Information Processing and Management 62 (2025) 104099

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Towards realistic evaluation of cultural value alignment in large language models: Diversity enhancement for survey response simulation

Haijiang Liu [a,b], Yong Cao [c], Xun Wu [d], Chen Qiu [a,b,*], Jinguang Gu [a,b], Maofu Liu [a,b], Daniel Hershcovich [e]

[a] School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, Hubei, 430065, China
[b] Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan, Hubei, 430065, China
[c] Tübingen AI Center, University of Tübingen, Tübingen, 72074, Germany
[d] Division of Social Science, The Hong Kong University of Science and Technology, Hong Kong, 511455, China
[e] Department of Computer Science, University of Copenhagen, Copenhagen, Denmark

ARTICLE INFO

Keywords:
Evaluation methods
Value investigation
Survey simulation
Large language models
U.S.-china cultures

ABSTRACT

Assessing Large Language Models (LLMs) alignment with human values has been a high priority in natural language processing. These models, praised as reservoirs of collective human knowledge, provoke an important question: Do they genuinely reflect the value preferences embraced by different cultures? We measure value alignment by simulating sociological surveys and comparing the distribution of preferences from model responses to human references. We introduce a diversity-enhancement framework featuring a novel memory simulation mechanism, which enables the generation of model preference distributions and captures the diversity and uncertainty inherent in LLM behaviors through realistic survey experiments. To better understand the causes of misalignment, we have developed comprehensive evaluation metrics. Our analysis of multilingual survey data illustrates that our framework improves the reliability of cultural value alignment assessments and captures the complexity of model responses across cultural contexts. Among the eleven models evaluated, the Mistral and Llama-3 series show superior alignment with cultural values, with Mistral-series models notably excelling in comprehending these values in both U.S. and Chinese contexts.[1]

1. Introduction

As artificial intelligence (AI) technology continues to advance (Christiano, Leike, Brown, Martic, Legg, & Amodei, 2017; Ouyang, Wu, Jiang, Almeida, et al., 2022; Schulman, Wolski, Dhariwal, Radford, & Klimov, 2017), various large language models (LLMs) such as Baichuan, WizardLM, ChatGLM, Llama, and Mistral have demonstrated impressive capabilities. These LLMs have opened up promising possibilities for building strong interactive human–machine systems (Dan, Lei, Gu, Li, et al., 2023; Li, Li, Zhang, Dan, & Zhang, 2023; Nascimento & Pimentel, 2023). However, their applications have also brought about unexpected social risks (Sheng, Chang, Natarajan, & Peng, 2021), which have raised questions about their trustworthiness (Huang, Sun, Wang, Wu, et al., 2024; Liu, Yao, Ton, Zhang, et al., 2023).

* Corresponding author at: School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, Hubei, 430065, China.
E-mail addresses: alecliu@ontoweb.wust.edu.cn (H. Liu), yong.cao@uni-tuebingen.de (Y. Cao), wuxun@hkust-gz.edu.cn (X. Wu), chen@wust.edu.cn (C. Qiu), simon@ontoweb.wust.edu.cn (J. Gu), liumaofu@wust.edu.cn (M. Liu), dh@di.ku.dk (D. Hershcovich).

[1] https://github.com/alexc-l/DEF-Survey-Sim

https://doi.org/10.1016/j.ipm.2025.104099
Received 29 October 2024; Received in revised form 6 February 2025; Accepted 7 February 2025
Available online 14 March 2025

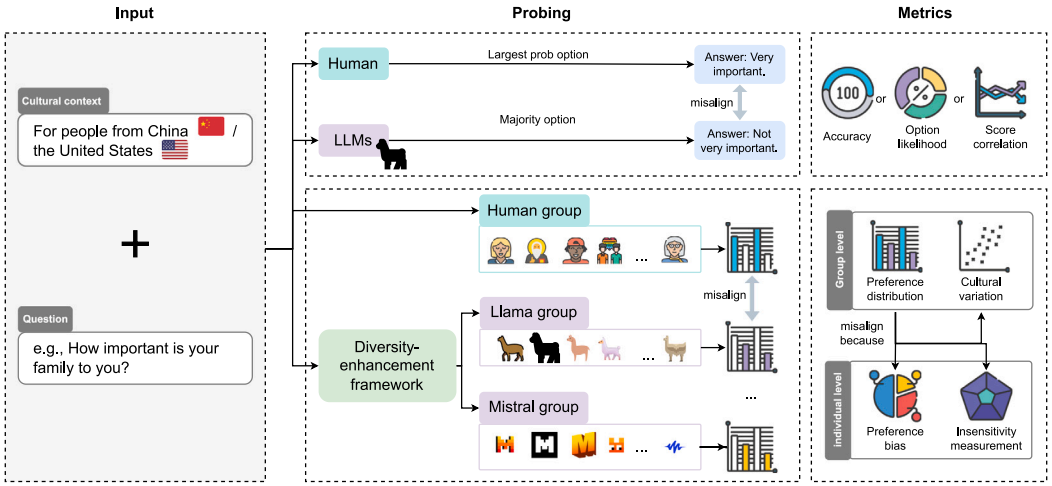


Fig. 1. Overview of our proposed assessment framework (lower panel) versus previous methods (upper panel), highlighting a novel perspective on value alignment evaluation through survey simulation with LLMs.

The growing concerns surrounding LLMs have drawn attention in the field, resulting in a series of evaluation studies proposed (Bang, Cahyawijaya, Lee, Dai, et al., 2023; Chang, Wang, Wang, Wu, et al., 2023; Xu et al., 2024). These studies target various societal features of models, ranging from norms (Ramezani & Xu, 2023; Scherrer, Shi, Feder, & Blei, 2023), bias (Dhamala, Sun, Kumar, Krishna, et al., 2021; Tao, Viberg, Baker, & Kizilcec, 2023), to values (Arora, Kaffee, & Augenstein, 2023; Cao, Zhou, Lee, Cabello, et al., 2023). Norms evaluations assess model reactions to practical situations to determine whether they align with human standards. Bias assessments investigate fairness, stereotypes, and other features presented in the model knowledge and responses.

Values, on the other hand, are higher-level and complex measurements that may provide more insights for model behavior (Frese, 2015). They are common goals influencing how models make decisions (Pawar, Park, Jin, et al., 2024). These evaluations reflect the considerable variation of beliefs and understanding between humans and LLMs with survey simulations, revealing that the beliefs and understandings exhibited are inconsistent and unstable like those of humans (AlKhamissi, ElNokrashy, Alkhamissi, & Diab, 2024; Arora et al., 2023; Cao et al., 2023). However, the evaluation metrics in these researches are typically straightforward, leading to limited insights into the causes of misalignment. The value inference strategies employed in their paper do not include creating a preference distribution for in-depth analysis, potentially overlooking the model's unpredictable behaviors.

For improving model response diversity so that we can generate distribution, Lahoti, Blumm, Ma, Kotikalapudi, et al. (2023) developed Collective-critique and Self-voting (CCSV), a four-step method that leverages large language models to critique and revise their responses. This strategy aims to enhance the diversity of various output objects, rather than being specifically designed for survey simulation or value assessment.

As illustrated in Fig. 1, we select several LLM representatives to simulate citizens from specific cultures and respond to a survey consisting of multiple-choice questions. Our diversity-enhanced framework (DEF), including prompt, configuration, and memory modifications, generates model groups based on different candidates to ensure accurate value preference distributions. We then compare the model responses to human preferences, conducting a multi-aspects cross-value assessment that measures preference distributions, cross-cultural variation, and character profiling to identify value distance and preference biases. Additionally, we design an insensitivity measurement based on Guilford's three-dimensional structure of intelligence model to quantify the value consistency of model responses.

To explore the differences between model and human values, we evaluate their preferences regarding two prominent countries: the United States and China. Based on insights from social science (Cao, Carstensen, Gao, & Frank, 2024), we recognize the United States and China as two distinct nations in terms of cognition, language, perception, and reasoning. Consequently, we select human value preferences from these two nations – extensively examined by social scientists – as the subjects of our research.

Experiments on Mistral-7B-Instruct suggest that our DEF system can capture the complexity and variability of model responses to generate preference distribution and lower the divergence to human value distribution. With the refined inference strategies and multifaceted assessment, **eleven representative models reveal notable limitations in their ability to align with the cultural value distribution of the United States and China**. Additionally, there are notable concerns regarding the lack of cross-cultural representation and preference biases related to gender and age across these models. Further analysis, however, indicates that the Mistral-series and Llama-3-series models demonstrate superior performance in value alignment, with the Mistral-series models exhibiting a more robust and comprehensive understanding of the underlying cultural values across both cultural contexts.

1.1. Research objectives

Our comprehensive research aims to transform the evaluation of cultural value alignment in Large Language Models (LLMs) by introducing a multifaceted inference-centric approach. Specifically, we seek to address the following critical research questions:

1. What innovative methodologies can we develop to capture the complexity and variability of model responses across different cultural contexts?
2. To what extent can refined inference strategies enhance the accuracy and reliability of cultural value alignment assessments?
3. Using the refined inference strategies, to what extent do the cultural value patterns exhibited by LLMs mirror those found in social survey responses from the United States and China?

1.2. Contributions

Our contributions can be summarized as follows:

- We introduce a Diversity-Enhanced Framework (DEF) that provides a novel approach to capturing the complexity and variability of model responses across cultural contexts.
- Our research demonstrates the potential of refined inference strategies to enhance the accuracy and reliability of cultural value alignment assessments.
- Through a comprehensive evaluation of eleven representative models, we conducted an in-depth analysis of how LLM responses align with social survey responses from the United States and China.

Moreover, recent research, such as Kirk, Whitefield, Röttger, Bean, et al. (2024) highlights the crucial role of value alignment in AI systems, particularly for complex, subjective topics. As information access shifts from traditional search engines to AI-driven integration, the risk of cultural misalignment grows. Our research addresses this challenge by identifying optimal models and methods for aligning AI systems with diverse values, contributing to the development of more culturally inclusive and respectful AI technologies.

In the following sections, we present some significant works in social surveys and evaluations of LLM cultural sensitivities in Section 2. We present the methodology design of how to generate diverse and realistic survey responses in Section 4 in terms of the dataset (Section 4.1), evaluation metrics (Section 4.2), and inference framework (Section 4.3). Our experimental settings, investigation results, and ablation studies are in Section 5. Finally, we present the implications of our investigation in Section 6 and conclude the paper in Section 7.

2. Related work

The concept of culture encompasses the distinct ways in which different groups of people think, feel, and behave, setting them apart from one another (Hershcovich, Frank, Lent, de Lhoneux, et al., 2022). Researchers have developed concrete frameworks for conducting social surveys to identify cross-cultural values among humans (Frese, 2015). Two of the most extensively studied cultures, due to their significant differences, are the United States of America and the People's Republic of China (Cao et al., 2024). A variety of surveys, such as the World Values Survey by Haerpfer, Inglehart, Moreno, Welzel, et al. (2022), the General Social Survey (Smith, Marsden, Hout, & Kim, 2012), and the Chinese General Social Survey, [2] have been employed to identify specific preferences within these two cultural backgrounds.

As artificial intelligence becomes more advanced, there are increasing concerns about its potential social risks. To address these concerns, the field of Large Language Models (LLMs) has established comprehensive evaluation for assessing a model's social stances (Xu, Sun, Ren, et al., 2024). These evaluations cover three aspects: norms (Fraser, Kiritchenko, & Balkir, 2022; Haemmerl, Deiseroth, Schramowski, Libovický, et al., 2023; Hendrycks, Burns, Basart, Critch, et al., 2021; Ramezani & Xu, 2023; Scherrer et al., 2023), bias (Dhamala et al., 2021; Kaneko, Imankulova, Bollegala, & Okazaki, 2022; Kurita, Vyas, Pareek, Black, & Tsvetkov, 2019; Miotto, Rossberg, & Kleinberg, 2022; Nadeem, Bethke, & Reddy, 2021; Ross, Katz, & Barbu, 2021; Tao et al., 2023), and values (Arora et al., 2023; Cao et al., 2023; Wang et al., 2023; Xu, Liu, Yan, Xu, et al., 2023; Yao, Yi, Wang, Gong, & Xie, 2023). All of these evaluations can contribute to building more trustworthy large language models under investigations explored by Huang et al. (2024) and Liu et al. (2023).

Norms. To determine if models align with human standards, evaluations of norms often use practical situations and assess model reactions. Recently, researchers (Hendrycks et al., 2021) have utilized various datasets and empirical studies to investigate model knowledge on concepts such as justice, well-being, duties, virtues, and commonsense morality. Some studies (Fraser et al., 2022) have found high correspondence between model moral principles and certain demographic groups, while others (Ramezani & Xu, 2023) have investigated how monolingual language models contain knowledge about moral norms in different countries. Additionally, researchers (Scherrer et al., 2023) have studied what moral beliefs are encoded in different models, particularly in cases where the right choice is not obvious. While some studies (Haemmerl et al., 2023) have found that models encode differing moral biases, it is important to note that these biases do not necessarily correspond to cultural differences or commonalities in human opinions.

