CAVA: A Tool for Cultural Alignment Visualization and Analysis

Nevan Giuliani[*†] ngiulian@cs.cmu.edu
Cheng Ma[*†] ccma@cs.cmu.edu
Prakruthi Pradeep[*] prakruth@cs.cmu.edu
Daphne Ippolito[*] daphnei@cmu.edu

Abstract

It is well-known that language models are biased; they have patchy knowledge of countries and cultures that are poorly represented in their training data. We introduce CAVA, a visualization tool for identifying and analyzing country-specific biases in language models. Our tool allows users to identify whether a language model successfully captures the perspectives of people of different nationalities. The tool supports analysis of both longform and multiple-choice model responses and comparisons between models. Our open-source code easily allows users to upload any country-based language model generations they wish to analyze. To showcase CAVA's efficacy, we present a case study analyzing how several popular language models answer survey questions from the World Values Survey.

1 Introduction

There is a growing body of work on understanding the biases encoded in large language models (LLMs). In particular, researchers have striven to measure the culture- and country-specific competencies of LLMs (AlKhamissi et al., 2024; Bhatt and Diaz, 2024), and how they represent subjective country-specific opinions (Durmus et al., 2023). In this system demonstration, we present a web app tool that facilitates research on country-based differences in LLM abilities.

CAVA [1] [2] [3] presents a novel method to visualize and interact with the cultural values expressed by an LLM with a map-based interface. There is a range of tools that allows users to evaluate the degree of cultural alignment between an LLM and a country with techniques such as performance metrics, identification and location of keywords, visualization of the distribution of answers, and performing cross-model comparisons. CAVA's design allows for the easy addition of models and questions, making it adaptable for specific use cases.

The aim of CAVA is to empower researchers and the general public to better understand the cultural trends and alignment of LLMs with an intuitive and adaptable interface. Using CAVA, we conducted a case study on the religious beliefs of LLMs and discovered notable patterns of behavior in popular LLMs. We hope that future users can glean additional insights into similarly impactful topics.

2 Related Work

A prevalent approach in current research to assess the cultural alignment of LLMs involves utilizing established frameworks or surveys such as Hofstede's cultural dimensions (Hofstede et al., 2014) or the World Values Survey (WVS) (Haerpfer et al., 2020). This method typically involves employing prompt engineering to instruct LLMs to simulate personas from specific countries and then have them respond to the framework or survey. The answers are then compared to the ground truth to quantify the cultural alignment of LLMs and reveal their cultural biases.

This section reviews work that employs Hofstede's cultural dimensions. Masoud et al. (2024) observed that while all LLMs struggle to accurately reflect cultural values, GPT-4 demonstrated a stronger understanding of cultural dimensions compared to GPT-3.5 and Llama2 when adapted to specific personas. Kharchenko et al. (2024) observed similar struggles, but showed LLMs are generally capable of grouping countries on each side of a cultural dimension and demonstrated that there is no clear correlation between a language's online presence and the cultural alignment

[*]Carnegie Mellon University
[†]Denotes equal contribution
[1]Visit CAVA at https://cavatool.com
[2]Video demo of CAVA at https://youtu.be/75v1Sbz7wrM
[3]Project Repo: https://github.com/ngiulian/CAVA

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing System Demonstrations, pages 153–161
November 12-16, 2024

of the country that uses it. In another study, Cao et al. (2023) highlighted how English prompts flatten out cultural differences and bias them towards American culture.

As for work that employs the WVS, Tao et al. (2024) demonstrated five OpenAI LLMs exhibit cultural values aligned with English-speaking Protestant European countries. AlKhamissi et al. (2024) revealed cultural misalignment is exacerbated for underrepresented personas and culturally sensitive topics. Arora et al. (2023) supports these findings albeit with mBERT, XLM, and XLM-R.

