ValuesRAG: Enhancing Cultural Alignment Through Retrieval-Augmented Contextual Learning

Wonduk Seo[1,2], **Zonghao Yuan**[3], **Yi Bu**[2,4*]
[1]Enhans, Seoul, South Korea
[2]Peking University, Beijing, China
[3]Tsinghua University, Beijing, China
[4]Peking University Chongqing Research Institute of Big Data, Chongqing, China
[1]wonduk@enhans.ai, [2]yzh23@mails.tsinghua.edu.cn, [3]buyi@pku.edu.cn

Abstract

Ensuring cultural values alignment in Large Language Models (LLMs) remains a critical challenge, as these models often embed Western-centric biases from their training data, leading to misrepresentations and fairness concerns in cross-cultural applications. Existing approaches such as role assignment and few-shot learning struggle to address these limitations effectively due to their reliance on pre-trained knowledge, limited scalability, and inability to capture nuanced cultural values. To address these issues, we propose ***ValuesRAG***, a novel and effective framework that applies Retrieval-Augmented Generation (RAG) with In-Context Learning (ICL) to integrate cultural and demographic knowledge dynamically during text generation. Leveraging the World Values Survey (WVS) dataset, ValuesRAG first generates summaries of values for each individual. We subsequently curate several representative regional datasets to serve as test datasets and retrieve relevant summaries of values based on demographic features, followed by a reranking step to select top-k relevant summaries. We evaluate ValuesRAG using 6 diverse regional datasets and show that it consistently outperforms baselines, both in main experiments and ablation settings. Notably, ValuesRAG achieves the best overall performance over prior methods, demonstrating its effectiveness in fostering culturally aligned and inclusive AI systems. We further conduct two qualitative case studies to illustrate how ValuesRAG retrieves demographically aligned value profiles, enabling more context-sensitive reasoning without relying on static prompts or stereotypes. Our findings underscore the potential of retrieval-based methods to bridge the gap between global LLM capabilities and localized cultural values.

Introduction

The rapid advancement of Large Language Models (LLMs) has revealed pressing challenges in cultural values alignment (Singh et al. 2024; Kharchenko et al. 2024; Hu et al. 2024). Predominantly trained on Western data sources (Achiam et al. 2023; Touvron et al. 2023; Jiang et al. 2023), LLMs inherently reflect Western cultural norms and social biases, raising concerns about their applicability in global contexts. These biases present significant challenges when deploying LLMs in cross-cultural environments, often resulting in misrepresentations and stereotypical outputs (Gallegos et al. 2024; Xie et al. 2024; Potter et al. 2024; Huang, Papyshev, and Wong 2024). Despite ongoing efforts to address these issues, existing strategies often fall short. While some countries have developed localized LLMs, such as China's ERNIE (Sun et al. 2021), ChatGLM (GLM et al. 2024), DeepSeek (Liu et al. 2024a), and South Korea's HyperCLOVA (Yoo et al. 2024), these models also exhibit biases inherited from their respective training datasets. As a result, cultural and social biases embedded in LLMs remain a critical concern, compelling researchers to explore more robust frameworks for cultural alignment (Gallegos et al. 2024; Xie et al. 2024; Potter et al. 2024).

Recent studies have proposed several approaches, such as *role-assignment* approaches (Tao et al. 2024) and *few-shot learning* techniques (Choenni and Shutova 2024), to mitigate these cultural biases. However, these methods still face several challenges: (1) Role-assignment approaches, relying solely on the model's pre-trained knowledge, provide predefined demographic information but fail to incorporate explicit values alignment text, which subsequently introduces stereotypes and biases rooted in Western-centric training data; (2) While offering example-based guidance, few-shot learning methods struggle to comprehensively capture the complex cultural values due to the limited correlation between different values dimensions, thus remain ineffective on values-related tasks that differ significantly from the examples; (3) In addition, these methods can only align with the values of a single individual, and singular values cannot represent the universal values of individuals with similar characteristics.

To address these challenges, we propose ***ValuesRAG***, a novel framework that utilizes Retrieval-Augmented Generation (RAG) and In-Context Learning (ICL) to dynamically incorporate cultural knowledge during text generation. Our framework leverages the World Values Survey (WVS) dataset (Haerpfer et al. 2022) as a culturally rich retrieval corpus, drawing on its comprehensive coverage of global values grounded in rigorous social science methodologies. Specifically, we first generate summaries for each topic, followed by generating individuals' summaries of values and demographic profiles in parallel. After constructing the knowledge base, we retrieve the top-100 relevant summaries based on demographic features, followed by a reranking step

*Yi Bu is the corresponding author for this work.

to ensure the most relevant top-k summaries are selected. Finally, we utilize a reasoning LLM that filters the most relevant demographic profiles and applies reasoning grounded in the retrieved values to generate the final answer to the question.

We evaluate the performance of ValuesRAG by comparing it against several baseline approaches, including: (1) *zero-shot inference*, (2) *role-assignment-only method* (Tao et al. 2024), (3) *few-shot learning* (Choenni and Shutova 2024), and (4) *a hybrid method combining both (1) and (2)*. To ensure a comprehensive evaluation, we curated diverse regional survey QA datasets which are designed to capture values-related question-answer pairs. Extensive experimental results show significant improvements in cultural and contextual understanding, demonstrating that ValuesRAG outperforms the baselines. Compared to previous methods that heavily depend on pre-trained knowledge or limited demonstrations, ValuesRAG dynamically retrieves and integrates multiple similar individual values summaries based on demographic features, enabling richer value representations and more context-aware responses compared to approaches relying on a single predefined prompt or role.

In addition, ablation studies on varying the number of retrieved summaries and on using only value-augmented generation confirm ValuesRAG's robust performance under different configurations. Adjusting the number of retrieved documents shows that moderate retrievals can balance diversity and relevance, also maintaining high accuracy across multiple benchmarks. Meanwhile, ValuesRAG surpasses the baselines through purely values-based generation. We also conduct two qualitative case studies using two datasets (U.S. and China), and show that ValuesRAG effectively retrieves demographically aligned value profiles that support nuanced and culturally grounded reasoning, even without relying on static prompts or predefined roles.

These findings highlight ValuesRAG's potential to foster inclusive AI systems, enhancing the reliability and fairness of AI-driven applications. Our study demonstrates ValuesRAG's robust capabilities on a global scale, also suggesting its applicability in aligning the values of diverse groups within a single country. ValuesRAG provides a cost-efficient tool for public policymakers and scientists from various disciplines to refine social simulations, enabling more precise predictions of policy outcomes (Li, Gong, and Jiang 2024). This, in turn, facilitates the creation of fairer and more effective policies. Moreover, NGOs can leverage ValuesRAG to develop LLMs that reflect specific value orientations while maintaining strong alignment with users' values, thereby increasing their persuasive impact. This approach benefits the promotion and spread of values that contribute to the planet's sustainable development and the long-term well-being of human society.