[2] http://cgss.ruc.edu.cn/English/Home.htm

Bias. The assessment of bias involves examining fairness, stereotypes, and other characteristics that are evident in the responses generated by models. Kurita et al. (2019) proposed a template-based method to quantify bias in BERT. Nadeem et al. (2021) have presented a large-scale natural English dataset to measure stereotypical biases in four domains. To benchmark bias across various domains, Dhamala et al. (2021) introduced an open-ended language generation dataset. Ross et al. (2021) introduced two generalizations quantifying social biases in texts to visually grounded embeddings and reduce biases in both vectors. To evaluate the personality, values, and self-reported demographics of GPT-3, Miotto et al. (2022) employed two validated measurement tools. Kaneko et al. (2022) proposed Multilingual Bias Evaluation score, to evaluate gender bias in various languages. Additionally, Tao et al. (2023) conducted an audit of large language models to identify cultural biases and assess the effectiveness of country-specific prompting as a mitigation strategy.

These research studies examine the social stance of a model from an observational perspective. However, there are certain limitations to these studies when it comes to probing the internal beliefs of the model and determining the consistency of the model's belief with the responses it generates.

Value. Values are different from behavior in that they operate at a higher, more complex level, providing deeper insights into model behavior than observational evaluations of specific actions (Frese, 2015). Researchers such as Xu et al. (2023) and Wang, Zhu, et al. (2023) have evaluated model values in Chinese and U.S. cultural contexts, respectively, but have not taken cross-cultural variations into account. Yao et al. (2023) have assessed model values using questions deemed risky in multilingual daily QA pairs and proposed an alignment paradigm. Arora et al. (2023) and Cao et al. (2023) have investigated model values using reliable social surveys from multiple cultural backgrounds, employing cloze-filling tasks and QA tasks. Despite these efforts, the consistency of value expressions and the portrayal of model behavior have yet to be fully explored.

In the realm of LLM inference studies, Lahoti et al. (2023) introduced a collective critique and self-voting (CCSV) approach to enhance the diversity of model responses. CCSV consists of a four-step process that harnesses the capabilities of large language models (LLMs) to critique and refine their responses. The steps are as follows:

(1) **Initial Response**: The LLM generates an initial response based on a given input prompt. (2) **Critique the Response**: The LLM is tasked with self-critique, providing suggestions for improvement on its generated response. (3) **Address the Critique and Rewrite**: The LLM receives the critiques from the previous step and is asked to address them by revising its initial response. (4) **Vote for the Best Response**: Finally, the LLM selects the best response from the revised drafts. This methodology is specifically aimed at enhancing the diversity of a list of entities, rather than being tailored for survey simulation or value assessment.

To explore cross-cultural diversity, we have included regional surveys to consider cultural discrepancies and streamlined our value metrics for improved comparison. Our diversity-enhanced framework enables us to more thoroughly and precisely measure model preferences and behaviors. Additionally, we have implemented multi-aspect evaluation metrics including insensitivity measurement to evaluate the coherence between model values and their corresponding behaviors.

3. Preliminaries: Value dimensions

In this section, we clarify the value dimensions obtained from previous surveys (Alexander, Inglehart, & Welzel, 2012; Dülmer, Inglehart, & Welzel, 2015; Hofstede, 2001; Hoftede, Hofstede, & Minkov, 2010; Welzel, Brunkert, Inglehart, & Kruse, 2019; Welzel & Inglehart, 2016; Welzel, Inglehart, & Kruse, 2017), which we utilize in the evaluation metrics for a better understanding of the paper. The **EVI** is developed by Haerpfer et al. (2022) and other dimensions are designed by Hofstede (2001).

- **Emancipative Values Index (EVI)** measures the extent to which society embraces values of freedom, self-expression, democracy, and human rights. It captures how much a culture emphasizes empowerment over coercion and conformity.
- **Individualism Index (IDV)** measures how people prioritize personal freedom, self-reliance, and independence over group loyalty, collective action, and social solidarity.
- **Indulgence versus Restraint Index (IVR)** measures how society allows relatively free gratification of natural human desires related to enjoying life.
- **Long-Term Orientation Index (LTO)** reveals time horizons and whether societal values are based on the past/present or oriented towards the future.
- **Masculinity Index (MAS)** measures beliefs about appropriate behaviors and attributes for each gender.
- **Power Distance Index (PDI)** provides insight into beliefs about social power, competition, and hierarchies.
- **Uncertainty Avoidance Index (UAI)** measures the degree to which people feel threatened by ambiguity, uncertainty, and unstructured situations.

4. Measuring alignment evaluation

4.1. Dataset

Introducing the simulation dataset - a comprehensive evaluation of the value profiles and consistency of LLMs in answering social survey questions from the U.S. and China. Our dataset is constructed in three steps, as shown in Fig. 2: 1. We carefully select multiple social questionnaires with global and regional features across various topics (see); 2. Social experts review each question and carefully exclude those that may intervene with the validity of the investigation for LLMs (see Section A.2); 3. Our dataset is expanded to accommodate the context sensitivity of the model (Tjuatja, Chen, Wu, Talwalkar, & Neubig, 2023) by incorporating closed-source LLMs like ChatGPT and Claude for rephrasing survey questions (see Appendix A.1).

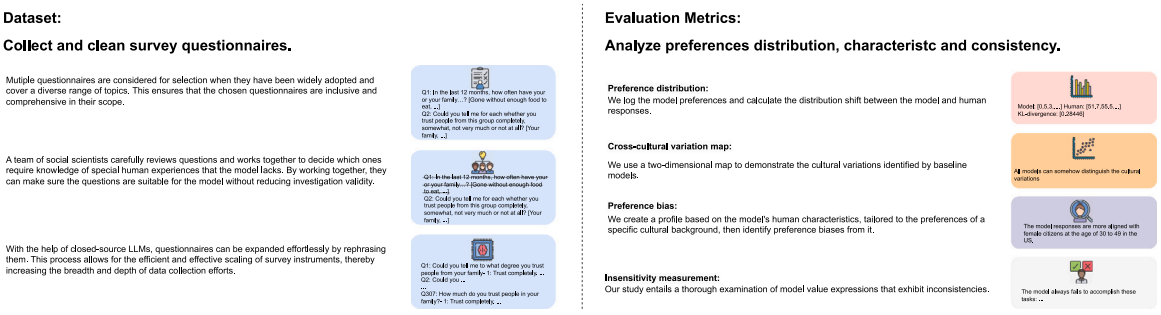


Fig. 2. The dataset involves three main steps (left): (1) collecting and cleaning survey questionnaires, (2) surveying LLMs using a diversity-enhanced framework, and (3) analyzing the distribution, characteristics, and consistency of preferences. We design multiple metrics (right) to measure the misalignment of model values using preference distribution, cross-cultural variation, preference bias, and insensitive measurement.

Data source. The experiment dataset consists of questionnaires spanning over five years of investigation, from 2018 to 2023. These surveys include: the seventh wave of the **World Value Survey**[3] (2023), a global network of social scientists studying changing values and their impact on social and political life. The **General Social Survey**[4] (2022), a survey of American adults monitoring trends in opinions, attitudes, and behaviors towards demographic, behavioral, and attitudinal questions, plus topics of special interest. The **Chinese General Social Survey**[5] (2018), the earliest nationwide and continuous academic survey in China collecting data at multiple levels of society, community, family, and individual.

Traditional topics such as attitudes & stereotypes, happiness & well-being, science & technology, and others are included in our dataset, it also addresses local culture and temporal issues, like attitudes towards policies during the COVID-19 era, energy restrictions, dietary practices during holidays, and other concerns.

The **Ipsos Understanding Society survey**,[6] is the preeminent online probability-based panel that accurately represents the adult population of the United States. Our dataset delves into topics such as the role of values in food habits and nutrition, as well as health and recreation. The Pew Research Center's **American Trends Panel**,[7] a nationally representative online survey panel, consisting of over 10,000 randomly selected adults from across the United States. Our dataset utilizes the panel to examine human attitudes towards democracy, social interactions, and social media. The **USA Today/Ipsos Poll** (Today USA, 2023), surveys a diverse group of 1023 adults aged 18 or older, including 311 Democrats, 290 Republicans, and 312 independents. Survey responses are collected on topics related to social opinions, including equality, ideology, and personal identities. The **Chinese Social Survey**,[8] longitudinal surveys focus on labor and employment, family and social life, and social attitudes. Our survey includes questions on public attitudes towards regulations and energy policy among other topics.

As a result, the dataset now encompasses a diverse range of survey topics (see Appendix A.4). Moreover, the dataset includes three question types: single-choice, multiple-choice, and scaling problems. Our comprehensive dataset contains over 1153 questions with up to 12,000 human annotations, featuring 613 U.S. queries (US survey) and 540 Chinese queries (CN survey). The value distribution of our dataset is illustrated in Fig. 3. With different research focusing on different cultures, the distribution of value dimensions and survey question topics are varied for US and CN survey tests.

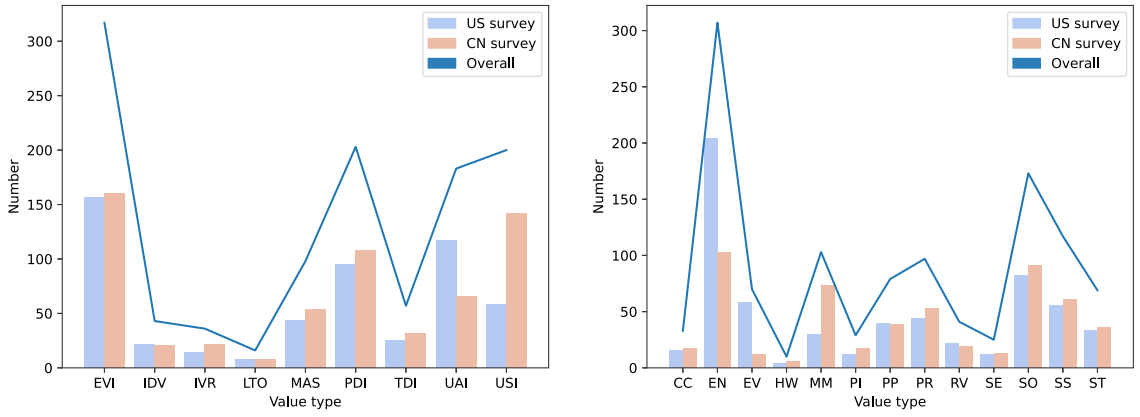
We also offer data analysis for quality control and contamination (Li, Geurin, & Lin, 2023), as detailed in Appendix A.

Prompt for LLMs. Following the design of the social survey, we frame our evaluation task in a multiple-choice question-and-answer format. To enhance the credibility of the model choice and quantify hallucination, we require the model to explain its chosen answer. We then assess the validity of the response using the insensitivity measurement (refer to Section 4.2.4).

Building upon previous explorations into zero-shot (Kojima, Gu, Reid, Matsuo, & Iwasawa, 2022) and multilingual (Huang, Tang, Zhang, Zhao, et al., 2023) prompting, we develop a Chain-of-Thought (CoT) cultural simulation prompt to assist the model in completing surveys. This prompt steers the model to answer questions and explain their choice step-by-step simulating a citizen from the U.S. or China. We form the instructions in English using the following prompt:

> I want you to act as a citizen from **[COUNTRY]**.
> You should obtain an answer from **[NUM]** choices.
> You should tell me the answer in the format 'Answer:' and explain afterward using ''Explanation:''.
> Request: **[QUESTION]**

[3] https://www.worldvaluessurvey.org/wvs.jsp
[4] https://gss.norc.org/About-The-GSS
[5] http://cgss.ruc.edu.cn/English/Home.htm
[6] https://www.ipsos.com/en-uk/understanding-society
[7] https://www.pewresearch.org/our-methods/u-s-surveys/the-american-trends-panel/
[8] http://css.cssn.cn/css_sy/



(a) Culture-based value distribution. (b) Culture-based survey topic distribution.

Fig. 3. The value-topic distribution of the dataset. This distribution is based on the human responses extracted from social surveys. For more details on value dimensions and survey topic, please refer to Appendix A.4 and C.

We replace keywords in the square brackets with information about the simulated citizen or queries to create a complete input, which is then fed into a diversity-enhanced model. Additional details on the Chinese version of the prompt and examples of variations created by our diversity-enhanced framework can be found in Appendix B.

4.2. Evaluation metrics

To assess model value preferences and the consistency of value expressions, we enhance the value dimensions utilized in current human social survey assessments based on our comprehensive multi-source dataset (referenced in Section 4.1). We then develop a multi-faceted measurement using these value dimensions to evaluate model value through alignment performance (Section 4.2.1), cross-cultural comprehension (Section 4.2.2), characteristics (Section 4.2.3), and expression consistencies (Section 4.2.4).