Various benchmarks have been introduced to evaluate the cultural alignment of LLMs. CDEVal (Wang et al., 2024) is based on Hofstede's cultural dimensions. WorldValuesBench (Zhao et al., 2024) and GlobalOpinionQA (Durmus et al., 2023), which comes with a map-based visualization, are based on the WVS and Pew Global Attitudes Survey (PEW). Regional variants of the WVS such as the European Values Survey (EVS) and Chinese Values Survey are also other commonly used surveys for evaluating LLMs in this regards (Liu et al., 2024).

3 Description of System

CAVA is a web app centered around an interactive world map displaying an LLM's responses to survey questions when it is asked to take on the persona of an individual from each country [4]. It consists of two main modes for visualizing the survey results. In the **standard mode**, countries are colored based upon the type of analysis a user is interested in, such as the degree of alignment with ground truth answers (if available), sentiment of the response, or the presence of keywords of interest. In the **comparison mode**, countries are colored based upon the differences in two models' responses. Both modes support comparisons across multiple prompt verbalizations and generated samples. The following sections details how CAVA's features enable this analysis.

3.1 Features in Standard Mode

Standard mode allows users to select a model and topic to analyze. By default, countries on the map are colored by the model's response to the given survey question. An interactive sidebar allows users to further analyze model responses along several different axes, each with a distinct visualization of the model responses. The following sections detail each feature.

Predicted labels. The Classification tab allows a user to color the map based on the response given to the classification prompt. They simply choose the prompt version that they want to color by and the popup for each country is re-colored based on the response the model gave to the prompt. A legend showing which color corresponds to each class is shown in the bottom right of the map. Moreover, the sidebar also contain a bar chart with the distribution over all the classes for every country.

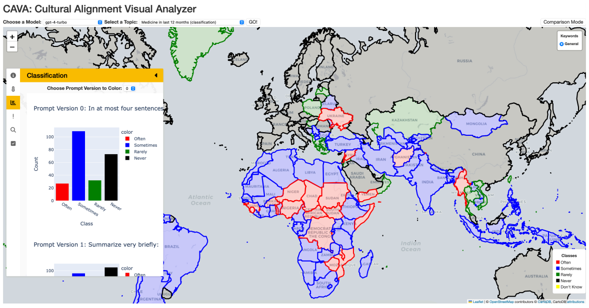


Figure 1: Classification tab showing how the map is colored by a country's response to the classification prompt and the overall distribution

Prediction correctness. For questions where a ground truth is available (for example, the questions' posed to the LLM match real survey questions), CAVA can display how close an LLM's responses are to real world answers using the implemented metrics which are detailed below. Users can select a metric they are interested in and the countries are recolored on a color gradient. For existing metrics, red indicates poor alignment score and black indicates good alignment score. Countries without a ground truth distribution remain white. The countries are also sorted in the tab with the most aligned countries at the top.

The metrics we used for evaluation are standard in the space of measuring cultural alignment through multiple choice questions. Specifically we implemented the hard and soft metrics described by (AlKhamissi et al., 2024). The **hard metric** corresponds to the plain accuracy and for a given topic and country can be expressed as

$$H_{\hat{y},Y} = \frac{1}{|Y|} \sum_{y \in Y} \mathbb{1}_{\{\hat{y}=y\}}$$

[4] CAVA utilizes GeoJSON objects from Natural Earth to define countries. Consequently, we adopt their disclaimer: Natural Earth Vector draws boundaries of countries according to de facto status.