Related Work

Evaluation of LLMs' Cultural Bias

Pre-trained models are facing growing criticism for their inherent social biases, with cultural bias emerging as a particularly nuanced and pervasive issue (Tao et al. 2024). While safety concerns and social discrimination in language models are typically explicit and well-recognized (Liu et al. 2024b), cultural biases often manifest subtly, reflecting dominant cultural perspectives embedded within training data. Studies have shown that LLMs often exhibit cultural biases aligned with the values of developed countries, resulting in the under-representation of perspectives from less developed regions (Manvi et al. 2024; Durmus et al. 2024). This imbalance not only perpetuates existing cultural hierarchies but also limits the global applicability of these models (Manvi et al. 2024). Various benchmarks and evaluation methods have been proposed to assess the cultural biases of pre-trained models (Gallegos et al. 2024). For instance, Caliskan et al. (2017) pioneered the use of word embeddings as quantitative measures of bias, while Webster et al. (2021) developed probability-based metrics to evaluate gender bias embedded in pre-trained models. More recently, Karinshak et al. (2024) introduced *LLM-GLOBE*, a benchmark where LLMs generate both quantitative and open-ended answers to values assessment questions, with subsequent evaluation using the LLM-as-a-Jury Protocol. These evaluation methods collectively highlight the complex nature of cultural bias in LLMs and the need for multifaceted assessment approaches.

Mitigation of LLMs' Cultural Bias

Techniques such as Reinforcement Learning from Human Feedback (RLHF) (Shen et al. 2023; Ji et al. 2024) are commonly employed for aligning LLMs with human values. However, these single-dimensional alignment methods are insufficient for mitigating cultural bias, as cultural values are inherently diverse, dynamic, and context-dependent, varying significantly across different regions and societies (Huang, Papyshev, and Wong 2024). Addressing cultural biases has become a critical area of research, with various strategies being proposed to enhance cultural sensitivity in LLMs. For instance, Tao et al. (2024) adopted national and cultural role assignments to adjust the cultural values of LLMs, while Masoud et al. (2024) developed a soft prompt tuning approach to mitigate bias. Moreover, Choenni and Shutova (2024) employed few-shot in-context learning to align cultural behaviors, demonstrating promising results in specific contexts. However, these approaches face significant limitations in fully capturing the complexity of cultural alignment. Tao et al.'s approach (Tao et al. 2024) mainly depends on national and cultural roles without explicitly integrating values assignments, causing an overreliance on latent internal representations. Meanwhile, Choenni and Shutova's few-shot learning approach (Choenni and Shutova 2024) similarly falls short of modeling cultural alignment in all its complexity. We therefore use these methods as baselines to benchmark our proposed approach.

Datasets

In this section, we first introduce the World Values Survey (WVS) as our retrieval corpus, highlighting its extensive coverage, global representativeness, and relevance for values-related studies. Subsequently, we describe six regional test datasets, which are carefully selected to ensure geographic, cultural, and demographic diversity.

Category	Dataset	Abbreviation	Region	Year	N	VQ
Retrieval Corpus	World Values Survey	WVS	Global	2017–2022	97.2k	259
Test Datasets	European Values Study	EVS	Europe	2017	59.4k	211
	The General Social Survey	GSS	North America	2021–2022	8.2k	44
	Chinese General Social Survey	CGSS	East Asia	2021	8.1k	58
	India Survey Dataset	ISD	South Asia	2019–2020	30.0k	33
	AmericasBarometer	LAPOP	Latin America	2021	59.1k	48
	Afrobarometer	Afrobarometer	Africa	2022	48.1k	144

Table 1: **Overview of the datasets utilized in our study.** The *Retrieval Corpus* (WVS) includes global data collected between 2017 and 2022, providing the basis for generating cultural summaries of values and validation for our method. The *Test Datasets* consist of six region-specific surveys, each capturing socio-cultural information from distinct geographic areas and time frames. N represents sample size in thousands (k). VQ represents the number of values-related questions.

Retrieval Corpus

WVS[1] (Haerpfer et al. 2022) is a globally recognized dataset that investigates human beliefs, values, and cultural norms through structured surveys conducted across multiple countries. WVS is selected as our retrieval corpus especially due to its numerous advantages:

1. ***Broad recognition and inclusiveness:*** WVS is widely recognized and frequently used by governments, social scientists, and major international organizations in comparative values studies. It currently covers 120 countries, representing 94.5% of the global population, ensuring broad geographic and cultural representation.
2. ***Expert-designed and accessible:*** The dataset is meticulously designed by leading domain experts to conduct comprehensive surveys of values, ensuring reliability, rigor, and relevance. It is publicly accessible, enabling reproducibility and transparency in research.
3. ***Effective structure and large scale:*** WVS has well-organized and comprehensive demographic questions, making it effective for retrieval tasks. Its large sample size (97,221 respondents) is also suitable for RAG tasks.

Since values evolve gradually over time, WVS is conducted in waves, with each wave occurring every five years. For our study, we utilize the most recent wave, spanning from 2017 to 2022.

The WVS codebook includes over 600 indicators, with 259 values-related and 31 demographic-related questions. The value questions span 13 topics, such as social trust, post-materialism, and political interest, as shown in Table 2. It covers most dimensions of values, allowing for a comprehensive and accurate measure of each respondent's values. We randomly select 20% (52 questions) per topic for validation and use the remaining 80% (207 questions) for summary generation. The 31 demographic features, including country, sex, age, education, social class, and employment status, are used to generate demographic summaries for retrieval tasks.

[1]https://www.worldvaluessurvey.org/wvs.jsp

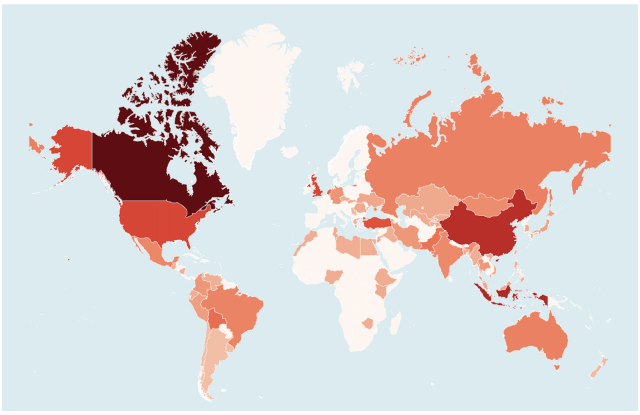

Figure 1: **Distribution of countries in the WVS dataset.** The WVS dataset spans surveys in over 120 countries across major regions, ensuring broad geographic and demographic coverage for building a reliable retrieval corpus in RAG-based frameworks. The color gradient shows response density, from light beige (under 500) to dark red (above 4,000).

