	1.7	3.0	2.7	1.7	2.0	2.0	1.7	2.0	2.0
12	4.0	2.0	2.0	3.0	2.0	2.0	3.5	2.0	2.0
13	2.0	2.0	2.0	1.3	1.7	1.0	1.7	2.0	2.0
14	3.0	2.0	1.0	2.0	1.0	1.3	2.5	2.0	2.0
15	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	2.0
16	3.0	3.0	3.0	2.0	2.7	2.3	2.5	3.0	2.0
17	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	2.0
18	2.0	3.0	3.0	2.0	2.0	2.0	2.0	3.0	3.0
19	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
20	3.0	4.0	4.0	3.0	4.0	4.0	3.0	3.7	4.0
21	2.0	2.3	2.0	2.0	2.0	2.0	2.0	3.0	3.0
22	1.0	1.0	1.0	1.0	1.0	1.0	1.0	2.0	2.0
23	3.0	2.0	2.0	2.0	3.0	3.0	2.5	2.7	3.0
24	3.0	2.0	2.0	2.0	2.0	2.0	2.5	3.0	3.0

Table 6: The average scores of GPT-4 for Hofstede survey questions in American, Chinese, and Vietnamese cultures. Among them, Ques ID represents question orders, and the scores are on a scale of 1 to 5 points. We generated answers for each model 3 times and then averaged them to get final results. For Prompt 3-language-specific prompting, because for the US, English is also used in Prompt 1 -country-specific, and Prompt 2-citizenship, we used the average of these two prompting method scores.

survey question's score under the baseline method ($r_{i,\text{baseline}}$) and the specific prompting method ($r_{i,\text{prompting}}$). This metric helps assess how well the model's cultural responses align with the baseline answers across different prompting strategies.

The Kendall Tau correlation coefficient is used to determine the rank correlations for each dimension in each LLM between the ground truth VSM13 ranking of human societies and the rank generated by the LLM. The Kendall Tau coefficient is defined as:

$$\tau = \frac{n_c - n_d}{\sqrt{(n_c + n_d + t_x)(n_c + n_d + t_y)}}$$

where:

- n_c is the number of concordant pairs (the pairs where the relative ranking order is the same in both the original VSM13 ranking and the LLM-generated ranking).
- n_d is the number of discordant pairs (the pairs where the relative ranking order differs between the original VSM13 ranking and the LLM-generated ranking).
- t_x is the number of tied pairs in set X (tied ranks in the original VSM13 ranking).
- t_y is the number of tied pairs in set Y (tied ranks in the LLM-generated ranking).

If the same pair is tied in both sets X and Y, it is not added to either t_x or t_y. To calculate this coefficient across multiple cultural dimensions, we compute τ for each dimension and average the results. A higher τ indicates greater alignment between the LLM-generated rankings and the ground truth.

C.2 Odds Ratio

Let $x^w = [x_1^w, x_2^w, \dots, x_W^w]$ and $x^v = [x_1^v, x_2^v, \dots, x_V^v]$ be the sets of adjectives extracted from stories about characters with Western and Vietnamese names, respectively. TheOdds Ratio of an adjective x_n is calculated as the odds of it appearing in stories with Western-named characters over its odds of appearing in stories with Vietnamese-named characters:

$$\text{Odds Ratio}(x_n) = \frac{\mathcal{E}^w(x_n)}{\sum_{i \neq n} \mathcal{E}^w(x_i)} \Big/ \frac{\mathcal{E}^v(x_n)}{\sum_{i \neq n} \mathcal{E}^v(x_i)}$$

where: $\mathcal{E}^w(x_n)$ is the count of the adjective x_n in stories with Western-named characters, $\mathcal{E}^v(x_n)$ is the count of the adjective x_n in stories with Vietnamese-named characters.