Effects of Language- and Culture-Specific Prompting on ChatGPT

Mustafa Tuna, Kristina Schaaff, Tim Schlippe
IU International University of Applied Sciences
Germany
{kristina.schaaff | tim.schlippe}@iu.org

***Abstract*—Advanced LLM-based chatbots like ChatGPT are immensely popular, engaging users from diverse cultural backgrounds. Previous studies indicate that when trained with a lot of English data, these chatbots predominantly reflect the nuances of English-speaking cultures, particularly in the US American context. Such a cultural bias may hinder effective communication with users and marginalize non-English cultures. In this paper, we investigate the cultural relevance of chatbots built on the LLMs GPT-3.5-turbo and GPT-4, specifically when used in languages other than English. Our analysis encompassed the five language areas English, German, Spanish, French, and Portuguese and the corresponding ten subcultures English-speaking Great Britain and the USA, German-speaking Germany and Austria, Spanish-speaking Spain and Mexico, French-speaking Canada and France, and Portuguese-speaking Brazil and Portugal. Our assessment of cultural appropriateness employed the ten dimensions from Inglehart and Welzel's cultural mapping framework. We benchmarked the chatbots' performances, elicited through specific prompts, against data from the World Value Survey. Subsequently, we developed a unique cultural map according to Inglehart and Welzel's model. Our findings indicate that GPT-3.5-turbo generally surpassed GPT-4 in cultural alignment, particularly within the German linguistic context. Conversely, the Spanish and Portuguese linguistic regions showed lower cultural alignment. At the subcultural level, German-speaking Germany demonstrated the highest cultural alignment, whereas Spanish-speaking Mexico and Portuguese-speaking Brazil demonstrated the least alignment.**

***Index Terms*—large language models, culture, language-specific prompting, culture-specific prompting, chatbots**

I. Introduction

In recent years, chatbots have become a prevalent tool in everyday life [1] as they are capable of engaging in conversations that closely mimic human interactions [2], providing assistance [3], information [4], and emotional support [5]. OpenAI's ChatGPT has emerged as one of the most widely utilized chatbots attracting over one million users within just one week of its launch [6].

In January 2023, nearly 20% of the worldwide traffic to the services at OpenAI.com originated from the USA [7]. Simultaneously, a consortium of nations including Germany, France, India, Brazil, the United Kingdom, Spain, and Canada accounted for almost 31% of the website traffic for the California-based artificial intelligence startup, while the remaining half was attributed to various other countries [7].

But despite ChatGPT being a chatbot used globally by people from diverse cultures, the texts generated do not equally consider the varied cultural values of users [8]. OpenAI acknowledges on its website that ChatGPT is oriented towards Western values in English conversations. In support of this, [9] explains that cultural values are best reflected within the English-speaking cultural domain. Additionally, other studies have found that American culture is predominantly represented in the English-speaking usage of ChatGPT [10], [11].

However, according to [12] and [13], it is important to understand cultural differences in communication. [12] emphasizes how variations in communication styles, particularly regarding space and time orientation, affect interpersonal interactions, while [13] highlights how cultural values influence work-related behaviors and attitudes, emphasizing the importance of recognizing and adapting to these differences for effective cross-cultural communication. Previous research has not addressed how appropriate cultures represented within a language area are represented by LLM-based chatbots. Consequently, given the widespread use of these chatbots globally, our work addresses the following research questions:

- To what extent does the behavior of ChatGPT align with cultural norms across various language areas?
- How well does ChatGPT demonstrate culturally appropriate behavior across various subcultures when subculture specifications are given in the prompt?

Culturally appropriate behavior refers to the correspondence between ChatGPT's answers and the answers of people from certain cultures. *Subculture specification* means that the user instructs the chatbot in the prompting to answer from the perspective of a person from the respective culture.

To answer these research questions, we investigated the effects of language- and culture-specific prompting on ChatGPT in the versions `GPT-3.5-turbo` and `GPT-4`. For the investigation of cultural appropriateness, results of the World Values Survey (WVS) are compared with the results of the chatbots. We examine the language areas English (*EN*), German (*DE*), Spanish (*ES*), French (*FR*) and Portuguese (*PT*) as well as the subcultures English-speaking Great Britain (*EN-GB*) and the USA (*EN-US*), German-speaking Austria (*DE-AT*) and Germany (*DE-DE*), Spanish-speaking Spain (*ES-ES*) and Mexico (*ES-MX*), French-speaking Canada (*FR-CA*) and French-speaking France (*FR-FR*) as well as Portuguese-speaking Brazil (*PT-BR*) and Portugal (*PT-PT*).

979-8-3503-5479-9/24/$31.00 ©2024 IEEE

II. Related Work

In this section, we will first present questionnaires for the investigation of cultural values that have been used extensively in cross-cultural research. Then, we will describe how other researchers investigate the cultural values of chatbots.

A. Questionnaires for the Investigation of Cultural Values

Geert Hofstede's *Cultural Dimensions Theory* [14] provides a significant framework for capturing and comparing cultural values across different cultures. Additionally, Hofstede developed a survey known as the Values Survey Module (VSM)[1] to measure cultural values in alignment with his theory. Hofstede's *Cultural Dimensions Theory* identifies six dimensions of national culture that can be used to compare and contrast cultural values between different countries. The dimensions include Power Distance, Individualism vs. Collectivism, Masculinity vs. Femininity, Uncertainty Avoidance, Long-Term Orientation vs. Short-Term Orientation, and Indulgence vs. Restraint. Over time, the VSM has undergone several revisions to enhance its effectiveness and relevance. But Hofstede's model is criticized in particular for the fact that the dimensions are predominantly based on the results of surveys that are not nationally representative, as these are essentially made up of the responses of employees and students [15].