Value dimensions. Our value dimension implemented during evaluation is based on research of Cieciuch and Schwartz (2012) and Hofstede (2001) utilized in social surveys. Specifically, we identify and incorporate seven dimensions of value from these researchers as presented in Section 3.

We expand these value dimensions with two additional aspects: The **tradition index (TDI)** and the **Universalism Index (USI)**. TDI measures the importance a society places on upholding cultural, family, or religious traditions. USI measures how people value understanding, appreciation, tolerance, and protection for humankind and nature.

The reflections of each value dimension in terms of high or low scores are introduced in Appendix C.

4.2.1. Metric 1: Preference distribution

To depict the information lost from model value preferences to humans, we calculate the Kullback–Leibler Divergence (KL-D) as an asymmetric measurement. This measurement has been proven effective for machine learning like VAE and GAN optimization, and the asymmetric feature of KL-D provides the direction of discrepancy between the distribution of model preferences to humans, which is more suitable than other metrics like Jensen–Shannon Divergence used by Durmus, Nyugen, Liao, Schiefer, et al. (2023).

In detail, we extract human preference distributions P_{human} for each query from the social report on the target region and compare the proximity of the model preference distribution P_{model} to the human preference distribution using $\mathbb{D}_{\text{KL}}(P_{\text{model}} \parallel P_{\text{human}})$.

$$\mathbb{D}_{\text{KL}}(P_{\text{model}} \parallel P_{\text{human}}) = \sum_{i}^{n_{\text{choices}}} p_{\text{model}}(i) \ln\left(\frac{P_{\text{model}}(i)}{P_{\text{human}}(i)}\right) \tag{1}$$

where $p_{\text{model}}(i)$ is the probability of choice i in model distribution, $P_{\text{model}}(i)$ is the cumulative probability up to choice i in model distribution, $P_{\text{human}}(i)$ is the cumulative probability up to choice i in human reference data, $n_{choices}$ indicates the number of choices.

The missing value from the social report is set as 0, and the infinite value for KL divergence, which indicates the divergences is too large, is recorded as $\max(\mathbb{D}_{\text{KL}} - \{\inf\}) + 1$.

4.2.2. Metric 2: Cultural variation

Inspired by the Inglehart-Welzel World Cultural Map (Haerpfer et al., 2022), which examines cross-cultural variations, we create a Cultural Variation Map through Principal Components Analysis (PCA) using the average scores of 9 value dimensions.

To achieve this, we first calculate the average scoring matrix for each model, denoted by $\mathbf{S}_{\text{AV}} \in \mathbb{R}_{n_{\text{D}}}^{n_{\text{subjects}}}$, using

$$\mathbf{S}_{\text{AV}} = \text{Concate}(\text{Avg}_{\text{D}}(\mathbf{S}_1), \ldots, \text{Avg}_{\text{D}}(\mathbf{S}_{n_{\text{subjects}}})) \tag{2}$$

where Avg_D calculates the average score of all value dimensions, and $\mathbf{S}_1$ to $\mathbf{S}_{n_{\text{subjects}}}$ represent the value-dimensional score vector of each subject.

We then obtain $\mathbf{S}'_{AV}$ by subtracting the total average score from the value dimension average-scoring matrix. The covariance matrix $\mathbf{C}$ of $\mathbf{S}_{AV}$ is then calculated, along with its corresponding eigenvalues and eigenvectors, using equation

$$\mathbf{C} = \frac{1}{n_D}\mathbf{S}'_{AV}\mathbf{S}^{\top}_{AV} \tag{3}$$

where $\mathbf{S}'_{AV}$ is calculated by

$$\mathbf{S}'_{AV} = \mathbf{S}_{AV} - \text{Avg}(\mathbf{S}_{AV})\mathbf{E} \tag{4}$$

where Avg calculates the matrix average, and $\mathbf{E} \in \mathbb{R}^{n_{\text{subjects}}}_{n_D}$ is the identity matrix, n_D is the number of value dimensions.

Two vectors with the largest eigenvalues are then extracted as matrix $\mathbf{P} \in \mathbb{R}^2_{n_{\text{subjects}}}$. Finally, the PCA result is obtained through

$$\mathbf{Y} = \mathbf{P} \cdot \mathbf{S}_{AV} \tag{5}$$

where $\mathbf{Y}$ is the PCA result.

The resulting two-dimensional metric allows us to evaluate the average positions of the model and human values, illustrating cross-cultural variations.

4.2.3. Metric 3: Preference bias

Following the social survey design, we create two profiling dimensions: gender (male, female) and age groups (up to 29, 30–49, and over 50).

To investigate the model alignment performance, we distinguish the model using matching profiles and mismatch profiles. First, we take the most preferred choice for each question and see if the model response matches the human preferences (“matching” or “mismatch” profile). Then, we record the character of humans that align with the model response as the model profile.

These profiles serve as a means of categorizing the model’s characteristics and presenting its impact on different groups. Valuable insights into the model’s values and biases are gained from these profiles.

4.2.4. Metric 4: Insensitivity measurement

To make full use of the model response that cannot form valid value preferences and analysis model limitations, we propose the 6-dimensional insensitivity measurement evaluation.

Our methodology involves recording these distinct instances and classifying them into six dimensions based on Guilford’s three-dimensional Structure of Intellect model (Guilford, 1988). The structure is comprised of operations, contents, and products.

Operations refer to general intellectual processes, where we evaluate the model’s responses for question capture ability (QC) and role-play failures (RP). Contents denote areas of information to which the human intellect applies operations, and we measure the model’s ability to understand instructions (IU). Products consist of results of applying particular operations to specific contents, and we gauge the model’s responses for conflict in value expression (CV), irrelevant expression output (IR), and false fact presentation phenomenon (FF).

We use automatic assessment by asking GPT-3.5 to evaluate the above problems on sampled responses, the evaluation prompt design and effectiveness of the automatic judges are presented in Appendix D.

4.3. Inference framework: Diversity-enhanced framework

To effectively capture the unpredictability of model behaviors when generating value preferences through social surveys, we present a Diversity-Enhanced Framework (DEF) to probe model cultural values and help our investigation follow social survey paradigms (see Algorithm 1). We conduct a primary experiment that shows that models tend to produce repetitive responses with CoT instructions (see Section 5.3.0.1). Consequently, the same backbone provides a onefold answer, which fails to capture the value distribution.

To address these issues, we design a Diversity-Enhanced Framework that creates different model participants to enlarge the evaluation scope and generate preference distributions. Our designed DEF system better simulates real survey conditions by introducing controlled randomness in responses, accounting for memory effects, and allowing for demographic variation. The framework conducts three aspect modifications on the model.

Memory manipulation. Memory is a unique feature for humans when completing the survey, human often finds it hard to remember previous questions when the length of the questionnaire is large. Remembering different answering histories can influence the current choice. To simulate this phenomenon and enhance the flexibility of model behavior, we propose a memory manipulation in LLMs, which sets a maximum memory length and randomly deletes conversation history. Moreover, we clean the memory by removing the responses that fail to present answers after each inference because we find that these memories will impact future reasoning, and we only keep the request queries and model answers in the memory to build a larger memory span.

Prompt background. Previous research reveals that adjusting instructions elicits varied responses from models (Lahoti et al., 2023). Based on this discovery, we introduce city-level simulation information to build different participants from a single LLM. In detail, we introduce state and city information sampling from the probabilities of human participants after the [**COUNRY**] keyword, thereby requesting the model to simulate a person from a specific location within the U.S. or China. It is critical to note, however, that the model’s responses may not necessarily reflect the views of the person from that particular location.

Algorithm 1: DEF inference over one survey question in one simulation

Input: Input prompt template P, cultural context C, survey question Q, and memory from previous inference $H = [(h_{q_1}, h_{a_1}), \dots]$

Output: Simulation result R, updated memory $H' = [(h_{q_1}, h_{a_1}), \dots]$

Hyperparameters:

1. Max memory length L_m
2. Max num_beams B_m
3. Decoding sample switch S

Initialization:

1. Simulation memory length $l \leftarrow$ `random(0, `L_m`)`
2. Simulation num_beams $b \leftarrow$ `random(0, `B_m`)`
3. Simulation sample switch $s \leftarrow$ `random([True, False])`

Prompt Modification: We incorporated randomly selected locations into the simulation prompt to increase the diversity of cultural backgrounds.

city $\leftarrow$ `random(`[city$_1$, ...], 1`)`

Model input $\leftarrow$ `replace(`P, "[COUNTRY]", C+"living in "+city`)`

Model input $\leftarrow$ `replace(`P, "[QUESTION]", Q`)`

Model input $\leftarrow$ `Conversation(`H, Model input`)`;

Configuration Modification: We employ the randomly sampled generation parameters l, b, s as configuration to guide model output.

$R \leftarrow$ `generate(`Model input, l, b, s`)`

Memory Modification: We clean the memory by removing the responses that fail to present answers and randomly delete conversation history if memory is at capacity.

$H' \leftarrow []$

for h_{a_i} in H **do**

  `// Cleaning the memory.`

  if "Answer: " in h_{a_i} **then**

    answer_start$\leftarrow$`find(`h_{a_i}, "Answer: "`)`

    exp_start$\leftarrow$`find(`h_{a_i}, "Explanation: "`)`

    answer_end$\leftarrow$`find(`h_{a_i}[exp_start:], "."`)`

    answer$\leftarrow h_{a_i}$[answer_start:answer_end]

    $H' \leftarrow$`append(`H', answer`)`

if `len(`H'`)>` l **then**

  `// Randomly delete conversation history if memory is at capacity.`

  $tmp\_H' \leftarrow []$

  sample_index$\leftarrow$`random.sample(range(len(`H'`)), `l`)`

  for index in sample_index **do**

    $tmp\_H' \leftarrow$`append(`tmp_H', H'[index]`)`

  $H' \leftarrow tmp\_H'$

return R, H'

Generation configuration. The output of generations in LLMs is largely influenced by the decoding strategies, which can be viewed as another approach to simulate different participant features. In particular, we tune the *num_beams* parameter to dramatically alter model responses. We posit that configuring the *num_beams* parameter enables the model to generate multiple possible answering trajectories and generate the final output based on route possibilities.

Specifically, we establish a modification range for each modification and allow the machine to randomly select from it. For instance, for our prompt modification, we include 50 states in the U.S. and 23 provinces in China. The configuration modification changes two parameters: "num_beams" (with a maximum of $n_{\text{max\_beam}}$) and "do_sample" (true or false). The memory modification is set with a maximum of $n_{\text{max\_memory}}$ lengths. Examples of inference outputs in English and Chinese are presented in Fig. 4.

Furthermore, we conduct n_{test} surveys for each culture. Therefore, the overall of variations of our experiment is $\frac{n_{\text{test}}}{50*n_{\text{max\_beam}}*2*n_{\text{max\_memory}}}$ for the U.S. survey and $\frac{n_{\text{test}}}{23*n_{\text{max\_beam}}*2*n_{\text{max\_memory}}}$ for the CN survey.

5. Experiments

This section presents the results of our experiments simulating eleven LLM candidates (refer to Section 5.1) using the multi-aspect measurements previously described. In Section 5.2.1, we evaluate the effectiveness of our refined inference strategies within the

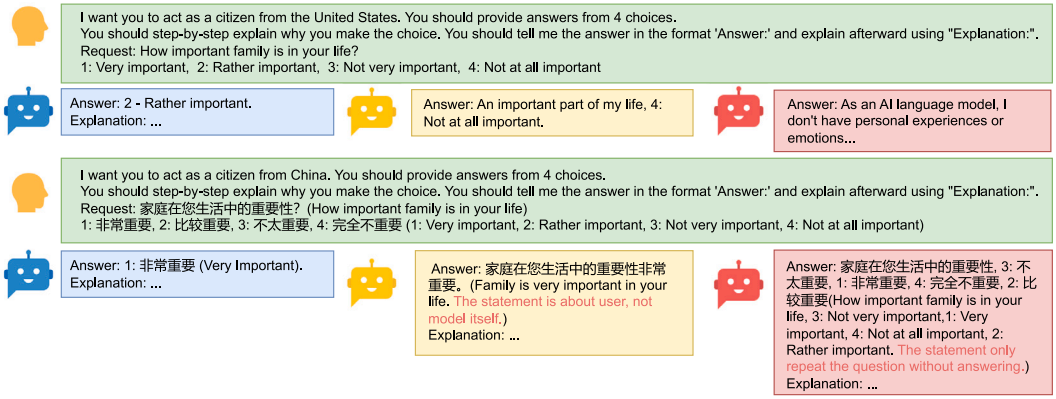


Fig. 4. After enhanced by DEF, a single model can produce different responses to the same question. Therefore, we can capture the model behavior during value investigations.