Topic	Count
Social Values, Norms, Stereotypes	45
Happiness and Wellbeing	11
Social Capital, Trust and Organizational Membership	47
Economic Values	6
Perceptions of Corruption	9
Perceptions of Migration	10
Perceptions of Security	21
Index of Postmaterialism	6
Perceptions about Science and Technology	6
Religious Values	12
Ethical Values	23
Political Interest and Political Participation	35
Political Culture and Political Regimes	25

Table 2: **Distribution of Values-related Questions in WVS.** The questions were categorized into 13 topics with a total of 259 questions covering most of the dimensions of values.

Test Datasets

We select six regional surveys to serve as test datasets based on the following criteria:

1. ***Demographic and values coverage:*** The datasets provide demographic features closely aligned with WVS's questions, along with sufficient values-related features to enable meaningful comparisons and analyses.
2. ***Temporal proximity:*** The datasets exhibit close temporal proximity to WVS Wave 7 (2017–2022), thereby allowing aligned comparisons and ensuring thorough consistency across diverse global evaluations.

The regions in our test datasets are meticulously chosen to encompass a wide range of geographic, cultural, and demographic diversity, ensuring that the data accurately reflects the majority of the global population. All of them are publicly accessible and are statistically representative at national or regional levels, which guarantees their reliability and validity. Also, these test sets include both values-related questions and demographic characteristics. The demographic characteristics are used to generate summaries, serving as retrieval targets for RAG. The values-related questions are utilized as test questions to calculate accuracy (ACC).[2] The specific datasets used in our evaluation, along with their characteristics, are described below:

EVS The first dataset comes from the *European Values Study*[3], a large-scale, cross-national, and longitudinal survey research program designed to explore values, beliefs, and attitudes across Europe. This dataset includes a total of 211 values-related questions and captures 34 demographic characteristics of the respondents. We select EVS 2017, conducted in 2017, ensuring alignment with the World Values Survey (WVS) in terms of the timeframe.

GSS We select GSS to represent the population of the United States. *The General Social Survey*[4] is a sociological survey that has been conducted since 1972 by the National Opinion Research Center (NORC) at the University of Chicago. Its primary purpose is to collect and analyze data on the opinions, behaviors, and demographic characteristics of adults in the United States, thereby monitoring societal change and the growing complexity of American society. Its questionnaire covers a comprehensive and wide range of topics, including many values-related questions. Specifically, within the GSS, we identify 44 questions as values-related and 33 questions as demographic characteristics.

CGSS *The Chinese General Social Survey*[5], initiated in 2003, is China's earliest national, comprehensive, and continuous academic survey project. Conducted by the National Survey Research Center at Renmin University of China, the CGSS systematically collects data at multiple levels, including society, communities, families, and individuals. CGSS only provided the questionnaire and data in Chinese, which we have translated into English to ensure its usability. We ultimately compile a total of 58 values-related questions and 13 demographic characteristics.

ISD To ensure that our experiment covers as much of the world's population as possible, we made efforts to include India within the scope of our test set. However, we were unable to obtain data from several government surveys in India, thus we used data published by the Pew Research Center instead. The Pew Research Center's *India Survey Dataset*[6] is a comprehensive resource that captures the perspectives of 29,999 Indian adults on various aspects of society, including religious beliefs and practices, identity, nationalism, and societal tolerance. Conducted through face-to-face interviews between November 17, 2019, and March 23, 2020, the survey encompassed participants from diverse religious backgrounds, such as Hindus, Muslims, Sikhs, Christians, Buddhists, Jains, and others. This dataset covers 33 values-related questions and 23 demographic characteristics.

LAPOP The Latin American Public Opinion Project (LAPOP)[7] is a research institute based at Vanderbilt University in Nashville, Tennessee. LAPOP's most notable survey is the *AmericasBarometer*, the most extensive survey of democratic public opinion and behavior covering the Americas, including North, Central, South America, and the Caribbean. This survey measures democratic values and behaviors through voter surveys, providing valuable insights into public sentiments across the region. We select this dataset to represent the population of Latin America. There are 48 values-related questions and 12 demographic characteristics.

Africa *Afrobarometer*[8] is a pan-African, non-partisan research network established in 1999 that conducts public attitude surveys on democracy, governance, economic conditions, and related issues across Africa. We selected data from the 8th round of Afrobarometer (collected in 2022), covering 34 African countries. After screening, a total of 144 values-related questions and 14 demographic characteristics are obtained.

Methodology

In this section, we present the *ValuesRAG* which is specifically designed to address cultural biases and enhance contextual alignment in LLM-driven scenarios through a Retrieval-Augmented Generation (RAG) approach. *ValuesRAG* consists of three key components: (1) *Values and Demographic Summary Generation*, which extracts and summarizes cultural values and demographic information from large-scale datasets; (2) *Values-Augmented Generation*, which incorporates these summaries into the generative process to align responses with the cultural context; and (3) *Retrieval-based Values Alignment*, which dynamically assigns relevant individual values to queries based on demographic profiles. An overview of the *ValuesRAG* framework is provided in Figure 2.

[2] A detailed description of the method for computing accuracy (ACC) is provided in *Section Experiments - Evaluation Metrics*.