Shalom Schwartz is known for developing the *Theory of Basic Human Values* [16], which encompasses a set of universal values that are believed to underlie human behavior across cultures. Schwartz listed a total of ten universal value dimensions, of which only seven were relevant to cultural differences. The *Schwartz Value Survey* (SVS) [16] presents respondents with a list of values and asks them to rate the importance of each value to themselves or to society. Similar to the criticism of Hofstede, Schwartz's model is also criticized for the fact that the sample is not nationally representative [15].

The *European Value Study* (EVS)[2] is a research project that aims to study the basic social, moral, and political values of European societies [17]. Since 1981, it has examined the beliefs, attitudes, and opinions of individuals across Europe regarding various social issues, including family, work, religion, politics, and society. The EVS project conducts surveys periodically to collect data on these values and analyzes trends and changes over time. The findings provide valuable insights into cultural diversity and commonalities within European societies, aiding policymakers, researchers, and the public in understanding societal dynamics and trends.

During the first implementation of the EVS in 1981, there was increasing interest in collecting data on cultural values outside Europe [18]. This led to a global questionnaire similar to the EVS—the *World Values Survey* (WVS)[3] which was developed by Ronald Inglehart and refined by Christian Welzel. To date, data has been collected in a total of 120 countries. This makes it one of the largest non-commercial social surveys. The relevance of the WVS is demonstrated by over 50,000 citations in scientific papers.

The *Global Leadership and Organizational Behavior Effectiveness* (GLOBE)[4] [19] study identifies nine cultural dimensions that are believed to influence leadership and organizational behavior across different cultures. The nine dimensions are partly based on Hofstede's cultural dimensions and are Uncertainty Avoidance, Power Distance, In-Group Collectivism, Institutional Collectivism, Gender Equalitarianism, Assertiveness, Future Orientation, Performance Orientation, and Human Orientation. The GLOBE cultural model is criticized for the fact that its scales are based on stereotypical assumptions about nationalities [20].

Since only the cultural dimensions of the WVS were developed on a data set with nationally representative data and the WVS has a large sample size that enhances the reliability and generalizability, we decided to use the WVS for our analyses.

B. Investigation of Cultural Values in LLM-based Chatbots

[10] investigated ChatGPT's cultural behavior with English, Chinese, German, Japanese, and Spanish questions from Hofstede's Values Survey Module (VSM) [14], considering American, Chinese, German, Japanese, and Spanish cultures. They prompted two scenarios: In the first scenario, all the questions were prompted in English with an explicit specification of the language areas. In the second scenario, the questions were prompted in the main languages. Responses significantly varied depending on the language used. Additionally, regardless of explicit language area specifications, ChatGPT in English predominantly represented American culture.

[11] analyzed the chatbots ChatGPT, GPT-4, and Bard [21], using questions from Hofstede's Values Survey Module (VSM) [14]. The chatbots were prompted in English, Arabic, Chinese, and Slovak with explicit cultural specifications covering the language area. Results from English prompts were compared with responses from individuals in the USA, Arabic prompts with Saudi Arabia, Chinese prompts with China, and Slovak prompts with Slovakia. [11] found that GPT-4 exhibited the highest cultural appropriateness, while Bard showed the lowest. Furthermore, cultural appropriateness was highest for the USA and lowest for Saudi Arabia.

[22] examined the chatbots ChatGPT and Bard. They conducted the prompting using questions from the GLOBE survey and compared the results with responses from individuals in the GLOBE study. Their results also show that both chatbots showed a higher alignment with English-speaking cultures.

[23] examined ChatGPT based on the LLM GPT-3.5 and 259 selective questions from the English WVS. They had ChatGPT answer these questions and obtained responses to 251 of the questions, while ChatGPT indicated its chatbot nature for the remaining eight questions. The researchers combined ChatGPT's responses with the results of the WVS and performed a principal component analysis. This revealed

[1]https://sjdm.org/dmidi/Values_Survey_Module.html

[2]https://europeanvaluesstudy.eu

[3]https://www.worldvaluessurvey.org

[4]https://globeproject.com

two principal components representing average social engagement and average distrust towards government and public organizations. The Euclidean distance between ChatGPT and various countries was calculated, with the smallest distances observed for Australia, Great Britain, and Northern Ireland, and the largest for Myanmar, China, and Egypt [23].

[24] investigated how closely chatbots based on OpenAI's LLMs GPT-3.5 and GPT-4 matched the cultural values of English-, Chinese-, Russian-, Indonesian-, Indian- and Arabic-speaking people. In contrast to [23], [24] focused exclusively on the selective questions of the WVS, which measured agreement with statements on a Likert scale from 1 (strongly disagree) to 5 (strongly agree). They analyzed the chatbots' responses using principal component analysis and reduced them to two principal components corresponding to Survival vs. Self-expression Values and Traditional vs. Secular-Rational Values. Like [23], [24] also calculated the Euclidean distances and found that the chatbots' responses had the smallest distance to the WVS results in English, no matter what language was investigated [24]. Even when the chatbots were instructed to answer the questions based on the culture of the respective language, the distance to English was the lowest, although the distances decreased for the other languages.

Compared to the related work, we are the first to analyze how well ChatGPT simulates values of subcultures—especially when specifying the subculture in the prompting.

III. Experimental Setup

A. Our Analyzed Chatbots

In the following subsection, we will describe the chatbots and their associated LLMs that we evaluate for our analysis regarding culturally appropriate behavior.

1) ChatGPT Version 3.5: ChatGPT is a state-of-the-art chatbot developed by OpenAI that can generate natural language text based on provided prompts or contextual cues. The chatbot is based on the LLM GPT-3.5 and was fine-tuned using reinforcement learning from human feedback [25]. This method enables the model to comprehend the meaning and intent behind user queries, resulting in responses that are pertinent and beneficial. To maintain safety and mitigate the generation of inappropriate or factually incorrect text, the training of ChatGPT was enhanced by incorporating a large dataset of human-human and human-chatbot conversations. OpenAI has not officially disclosed the precise size of the training data used for ChatGPT. However, the predecessor model GPT3 with 175 billion parameters was already significantly larger than other language models like BERT, RoBERTA, or T5 and was trained with 499 billion crawled tokens (i.e., subword units) [26]. Through extensive exposure to human language nuances and intricacies within this extensive dataset, ChatGPT is capable of generating text that closely resembles human writing, rendering it highly realistic and difficult to discern from human-authored content [27].