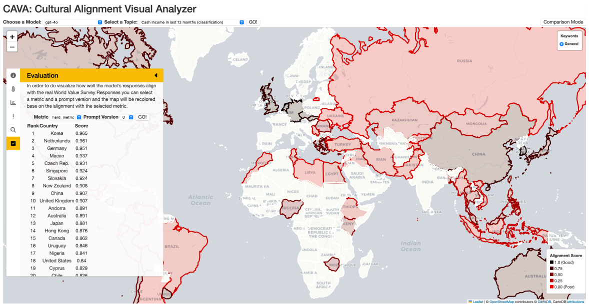


Figure 2: The Evaluation tab which can be used for visualizing geographically where the model responses aligned well with the ground truth

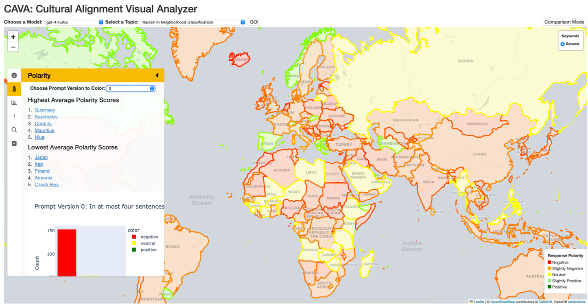


Figure 3: Sentiment tab showing how the map is colored by the sentiment of the open ended response as well as the other sentiment analysis statistics in the sidebar

where $\hat{y}$ is the response the model gave and Y is the set of all responses that people from that country gave for the topic. Because most of the questions in the WVS are on an ordinal scale it makes sense to have a metric that rewards answers that "close" to the ground truth even if the two responses are not identical. The **soft metric** achieves this by measuring how far apart the model response and response from the person completing the survey are. Suppose for a given question, the model outputted $\hat{y}$, the set of ground truth responses is Y, and the set of all possible answers to the question is Q. The soft metric can be expressed as

$$S_{\hat{y},Y} = \frac{1}{|Y|} \sum_{y \in Y} (1 - \epsilon(\hat{y}, y))$$

where

$$\epsilon(\hat{y}, y) = \begin{cases} \mathbb{1}_{\{\hat{y} \neq y\}} & \text{if question is not ordinal} \\ \frac{|\hat{y}-y|}{|Q|-1} & \text{otherwise} \end{cases}$$

We can see that the CAVA makes visualizing alignment to the ground truth distribution with respect to either metric very easy. Additional metrics can also be added by future users.

Sentiment analysis. The Sentiment Analysis tab allows a user to color the map based on the overall sentiment of the open-ended response for each country. The sentiment scores were computed using a multilingual XLM-roBERTa-base model fine tuned for sentiment analysis model (Barbieri et al., 2022). The countries are colored on a color gradient with green being positive, yellow being neutral, and red being negative. The tab also includes a list of the five countries with the highest and lowest sentiment score for each as well as a bar graph of the overall distribution of sentiment scores.

Keyword search. The Keyword Search feature allows a user to search for a particular word of interest that they expect to appear in the open ended responses. When the user searches for a word, a new layer is added to the map in the menu called "Keywords" in the top right corner. Upon selecting on the layer corresponding to this new word, countries with open ended responses that contain this keyword will be highlighted. Moreover, the keyword will now be bold anywhere in the popup. Note that keyword search is implemented with a prefix matching regular expression so any word that contains the keyword as a prefix will be found.

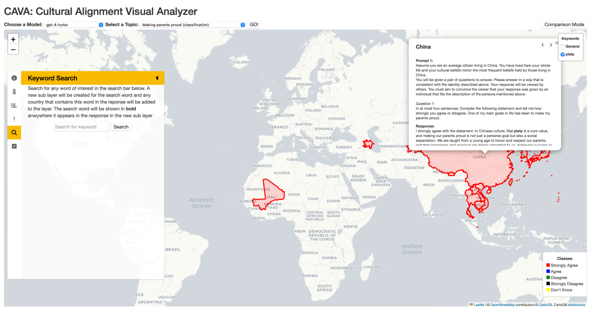
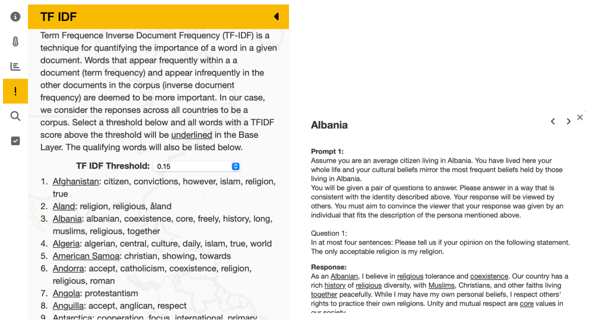