DEF system, focusing on accuracy and diversity. We then discuss the insights gathered from the grouped model responses analyzed through DEF across our multifaceted assessment (Sections 5.2.2 to 5.2.5). Additionally, we provide an ablation study examining the role of linguistic prompting and the contributions of each modification made to the DEF system (Section 5.3). To enhance understanding of model behaviors during survey completion, we present several cases in Appendix E that may illuminate potential improvements in value alignment and expression consistency.

5.1. Experiment settings

Model candidates. To accurately represent the parameter scales and data distribution of language models on English and Chinese corpora, we identify 11 large language model (LLM) candidates representing different architectures (encoder–decoder, decoder-only), covering major training approaches (SFT, RLHF, FFT), including both specialized (Chinese-focused) and general models for testing. These models are: Baichuan2-13B-Chat Yang, Xiao, Wang, Zhang, et al. (2023),[9] ChatGLM2-6B (Team, Zeng, Xu, Wang, et al., 2024),[10] WizardLM-13B (Xu et al., 2024),[11] Mistral-7B-Instruct (Jiang, Sablayrolles, Mensch, Bamford, et al., 2023),[12] Dolphin-2.2.1-Mistral-7B,[13] Mixtral-8x7B-Instruct (Jiang, Sablayrolles, Roux, Mensch, et al., 2024),[14] Llama-3-8B (Grattafiori, Dubey, Jauhri, Pandey, et al., 2024),[15] Llama-3-8B-Instruct,[16] Dolphin-2.9.1-Llama-3-8B,[17] Llama-3-Chinese-8B-Instruct,[18] Claude-3.5-Sonnet.[19] Particularly, we use Mistral-7B-Instruct and Dolphin-2.9.1-Llama-3-8B as a showcase to present the effectiveness of our DEF system.

Inference strategy baselines. As our probing framework is rooted in survey QA and designed to increase response diversity, we choose two baseline methods that fall into this category: (1) CCSV (Lahoti et al., 2023, see Section 2). This prompting technique uses collective critique and self-voting to self-improve LLM diversity reasoning capabilities without relying on hand-crafted examples or prompt tuning. (2) Cao et al. (2023) investigates the values of the model using social surveys and asks the model to infer the average human preference in a specific culture with prompt design.

Implementation details. For all datasets and the baselines, we set $n_{\text{max\_beam}} = 5, n_{\text{max\_memory}} = 15, n_{\text{test}} = 20$, do_sample=true, generation temperature as 1, top_p=1 (if effective), and max_new_tokens=512. We use packages like transformers to load the model, and anthropic to handle the Claude API. We employ 4-bit quantization using AutoGPT for 10B+ models, as past studies indicate that it has a minimal effect on model generation (Jin, Du, Huang, Liu, et al., 2024). All the experiments run on an NVIDIA Tesla A100 40G GPU for about 1400 h.

We conduct the inference baselines for $n_{\text{test}} = 20$ times for fair comparison. We set the number of decodes as 5 and used multilingual prompts (prompts 1 and 2) suggested in their research (Cao et al., 2023; Lahoti et al., 2023). CCSV is designed to generate lists of objects and enhance response diversity, which does not align with our simulation framework. To ensure a

[9] https://huggingface.co/baichuan-inc/Baichuan2-13B-Chat-4bits
[10] https://huggingface.co/THUDM/chatglm2-6b
[11] https://huggingface.co/WizardLMTeam/WizardLM-13B-V1.2
[12] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1
[13] https://huggingface.co/cognitivecomputations/dolphin-2.2.1-mistral-7b
[14] https://huggingface.co/TheBloke/Mixtral-8x7B-Instruct-v0.1-GPTQ
[15] https://huggingface.co/meta-llama/Meta-Llama-3-8B
[16] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[17] https://huggingface.co/cognitivecomputations/dolphin-2.9.1-llama-3-8b
[18] https://huggingface.co/FlagAlpha/Llama3-Chinese-8B-Instruct
[19] https://www.anthropic.com/news/claude-3-5-sonnet

Table 1
The result for diversity score and value divergence of DEF across countries. "Avg.↓" represents the average KL-divergence of all value dimensions. Our DEF framework can induce diverse responses and map model preference distribution with lower divergence to human value distribution against all baseline methods.

Probing framework	Diversity %↑	EVI	IDV	IVR	LTO	MAS	PDI	TDI	UAI	USI	Avg.↓
Mistral-7B-Instruct											
Cao et al. (2023)	0	6.34	6.08	5.33	6.34	6.44	6.48	6.39	6.32	6.36	6.36
CCSV-single	0	3.86	4.11	3.97	4.44	4.01	3.49	3.60	3.91	3.87	3.78
CCSV-list	19.20	0.59	0.68	0.41	0.42	0.65	0.76	0.39	0.37	0.48	0.53
DEF (Ours)	**89.80**	0.85	**0.65**	0.73	0.65	0.92	0.83	0.70	0.86	1.06	0.81
Dolphin-2.9.1-Llama-3-8B											
Cao et al. (2023)	0	5.99	5.17	5.89	5.76	5.25	5.58	6.17	5.45	5.37	5.63
CCSV-single	0	3.19	2.20	2.46	3.12	2.67	2.90	3.42	2.83	3.38	3.01
CCSV-list	19.20	0.59	0.68	0.41	0.42	0.65	0.76	0.39	0.37	0.48	0.53
DEF (Ours)	**89.86**	0.92	**0.58**	0.57	0.82	0.78	0.89	0.64	0.81	0.92	0.77

fair comparison, we implement two variations of its baseline methodologies: CCSV-single, which utilizes its recurrent generation approach to improve simulation diversity by focusing on one response at a time, which is tested 20 times during the experiment, and CCSV-list, which employs the same recurrent generation method to enhance the diversity of a total of 20 listed responses.

5.2. Result analysis

In this section, we seek to assess the effectiveness of our proposed probing framework, DEF, by comparing it with existing probing methods (as detailed in Section 5.2.1). We will also demonstrate how DEF can more comprehensively extract cultural value preferences by showcasing model alignment performance across each of the four evaluation metrics (from Section 5.2.2 to 5.2.5).

5.2.1. Probing effectiveness

To evaluate the diversity of model response induced by our DEF, we count the number of different responses n_{diff}, and calculate the diversity using the following metrics:

$$D = \frac{n_{\text{diff}} - n_{\text{ques}}}{n_{\text{test}} * n_{\text{ques}}} \tag{6}$$

where n_{diff} is the number of different responses observed, n_{ques} is the number of questions, n_{test} is the number of test iterations. Specifically, $0 \leq D \leq 1$, $D = 0$ indicates no diversity, $D = 1$ indicates maximum diversity.

According to the findings in Table 1, our DEF framework can significantly boost response diversity by 89% for both models in terms of DEF. Our DEF probing framework consistently achieves the lowest KL divergence (other than CCSV-list) across all value dimensions on the Mistral-7B-Instruct and Dolphin-2.9.1-Llama-3-8B compared to around 3 for CCSV-single and over 5 for Cao. This indicates that our Diversity-Enhanced Framework not only increases the model response diversity to obtain a distribution of the model preferences but also reduces the preferences divergence between model and human values.

Interestingly, the CCSV-single method, despite aligning with our task formulation and diverging from its original design (Lahoti et al., 2023), fails to deliver any degree of diversity within our dataset. This results in a significant disparity between model predictions and human preferences with KL-divergence around 3. In contrast, the compatible design of CCSV-list successfully achieves low-level diversity and minimizes divergence. Experimental results suggest that this inference strategy tends to repeat the options listed in the survey questions approximately 20 times, lacking differentiation in terms of cultural context and the model itself. However, our DEF provides the simulation answer along with corresponding explanations, which enhances transparency and is more conducive to human oversight. Moreover, our preference distribution reveals a distinguished peak of preferred options, yielding more meaningful statistical insights for further investigation.

In the following sections, we will present the discoveries in terms of induced alignment performance and captured model behaviors from four metrics.

5.2.2. Preference distribution

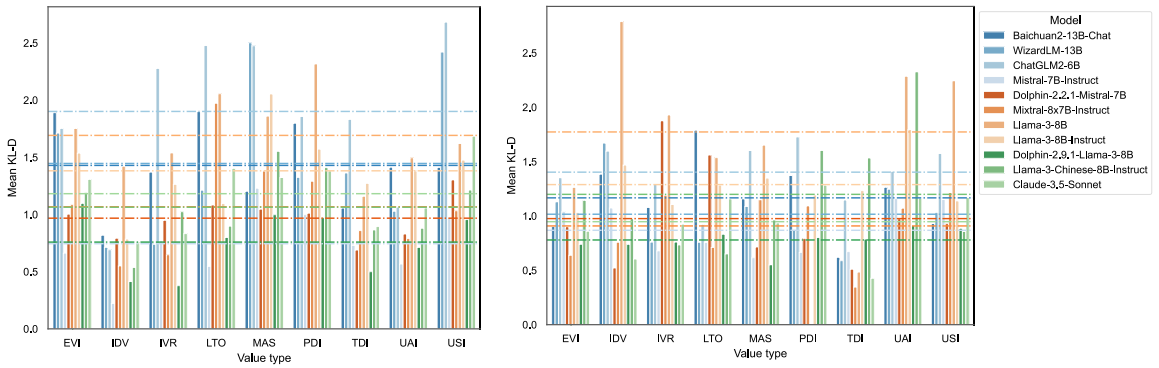
This section presents an analysis of the effectiveness of our Diversity-Enhanced Framework (DEF) in reducing the distribution shift between model and human preferences. We investigate how DEF affects the alignment of responses across a range of Large Language Models (LLMs), demonstrating its ability to improve performance on various value dimension inferences. We also present the best and worst aligned model configuration in Appendix F for capturing their unpredicted behaviors.

General performance: Our analysis reveals that DEF refinement mitigates the distributional gap between LLMs' and humans' responses in US and CN surveys, as shown in Table 2 and Fig. 5. Notably, various LLMs benefit from DEF, including Mistral-7B-Instruct, which achieves the smallest difference, especially in the U.S. culture, and Dolphin-2.9.1-Llama-3-8B, which presents lower divergence in Chinese cultural contexts.

Regarding cross-cultural alignment consistency, *while most models show coherent alignment with both U.S. and Chinese values*, some LLMs exhibit a stronger affinity for specific cultural contexts. For instance, Llama-3-8B-Instruct aligns more with U.S. values,

Table 2
Comparison of cultural value preference extraction accuracy across language models. The KL divergence score for each column is the average divergence of U.S. and Chinese culture. Mistral-7B-Instruct and Dolphin-2.9.1-Llama-3-8B achieve the lowest KL-D.

Model	EVI	IDV	IVR	LTO	MAS	PDI	TDI	UAI	USI	Avg.↓
Baichuan2-13B-Chat	1.40	1.10	1.22	1.84	1.18	1.59	0.84	1.34	1.17	1.30
ChatGLM-6B	1.55	1.14	1.79	1.70	2.04	1.79	1.49	1.24	2.13	1.65
Claude-3.5-Sonnet	1.08	0.68	0.88	1.28	1.13	1.33	0.66	1.11	1.43	1.06
Dolphin-2.9.1-Llama-3-8B	0.92	0.58	0.57	0.82	0.78	0.89	0.64	0.81	0.92	0.77
Dolphin-2.2.1-Mistral-7B	0.95	0.66	1.41	1.32	0.88	0.90	0.60	0.90	1.12	0.97
Llama-3-Chinese-8B	1.17	0.76	0.88	0.77	1.26	1.51	1.20	1.60	1.04	1.13
Llama-3-8B	1.51	2.11	1.73	1.80	1.76	2.32	0.82	1.89	1.93	1.76
Llama-3-8B-Instruct	1.28	1.13	1.18	1.19	1.70	1.38	1.25	1.59	1.31	1.34
Mistral-7B-Instruct	0.58	0.65	0.73	0.65	0.92	0.83	0.70	0.86	1.06	0.81
Mixtral-8 × 7B-Instruct	0.86	0.65	0.92	1.34	1.27	1.19	0.60	0.93	1.12	0.99
WizardLM-13B	1.42	1.19	0.75	0.99	1.80	1.10	0.98	1.14	1.73	1.23



(a) US survey (b) CN survey

Fig. 5. Mean KL-Divergence of LLM candidates on US and CN 9-value dimensions. Horizontal line: average KL-D. Dolphin-2.9.1-Llama-3-8B and Mistral-7B-Instruct show consistent alignment across both cultures.

and Llama-3-Chinese-8B-Instruct aligns more with Chinese values. Similarly, most language models, except ChatGLM2-6B and Llama-3-8B, align more strongly with U.S. cultural values.

The *base Llama-3-8B model shows consistent but low performance in both U.S. and Chinese contexts*, indicating our DEF system does not ensure strong alignment on pre-trained only models. While reduced misalignment was observed on additional trained models (like SFT and RLHF), with a greater effect in the U.S. context (KL-D of −0.49), the most significant improvements across both cultures (KL-D around −1) are seen with models using full-parameter fine-tuned (FFT) on unbiased datasets, as in Dolphin-2.9.1-Llama-3-8B.