2) ChatGPT Version 4: GPT-4 has been available in ChatGPT since March 2023. It was trained on a text corpus of about 13 trillion tokens. Some of these tokens come from well-known datasets such as *CommonCrawl* and *RefinedWeb*, while others come from undisclosed sources [28], [29]. GPT-4 was first fine-tuned with data sourced from ScaleAI plus text data from OpenAI. Subsequently, it was fine-tuned using the reward model (Reinforcement Learning from Human Feedback) and the Proximal Policy Optimization algorithm [29], [30]. The model is believed to possess approximately 1.8 trillion parameters [28], [29].

B. WVS Questions for the Creating the Cultural Map

To evaluate to what extent the behavior of ChatGPT aligns with cultural norms across various language regions and subcultures, we used the questions defined by [31] for creating the *cultural map*, a subset of the WVS questions covering the topics *Abortion*, *Autonomy*, *Faith*, *Happiness*, *Homosexuality*, *Patriotism*, *Politicization*, *Respect*, *Postmaterialism*, and *Trust*. This map visualizes cultural orientations along two main axes that are determined by a dimensional reduction. These axes form a two-dimensional space where each society or culture can be plotted based on its position relative to these dimensions. From the entire WVS, we have focused only on the questions of the cultural map due to the following reasons:

- The questions of the cultural map have proven to be important questions for discriminating cultures and subcultures.
- In the WVS, some of the questions differ for certain languages and the respective subcultures. Therefore, our goal was to find questions/topics present in all languages and subcultures. The questions of the cultural map fulfill this requirement.
- The questions of the cultural map enable us to visualize the distances between the WVS values of the human participants and the chatbots in the cultural map.

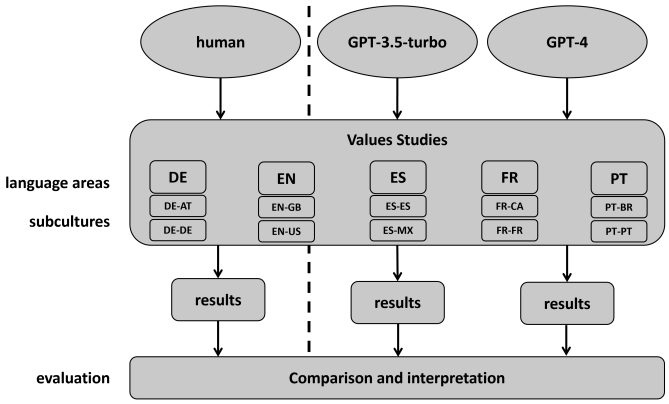

Fig. 1. Language- and Culture-Specific Prompting Strategy.

C. Language- and Culture-Specific Prompting Strategy

The setup of our experiments is visualized in Figure 1. The analysis encompassed the five language areas German (DE), English (EN), Spanish (ES), French (FR),

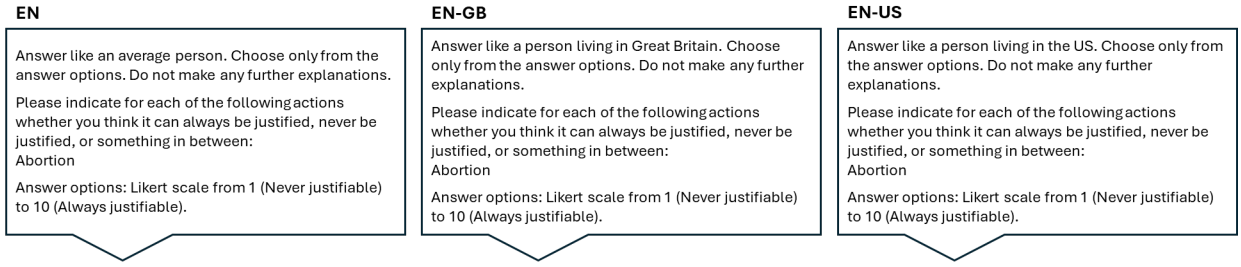

Fig. 2. Prompting of *EN*, *EN-GB* and *EN-US*.

TABLE I
DISTANCE BETWEEN CHATBOTS' AND HUMANS' SCORES FOR EN, EN-US AND EN-GB.

Item	**Distance to *human* (EN)**		**Distance to *human* (EN-US)**		**Distance to *human* (EN-GB)**	
	GPT-3.5-turbo	**GPT-4**	**GPT-3.5-turbo**	**GPT-4**	**GPT-3.5-turbo**	**GPT-4**
Abortion	0.22%	2.44%	0.89%	16.44%	12.22%	3.33%
Autonomy	28.00%*	33.00%**	33.25%**	33.25%**	7.75%	32.75%**
Faith	0.78%	8.11%	5.11%	14.89%	5.33%	10.22%
Happiness	5.33%	5.33%	3.67%	3.67%	9.00%	9.00%
Homosexuality	29.33%	16.00%	27.00%	38.11%*	19.11%	25.78%*
Patriotism	11.67%	11.67%	8.33%	28.33%*	8.33%	1.67%
Politicization	2.50%	22.50%*	12.00%	22.00%	5.00%	15.00%
Postmaterialism	48.50%**	38.50%**	23.50%	33.50%*	3.50%	36.50%**
Respect	22.50%	12.50%	21.50%	28.50%	25.50%*	24.50%
Trust	39.00%	41.00%	59.00%**	41.00%	55.00%*	45.00%*
Average	18.78%	19.11%	19.43%	25.97%	15.08%	20.38%

and Portuguese (PT) and the corresponding ten subcultures German-speaking Germany (DE-DE) and Austria (DE-AT), English-speaking Great Britain (EN-GB) and the USA (EN-US), Spanish-speaking Spain (ES-ES) and Mexico (ES-MX), French-speaking France (FR-FR) and Canada (FR-CA), and Portuguese-speaking Portugal (PT-PT) and Brazil (PT-BR).