[3] https://europeanvaluesstudy.eu

[4] https://gss.norc.org

[5] http://cgss.ruc.edu.cn

[6] https://www.pewresearch.org/dataset/india-survey-dataset

[7] https://www.vanderbilt.edu/lapop

[8] https://www.afrobarometer.org

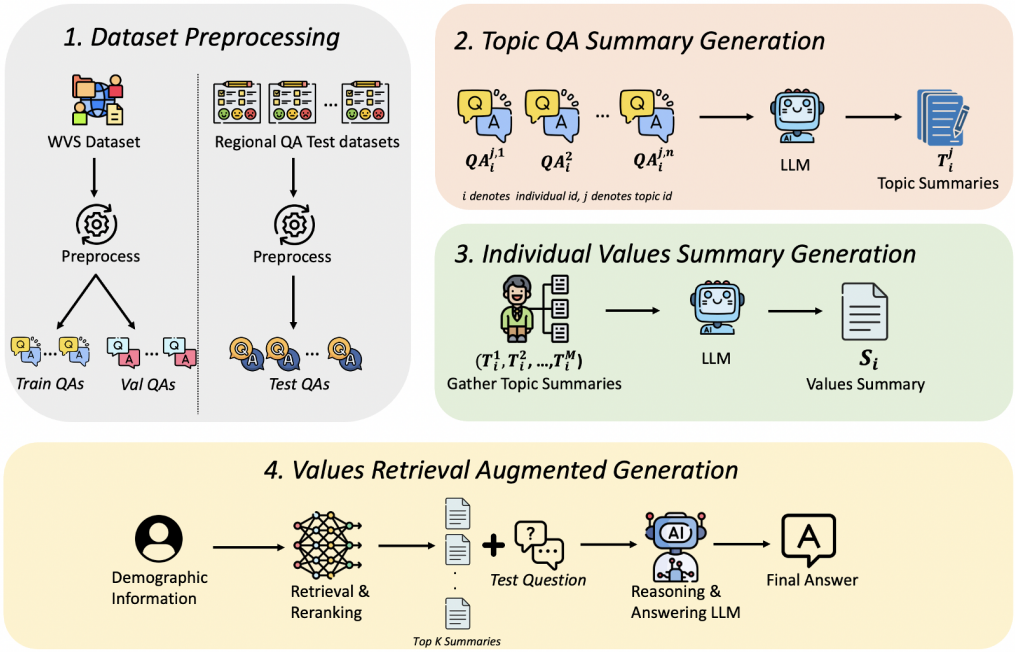

Figure 2: **Overview of the proposed ValuesRAG framework for cultural alignment.** The framework comprises four key stages: (1) Dataset preprocessing to separate training and test QA pairs from WVS and regional datasets, (2) Topic-wise summary generation using LLMs for each individual, (3) Aggregation of topic summaries into comprehensive individual values profiles, and (4) Retrieval-Augmented Generation that retrieves and reranks relevant value summaries based on demographic similarity to guide final response generation.

Values and Demographic Summary Generation

To systematically generate concise summaries of values and demographics for each individual, we process the dataset in three stages. First, the dataset is stratified by topics and split into train and validation sets, ensuring that the distribution of each topic is preserved, as described in previous section. In parallel, topic-based summaries and demographic summaries are generated separately. For topic-based summaries, values-related QA sets are used to produce summaries for each topic, while demographic summaries are generated using demographic-related QA sets:

$$\begin{aligned} T_i^j &= f_{\text{gen}}(\text{QA}_i^{j,1}, \text{QA}_i^{j,2}, \ldots, \text{QA}_i^{j,N_j}), \\ D_i &= f_{\text{gen}}(\text{QA}_i^{\text{demo},1}, \text{QA}_i^{\text{demo},2}, \ldots, \text{QA}_i^{\text{demo},K}). \end{aligned} \tag{1}$$

where f_{gen} denotes the generative model, T_i^j is the summary for topic j of individual i, based on N_j values-related QA pairs, and D_i represents the demographic summary derived from K demographic-related QA pairs. Finally, individual summaries are constructed by combining all topic summaries:

$$S_i = f_{\text{gen}}(T_i^1, T_i^2, \ldots, T_i^M), \tag{2}$$

with M denoting the total number of topics. The result, denoted as S_i, forms a comprehensive values summary for individual i. These generate summaries serve as structured references for retrieval in later stages and are used to augment the validation set for evaluation as well.

Values Augmented Generation

Once the comprehensive summaries for each individual are generated in the previous step, we construct an augmented generation process for evaluating on the validation question-answer data. For each validation question, we concatenate the corresponding individual's values summary with the question itself, forming a context-rich input for the LLM:

$$C_i = \text{concatenate}\left(S_i, \text{Q}_i^{\text{val},k}\right) \tag{3}$$

where C_i represents the combined context, S_i is the values summary for individual i, and $\text{Q}_i^{\text{val},k}$ is the k-th validation question. We subsequently concatenate C_i with the demographic summary D_i to further enhance the context, enabling the generation of responses based on both values and demographic information:

$$A_i = f_{\text{gen}}\left(C_i, D_i\right) \tag{4}$$

Here, A_i represents the answer generated by the function f, and (C_i, D_i) embeds the augmented context C_i and demographic information D_i into a structured input format. Additionally, we utilize Chain-of-Thought (CoT) prompting (Wei et al. 2022) to enhance reasoning and emulate the behavior of the corresponding individual, ensuring responses that are contextually aligned with the values captured in the summaries and demographic characteristics.

Retrieval-based Values Alignment

To dynamically assign relevant values to test individuals, we leverage demographic information as documents for retrieval. The demographic data from both the train (retrieval corpus) and test datasets are preprocessed into a structured context format, as described earlier, and embeddings are generated for each demographic context using a sentence-transformer-based model (Reimers and Gurevych 2019).[9] We first retrieve the top-100 most similar summaries of values for each test individual by computing the cosine similarity between the embeddings of the test and train demographics:

$$\text{Sim}(E_{\text{test}}, E_{\text{train}}) = \frac{E_{\text{test}} \cdot E_{\text{train}}}{\|E_{\text{test}}\| \|E_{\text{train}}\|} \tag{5}$$

E_{test} and E_{train} specifically represent the embeddings of the test and training (retrieval corpus) demographic contexts, respectively, and $\text{Sim}(\cdot, \cdot)$ denotes the cosine similarity score. The top-100 embeddings with the highest similarity scores are initially selected as candidates. We subsequently apply a reranking step to refine the selection and identify the most relevant summaries among the retrieved candidates.

The reranking process evaluates the semantic relevance by passing each candidate summary embedding E_{C_j} along with the test individual's embedding E_{test} through a neural reranker f_{rerank}, which outputs a relevance score s_j:

$$s_j = f_{\text{rerank}}\left(E_{\text{test}}, E_{C_j}\right), \quad j \in \{1, 2, \ldots, 100\} \tag{6}$$

We then sort the candidate summaries based on these scores in descending order and select the top-k summaries with the highest scores as the final reranked set R'_k:

$$R'_k = \text{Top-}k\left(\{C_j\}_{j=1}^{100}, \{s_j\}_{j=1}^{100}\right) \tag{7}$$

The reranked top-k summaries R'_k are incorporated into the prompts, enriching the contextual alignment of the generated responses. In detail, for each test individual, the retrieved and reranked summaries are combined into the final prompt, and the answer is subsequently generated using the function f_{gen}:

$$\begin{aligned} P_{\text{test}} &= \left(D_{\text{test}}, R'_1, R'_2, \ldots, R'_K, Q_{\text{test}}\right), \\ A_{\text{test}} &= f_{\text{gen}}\left(P_{\text{test}}\right). \end{aligned} \tag{8}$$

Here, P_{test} is the final prompt, D_{test} is the demographic information of the test individual, $\{R'_1, R'_2, \ldots, R'_K\}$ represents the top-k reranked summaries, and Q_{test} is the test question. A_{test} denotes the generated answer for the test question, and f_{gen} represents the generation function.