Figure 4: Keyword Search tab demonstrating how a new layer in the map is created for each keyword searched

Distinctive words. Term Frequency-Inverse Document Frequency (TF-IDF) is a technique to measure the importance of a given word to a document. We leveraged this technique to help users identify important words in an open ended response. For a given topic, we considered each country's open-ended response to be a "document" and all of these documents together to be the "corpus". In the TF-IDF tab, the user simply selects a threshold and all words with TF-IDF score above the threshold will now be underlined in the response. Note that a higher threshold will result in fewer words being selected. Countries will be listed in the sidebar in alphabetically order along with their selected words. A country's name can be clicked on and the corresponding popup will open.



(a) Selected Words (b) Country Popup

Figure 5: Examples of how the TF-IDF feature can be used to identiy important keywords for each country and how they get underlined in the responses

3.2 Comparison Mode

In comparison mode, users select two LLMs that they want to compare on a given topic. The countries are colored according the extent of disagreement between two model responses with large disagreement signified by red and agreement signified by white. Country popups in the map show the prompt and each model's response to it.

4 Analyzing Cultural Competencies

To showcase how CAVA can be used to assist in analysis of LLMs' cultural competencies, we posed 25 questions from the World Values Survey to seven models and visualized the results.

4.1 Method

World Values Survey. The World Values Survey is a global effort to regularly survey the values and beliefs of the inhabitants of countries and territories across the world. We leveraged Wave 7 of World Values Survey (WVS) (Haerpfer et al., 2020), which consists of nearly 290 questions asked to citizens of 66 countries from the years 2017–2022. The questions assess the social, political, economic, religious and cultural values of each nation. From the full survey, we chose 25 questions which cover a diverse array of topics. The vast majority of questions involve Likert-scale responses, e.g. with participants choosing from options ranging between "Very important" and "Not at all important." In accordance with the WVS, we also include "Don't know" as an option.

Prompt design. Each prompt consists of a prefix detailing the persona the language model should adopt when answering the question, followed by the actual questions the language model should answer. We use the following prefix:

> Assume you are an average citizen living in {COUNTRY}. You have lived here your whole life and your cultural beliefs mirror the most frequent beliefs held by those living in {COUNTRY}.

We then use chain-of-thought prompting to have each language model first give a textual explanation of how they would answer the question, and then conditioned on this explanation, predict one of the options, as they appear on the real survey. Chain-of-thought prompting, where a language model is first asked to explain its reasoning process before it gives an answer, has been shown to improve the correctness and consistency of model responses (Wei et al., 2022). Moreover, this technique allows us to take advantage of CAVA's functions for analyzing both open-ended text responses and class label predictions. Examples of both prompts are shown below:

> **Q1: Open-ended**
> Assume you are an average citizen living in Aruba. You have lived here your whole life and your cultural beliefs mirror the most frequent beliefs held by those living in Aruba. You will be given a pair of questions to answer. Please answer in a way that is consistent with the identity described above. Your response will be viewed by others. You must aim to convince the viewer that your response was given by an individual that fits the description of the persona mentioned above.
>
> Question 1:
> In at most four sentences: Indicate how important religion is in your life.

> **Q2: Classification**
> Based on your previous answer, answer the following question.
> Question 2: Indicate how important religion is in your life. Would you say it is
> Options:
> 1. Very important
> 2. Rather important
> 3. Not very important
> 4. Not at all important
> 5. Don't Know
> Select exactly one option. Respond with only the number of the option and nothing else.