Figure 2 shows an example of the prompting of the question which covers the item *Abortion* for the language area EN and subcultures EN-GB and EN-US. We prompted the other items covering the topics *Autonomy*, *Faith*, *Happiness*, *Homosexuality*, *Patriotism*, *Politicization*, *Respect*, *Postmaterialism*, and *Trust* in the same way using the original questions from the respective languages. For the language-specific prompting, the text "Answer like an average person." was added in the corresponding language at the beginning of each prompt. For the culture-specific prompting, the text "Answer like a person living in <country>." was added in the corresponding language at the beginning of each prompt, whereas in <country> we added the corresponding country name. Furthermore, we added "Choose only from the answer options", the hint " Do not make any further explanations"', and a brief description of the scale. After the question, we defined the answer options. Since ChatGPT does not always give the same answers despite setting the parameter *temperature* to 0, we asked each question five times and calculated the average scores of the results.

IV. OUR RESULTS

Our paper focuses on raising awareness that subcultural differences must be considered when developing and using chatbots. We will first show how well the individual items of the language areas and subcultures are represented by the chatbots. Then, we will visualize the distances between the WVS scores of the human participants and the chatbots on a cultural map.

A. Cultural Norms across Language Areas and Subcultures

Tables I—V show the Euclidean distances in percent between the chatbots' and the humans's average WVS scores (*human*) for our language areas and subcultures. As described in the related work in Section II, distance is frequently used to express discrepancies in cultures, e.g., [23], and other works using the WVS. Since the ranges (max, min) in the WVS items are not normalized and are therefore difficult to compare and interpret, we decided to normalize the values by expressing them as percentages. The closer the scores of humans and a chatbot are to each other, i.e. the closer the percentage value to 0, the better the chatbot represents the human language area or subculture. Significant distances between the means of humans and chatbots based on Mann-Whitney U tests [32] are marked with "*" in the tables, highly significant differences with "**". The Mann-Whitney U test is a statistical procedure for checking a significant difference between two observation groups. In contrast to the t-test, the Mann-Whitney U test does not assume a normal distribution or homogeneity of variance, This fits our experimental setup where we asked the chatbots each question five times, as described in Section III-C, and have thousands of answers from the human WVS participants.

TABLE II
DISTANCE BETWEEN CHATBOTS' AND HUMANS' SCORES FOR DE, DE-DE AND DE-AT.

	Distance to *human* (DE)		**Distance to *human* (DE-DE)**		**Distance to *human* (DE-AT)**	
Item	**GPT-3.5-turbo**	**GPT-4**	**GPT-3.5-turbo**	**GPT-4**	**GPT-3.5-turbo**	**GPT-4**
Abortion	6.56%	4.33%	1.44%	14.11%	3.56%	8.00%
Autonomy	17.00%	2.00%	2.75%	7.75%	18.75%	3.75%
Faith	0.22%	0.22%	18.44%	1.56%	22.00%	0.22%
Happiness	7.00%	7.00%	4.33%	4.33%	15.00%	8.33%
Homosexuality	7.33%	7.33%%	23.11%	23.11%	33.78%*	33.78%*
Patriotism	6.67%	6.67%	5.33%	5.33%	2.00%	8.67%
Politicization	1.00%	29.00%**	19.50%	19.50%	30.00%	30.00%
Postmaterialism	22.00%	32.00%*	15.00%	35.00%**	1.00%	49.00%**
Respect	2.50%	2.50%	16.50%	36.50%	17.00%	47.00%*
Trust	27.00%	53.00%*	54.00%*	46.00%*	29.00%	49.00%*
Average	9.73%	14.41%	16.04%	19.32%	17.21%	23.78%

1) English Language Area and Subcultures: Table I lists the distances of each of the 10 items between the chatbots' and humans' average scores for the language area EN and the subcultures EN-US and EN-GB. Considering the language area EN, GPT-3.5-turbo (18.87%) shows a slightly lower distance to *human* than GPT-4 (19.11%). There are no significant differences to *human* for eight out of the ten WVS items (80%) for GPT-3.5-turbo. In contrast, there are no significant differences between GPT-4 and *human* in seven out of ten items (70%). This demonstrates that GPT-3.5-turbo can be used well in communications covering the items *Abortion*, *Faith*, *Happiness*, *Homosexuality*, *Patriotism*, *Politicization* and *Postmaterialism*. In comparison, GPT-4 can be used well in the areas of *Abortion*, *Faith*, *Happiness*, *Homosexuality*, *Patriotism*, *Respect* and *Trust*. Looking at the subculture EN-US indicates that again GPT-3.5-turbo (19.43%) shows a slightly lower distance to *human* than GPT-4 (25.97%). There are no significant differences to *human* for eight out of the ten WVS items (80%) for GPT-3.5-turbo. However, the two topics of *Autonomy* and *Trust* should be treated with caution when using GPT-3.5-turbo. When employing GPT-4 in EN-US, conversations on the four topics of *Autonomy*, *Homosexuality*, *Patriotism* and *Postmaterialism* could lead to problems. In the evaluation of EN-GB, we also see that GPT-3.5-turbo (15.08%) is on average closer to the behavior of humans than GPT-4 (20.38%). From the significant distances, we see that GPT-3.5-turbo (80%) does not perfectly cover the items of *Respect* and *Trust*. Furthermore, GPT-4 (60%) does not manage to show culturally appropriate behavior for the items *Autonomy*, *Homosexuality*, *Postmaterialism* and *Trust*.