This retrieval-based approach, followed by reranking, enhances reasoning by explicitly guiding the LLM to critically evaluate which retrieved values best align with the test individual's demographic characteristics. The final prompts are then used to generate answers following the chain-of-thought prompting strategy, ensuring that the responses are contextually coherent and culturally aligned with the test individual's profile. For the comprehensive implementation of the *ValuesRAG*, we provide the **Algorithm 1**, which systematically outlines the processes of values and demographic summary generation, values-augmented generation, and retrieval-based values alignment, as shown below:

Algorithm 1: *Values Generation and Retrieval Process*

Require: Dataset $\mathcal{D}$ with topics and demographic QA pairs, Generative Model f_{gen}, Embedding Model f_{embed}, Reranking Model f_{rerank}, Retrieval Top-K

// Values and Demographic Summary Generation

for each individual i in $\mathcal{D}$ **do**
  Generate topic-based values summaries T_i^j for each topic j
  Generate demographic summaries D_i
  Combine T_i^j into comprehensive values summary S_i
end for

// Values Augmented Generation

for each validation question $Q_i^{\text{val},k}$ of individual i **do**
  Construct context $C_i = \text{concat}(S_i, Q_i^{\text{val},k})$
  Augment context with demographic summary D_i
  Generate answer $A_i = f_{\text{gen}}(C_i, D_i)$
end for

// Retrieval-based Values Alignment

for each test individual i **do**
  Compute embeddings $E_{\text{test}} = f_{\text{embed}}(D_{\text{test}})$
  Retrieve top-100 values summaries by similarity: $\text{Sim}(E_{\text{test}}, E_{\text{train}})$
  Rerank top-K summaries: $R'_k = f_{\text{rerank}}(E_{\text{test}}, E_{C_j})$
  Final prompt $P_{\text{test}} = (D_{\text{test}}, R'_1, \ldots, R'_K, Q_{\text{test}})$
  Generate answer $A_{\text{test}} = f_{\text{gen}}(P_{\text{test}})$
end for

Experiments

Setup

Models Used. We utilize *GPT-4o-mini* (Achiam et al. 2023) and *Gemini-1.5-Flash* (Team et al. 2024) for our generation tasks, which are accessed via APIs. We set the temperature parameter of these models to 0.7 to achieve a balance between coherence and creativity. For the retrieval task, we employed the *E5 (base)* model (Wang et al. 2022), which generates embeddings and retrieves the top 100 most relevant summaries of values based on cosine similarity.

To refine this retrieval, we apply a reranker using the *GTE-multilingual-reranker-base* model (Zhang et al. 2024). Each retrieved candidate is paired with the test individual's demographic summary and passed as a sentence pair to the reranker. The model outputs a relevance score for each pair, which we use to rank all candidates. The top-k summaries with the highest scores are then selected as the final set used for generation. This two-stage process: retrieval followed by reranking, ensures both semantic similarity and contextual relevance.

Prompts Used We provide the prompts designed for various components of the ValuesRAG, including prompts for

[9] A detailed description of the model used is provided in *Section Experiments - Models Used*.

performing question answering tasks, as well as for generating values and demographic summaries below:

Prompt for Question Answering

Task: Respond to the question as the target individual, selecting the answer that aligns with their values and demographic context.
Rules:

- Step-by-step analysis using retrieved demographics and values data.
- Maintain the target individual's perspective throughout the analysis.
- Provide the response in JSON format, without additional explanation.

Steps for Inferring:

1. Analyze the demographics (age, gender, cultural background, social class, religion, and economic class) of retrieved individuals. Compare them with the target individual.
2. Identify individuals whose demographics most closely match the target individual. Note their IDs.
3. Based on the matched individual's values, infer how the target would respond.
4. Select the response that best aligns with the inferred values, and return only the integer representing the selected option.

Prompt for Values Summary Generation

You are a summarization expert with expertise in extracting key insights from complex data. Based on the provided context, summarize this person's values in one paragraph.

Prompt for Demographic Summary Generation

You are a summarization expert with expertise in extracting key insights from complex data. Based on the provided context, summarize this person's demographics in one paragraph.

Baseline Methods and Implementation. Our baseline methods include: (1) *Zero-shot inference*, (2) *the role-assignment-only approach* (Tao et al. 2024), (3) *a few-shot learning method* (Choenni and Shutova 2024), and (4) *a hybrid method that combines both (1) and (2)*. For the role-assignment baseline, we specifically use the same demographic summaries as in ValuesRAG to ensure fairness by assigning roles based on demographic information from the survey data. For the few-shot method, we follow the approach outlined in the previous work, where we randomly select five examples from the test set as prompts. The hybrid method combines both strategies, assigning roles based on demographic summaries and augmenting the prompts with five randomly selected few-shot examples from the test set. Additionally, We use *ValuesRAG* with $k = 3$ retrieved summaries—chosen to provide a good balance between retrieval diversity and contextual relevance.[10]

Evaluation Metrics. We utilize accuracy as the primary evaluation metric, following previous work (Choenni and Shutova 2024), by converting multiple-choice responses into a binary format for consistency and simplicity. Specifically, we transform each original response r_i, ranging on a scale (e.g., Likert scale from 1 to 10), into a binary value b_i indicating disagreement or agreement as follows. Here, m denotes the midpoint of the response scale:

$$b_i = \begin{cases} 0, & \text{if } r_i \leq m \quad \text{(disagree)} \\ 1, & \text{if } r_i > m \quad \text{(agree)} \end{cases} \tag{9}$$

We adopt the midpoint value m for binarization since each question may use a different scoring range. This categorization metrics effectively captures distinct answer patterns and aligns naturally with values-related questions often posed through Likert-scale formats.

Experimental Analysis

In our experiments, we compare *ValuesRAG* with four baseline methods: (1) zero-shot inference, (2) role-assignment, (3) few-shot learning, and (4) a hybrid approach that combines role-assignment and few-shot learning. Following previous work, as detailed in the Evaluation Metrics section, we specifically evaluate model performance using accuracy. All responses are binarized based on contrasting answer patterns to ensure consistency across different values-related questions. Experimental results are shown in Table 3.

We find that the role-assignment method generally surpasses both zero-shot and few-shot approaches. By grounding the agent's responses in a clearly defined demographic context, it ensures more consistent performance. Yet, role assignment can sometimes lead to overly narrow representations when demograhpic roles are interpreted stereotypically. Meanwhile, few-shot learning can incorporate example-driven context, but its limited number of prompts may not consistently address the intricate ways individuals' belifs diverge within similar social settings. As a result, it struggles to generalize to the multifaceted nature of human values, particularly when faced with unexpected or complex cultural scenarios. The hybrid method, which merges role assignment and few-shot prompts, does offer a partial improvement in contextual diversity. However, it still remains insufficient for capturing the full spectrum of nuances that can arise from overlapping demographic factors and idiosyncratic personal persepectives.

In contrast, ValuesRAG overcomes these challenges by dynamically retrieving and integrating specific values-related cultural data for each agent. By focusing on values as the primary retrieval targets, this retrieval-augmented

[10] A detailed analysis of varying k and its implications on model performance is provided in *Ablation Study - Varying the Number of Retrieved Summaries*.