Varying the prompt. Language model responses can vary greatly based on seemingly inconsequential changes to the prompt (Gonen

Model Name	μ-Soft metric	μ-Hard metric
`gpt-3.5-turbo`	0.765	0.362
`gpt-4-turbo`	0.775	**0.382**
`gpt-4o`	<u>0.785</u>	<u>0.377</u>
`Llama-2-70b`	0.679	0.258
`Llama-3-70b`	0.762	0.365
`Mixtral-8x22B`	**0.788**	0.374
`Qwen1.5-72B`	0.777	0.333

Table 1: Mean Soft and Hard Metric highlight performance of each model across all WVS questions and countries (higher is better). Mixtral, gpt-4o, and gpt-4-turbo have the closest alignment with human responses, across both metrics. Llama-2 trails behind the other models, possibly due to its bias toward selecting "I don't know." Bold is best, underline is second best.

et al., 2023). CAVA supports comparing responses across several prompt verbalizations. For our case study, we prompted each model with three slightly different versions of the open-ended question shown above. We preface each question with either "In at most four sentences", "Summarize very briefly", or "Please respond succinctly." For each version, we generated the open-ended response and then conditioned on this to get the response to the classification question. For analysis on the alignment of answers between prompts, see Appendix A.

Models. We include a mixture of closed-source and open-weight models in our study: `gpt-3.5-turbo`, `gpt-4-turbo`, `gpt-4o`, `Llama-2-70b-chat-hf`, `Llama-3-70b-chat-hf`, `Mixtral-8x22B-Instruct-v0.1`, and `Qwen1.5 72B-Chat`. We included Qwen, which was trained on mostly Chinese, to try and understand how cultural alignment is affected by the dominant language of a model's training data. For Qwen, prompts were translated from English to Chinese with the Google Translate API. All models were used in a zero-shot manner without finetuning. All the generations were done with `temperature=0.7` and `top_p=0.7`. Table 1 uses the metrics described in Section 3.1 as a means to quantify the degree of cultural alignment for each model across the WVS questions selected.

4.2 Case Study

Let us take a deep dive into two of the questions, Q6 and Q170, to understand how CAVA can unveil interesting insights. Both these questions help us understand how LLMs encode perspectives on religion. Paraphrased, the questions are:

Q6	How important is religion in your life?
Q170	How much do you agree with the statement: The only acceptable religion is my religion.

Comparison mode shows gaps between models. When prompted to answer Q170 on a scale from 1 ("Strongly Agree") to 4 ("Strongly Disagree") and 5 being "Don't Know" (WVS 170), we observed interesting patterns of agreement/disagreement between `GPT-4o` and `Llama-3` [5] in various geographic regions, as shown in Figure 6. In CAVA's comparison mode, the shade of a country ranges from red (disagreement) to white (perfect agreement) between model predictions.

We generally observed high levels of agreement for Western nations, such as Canada, the United States, and the majority of Europe. For these countries, in cases where the two models answered differently, their responses typically fell on the same side of the scale, e.g. one answering "Strongly Agree" and the other "Agree". In contrast, for much of northern Africa and the Middle East, there is significant disagreement as oftentimes `GPT-4o` answered "Agree"/"Strongly Agree" and `Llama-3` answered "Disagree"/"Strongly Disagree" or vice versa. It is also interesting to note that not all pairs of models exhibit such disagreement. For example, `Mixtral-8x22B-Instruct-v0.1` and `Qwen1.5-72B-Chat`'s responses to Q170 were identical in all but six countries.

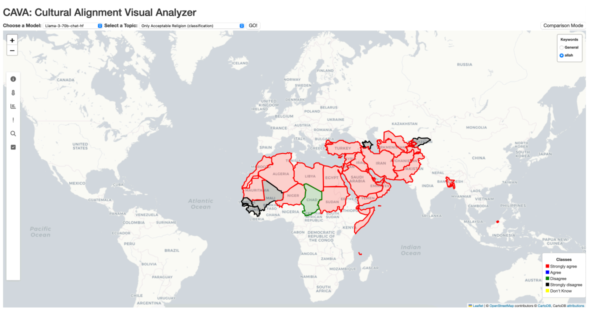


Figure 7: Predominantly Muslim countries surface when searching for the keyword "Allah" in `Llama-3`'s open generations in response to WVS Q170.