2) German Language Area and Subcultures: Table II shows the differences across the 10 items within the language area DE and its subcultures DE-DE and DE-AT. Within DE, GPT-3.5-turbo (9.73%) demonstrates a slightly smaller distance to *human* compared to GPT-4 (14.41%). GPT-3.5-turbo exhibits no statistically significant distance to *human* across all 10 WVS items (100%), while GPT-4 shows no significant differences in seven of the ten items (70%). GPT-3.5-turbo generally aligns well with human cultural attitudes across all items within DE, whereas GPT-4 performs better in specific items such as *Abortion*, *Autonomy*, *Faith*, *Happiness*, *Homosexuality*, *Patriotism*, and *Respect*. Our DE-DE analyses reveals a similar pattern, with GPT-3.5-turbo (16.04%) exhibiting a slightly smaller disparity from *human* compared to GPT-4 (19.32%). GPT-3.5-turbo shows no significant deviations from *human* in nine out of the ten WVS items (90%), but caution is advised regarding the topic of *Trust*. Conversely, GPT-4 may face challenges in conveying culturally appropriate responses for the items of *Postmaterialism* and *Trust* within DE-DE. In the evaluation of DE-AT, GPT-3.5-turbo (17.21%) demonstrates closer alignment with human behavior compared to GPT-4 (23.78%). However, significant disparities reveal that GPT-3.5-turbo (90%) does not fully capture *Homosexuality*, while GPT-4 (60%) struggles with *Homosexuality*, *Postmaterialism*, *Respect*, and *Trust*.

3) Spanish Language Area and Subcultures: Table III displays the distances across the 10 items within the language area ES and its subcultures ES-ES and ES-MX. Within ES, unlike EN and DE, GPT-3.5-turbo (37.29´%) exhibits a greater disparity from *human* compared to GPT-4 (27.95%). Only four out of the ten WVS items (40%) show no significant differences between GPT-3.5-turbo and *human*, while for GPT-4, this number is six out of ten items (60%). GPT-3.5-turbo demonstrates proficiency primarily in conveying *Happiness*, *Homosexuality*, *Politicization*, and *Postmaterialism*, whereas GPT-4 excels in areas such as *Faith*, *Happiness*, *Homosexuality*, *Postmaterialism*, *Respect*, and *Trust*. Our analysis of the subculture ES-ES reveals that GPT-3.5-turbo (17.06%) exhibits a slightly smaller disparity from *human* compared to GPT-4 (26.96%). GPT-3.5-turbo shows no significant deviations from *human* in seven out of the ten WVS items (70%), while for GPT-4, this is the case for only five out of ten items (50%). Caution is advised particularly regarding the topics of *Autonomy*, *Homosexuality*, *Politicization*, and *Postmaterialism* when utilizing GPT-3.5-turbo. On the other hand, employing GPT-4 in ES-ES may encounter challenges in conversations centered around *Autonomy*, *Homosexuality*, *Patriotism*, *Politicization*, and *Postmaterialism*. In the assessment of ES-MX, GPT-4 (28.46) demonstrates a slightly closer alignment with human behavior compared to GPT-3.5-turbo

TABLE III
DISTANCE BETWEEN CHATBOTS' AND HUMANS' SCORES FOR ES, ES-ES, AND ES-MX.

	Distance to *human* (ES)		**Distance to *human* (ES-ES)**		**Distance to *human* (ES-MX)**	
Item	**GPT-3.5-turbo**	**GPT-4**	**GPT-3.5-turbo**	**GPT-4**	**GPT-3.5-turbo**	**GPT-4**
Abortion	25.89%*	21.44%*	15.22%	7.00%	41.56%**	19.33%
Autonomy	57.25%**	57.25%**	33.00%**	43.00%**	33.00%**	48.00%**
Faith	39.00%**	34.56%	12.89%	1.78%	8.89%*	15.56%
Happiness	10.00%	10.00%	2.33%	2.33%	16.67%*	16.67%*
Homosexuality	57.44%	61.89%	33.11%**	33.11%**	62.89%**	62.89%**
Patriotism	21.33%**	21.33%**	9.00%	24.33%*	14.00%*	12.67%
Politicization	13.00%	33.00%*	7.00%	47.00%*	52.50%**	62.50%**
Postmaterialism	24.50%	14.50%	27.50%	47.50%**	34.00%*	14.00%
Respect	36.50%**	13.50%	28.50%**	21.50%	27.00%*	23.00%
Trust	88.00%**	12.00%	2.00%	42.00%	10.00%	10.00%
Average	37.29%	27.95%	17.06%	26.96%	30.05%	28.46%

TABLE IV
DISTANCE BETWEEN CHATBOTS' AND HUMANS' SCORES FOR FR, FR-FR AND FR-CA.

Item	**Distance to *human* (FR)**		**Distance to *human* (FR-FR)**		**Distance to *human* (FR-CA)**	
	GPT-3.5-turbo	**GPT-4**	**GPT-3.5-turbo**	**GPT-4**	**GPT-3.5-turbo**	**GPT-4**
Abortion	19.78%	20.22%	5.78%	25.33%*	18.33%*	23.89%*
Autonomy	12.25%	32.25%**	12.25%	37.25%**	9.00%	24.00%**
Faith	13.33%	8.89%	4.44%	8.89%	3.78%	14.89%
Happiness	7.33%	7.33%	8.33%	8.33%	4.00%	9.33%
Homosexuality	21.67%	19.44%	22.56%	29.22%*	17.89%	17.89%
Patriotism	11.33%	2.00%	14.00%	6.00%	0.33%**	27.00%
Politicization	24.00%*	16.00%	10.50%	20.50%	18.50%*	11.50%
Postmaterialism	26.50%	36.50%*	27.50%	37.50%*	42.50%**	32.50%*
Respect	29.50%**	20.50%	36.50%**	13.50%	22.50%	27.50%*
Trust	61.00%**	39.00%	13.00%	27.00%	51.00%*	9.00%
Average	22.67%	20.21%	15.49%	21.35%	18.78%	19.75%

(30.05%). Notably, significant disparities indicate that while GPT-3.5-turbo fails to accurately capture all items except for *Trust* (10%), GPT-4 (60%) struggles to exhibit culturally appropriate behavior across the items of *Autonomy*, *Happiness*, *Homosexuality*, and *Politicization*.