Model	Methods	EVS	GSS	CGSS	ISD	LAPOP	Africa	Avg. Accuracy
GPT-4o mini	Zero-shot Inference	0.5566	0.6026	0.4019	0.6109	0.4195	0.3923	0.4973
	Role-Assignment (2024)	0.5738	0.7564	0.4813	0.6164	0.4742	0.5563	0.5764
	Few-Shot Learning (2024)	0.5271	0.6538	0.4631	0.5804	0.4220	0.4258	0.5120
	Hybrid Method	0.5938	0.7292	0.5048	0.6330	0.4414	0.5305	0.5721
	ValuesRAG†	**0.6021***	**0.7781***	**0.5387***	**0.7001***	**0.5030***	**0.5953***	**0.6195***
Gemini 1.5 Flash	Zero-shot Inference	0.5419	0.6408	0.4502	0.6017	0.4149	0.4181	0.5113
	Role-Assignment (2024)	0.5598	0.7493	0.4770	0.6048	**0.4747**	0.5262	0.5653
	Few-Shot Learning (2024)	0.5225	0.6376	0.4559	0.5782	0.4194	0.4758	0.5149
	Hybrid Method	0.5845	0.7193	0.5026	0.6253	0.4448	0.5166	0.5655
	ValuesRAG†	**0.5869**	**0.7686***	**0.5337***	**0.6789***	0.4705	**0.5473***	**0.5977***

Table 3: **Accuracy scores for various methods compared with multiple baselines across six regional datasets.** k indicates the number of summaries to be retrieved. **Bold text** indicates the best performance, underlined text indicates the second-best performance. * denotes significant improvements (paired t-test with Holm-Bonferroni correction, $p < 0.05$) over all baseline model(s). † denotes our proposed method.

Model	Num(K)	EVS	GSS	CGSS	ISD	LAPOP	Africa	Avg. Accuracy
GPT-4o mini	1	0.5960	0.7722	0.5347	0.6853	0.4682	0.5905	0.6078
	3	0.6021	**0.7781**	**0.5387**	0.7001	0.5030	**0.5953**	**0.6195**
	5	**0.6052**	0.7706	0.5301	**0.7016**	**0.5061**	0.5905	0.6174
	10	0.6020	0.7380	0.5317	0.7014	0.5030	0.5680	0.6074
Gemini 1.5 Flash	1	0.5753	0.7668	0.5272	0.6646	0.4548	0.5369	0.5876
	3	**0.5869**	0.7686	**0.5337**	**0.6789**	**0.4705**	0.5473	**0.5977**
	5	0.5868	**0.7690**	0.5303	0.6734	0.4661	**0.5498**	0.5959
	10	0.5852	0.7665	0.5279	0.6773	0.4509	0.5464	0.5924

Table 4: **Accuracy scores across six regional datasets for the ablation study:** ***Varying the Number of Retrieved Summaries***. Num(K) ($k \in \{1, 3, 5, 10\}$) indicates the number of demographic summaries retrieved. **Bold text** indicates the best performance, underlined text indicates the second-best performance.

framework enables the model to include an expansive set of contextual clues, helping it reflect the depth and breadth of each individual's background and values. Crucially, our ValuesRAG provides a more adaptive mechanism for representing the subtle interplay of personal beliefs and cultural norms via avoiding the limitations of rigid demographic labels or small-sample prompts. ValuesRAG more effectively captures the complex dynamics that can shape a respondent's stance on different questions. Evaluations across diverse test datasets demonstrate that ValuesRAG with $k = 3$ consistently outperforms baseline methods, highlighting its ability to better represent cultural diversity, improve contextual alignment, and enhance overall model performance.

Ablation Study

We conduct two ablation studies to analyze the configuration and robustness of ValuesRAG. First, we vary the number of retrieved summaries (k) to examine how retrieval depth affects the model's performance. We subsequently isolate the effect of using only values-based generation.

Varying the Number of Retrieved Summaries To quantify how the number of retrieved summaries k impacts *ValuesRAG*'s performance, we evaluate $k \in \{1, 3, 5, 10\}$. Table 4 reports accuracy on six regional datasets for both GPT-4o-mini and Gemini 1.5 Flash.

For GPT-4o-mini, increasing k from 1 to 3 boosts accuracy on five of six datasets, rising from an average of 0.6078 to 0.6195, and yields the best scores on GSS (0.7781), CGSS (0.5387), and Africa (0.5953). Although $k = 5$ further improves EVS (0.6052) and ISD (0.7016), it degrades GSS and CGSS, dropping the average to 0.6174. At $k = 10$, performance falls across most datasets (average 0.6074), indicating that an excessive number of summaries dilute relevance. Similarly, for Gemini 1.5 Flash, $k = 3$ attains the highest overall accuracy (0.5977), with top scores on EVS (0.5869), CGSS (0.5337), ISD (0.6789), and LAPOP (0.4705). While $k = 5$ slightly edges out $k = 3$ on GSS (0.7690) and Africa (0.5498), it lowers EVS and ISD and yields a lower average (0.5959). Retrieval at $k = 10$ further declines.

These results reveal a clear trade-off: $k = 1$ constrains diversity, whereas $k > 3$ introduces marginally relevant or noisy summaries and increases latency. Consequently, $k = 3$ achieves the best balance between contextual breadth, accuracy gains, and computational cost. We thus adopt $k = 3$ as our default retrieval depth across all experiments.

Impact of Values-Only Generation To validate the robustness of ValuesRAG, we perform an ablation study using only values context augmented generation, thereby exclud-

ing the impact of demographic summaries on the model's performance. We use the WVS validation set: separated from the training data, which served as the retrieval corpus (as outlined in Section: *Datasets*), to evaluate the models. Table 5 presents a comparison of our method, using *only summaries of values*, against four baseline methods. Notably, ValuesRAG consistently outperforms all baselines across this validation data, achieving the highest accuracy despite relying exclusively on the values summaries.

Methods	GPT-4o-mini	Gemini-1.5-flash
Zero-Shot	0.6176	0.6041
Role-Assignment	<u>0.6747</u>	<u>0.6505</u>
Few-Shot Learning	0.6359	0.6086
Hybrid Method	0.6670	0.6354
Values Augmented	**0.6894**	**0.6583**

Table 5: Accuracy comparison between baseline methods and Values Augmented Generation method using the WVS validation set. **Bold text** indicates the best performance, <u>underlined text</u> indicates the second-best performance.

Compared to zero-shot inference (0.6176 for GPT-4o-mini; 0.6041 for Gemini), Values-Only achieves 0.6894 with GPT-4o-mini and 0.6583 with Gemini, demonstrating structured value context provides substantially richer guidance than unconstrained prompts. Against the role-assignment (0.6747 for GPT-4o-mini; 0.6505 for Gemini), Values-Only attains higher scores as well, confirming that our generated value summaries capture more nuanced cultural patterns than static demographic labels. Similarly, Values-Only outperforms few-shot learning (0.6359 for GPT-4o-mini; 0.6086 for Gemini) and the hybrid method (0.6670 for GPT-4o-mini; 0.6354 for Gemini), further illustrating even without explicit examples or demographic anchors, our values-driven prompts produce more context-aligned predictions.