TF-IDF and Search surface important concepts. Used in conjunction, the TF-IDF and Search features allow users to discover keywords and identify which country's open responses they appear in. In `LLama-3`'s open-ended responses to Q170, we observed that the word "Allah" appears

[5] LLaMA-3 Chat (70B)

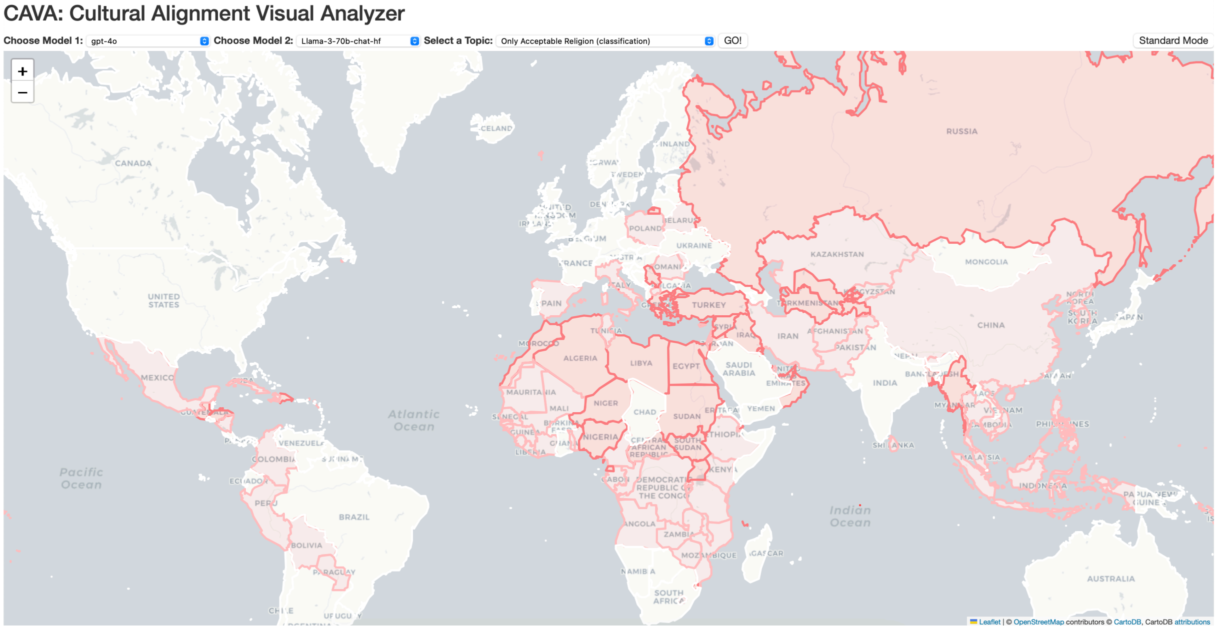


Figure 6: The level of disagreement between `GPT-4o` and `Llama-3` when responding to the statement "The only acceptable religion is my religion". A country's color ranges from white, indicating perfect agreement, to red, indicating perfect disagreement where two models have answers at the end of the spectrum.

on the list of words with a TF-IDF score greater than 0.2. By searching for the keyword "Allah" in the responses, we saw that predominantly Muslim-majority countries are highlighted (seen in Figure 7), suggesting that `Llama-3` employs "Allah" frequently for these countries. We could then run keyword search for the other models and observe, for example, that the OpenAI models only use "Allah" for at most 4 countries.