4) French Language Area and Subcultures: Table IV displays the disparities observed across the 10 items within the FR language area and its subcultures FR-FR and FR-CA. Within FR, GPT-3.5-turbo (22.67%) exhibits a slightly higher deviation from *human* compared to GPT-4 (20.21%). GPT-3.5-turbo shows no statistically significant deviations from *human* in seven of the 10 WVS items (70%), while GPT-4 does so in eight of the 10 items (80%). GPT-3.5-turbo generally aligns well with human cultural attitudes in the areas of *Abortion*, *Autonomy*, *Faith*, *Happiness*, *Homosexuality*, *Patriotism*, and *Postmaterialism* within FR, whereas GPT-4 performs well in the topics of *Abortion*, *Faith*, *Happiness*, *Homosexuality*, *Patriotism*, *Politicization*, *Respect*, and *Trust*. Within FR-FR, GPT-3.5-turbo (15.49%) exhibits a slightly higher deviation from *human* compared to GPT-4 (21.35%). GPT-3.5-turbo shows no significant deviations from *human* in nine out of the ten WVS items (90%), but caution is advised regarding the topic of *Respect*. Conversely, GPT-4 (60%) may encounter challenges in conveying culturally appropriate responses for the items of *Abortion*, *Autonomy*, *Homosexuality*, and *Postmaterialism* within FR-FR. In the evaluation of FR-CA, GPT-3.5-turbo (18.78%) demonstrates slightly closer alignment with human behavior compared to GPT-4 (19.75%). However, significant disparities reveal that GPT-3.5-turbo (50%) does not fully capture the nuances of the five items *Abortion*, *Patriotism*, *Politicization*, *Postmaterialism*, and *Trust*, while GPT-4 (60%) struggles with *Abortion*, *Autonomy*, *Postmaterialism*, and *Respect*.

5) Portuguese Language Area and Subcultures: Table V illustrates the differences observed across the 10 items within the PT language area and its subcultures PT-PT and PT-BR. Within PT, GPT-3.5-turbo (28.36%) exhibits a slightly smaller deviation from *human* compared to GPT-4 (32.89%). GPT-3.5-turbo shows no statistically significant deviations from *human* in only half of the 10 WVS items (50%), while GPT-4 does so in only four of the 10 items (40%). GPT-3.5-turbo generally aligns well with human cultural attitudes in the areas of *Faith*, *Happiness*, *Homosexuality*, *Patriotism*, and *Postmaterialism* within PT, whereas GPT-4 performs well only in the topics of *Happiness*, *Patriotism*, *Respect*, and *Trust*. Within PT-PT, GPT-3.5-turbo (22.87%) exhibits a slightly larger deviation from *human* compared to GPT-4 (33.37%). GPT-3.5-turbo shows no significant deviations from *human* in six out of the ten WVS items (60%), but caution is advised regarding the topics of *Homosexuality*, *Patriotism*, *Politicization*, and

TABLE V
DISTANCE BETWEEN CHATBOTS' AND HUMANS' SCORES FOR PT, PT-PT AND PT-BR.

Item	Distance to *human* (PT)		Distance to *human* (PT-PT)		Distance to *human* (PT-BR)	
	GPT-3.5-turbo	GPT-4	GPT-3.5-turbo	GPT-4	GPT-3.5-turbo	GPT-4
Abortion	25.44%*	63.22%**	24.11%	50.78%**	39.56%**	72.89%**
Autonomy	30.25%**	50.25%**	16.00%	41.00%**	32.25%**	57.25%**
Faith	10.11%	36.78%**	12.78%	19.44%	29.00%**	8.78%
Happiness	3.67%	3.67%	0.33%	0.33%	6.00%	6.00%
Homosexuality	35.44%	57.67%**	47.33%**	60.67%**	46.56%*	55.44%**
Patriotism	13.67%	6.33%	16.67%*	10.00%	6.00%	39.33%**
Politicization	37.00%*	37.00%*	43.00%*	63.00%**	14.50%	34.50%*
Postmaterialism	21.00%	51.00%**	23.50%	63.50%**	29.00%*	49.00%**
Respect	18.00%*	12.00%	23.00%**	7.00%	34.50%**	15.50%
Trust	89.00%**	11.00%	22.00%	18.00%	54.00%**	6.00%
Average	28.36%	32.89%	22.87%	33.37%	29.14%	34.47%