These results confirm the effectiveness and robustness of the values-augmented generation approach. ValuesRAG leverages structured values summaries to generate contextually rich and culturally aligned responses. Even without demographic augmentation, ValuesRAG achieves superior performance by dynamically capturing the underlying value patterns, demonstrating its ability to generalize across diverse cultural contexts without requiring predefined QA examples or demographic anchors. The results demonstrate the framework's scalability and adaptability, effectively mitigating biases and generating culturally coherent outputs with minimal dependence on external context.

Case Study

To further illustrate the effectiveness and practical utility of *ValuesRAG*, we present two qualitative case studies derived from the GSS (United States) and CGSS (China) datasets. These case studies clearly demonstrate how our method dynamically forms a culturally and demographically relevant retrieval corpus, enhances values alignment, and consistently outperforms baseline approaches by providing richer, context-sensitive reasoning for the generation process.

Case 1: Cultural Attitudes Toward Premarital Sex (United States)

Question: Is premarital sex always wrong, almost always wrong, sometimes wrong, or not wrong at all?
Original Demographics: 63-year-old White male, U.S., English-speaking, high school graduate, married, no children, household income $90k–109k, middle class, Protestant, United Presbyterian Church, stable finances.
Retrieved Text 1: *Demographics:* 60-year-old White non-Hispanic male, U.S.-born, bachelor's degree, employed full-time in private sector, divorced, two children, Protestant, high household income. *Values:* Family, empowerment, financial responsibility, social equity, progressive politics, civil rights, environmental sustainability, faith amid diverse beliefs.
Retrieved Text 2: *Demographics:* 63-year-old White non-Hispanic male, married, two children, high school diploma, employed full-time in semi-skilled private sector job, Protestant, middle-income, financially stable. *Values:* Family, work as life's center, moderation in politics, tolerance, dedication, balance between fairness, economic security, civic responsibility, and pragmatic governance.
Retrieved Text 3: *Demographics:* 65-year-old White male, U.S.-born, English-speaking. (Truncated in dataset.)

Table 6: **Case Study (United States): Cultural Attitudes Toward Premarital Sex.** Retrieved summaries provide demographically similar but value-diverse profiles, supporting nuanced reasoning.

The first case (Table 6) involves a 63-year-old middle-class Protestant male from the U.S. *ValuesRAG* constructs the retrieval corpus by embedding demographic information and retrieving topically relevant profiles with similar backgrounds (e.g., age, ethnicity, and religious affiliation). Specifically, the top three retrieved profiles closely align demographically with the original individual yet introduce nuanced ideological variations. Each retrieved document not only reflects the target individual's demographic features but also articulates distinctive value orientations, highlighting the contextual breadth and depth of the retrieved corpus.

ValuesRAG enables more nuanced cultural reasoning by retrieving values profiles that are demographically aligned yet diverse in perspectives, allowing the LLM to consider various contextually valid stances rather than reinforcing stereotypes typically found in static demographic-based methods. Compared to role-assignment or few-shot approaches, ValuesRAG dynamically integrates subtle but critical distinctions across retrieved summaries. As a result, the generated response reflects realistic and culturally plausible reasoning that respects both traditional values and evolving societal norms.

Case 2: Family Planning Preferences (China)

The second case (Table 7) examines a 27-year-old secular male with postgraduate education from urban China ad-

Question: How many children to have is a personal matter, and the government should not interfere. Do you agree?
Original Demographics: 27-year-old Han male, China, postgraduate education, unemployed, middle class, no religion, no income in past year.
Retrieved Text 1: *Demographics:* 28-year-old Chinese male, bachelor's degree, clerical private-sector worker, married, no children, parents in household, middle income, no religion. *Values:* Personal autonomy, equality, skepticism toward institutions, environmental responsibility, scientific advancement, democratic participation, balance between family loyalty and social order.
Retrieved Text 2: *Demographics:* 29-year-old Chinese male, bachelor's degree, employed in sales sector, single, no children, working-class background, parents with lower education, middle income, no religion. *Values:* Importance of family, fairness, tolerance, dedication, pragmatism, economic security, cautious governance, opposition to corruption, recognition of complex gender and social dynamics.
Retrieved Text 3: *Demographics:* 25-year-old Chinese male, Standard Chinese speaker. (Truncated in dataset.)

Table 7: **Case Study (China): Family Planning Preferences.** Retrieved profiles highlight social shifts toward autonomy and equality while reflecting demographic similarities.

dressing the question. In this scenario, *ValuesRAG* retrieves profiles of individuals with closely matching demographics, such as age, education, urban residence, etc. Each retrieved summary reflects values related to personal agency, social responsibility, and family planning, with slight variations in economic status or professional roles. The constructed retrieval corpus is well-aligned with the original profile and relevant to the thematic context of the prompt.

By incorporating these diverse yet demographically coherent value profiles, *ValuesRAG* enables the LLM to generate responses that account for a variety of socially grounded perspectives on family decision-making. In contrast to static prompting methods that may overlook such intra-cultural nuances, ValuesRAG captures shifts in value expression across population segments in a context-aware manner.

Conclusion

We propose *ValuesRAG*, a novel framework designed to advance cultural values alignment through context-aware reasoning. In contrast to prior methods that rely on fixed demographic labels or limited few-shot prompting, *ValuesRAG* dynamically retrieves and integrates contextually rich value summaries using adaptive retrieval, reranking, and In-Context Learning (ICL). Our contributions extend beyond the framework design to include a comprehensive evaluation across diverse, culturally and geographically representative test datasets. Extensive experiments demonstrate that *ValuesRAG* consistently outperforms existing approaches, effectively capturing complex cultural nuances, reducing biases, and generating contextually aligned responses.

Limitations

While baseline comparisons show that ValuesRAG generally outperforms other methods, it does not always precisely capture an individual's true values. Because our method uses dynamically retrieved summaries from a fixed corpus (WVS), mismatches can occur when applying these summaries to unfamiliar datasets. Relying solely on WVS may also reinforce existing biases and limit representativeness. Future work should develop more adaptive retrieval strategies to improve contextual generalizability. Additionally, converting the Likert scale to binary for evaluation may obscure nuanced value differences. Future research could explore metrics that better capture the full range of responses.

Ethical Considerations Statement

While ValuesRAG effectively mitigates the common stereotype reinforcement seen in static methods, its reliance on demographic features inherently carries potential ethical risks related to profiling, bias, and fairness. If deployed in sensitive or high-stakes contexts such as public policy or persuasive communication, improper handling of demographic information could unintentionally perpetuate stereotypes or biases, adversely affecting vulnerable groups. We emphasize that our framework is intended not to reinforce specific demographic-driven stereotypes or biases, but rather to surface contextually relevant values explicitly for scrutiny. Moreover, the use of general-purpose models such as GPT-4o, which are known to reflect societal biases, requires careful consideration. Future implementations should consider using language models that have undergone explicit bias mitigation to minimize potential harms in generating values.