Value-specific performance: The analysis in Fig. 6 shows the KL-D distribution across 9 value dimensions for all queries. Our research suggests a varied performance of the model over these values, but these variations are consistent with the general performance. In the Chinese context, most KL-D values cluster in the lower range (0 to 2), leading to a lower mean KL-D. The *Individualism (IDV) dimension is particularly challenging for most models in the Chinese setting*, as indicated by the long tail in the distribution. In contrast, *while the U.S. context shows greater divergence in model performance, models generally align better with values like Uncertainty Avoidance (UAI)* compared to their performance in the Chinese context.

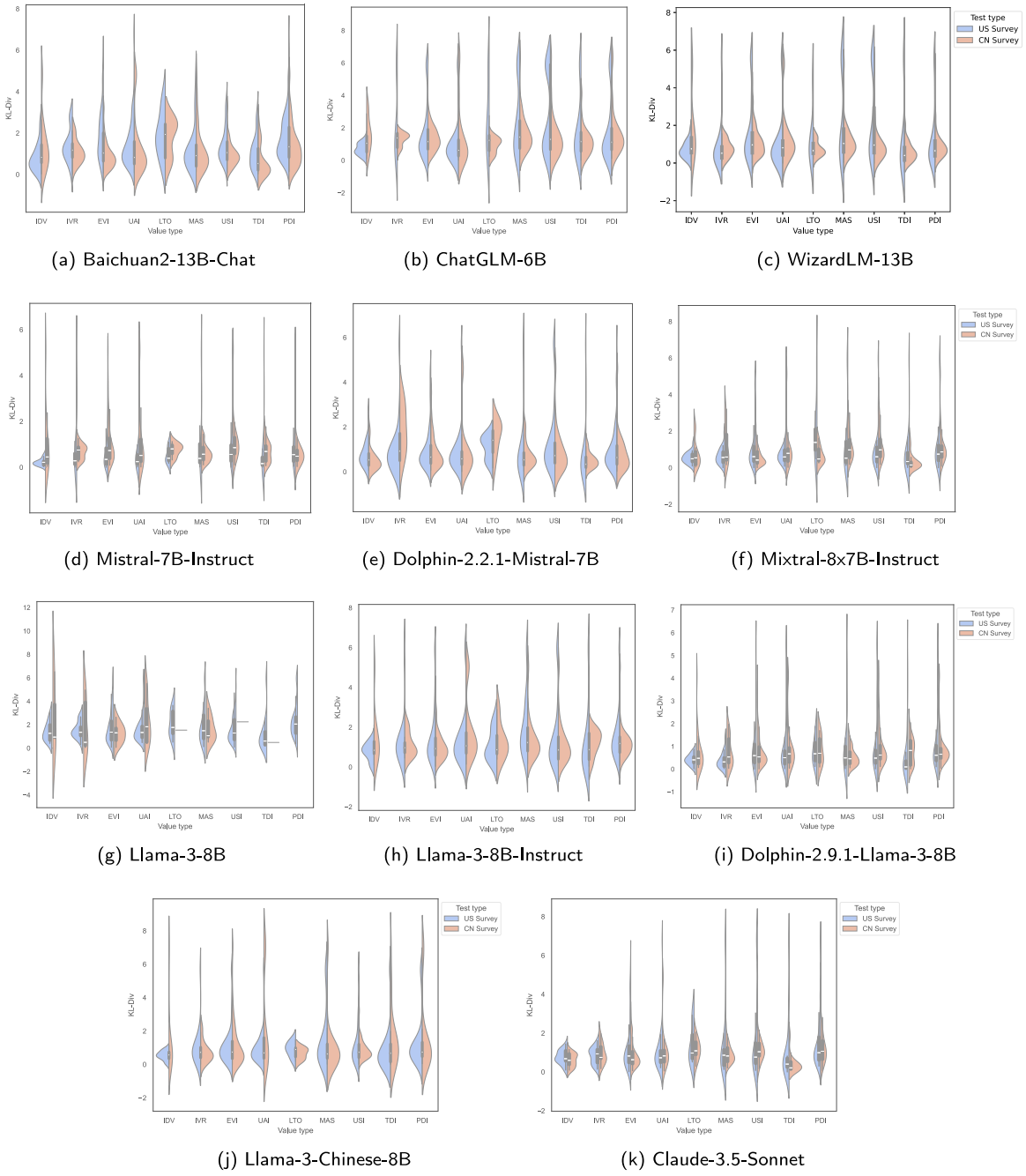
Best-aligned models augmented by DEF, Mistral-7B-Instruct and Dolphin-2.9.1-Llama-3-8B (Figs. 6(d) and 6(i)), have a skewed KL-D distribution (under 7), excelling in the Long-Term Orientation (LTO) value across both U.S. and Chinese contexts. These models also perform well on IDV in the U.S. and align better with Masculinity (MAS) and Power Distance (PDI) in the Chinese context. Dolphin-2.9.1-Llama-3-8B outperforms Mistral-7B-Instruct on Indulgence vs. Restraint (IVR) and Traditional (TDI) in the U.S., while Mistral-7B-Instruct excels in these dimensions in the Chinese context.

For U.S. culture, Mixtral-8x7B-Instruct and Claude-3.5-Sonnet outperform others in aligning with IDV and IVR. In the Chinese context, Llama-3-Chinese-8B-Instruct shows better performance on Emancipative Values (EVI), MAS, and Universalism (USI).

The analysis suggests that DEF performs differently across multiple training stages. *It can reduce KL-D distributions across most value dimensions on additional trained models*, particularly for IDV in the U.S. context. In the Chinese context, the improvement presents the model's ability to generate meaningful preference distributions rather than null, as reflected "-" in Fig. 6(g).

Furthermore, DEF performance on the FFT models, like Dolphin-2.9.1-Llama-3-8B model (Fig. 6(i)), generally outperforms the SFT combined with the RLHF approach. This is especially true for alignment with the IVR and LTO value dimensions.

Which is better? Mistral-series or Llama-3-series? As outlined in Table 2, models from the Mistral and Llama-3 series demonstrate the closest alignment in terms of value preferences within U.S. and Chinese cultural contexts. Understanding which model exhibits greater cultural sensitivity can aid researchers in selecting the appropriate model for specific applications and future studies.



(a) Baichuan2-13B-Chat (b) ChatGLM-6B (c) WizardLM-13B

(d) Mistral-7B-Instruct (e) Dolphin-2.2.1-Mistral-7B (f) Mixtral-8x7B-Instruct

(g) Llama-3-8B (h) Llama-3-8B-Instruct (i) Dolphin-2.9.1-Llama-3-8B

(j) Llama-3-Chinese-8B (k) Claude-3.5-Sonnet

Fig. 6. KL-Divergence distribution across 9 value dimensions for 11 LLM candidates (US and CN surveys). Better-aligned models excel in LTO for both cultures. DEF improves performance on IDV, IVR, and LTO with additional trained models. "–" denotes null KL-D distribution: value preferences unextractable from model responses.

To achieve this, we compile the distribution of value preferences for all models in the Mistral and Llama-3 series into two distinct groups and perform a t-test validation using Eq. (7) to assess whether there are significant differences in performance between these two series of models.

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}} \tag{7}$$

where $\bar{x}_1, \bar{x}_2$ are means of two model groups, s_1^2, s_2^2 are variances, and n_1, n_2 are sample sizes.

Table 3
The statistical result of t-test validation between Mistral- and Llama-3-series models. The result suggests that the Mistral-series models generally have significantly better value alignment performance than the Llama-3-series models.

Culture	Model	Mean	SD	SEM	t-value	DF	SED	p-value
U.S.	**Llama-3-series**	1.28	0.50	0.10	3.01	52	0.12	0.0041
	Mistral-series	0.93	0.34	0.07				
Chinese	**Llama-3-series**	1.35	0.74	0.14	2.77	52	0.16	0.0079
	Mistral-series	0.92	0.33	0.06				

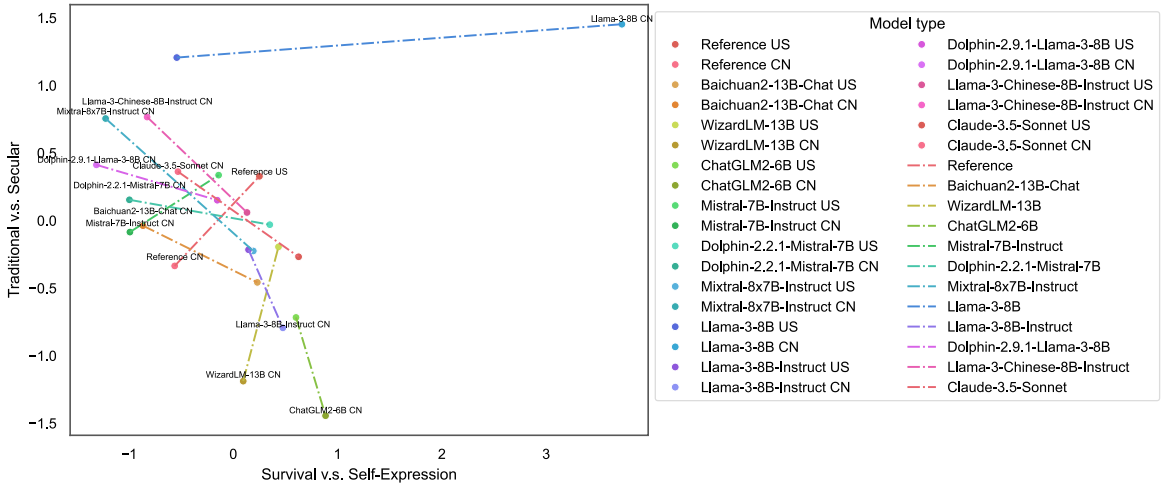


Fig. 7. The cross-cultural variation map of model candidates. All models can somehow distinguish the cultural variations. Mistral-7B-Instruct preserves better cross-cultural variations.

As presented in Table 3, the Mistral-series models generally exhibit significantly better value alignment performance than the Llama-3-series models. For the U.S. cultural context, the t-test result shows a t-value of 3.01 with 52 degrees of freedom and a p-value of 0.0041. This p-value is well below the conventional 0.05 significance threshold, indicating a statistically significant difference between the two model groups. Similarly, for the Chinese cultural context, the t-test yields a t-value of 2.77 with 52 degrees of freedom and a p-value of 0.0079. Again, this p-value is less than 0.05, providing strong statistical evidence that the Mistral-series models (mean of 0.92) outperform the Llama-3-series models (mean of 1.35) in value alignment, with a SED of 0.16.

These findings suggest that the **Mistral-series models have a more robust and comprehensive understanding of the underlying cultural values in both the U.S. and Chinese contexts**, resulting in superior performance on the value alignment benchmarks compared to the Llama-3-series models. The statistical significance of the results lends credibility to the conclusion that this difference is unlikely to have occurred by chance.

5.2.3. Cultural variation map

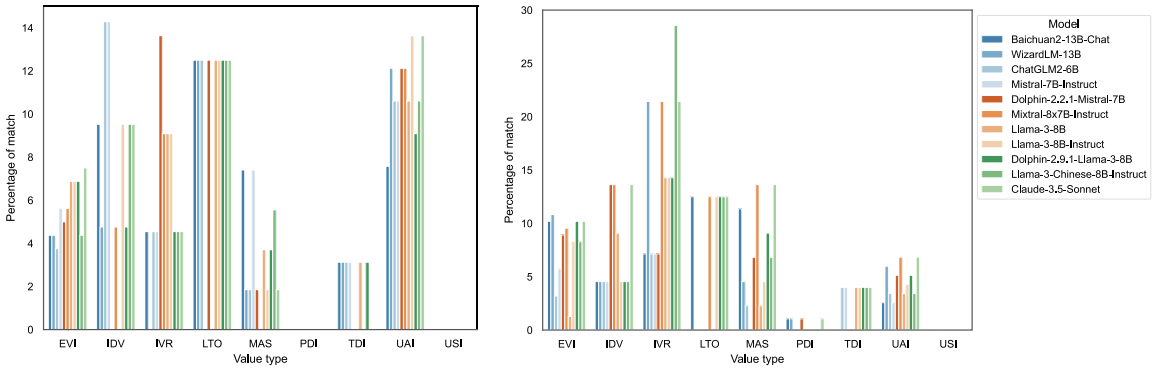
Our DEF framework can also be dynamically adjusted to different cultural contexts, which is captured by the Cultural Variation Maps metric. The experiment results are depicted in Fig. 7. This analytical technique is critical in understanding and managing the complex intricacies of cultural diversity in an open-domain environment.