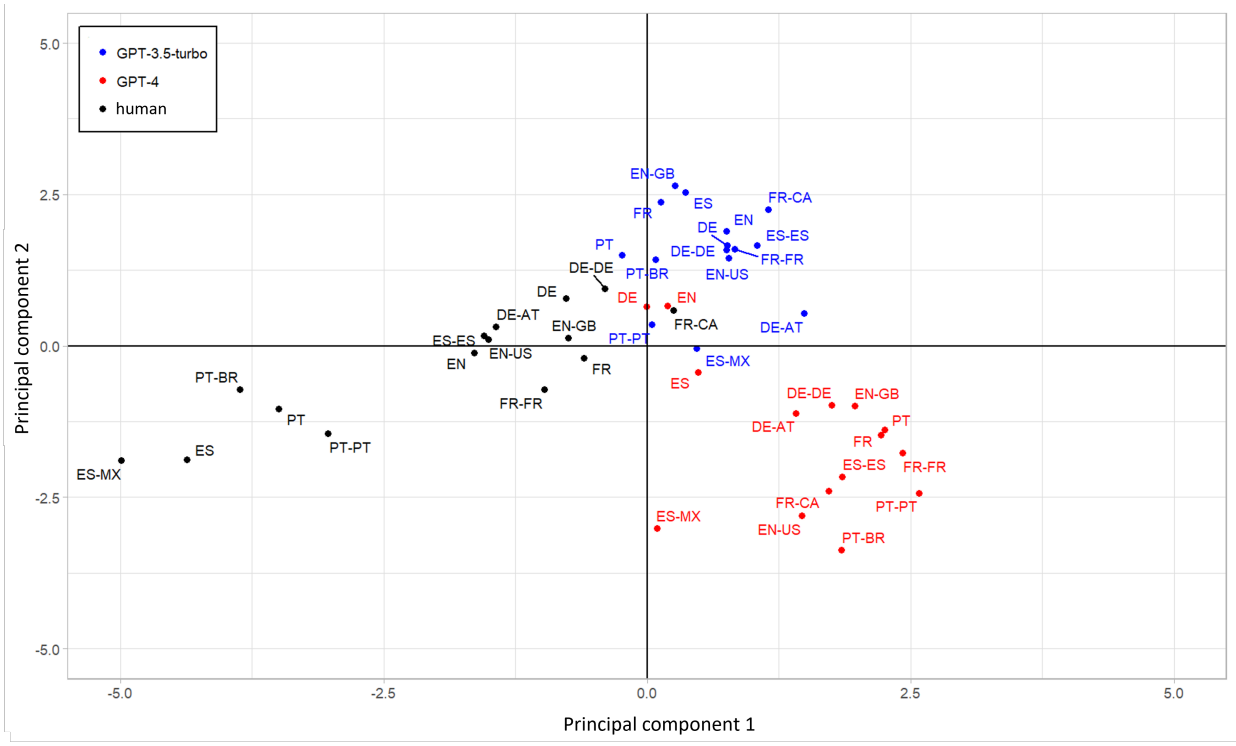

Fig. 3. Cultural Map.

Respect. Conversely, GPT-4 (50%) may encounter challenges in conveying culturally appropriate responses for the items of *Abortion*, *Autonomy*, *Homosexuality*, *Politicization*, and *Postmaterialism* within PT-PT. In the evaluation of PT-BR, GPT-3.5-turbo (29.14%) demonstrates closer alignment with human behavior compared to GPT-4 (34.47%). However, significant disparities reveal that GPT-3.5-turbo (30%) does not fully capture the nuances of the five items *Abortion*, *Autonomy*, *Faith*, *Homosexuality*, *Postmaterialism*, *Respect*, and *Trust*, while GPT-4 (40%) struggles with *Abortion*, *Autonomy*, *Homosexuality*, *Patriotism*, *Politicization*, and *Postmaterialism*.

B. Cultural Map

Figure 3 shows the cultural map with the WVS scores of the language areas and the subcultures transferred in a two-dimensional space. The factor values of the first principal component are displayed on the x-axis, while the factor values of the second principal component are shown on the y-axis. The black dots visualize the scores achieved by the human WVS participants of the tested language areas and subcultures (*human*). The blue dots indicate the scores obtained by GPT-3.5-turbo. The scores of GPT-4 are shown with the red dots.

All except one (FR-CA) *human* scores (black dots) are distributed exclusively in the upper and lower left quadrants. In contrast, the scores of GPT-3.5-turbo (blue dots) are mainly

located in the upper right quadrant, while GPT-4 (red dots) can mainly be found in the lower right quadrant. Further, we notice that the language areas and subcultures, based on the WVS data from *human*, are generally further apart from each other. However, in the case of the chatbots, the individual dots are closer to each other. Calculating the average Euclidean distance between the dots confirms this observation: The average distance between the *human* dots is 2.20, while for GPT-3.5-turbo, it is 1.23, and for GPT-4, it is 1.85.

V. Conclusion & Future Work

We explored the cultural relevance of GPT-3.5-turbo and GPT-4 chatbots in non-English languages across five linguistic regions and ten subcultures. Our results demonstrate that chatbots do not always behave in culturally appropriate ways. By benchmarking their performance against cultural dimensions from Inglehart and Welzel's framework and WVS data, we found that GPT-3.5-turbo generally outperforms GPT-4, particularly in the German context. However, Spanish-speaking Mexico and Portuguese-speaking Brazil exhibit lower cultural congruence compared to other regions. The overall poorer performance of GPT-4 could be due to OpenAI's debiasing efforts [30]. To contribute to the improvement of LLM-based chatbots' culturally appropriate behavior, we share the chatbots' answers with the research community[5].

While our paper focused on raising awareness that subcultural differences must be considered when using chatbots, future work should investigate how to address this problem, e.g. by fine-tuning the chatbot or providing specific instructions. Furthermore, it could be investigated how debiasing efforts counteract cultural alignment. Future work could also include extending our research to other language areas and subcultures and investigating the impact of the cultural appropriateness of chatbots for specific use cases. Using GPT-3.5-turbo and GPT-4, we have shown a way how cultural appropriateness can be analyzed at a topic level. This method can be used to examine other state-of-the-art chatbots.

Acknowledgment

This research was supported within the framework of the internal initial funding by the IU International University of Applied Sciences (*IU Incubator*) for the period from October 2023 to September 2025.

Ethical Impact Statement

No user study was conducted and no private data was collected. Only previously published numerical values from the WVS were used and new numerical values were produced by our evaluations of the chatbot responses. The items or subject areas of the World Value Survey may not correspond with the opinions of some readers and it could also be that—as with other cultural surveys—certain cultural aspects are not covered in the WVS questions or need to be asked differently in relation to certain cultures.

[5]https://github.com/iu-ai-research/CrossCulture_LLM

References

[1] C. Pelau, D.-C. Dabija, and I. Ene, "What Makes an AI Device Human-Like? The Role of Interaction Quality, Empathy and Perceived Psychological Anthropomorphic Characteristics in the Acceptance of Artificial Intelligence in the Service Industry," *Computers in Human Behavior*, vol. 122, p. 106855, 2021.

[2] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, and Q. V. Le, "Towards a Human-Like Open-Domain Chatbot," *ArXiv Preprint ArXiv:2001.09977*, 2020.