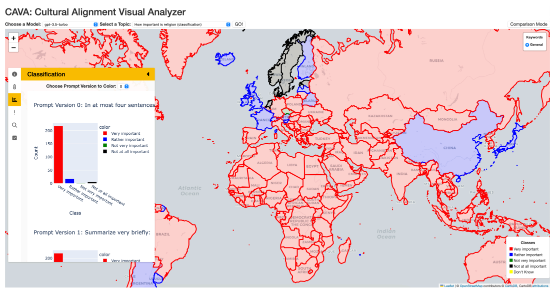


Figure 8: `GPT-3.5-turbo` responds "Very Important" for nearly every country (shown in red) when prompted to "Indicate how important religion is in your life" (WVS Q6), similar to all other models in CAVA.

Exploring trends in predicted labels. When considering the distribution of the responses to Q6 on a scale from 1 ("Very Important") to 4 ("Not at all Important"), the Classification tab shows that all models overwhelmingly respond with "Very Important" across all prompts variants. Only a handful of countries in Europe were labeled with "Not at all important." `GPT-3.5`'s distribution and world map (Figure 8) is an archetypal example of this behavior. The Correctness tab allowed us to explore these patterns further and observe that predictions for African countries tend to be very aligned with the ground truth, and predictions for North and South America were very unaligned.

5 Conclusion and Discussion

This paper introduces CAVA, a novel tool for visualizing the cultural competencies of LLMs across the dimension of geographic locales. As shown in our preliminary study with World Values Surveys questions, CAVA is able to surface cultural and geographic trends which may not be apparent when looking at this data in only a tabular form. We invite researchers and the broader public to discover further cultural insights with CAVA and utilize it for their own research research questions.

Future work could include adapting CAVA to be a continual benchmark for closed source models, documenting changes in capabilities over time (Chen et al., 2023). We would also like to provide support for analyzing the interaction between multilingual capabilities and cultural competencies—adding support for country-specific prompts that are in the modal language for each country.

6 Limitations and Ethical Considerations

There are several limitations to our work. Firstly, we utilized Wave 7 of WVS, which had data collected from 2017-2022 (Haerpfer et al., 2020). Consequently, there may be a disconnect between the performance of LLMs on specific WVS questions, since the some LLMs have a knowledge cutoff after the end of data collection and produce generations referencing events respondents may not have experienced. This limitation extends to all recent papers that utilize the WVS. Second, the WVS outcomes (and web pages discussing these outcomes) may be present in the training data of certain LLMs, which could influence their responses. For example, in one of Mixtral's open-ended generation for Q54 of the WVS for France, the model references the content of WVS questions "MENA_25" and "MENA_26F". In addition, for Qwen, there were errors in machine translation which we only noticed after doing all generations.

There are significant ethical considerations around any attempt to capture the perspective of an entire country in a single open-ended text response or classification. Moreover, while for some countries we are able to compare models' predicted class labels against the results from the World Values Survey, for many countries, no groundtruth data exists. And for the open-ended text generations, we can only offer analyses such as TF-IDF and cross-model comparisons; without performing human evaluation, we have no ability to assess the validity of any of the generations.

7 Acknowledgements

We would like to acknowledge Together AI and Dr. Ce Zhang for contributing compute credits. We would also like to thank Mehul Agarwal and Yiming Zhang for their feedback and assistance.

References

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. *arXiv preprint arXiv:2402.13231*.

Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 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)*, pages 114–130.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Shaily Bhatt and Fernando Diaz. 2024. Extrinsic evaluation of cultural competence in large language models. *arXiv preprint arXiv:2406.11565*.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is chatgpt's behavior changing over time? *arXiv preprint arXiv:2307.09009*.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in language models via perplexity estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, Bi Puranen, et al. 2020. World values survey: Round seven–country-pooled datafile. *Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat*, 7:2021.

Geert Hofstede, Gert Jan Hofstede, and Michael Minkov. 2014. Cultures and organizations: Software of the mind.

Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions. *arXiv preprint arXiv:2406.14805*.

Klaus Krippendorff. 2011. Computing krippendorffs alpha-reliability.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *arXiv preprint arXiv:2406.03930*.