The analysis reveals distinct cultural differences in the performance of the models. The reference value in the Chinese cultural context is generally positioned at the lower end of the spectrum, while the U.S. cultural context tends to be at the higher end. Notably, models like Mistral-7B-Instruct and WizardLM-13B stand out for their effectiveness in managing cross-cultural variations. These models align more closely with the reference direction, indicating a better ability to navigate cultural nuances.

Interestingly, despite ChatGLM2's lower overall alignment and greater deviation from reference points, it excels in capturing cross-cultural variations. The proximity of its mapped data points suggests a strong ability to encode the multidimensional nature of cultural differences.

Best-aligned model, *Mistral-7B-Instruct, not only preserves the distribution of human preferences but also effectively encodes cross-cultural variations*. In contrast, Dolphin-2.9.1-Llama-3-8B aligns closely with U.S. culture but, like most models, fails to preserve these variations.

The analysis of Llama-3-based models reveals several key insights regarding the impact of DEF over models using different training stages on encoding cultural variations. For the pre-trained model, DEF primarily encodes the direction of cultural variations, rather than the distance between cultural preferences. DEF tends to over-fixate on cross-cultural variations to models with SFT and RLHF, affecting both the direction and distance. In contrast, for models using Full-Parameter Fine-Tuning (FFT) on unbiased datasets, DEF reduces the overall variation while still impacting the variation directions, observed across both the 2.2.1 and 2.9.1 versions of the datasets.



(a) US survey (b) CN survey

Fig. 8. The distribution of matching profiles for model candidates on each value dimension. There is a limited proportion of matching cases for all models in both US and CN surveys, with advantages on EVI and UAI, while failing on USI, PDI, and TDI.

5.2.4. Preference bias

In addition to identifying differences existing between the model and human preferences, our evaluation is also capable of exploring the potential biases that might be present in the model's preferences based on factors such as gender and age. We include a visual representation of the number distribution of matching profiles across each value dimension in Fig. 8.

Matching preference: In the survey tests conducted in both the US and CN, performances of DEF with different models demonstrate limited success in aligning with human preferences, with matching percentages below 15% in the US survey and around 30% in the CN survey. These findings are consistent with the preference distribution observed. Value types with more matching cases exhibit lower KL-D scores. *Across all models, performance on EVI and UAI surpasses that of other value types.* However, they encounter difficulties in aligning with the USI value, achieve disappointing matching percentages for PDI and TDI, and exhibit inconsistent performance in matching with LTO across different cultures. Furthermore, challenges are encountered in aligning with value types like IDV and IVR in the US survey test, and with UAI in the CN survey test.

The ranking of the overall matching cases aligns with the alignment ranking consistent with the KL-D evaluation, from best Mistral-7B-Instruct and Dolphin-2.9.1-Llama-3-8B to worst ChatGLM2-6B. This indicates that understanding human values is a complex task requiring strong cross-cultural sensitivities and unbiased preferences. It is worth noting that while Baichuan2-13B-Chat may not demonstrate the topmost alignment performance in terms of KL-D, it consistently represents value types, with 7 out of 9 values. This could explain its moderate distributional performance while encoding cross-cultural preferences in a way that differs from the references.

In U.S. culture, *Llama-3-based models, particularly Llama-3-8B-Instruct and Dolphin-2.9.1-Llama-3-8B, consistently outperform other models*. Following closely behind are the Claude-3.5-Sonnet and Mistral-Based models, which generally exhibit better performance on metrics such as EVI, MAS, and UAI. Meanwhile, in Chinese culture, models with Chinese specialties, such as Llama-3-Chinese-8B-Instruct, stand out in IVR value, with Claude-3.5-Sonnet, Mistral-based models, and other Llama models trailing behind. Our distributional top-performance model, *Mistral-7B-Instruct, exhibits significantly better performance than other models in terms of IDV and MAS within the U.S. culture*.

Bias of mismatched preferences: As in Fig. 11, it is apparent that there is a distinct gender bias in the US and CN surveys for matching profiles. *The US survey shows a relatively smaller gender bias than the CN survey*. Notably, Llama-3-8B-Instruct and Llama-3-Chinese-8B-Instruct exhibit less bias than other models, despite not being the best-aligned models. DEF on the pre-trained model establishes a gender balance on mismatching profiles, while on SFT + RLHF trained models, a significant reduction in the preference bias is observed.

In Fig. 12, it is clear that *middle-aged characters are significantly predominant in both the matching and mismatched profiles*, with the latter showing relatively less bias. However, the models *inadequately represent preferences for characters under the age of 29*. Interestingly, results on the pre-trained-only model, Llama-3-8B, exhibit reduced bias. Models with less gender bias, such as Llama-3-8B-Instruct and Llama-3-Chinese-8B-Instruct, also demonstrate less age bias among age groups of 30 to 49 and over 50. The reduction of bias among age groups over 29 years old on SFT + RLHF models is greater than that on FFT using unbiased data.

These discoveries, which are not closely consistent with the findings of other tasks in value alignment, suggest that preference bias and value alignment might entangled in a complex way that needs to be balanced in future investigations.

For a more thorough understanding of the mismatch profiles for each model, please refer to Appendix G. Additionally, you can find the configurations for the best and worst aligned model in the last two columns of Appendix F for further research.

5.2.5. Insensitivity measurement

The inherent uncertainty in model responses captured by our DEF system is evaluated by insensitive measurement. In the following section, we will share our observations on the consistency of responses using insensitivity measurement. To gather data, we sampled 1% of responses from each model on both US and CN surveys. The overall results are presented in Fig. 9.

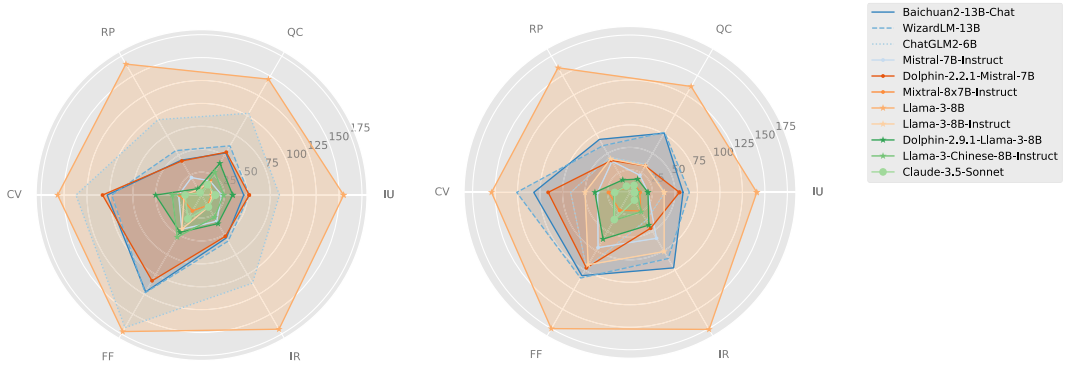


Fig. 9. The overall comparison of the insensitivity measurement among models on each survey. Model shows significant issues with FF and CV, with larger models having fewer problems on the US survey, and smaller models on the CN survey.

Table 4
The ablation result for verifying DEF in terms of the average KL-Divergence and Diversity across countries.

Model	Diversity	KL-D
None	0	2.2341
DEF-w/o memory	3.3%	2.0607
DEF-w/o config	67.3%	2.0431
DEF-w/o prompt	78.3%	2.1612
DEF	**79.8%**	**2.0030**

General performance: As the model-specific result illustrated in Figure 19, we notice that *most models have fewer problems on the US survey* test except for ChatGLM2-6B and Llama-3-Chinese-8B-Instruct. Our analysis indicates that *all model candidates experienced significant issues with* FF *and* CV.

Issue-specific performance. On both survey tests, the *Mixtral-8x7B-Instruct with larger-scale MoE shows the fewest issues*, followed by the Claude-3.5-Sonnet. On the other hand, the Llama-3-8B exhibits the poorest performance. Most models have difficulties with CV and FF in both tests, but they show improvements in IR and IU for the US survey test. For detailed results on each model, please refer to Figure 19.

It is noted that inferencing on pre-trained only models presents more significant challenges compared to other models. However, these challenges are reduced with additional trained models. In line with previous research (Ouyang et al., 2022), it is found that *models with SFT + RLHF yield greater improvements than with FFT in both US and CN survey tests, having a more noticeable impact on the US test.*

5.3. Ablations

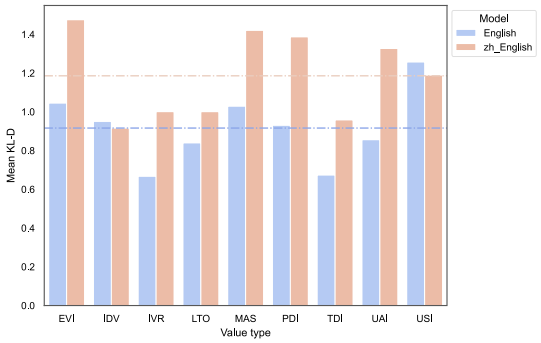
Language ablation. We conducted a language ablation study on the Mistral-7B-Instruct model to explore how it aligns with or differs in values when presented with questions from surveys in different languages. To achieve this, we utilized the benchmark dataset and translated it into different languages (Chinese to English, English to Chinese) using the Google translation tool. We then employed the DEF framework to examine value preferences using the translated data. The experiment results, including preference distribution and cultural variation map, are illustrated in Fig. 10.

Our research shows that Mistral effectively incorporates cultural values when using the native language. However, we observe that the model only aligns well with a handful of value dimensions such as USI. The cross-cultural variation map reveals that *utilizing different languages for analysis may significantly disrupt cross-cultural differences as the model struggles to differentiate between cultural variances* in terms of both direction and distance on the map.

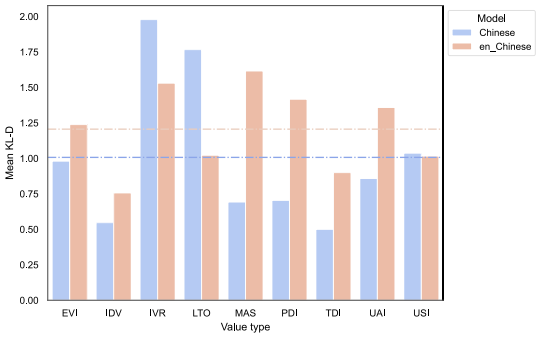
5.3.0.1. Diversity ablation. Our Diversity-enhanced Framework is created to treat LLMs as combinations of culture groups and modify various features of the model to generate diverse survey participants for value analysis. To understand how much influence each modification contributes to the response diversity, we conduct the following ablation study. Following the experiment design in Section 5.2.1, we disable each modification in the diversity-enhanced framework to create ablation experiments (DEF-w/o prompt, DEF-w/o config, DEF-w/o memory), and record their diversity score.

According to the findings in Table 4, three modifications can significantly boost response diversity by 79.8% in terms of DEF and obtain the lowest KL-D. Our memory mechanism, which is a new approach, can generate a diversity of up to 76.5%.

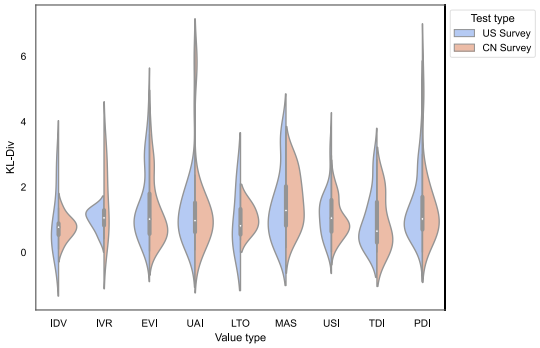
This indicates that our Diversity-Enhanced Framework not only increases the model response diversity to obtain a distribution of the model preferences but also reduces the preferences divergence between model and human values.



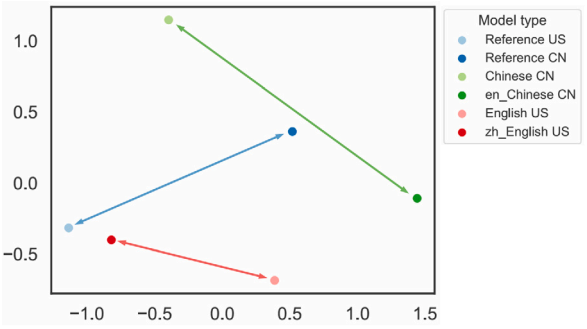
(a) Mean KL-Divergence (KL-D) over 9 value dimensions using original data (English) and translated data (zh_English).



(b) Mean KL-Divergence (KL-D) over 9 value dimensions using original data (Chinese) and translated data (en_Chinese).



(c) KL-Divergence (KL-D) distribution over 9 value dimensions using translated data.



(d) The cross-cultural variation map of language ablation study.

Fig. 10. Language ablation study results. Using alternative languages can impact model value preferences as well as cross-cultural sensitivities.