[3] M. Dibitonto, K. Leszczynska, F. Tazzi, and C. M. Medaglia, "Chatbot in a Campus Environment: Design of LiSA, a Virtual Assistant to Help Students in Their University Life," in *Human-Computer Interaction. Interaction Technologies: 20th International Conference, HCI International 2018, Las Vegas, NV, USA, July 15–20, 2018, Proceedings, Part III 20*. Springer, 2018, pp. 103–116.

[4] D. Arteaga, J. Arenas, F. Paz, M. Tupia, and M. Bruzza, "Design of Information System Architecture for the Recommendation of Tourist Sites in the City of Manta, Ecuador through a Chatbot," in *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, 2019, pp. 1–6.

[5] C. Falala-Séchet, L. Antoine, I. Thiriez, and C. Bungener, "OWLIE: A Chatbot that Provides Emotional Support for Coping With Psychological Difficulties," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 2019, pp. 236–237.

[6] V. Taecharungroj, ""What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter," *Big Data and Cognitive Computing*, vol. 7, no. 1, p. 35, 2023.

[7] T. Bianch, "Distribution of OpenAI.com worldwide traffic in January 2023, by region," 2023. [Online]. Available: https://www.statista.com/statistics/1386581/openaicom-traffic-distribution-region

[8] OpenAI, "Is ChatGPT Biased? - Bias in ChatGPT," 2023. [Online]. Available: https://help.openai.com/en/articles/8313359-is-chatgpt-biased

[9] W. Wang, W. Jiao, J. Huang, R. Dai, J. tse Huang, Z. Tu, and M. R. Lyu, "Not All Countries Celebrate Thanksgiving: On the Cultural Dominance in Large Language Models," 2024.

[10] Y. Cao, L. Zhou, S. Lee, L. Cabello, M. Chen, and D. Hershcovich, "Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study," in *The First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, S. Dev, V. Prabhakaran, D. Adelani, D. Hovy, and L. Benotti, Eds. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 53–67. [Online]. Available: https://aclanthology.org/2023.c3nlp-1.7

[11] R. I. Masoud, Z. Liu, M. Ferianc, P. Treleaven, and M. Rodrigues, "Cultural Alignment in Large Language Models: An Explanatory Analysis Based on Hofstede's Cultural Dimensions," 2023.

[12] E. Hall, *Beyond Culture*, ser. Anchor Books. Knopf Doubleday Publishing Group.

[13] G. Hofstede, *Culture's Consequences: International Differences in Work-Related Values*, ser. Cross Cultural Research and Methodology. SAGE Publications, 1984.

[14] G. Hofstede and M. Minkov, "Values Survey Module 2013 Manual," May 2013. [Online]. Available: https://geerthofstede.com/wp-content/uploads/2016/07/Manual-VSM-2013.pdf

[15] L. Allison, C. Wang, and J. Kaminsky, "Religiosity, Neutrality, Fairness, Skepticism, and Societal Tranquility: A data Science Analysis of the World Values Survey," *PLOS ONE*, vol. 16, no. 1, pp. 1–22, 01 2021.

[16] S. H. Schwartz, "An Overview of the Schwartz Theory of Basic Values," *Online Readings in Psychology and Culture*, vol. 2, p. 11, 2012.

[17] P. Bréchon and F. Gonthier, *European Values: Trends and Divides Over Thirty Years*, ser. European values studies. Brill, 2017.

[18] R. Inglehart, *Human Values and Social Change: Findings from the Values Surveys*, ser. Brill Book Archive Part 1, ISBN: 9789004472495. Brill, 2003.

[19] W. Thom, "Culture, Leadership, and Organizations: The GLOBE Study of 62 Societies [review] / House, R. J., Hanges, P.J., & Javidan, M., Eds." *The Journal of Applied Christian Leadership*, vol. 1, pp. 55–71, 2006. [Online]. Available: https://digitalcommons.andrews.edu/jacl/vol1/iss1/6

[20] M. Minkov and V. Blagoev, "What do project globe's cultural dimensions reflect? an empirical perspective," *Asia Pacific Business Review*, vol. 18, no. 1, pp. 27–43, 2012.

[21] S. Narang and A. Chowdhery, "An Overview of Bard: An Early Experiment with Generative AI," October 2023. [Online]. Available: https://ai.google/static/documents/google-about-bard.pdf

[22] W. Messner, T. Greene, and J. Matalone, "From Bytes to Biases: Investigating the Cultural Self-Perception of Large Language Models," 2023.

[23] C. Lindahl and H. Saeid, "Unveiling the Values of ChatGPT : An Explorative Study on Human Values in AI Systems," 2023.

[24] W. Wang, W. Jiao, J. Huang, R. Dai, J.-t. Huang, Z. Tu, and M. R. Lyu, "Not All Countries Celebrate Thanksgiving: On the Cultural Dominance in Large Language Models."

[25] Natalie. (2023) What is ChatGPT? [Online]. Available: https://help.openai.com/en/articles/6783457-what-is-chatgpt

[26] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," *CoRR*, vol. abs/2005.14165, 2020.

[27] S. Mitrović, D. Andreoletti, and O. Ayoub, "ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-Generated Text," *arXiv preprint arXiv:2301.13852*, 2023.

[28] D. Patel and G. Wong, "GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE," https://github.com/llv22/gpt4_essay/blob/master/GPT-4-4.JPG, July 2023, accessed: 30-09-2023.

[29] D. Yalalov and D. Myakin, "GPT-4's Leaked Details Shed Light on its Massive Scale and Impressive Architecture," *Metaverse Post*, July 2023. [Online]. Available: https://mpost.io/gpt-4s-leaked-details-shed-light-on-its-massive-scale-and-impressive-architecture/#gpt-4s-massive-parameters-count

[30] OpenAI, "GPT-4," March 2023. [Online]. Available: https://openai.com/research/gpt-4

[31] R. Inglehart and C. Welzel, *Modernization, Cultural Change, and Democracy*. Cambridge University Press, 2005.

[32] H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50 – 60, 1947. [Online]. Available: https://doi.org/10.1214/aoms/1177730491