Researchers and practitioners utilizing ValuesRAG must proactively examine and address ethical implications surrounding demographic-based retrieval. It is critical to integrate ValuesRAG into broader evaluative frameworks designed for continuous monitoring of potential societal impacts, ensuring its responsible deployment. Additionally, the use of demographic similarity as a retrieval mechanism can risk reinforcing normative group boundaries; a more nuanced treatment that accounts for intra-group diversity is essential to avoid essentializing cultural identities. Future work should explicitly investigate the ethical dimensions of demographic profiling within retrieval-augmented methods, further promoting fairness, accountability, and transparency in AI-driven cultural alignment tasks.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics Derived Automatically from Language Corpora Contain Human-like Biases. *Science*, 356(6334): 183–186.

Choenni, R.; and Shutova, E. 2024. Self-Alignment: Improving Alignment of Cultural Values in LLMs via In-Context Learning. arXiv:2408.16482.

Durmus, E.; Nguyen, K.; Liao, T. I.; Schiefer, N.; Askell, A.; Bakhtin, A.; Chen, C.; Hatfield-Dodds, Z.; Hernandez, D.; Joseph, N.; Lovitt, L.; McCandlish, S.; Sikder, O.; Tamkin, A.; Thamkul, J.; Kaplan, J.; Clark, J.; and Ganguli, D. 2024. Towards Measuring the Representation of Subjective Global Opinions in Language Models. arXiv:2306.16388.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3): 1097–1179.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Haerpfer, C.; Inglehart, R.; Moreno, A.; Welzel, C.; Kizilova, K.; Diez-Medrano, J.; Lagos, M.; Norris, P.; Ponarin, E.; and Puranen, B. 2022. World values survey wave 7 (2017-2022) cross-national data-set. *(No Title)*.

Hu, T.; Kyrychenko, Y.; Rathje, S.; Collier, N.; van der Linden, S.; and Roozenbeek, J. 2024. Generative language models exhibit social identity biases. *Nature Computational Science*, 1–11.

Huang, L. T.-L.; Papyshev, G.; and Wong, J. K. 2024. Democratizing Value Alignment: From Authoritarian to Democratic AI Ethics. *AI and Ethics*, 1–8.

Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Zhou, J.; Zhang, Z.; Zeng, F.; Ng, K. Y.; Dai, J.; Pan, X.; O'Gara, A.; Lei, Y.; Xu, H.; Tse, B.; Fu, J.; McAleer, S.; Yang, Y.; Wang, Y.; Zhu, S.-C.; Guo, Y.; and Gao, W. 2024. AI Alignment: A Comprehensive Survey. arXiv:2310.19852.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Karinshak, E.; Hu, A.; Kong, K.; Rao, V.; Wang, J.; Wang, J.; and Zeng, Y. 2024. LLM-GLOBE: A Benchmark Evaluating the Cultural Values Embedded in LLM Output. arXiv:2411.06032.

Kharchenko, J.; Roosta, T.; Chadha, A.; and Shah, C. 2024. How Well Do LLMs Represent Values Across Cultures? Empirical Analysis of LLM Responses Based on Hofstede Cultural Dimensions. arXiv:2406.14805.

Li, H.; Gong, R.; and Jiang, H. 2024. Political Actor Agent: Simulating Legislative System for Roll Call Votes Prediction with Large Language Models. *arXiv preprint arXiv:2412.07144*.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024a. DeepSeek-V3 Technical Report. *arXiv preprint arXiv:2412.19437*.

Liu, Y.; Yao, Y.; Ton, J.-F.; Zhang, X.; Guo, R.; Cheng, H.; Klochkov, Y.; Taufiq, M. F.; and Li, H. 2024b. Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment. arXiv:2308.05374.

Manvi, R.; Khanna, S.; Burke, M.; Lobell, D.; and Ermon, S. 2024. Large Language Models Are Geographically Biased. arXiv:2402.02680.

Masoud, R. I.; Ferianc, M.; Treleaven, P. C.; and Rodrigues, M. R. 2024. LLM Alignment Using Soft Prompt Tuning: The Case of Cultural Alignment. In *Workshop on Socially Responsible Language Modelling Research*.

Potter, Y.; Lai, S.; Kim, J.; Evans, J.; and Song, D. 2024. Hidden Persuaders: LLMs' Political Leaning and Their Influence on Voters. arXiv:2410.24190.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Shen, T.; Jin, R.; Huang, Y.; Liu, C.; Dong, W.; Guo, Z.; Wu, X.; Liu, Y.; and Xiong, D. 2023. Large Language Model Alignment: A Survey. arXiv:2309.15025.

Singh, S.; Romanou, A.; Fourrier, C.; Adelani, D. I.; Ngui, J. G.; Vila-Suero, D.; Limkonchotiwat, P.; Marchisio, K.; Leong, W. Q.; Susanto, Y.; Ng, R.; Longpre, S.; Ko, W.-Y.; Smith, M.; Bosselut, A.; Oh, A.; Martins, A. F. T.; Choshen, L.; Ippolito, D.; Ferrante, E.; Fadaee, M.; Ermis, B.; and Hooker, S. 2024. Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation. arXiv:2412.03304.

Sun, Y.; Wang, S.; Feng, S.; Ding, S.; Pang, C.; Shang, J.; Liu, J.; Chen, X.; Zhao, Y.; Lu, Y.; et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Tao, Y.; Viberg, O.; Baker, R. S.; and Kizilcec, R. F. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9): pgae346.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, L.; Yang, N.; Huang, X.; Jiao, B.; Yang, L.; Jiang, D.; Majumder, R.; and Wei, F. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Webster, K.; Wang, X.; Tenney, I.; Beutel, A.; Pitler, E.; Pavlick, E.; Chen, J.; Chi, E.; and Petrov, S. 2021. Measuring and Reducing Gendered Correlations in Pre-trained Models. arXiv:2010.06032.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Xie, C.; Chen, C.; Jia, F.; Ye, Z.; Lai, S.; Shu, K.; Gu, J.; Bibi, A.; Hu, Z.; Jurgens, D.; Evans, J.; Torr, P.; Ghanem, B.; and Li, G. 2024. Can Large Language Model Agents Simulate Human Trust Behavior? arXiv:2402.04559.

Yoo, K. M.; Han, J.; In, S.; Jeon, H.; Jeong, J.; Kang, J.; Kim, H.; Kim, K.-M.; Kim, M.; Kim, S.; et al. 2024. HyperCLOVA X Technical Report. *arXiv preprint arXiv:2404.01954*.

Zhang, X.; Zhang, Y.; Long, D.; Xie, W.; Dai, Z.; Tang, J.; Lin, H.; Yang, B.; Xie, P.; Huang, F.; Zhang, M.; Li, W.; and Zhang, M. 2024. mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval. arXiv:2407.19669.