6. Discussion

Value alignment has shown to be an effective approach to addressing the societal issues arising from the employment of LLMs (Pawar et al., 2024). Our research paper introduces the realistic evaluation framework of cultural value alignment, which utilizes social surveys as a practical example for generating survey responses through LLMs. To assess the value discrepancies and behaviors of the models, we designed comprehensive evaluation metrics that encompass multiple aspects of value alignment. Additionally, we introduce a diversity-enhanced framework that serves as a corresponding evaluation method, enabling us to explore the model's value preferences and unpredictable behaviors.

6.1. Insights for model construction

Understanding factors that influence the model alignment performances in terms of construction stages is one of the largest concerns in the field. Beyond the robust quantitative results, we also observe interesting tendencies that need to be confirmed by future research:

While conducting observations of limited models, we noticed that a combination of automatically enhanced data, multilingual data, and human-annotated data can lead to better-aligned large language models than the base model. Plus, increasing the percentage of multilingual data (like Chinese data) can lower the average distributional gaps and improve performance on insensitivity measurement for Chinese culture.

In terms of model size, the larger model shows more matching profiles and better response consistency, but they do not necessarily lead to better alignment or cultural representation. The larger model also displayed more bias in certain demographic representations.

Further research is needed to confirm or refute this finding, including a controlled experiment to manipulate the alignment techniques, multilingual data proportion, and model size. This could provide valuable insights into the relationship between different model construction decisions and value alignment performance and could have important implications for our understanding of how to build a better-aligned model.

We provide examples of analysis to support our findings in Appendix I, but we acknowledge that limitations exist and further research may make different discoveries.

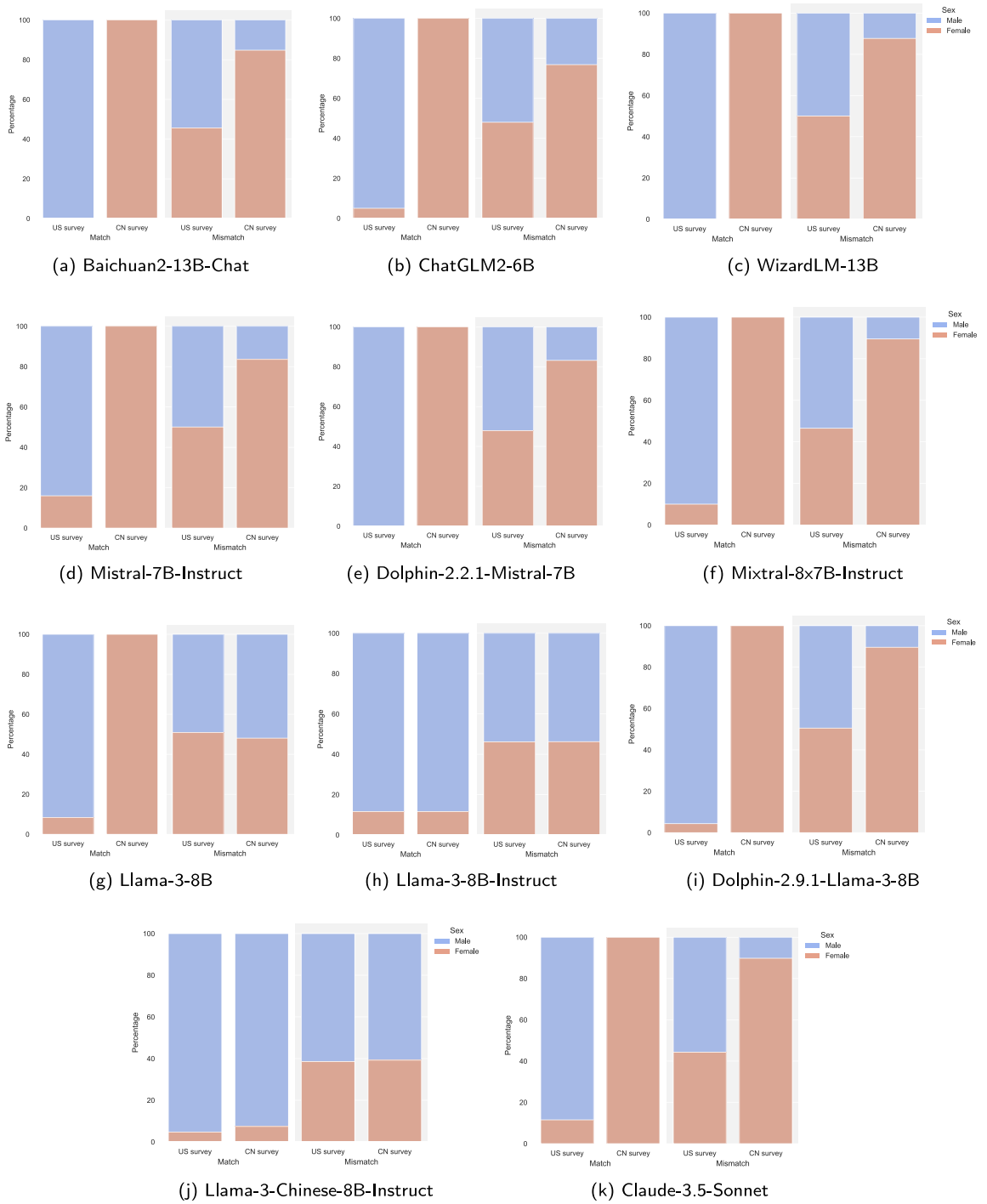


Fig. 11. Mismatch proportions by gender for all candidates. There is a heavy bias on male profiles in the US survey, and female profiles in the CN survey. But in general, gender bias on the mismatch profiles in the US survey is less significant than in the CN survey.

6.2. Theoretical implications

Our refined inference strategy addresses the first research question (**RQ1**: *What innovative methodologies can we develop to capture the complexity and variability of model responses across different cultural contexts?*). We present a Diversity-Enhanced Framework (DEF) that offers a novel approach to capturing the intricacies and variability of model responses across various cultural settings.

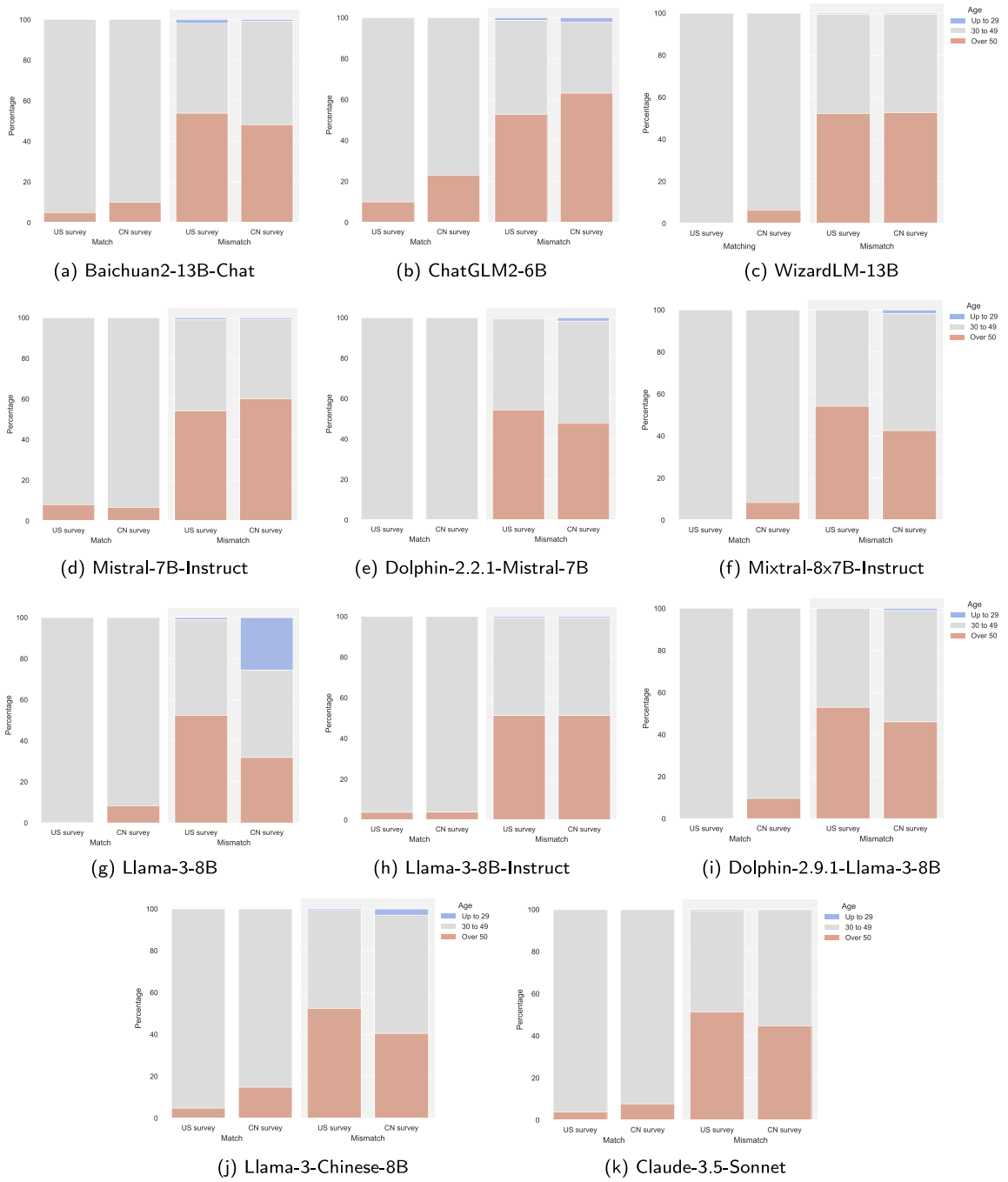


Fig. 12. Proportion of each age group in mismatch profiles across all candidates, following the social survey design. Both matching and mismatch profiles show a strong bias toward middle-aged characters, with under-representing preferences for characters under 29 years old.

Specifically, we implement a comprehensive system that includes modifications to prompts, configurations, and memory. This refined inference strategy generates groups of models based on different candidates to ensure accurate distributions of value preferences. Furthermore, we develop a multifaceted assessment to capture the diversity and uncertainty inherent in the behaviors of large language models.

In our experiment, we investigate two existing inference baselines to address the second research question (**RQ2**: *To what extent can refined inference strategies enhance the accuracy and reliability of cultural value alignment assessments?*). We find that our DEF system

is effective in improving response diversity and minimizing preference divergence between models and human responses. Notably, our proposed memory modification techniques significantly enhance response diversity and substantially reduce divergence.

In our analysis, which examined eleven model candidates across four evaluation metrics, we specifically addressed the third research question (**RQ3**: *Using the refined inference strategies, to what extent do the cultural value patterns exhibited by LLMs mirror those found in social survey responses from the United States and China?*). We identified significant limitations in the models' ability to align with the cultural value distributions of both nations, as well as observable biases in gender and age representation. Furthermore, the analysis indicates that certain models, particularly the Mistral-series and Llama-3-series, demonstrate superior performance in value alignment, with Mistral-series models showcasing a more robust understanding of cultural values across various contexts.

6.3. Practical implications

Our evaluation framework is a valuable tool for analyzing cultural sensitivity and identifying preference bias in LLM responses. Our diversity-enhanced framework can prompt the model to generate value preference distribution, while our multi-aspect measuring tasks provide a comprehensive view of value alignment, cross-cultural understanding, preference bias, and value expression consistency. This information can enable computer researchers to develop culturally sensitive LLM agents and reduce bias in future model development. Models demonstrating superior performance on our benchmarks exhibit a more nuanced understanding of the foundational aspects of cultural preferences. Theoretically, this should translate to improved performance on related tasks involving norms and bias. However, due to the inherent unpredictability of Large Language Models (LLMs), these correlations cannot be definitively established.

Furthermore, recent studies Kirk et al. (2024) emphasize the critical role of value alignment in enhancing user experiences, particularly within complex, value-laden domains. This focus underscores the significance of our evaluation research in identifying optimal model construction approaches for these challenging scenarios. Concurrently, a paradigm shift in information access is occurring, transitioning from traditional search engine methodologies to AI-driven integration systems. These emerging frameworks, while potentially enhancing information retrieval efficiency, introduce risks of cultural mediation and misalignment with diverse user values. Consequently, the imperative for rigorous investigation of model value preferences becomes paramount.

Our research contributes to this critical domain by offering insights into optimal model selection and methodological strategies, thereby advancing the development of more culturally sensitive and value-aligned AI systems capable of navigating the complex landscape of multicultural user bases while mitigating potential biases and misalignments.