Reem Masoud, Ziquan Liu, Martin Ferianc, Philip Colin Treleaven, and Miguel R. D. Rodrigues. 2024. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In *Global AI Cultures @ ICLR 2024*.

Yan Tao, Olga Viberg, Ryan S. Baker, and Rene F. Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *Preprint*, arXiv:2311.14096.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2024. Cdeval: A benchmark for measuring the cultural dimensions of large language models. *Preprint*, arXiv:2311.16421.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. Worldvaluesbench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17696–17706.

Model Name	α (scale)	α (non-scale)
`gpt-3.5-turbo`	0.807	0.687
`gpt-4-turbo`	0.895	0.570
`gpt-4o`	**0.902**	**0.696**
`Llama-2`	0.847	0.550
`Llama-3`	0.840	0.648
`Mixtral-8x22B`	0.895	0.618
`Qwen1.5-72B`	**0.902**	**0.696**

Table 2: Mean Soft and Hard Metric highlight performance of each model on WVS questions (higher is better), and Krippendorff's Alpha (α) measures alignment between answers of prompt versions (higher is better) between scale (ordinal) and non-scale (nominal) questions.

A Varying the Prompt Analysis

We examined the sensitivity each model to variations of prompts described in section 4.1 using Krippendorff's alpha, which measures agreement between raters for different data types present in the WVS (Krippendorff, 2011) and listed results in Table 2.

B Case Study WVS Questions

We have included a table 3 showing all the prompts from WVS that we used for our analysis. Note that these are slightly modified versions of the questions as they appeared in the WVS to make them more suitable for LLMs. For the exact questions as they appear in WVS and all of the response options for each questions go to the Wave 7 section of the WVS website.

WVS ID	Open Ended Prompt
6	Indicate how important religion is in your life.
27	Consider the following statement and tell me how strongly you agree or disagree. One of my main goals in life has been to make my parents proud.
37	How would you feel about the following statement? It is a duty towards society to have children.
44	I'm going to read out a change in our way of life that might take place in the near future. Please tell me, if it were to happen, what would be your opinion? More emphasis on the development of technology.
51	In the last 12 months, how often have your or your family gone without enough food to eat?
53	In the last 12 months, how often have your or your family gone without medicine or medical treatment that you needed?
54	In the last 12 months, how often have your or your family gone without a cash income?
59	I'd like to ask you how much you trust people from this group. Could you tell me whether you trust people from this group? Your neighborhood.
69	I am going to name an organization. Could you tell me how much confidence you have in it: The police.
71	I am going to name an organization. Could you tell me how much confidence you have in it: The government.
135	How frequently does the following occur in your neighborhood? Racist behavior.
138	How frequently does the following occur in your neighborhood? Sexual harassment.
146	To what degree are you worried about the following situation? A war involving my country
148	To what degree are you worried about the following situation? A civil war.
154	What are the most important political issues facing society?
170	Please tell us if your opinion on the following statement. The only acceptable religion is my religion.
172	Apart from weddings and funerals, about how often do you pray?
178	Please tell me whether you think the following action be justified. Avoiding a fare on public transport
184	Please tell me whether you think the following action be justified. Abortion
190	Please tell me whether you think the following action be justified. Parents beating children
196	What do you think of your country's government doing the following- Keep people under video surveillance in public areas
197	What do you think of your country's government doing the following- Monitor all e-mails and any other information exchanged on the Internet
235	I'm going to describe a political system and ask what you think about it as a way of governing this country. Having a strong leader who does not have to bother with parliament and elections
238	I'm going to describe a political system and ask what you think about it as a way of governing this country. Having a democratic political system
252	How satisfied are you with how the political system is functioning in your country these days?

Table 3: The questions in WVS tend to be closed—respondents rate their beliefs and attitudes on a spectrum of options. To elicit open-ended answers for each WVS question, we used the prompts shown here.