7. Conclusion

In conclusion, this study examined the cultural value alignment of Large Language Models (LLMs) through an innovative, diversity-enhanced methodological framework that simulated citizen responses across the United States and Chinese cultural contexts. By developing refined inference strategies and implementing a comprehensive multi-aspects cross-value assessment, our research revealed significant limitations in current LLM models' ability to authentically represent and align with diverse cultural value distribution. The systematic analysis exposed notable concerns regarding cross-cultural representation, potential preference biases related to gender and age demographics, and inconsistency of value expressions. Notably, the Mistral-series models particularly distinguish themselves through a more nuanced and robust understanding of underlying cultural complexities. These findings not only underscore the critical importance of developing culturally sensitive AI technologies but also provide a methodological blueprint for future research aimed at enhancing the cross-cultural competence of large language models, ultimately contributing to more inclusive and contextually aware artificial intelligence systems.

Limitations. The limited scope of the current dataset, which only includes a moderate number of cultures and a lack of better alignment methods for our task, presents a constraint for more practical applications. However, the promising results observed thus far demonstrate the high quality of the dataset and the potential of the proposed paradigm, which can be further advanced by incorporating broader multicultural contexts with social survey data and investigating proper methods for alignment on the task.

The individuals or groups tasked with designing and administering the surveys may inadvertently introduce their own biases and subjective interpretations, which can then influence the evaluation of LLM values alignment. Future investigations should carefully design survey questions and response scales to minimize the introduction of researcher biases.

The lack of transparency regarding the actual data sources used for pretraining, content domains, and cultural absence in many LLMs, such as Mistral and Llama-3, presents a significant limitation. To address this, future research should aim to increase the interpretability and explainability of these black-box models, which not only constrains the ability to comprehensively understand their behavior but also has ethical implications downstream.

CRediT authorship contribution statement

Haijiang Liu: Writing – review & editing, Writing – original draft, Software, Methodology, Data curation, Conceptualization. **Yong Cao:** Writing – review & editing, Conceptualization. **Xun Wu:** Writing – review & editing, Conceptualization. **Chen Qiu:** Writing – review & editing, Funding acquisition. **Jinguang Gu:** Writing – review & editing, Supervision, Funding acquisition. **Maofu Liu:** Writing – review & editing, Conceptualization. **Daniel Hershcovich:** Writing – review & editing, Conceptualization.

Acknowledgments

We present our gratitude to Jeff Pan, Chao Gao, and Qiyuan Li for their insightful suggestions. The study is sponsored by the National Key Research and Development Program of China under Grants 2022YFC3300801 and the Knowledge Innovation Program of Wuhan-Shuguang Project under Grants 2023010201020409. We thank the Wuhan Supercomputing Center's assistance and the following esteemed researchers for manual assessment: Zihui Shi, Jian Chen, Yiran Hou, Zhongxin Li, Weimin Wu, Hao Ren, and Tong Fang. We also thank the anonymous reviewers and associate editors for their valuable feedback.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ipm.2025.104099.

Data availability

Data will be made available on request.

References

Alexander, A. C., Inglehart, R., & Welzel, C. (2012). Measuring effective democracy: A defense. *International Political Science Review*, *33*(1), 41–62.

AlKhamissi, B., ElNokrashy, M., Alkhamissi, M., & Diab, M. (2024). Investigating cultural alignment of large language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 12404–12422). Bangkok, Thailand: Association for Computational Linguistics.

Arora, A., Kaffee, L.-a., & Augenstein, I. (2023). Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the first workshop on cross-cultural considerations in NLP (c3NLP)* (pp. 114–130). Dubrovnik, Croatia: Association for Computational Linguistics.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., et al. (2023). A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th international joint conference on natural language processing and the 3rd conference of the Asia-Pacific chapter of the association for computational linguistics, (IJCNLP) 2023 -volume 1: long papers* (pp. 675–718). Nusa Dua, Bali: Association for Computational Linguistics.

Cao, A., Carstensen, A., Gao, S., & Frank, M. C. (2024). United States-China differences in cognition and perception across 12 tasks: Replicability, robustness, and within-culture variation. *Journal of Experimental Psychology: General (Washington, DC)*.

Cao, Y., Zhou, L., Lee, S., Cabello, L., et al. (2023). Assessing cross-cultural alignment between chatGPT and human societies: An empirical study. In *Proceedings of the first workshop on cross-cultural considerations in NLP (c3NLP)* (pp. 53–67). Dubrovnik, Croatia: Association for Computational Linguistics.

Chang, Y., Wang, X., Wang, J., Wu, Y., et al. (2023). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in neural information processing systems 30: annual conference on neural information processing systems 2017* (pp. 4299–4307). Long Beach, CA, USA.

Cieciuch, J., & Schwartz, S. H. (2012). The number of distinct basic values and their structure assessed by PVQ-40. *Journal of Personality Assessment*, *3*(94), 321–328.

Dan, Y., Lei, Z., Gu, Y., Li, Y., et al. (2023). EduChat: A large-scale language model-based chatbot system for intelligent education. CoRR abs/2308.02773.

Dhamala, J., Sun, T., Kumar, V., Krishna, S., et al. (2021). BOLD: dataset and metrics for measuring biases in open-ended language generation. In *2021 ACM conference on fairness, accountability, and transparency* (pp. 862–872). Toronto, Canada: ACM.

Dülmer, H., Inglehart, R., & Welzel, C. (2015). Testing the revised theory of modernization: measurement and explanatory aspects. *World Values Research*, *8*(2), 68–100.

Durmus, E., Nyugen, K., Liao, T. I., Schiefer, N., et al. (2023). Towards measuring the representation of subjective global opinions in language models. CoRR abs/2306.16388.

Fraser, K. C., Kiritchenko, S., & Balkir, E. (2022). Does moral code have a moral code? Probing delphi's moral philosophy. In *Proceedings of the 2nd workshop on trustworthy natural language processing* (pp. 26–42). Seattle, U.S.A.: Association for Computational Linguistics.

Frese, M. (2015). Cultural practices, norms, and values. *Journal of Cross-Cultural Psychology*, *46*(10), 1327–1330.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., et al. (2024). The llama 3 herd of models.

Guilford, J. (1988). Some changes in the structure-of-intellect model. *Educational and Psychological Measurement*, *48*(1), 1–4.

Haemmerl, K., Deiseroth, B., Schramowski, P., Libovický, J., et al. (2023). Speaking multiple languages affects the moral bias of language models. In *Findings of the association for computational linguistics:* (pp. 2137–2156). Toronto, Canada: Association for Computational Linguistics.

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., et al. (2022). World values survey: Round seven-country-pooled datafile version 5.0. *Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat*, *12*(10), 8.

Hendrycks, D., Burns, C., Basart, S., Critch, A., et al. (2021). Aligning AI with shared human values. In *9th International conference on learning representations:*. Austria: OpenReview.net.

Hershcovich, D., Frank, S., Lent, H. C., de Lhoneux, M., et al. (2022). Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers):* (pp. 6997–7013). Dublin, Ireland: Association for Computational Linguistics.

Hofstede, G. (2001). Culture's consequences: Comparing values, behaviors, institutions and organizations across nations. Thousand Oaks.

Hoftede, G., Hofstede, G. J., & Minkov, M. (2010). *Cultures and organizations: software of the mind: intercultural cooperation and its importance for survival*. McGraw-Hill.

Huang, Y., Sun, L., Wang, H., Wu, S., et al. (2024). Trustllm: Trustworthiness in large language models. In *Forty-first international conference on machine learning*.

Huang, H., Tang, T., Zhang, D., Zhao, X., et al. (2023). Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the association for computational linguistics:* (pp. 12365–12394). Singapore: Association for Computational Linguistics.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., et al. (2023). Mistral 7B.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., et al. (2024). Mixtral of experts.

Jin, R., Du, J., Huang, W., Liu, W., et al. (2024). A comprehensive evaluation of quantization strategies for large language models.

Kaneko, M., Imankulova, A., Bollegala, D., & Okazaki, N. (2022). Gender bias in masked language models for multiple languages. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*. Seattle: Association for Computational Linguistics.

Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., et al. (2024). The PRISM alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. arXiv preprint arXiv:2404.16019.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In *Advances in neural information processing systems 35: annual conference on neural information processing systems 2022: (neurIPS), vol. 35* (pp. 22199–22213).

Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. In *Proceedings of the first workshop on gender bias in natural language processing* (pp. 166–172). Florence, Italy: Association for Computational Linguistics.

Lahoti, P., Blumm, N., Ma, X., Kotikalapudi, R., et al. (2023). Improving diversity of demographic representation in large language models via collective-critiques and self-voting. In *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 10383–10405). Singapore: Association for Computational Linguistics.

Li, Y., Geurin, F., & Lin, C. (2023). Avoiding data contamination in language model evaluation: Dynamic test construction with latest materials. CoRR, abs/2312.12343.

Li, Y., Li, Z., Zhang, K., Dan, R., & Zhang, Y. (2023). ChatDoctor: A medical chat model fine-tuned on LLaMA model using medical domain knowledge. CoRR, abs/2303.14070.

Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., et al. (2023). Trustworthy LLMs: a survey and guideline for evaluating large language models' alignment. In *Socially responsible language modelling research*.

Miotto, M., Rossberg, N., & Kleinberg, B. (2022). Who is GPT-3? An exploration of personality, values and demographics. In *Proceedings of the fifth workshop on natural language processing and computational social science (nLP+cSS)* (pp. 218–227). Abu Dhabi, UAE: Association for Computational Linguistics.

Nadeem, M., Bethke, A., & Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)* (pp. 5356–5371). Online: Association for Computational Linguistics.

Nascimento, C. M. C., & Pimentel, A. S. (2023). Do large language models understand chemistry? A conversation with ChatGPT. *J. Chem. Inf. Model.*, *63*(6), 1649–1655.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., et al. (2022). Training language models to follow instructions with human feedback. In *NeurIPS*.

Pawar, S., Park, J., Jin, J., et al. (2024). Survey of cultural awareness in language models: Text and beyond.

Ramezani, A., & Xu, Y. (2023). Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 428–446). Toronto, Canada: Association for Computational Linguistics.

Ross, C., Katz, B., & Barbu, A. (2021). Measuring social biases in grounded vision and language embeddings. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 998–1008). Online: Association for Computational Linguistics.

Scherrer, N., Shi, C., Feder, A., & Blei, D. M. (2023). Evaluating the moral beliefs encoded in LLMs. In *Advances in neural information processing systems 36: annual conference on neural information processing systems 2023, (neurIPS)*.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. CoRR abs/1707.06347.

Sheng, E., Chang, K., Natarajan, P., & Peng, N. (2021). Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing: (ACL/IJCNLP) 2021, (volume 1: long papers)* (pp. 4275–4293). Virtual Event: Association for Computational Linguistics.

Smith, T. W., Marsden, P., Hout, M., & Kim, J. (2012). *General social surveys*.

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2023). Auditing and mitigating cultural bias in LLMs. CoRR abs/2311.14096.

Team, G., Zeng, A., Xu, B., Wang, B., et al. (2024). Chatglm: A family of large language models from GLM-130b to GLM-4 all tools. arXiv:2406.12793.

Tjuatja, L., Chen, V., Wu, S. T., Talwalkar, A., & Neubig, G. (2023). Do LLMs exhibit human-like response biases? A case study in survey design. CoRR abs/2311.04076.

Today USA (2023). *USA Today/Ipsos poll*. Roper Center for Public Opinion Research, Cornell University, Ithaca, NY, Version 2..

Wang, Y., Zhu, Y., Kong, C., Wei, S., et al. (2023). Cdeval: A benchmark for measuring the cultural dimensions of large language models. CoRR abs/2311.16421.

Welzel, C., Brunkert, L., Inglehart, R. F., & Kruse, S. (2019). Measurement equivalence? A tale of false obsessions and a cure. *World Values Research*, *11*(3), 54–84.

Welzel, C., & Inglehart, R. F. (2016). Misconceptions of measurement equivalence: Time for a paradigm shift. *Comparative Political Studies*, *49*(8), 1068–1094.

Welzel, C., Inglehart, R., & Kruse, S. (2017). Pitfalls in the study of democratization: Testing the emancipatory theory of democracy. *British Journal of Political Science*, *47*(2), 463–472.

Xu, G., Liu, J., Yan, M., Xu, H., et al. (2023). Cvalues: Measuring the values of Chinese large language models from safety to responsibility. CoRR abs/2307.09705.

Xu, R., Sun, Y., Ren, M., Guo, S., et al. (2024). AI for social science and social science of AI: A survey. *Information Processing & Management*, *61*(3), Article 103665.

Xu, C., Sun, Q., Zheng, K., Geng, X., et al. (2024). WizardLM: Empowering large pre-trained language models to follow complex instructions. In *The twelfth international conference on learning representations*.

Yang, A., Xiao, B., Wang, B., Zhang, B., et al. (2023). Baichuan 2: Open large-scale language models. CoRR abs/2309.10305.

Yao, J., Yi, X., Wang, X., Gong, Y., & Xie, X. (2023). Value FULCRA: mapping large language models to the multidimensional spectrum of basic human values. CoRR abs/2311.10766.