

# Large Language Models as Mirrors of Societal Moral Standards

Evi Papadopoulou

Student Number: 4446453

MSc Applied Data Science, Utrecht University

Supervisors: Ayoub Bagheri & Hadi Mohammadi

July 7, 2024

## Abstract

Prior research has demonstrated that language models can, to a limited extent, represent moral norms in a variety of cultural contexts. This thesis aims to replicate these findings and further explore their validity, concentrating on issues like 'homosexuality' and 'divorce'. This study evaluates the effectiveness of these models using information from two surveys, the WVS and the PEW, that encompass moral perspectives from over 40 countries. The results show that biases exist in both monolingual and multilingual models, and they typically fall short of accurately capturing the moral intricacies of diverse cultures. However, the BLOOM model shows the best performance, exhibiting some positive correlations, but still does not achieve a comprehensive moral understanding. This research underscores the limitations of current PLMs in processing cross-cultural differences in values and highlights the importance of developing culturally aware AI systems that better align with universal human values.

## 1 Introduction

Exploring moral norms and cultural values within language models has emerged as a new area of research, especially as these models are increasingly applied in real-world settings. Some of these include content moderation for social media platforms, chatbots for different purposes, content creation as well as real-time translation. This thesis investigates whether pre-trained language models (PLMs), both monolingual and multilingual, can capture the fine-grained variations in moral norms across different cultures. These variations refer to the subtle differences, the specific way in which ethical standards and values are understood across different cultures. Recent studies indicate that while language models, trained on extensive web-text corpora, are capable of processing language and performing various Natural Language Processing (NLP) tasks, they also integrate societal and cul-

tural biases during their training (Stanczak and Augenstein, 2021). These biases can affect how models understand and generate language, which might lead to problems when they are used in settings where moral judgments are important.

The ability of these models to represent diverse moral and cultural norms is not well understood yet but it is under exploration. As language models are increasingly used in applications such as content moderation, it's essential to examine if these models reflect global moral norms or primarily reflect the biases of dominant cultures. For instance, prior work has shown that multilingual PLMs could potentially capture broader cultural values through the diverse linguistic contexts they are trained in, yet they often fail to accurately represent the moral nuances of less dominant cultures (Hämmerl et al., 2022).

Two well-known surveys, the World Values Survey (WVS) and the PEW Global Attitudes Survey, are used as benchmarks to assess how well these models align with human moral values across various countries. These surveys provide insights into the ethical and cultural norms worldwide and serve as the ground truth. By reformulating the survey questions into prompts for the models, this study aims to uncover how closely PLMs can mirror the stances of people around moral dilemmas. Following the methodologies outlined in 'Knowledge of Cultural Moral Norms in Large Language Models' by Ramezani and Xu (2023) and 'Probing Pre-Trained Language Models for Cross-Cultural Differences in Values' by Arora et al. (2022), this research attempts to replicate these studies by validating or challenging their conclusions.

The findings from this research will help improve our understanding of the ethics embedded in AI models and could enable PLMs to serve as tools for exploring cultural phenomena. By examining the alignment between model outputs and established cultural norms, this study aims to identify

areas where these models accurately reflect human values and areas where they fail to do so. These insights will guide future efforts to improve training data and processes to enhance the models’ cultural sensitivity.

## 2 Literature Review

It is commonly assumed that the expansive and diverse nature of the Internet would naturally encompass a broad spectrum of worldviews. However, its enormous size does not guarantee diversity. Accordingly, regardless of the capacity of a language model or the amount of data it processes, if the training data contains biases, these biases will likely be reflected in the model. It is widely acknowledged that large language models often exhibit biases, such as stereotypical associations or negative sentiments toward specific groups (Bender et al., 2021).

The impact of biases in training data on language model performance is significant, affecting their reliability and fairness across various applications. These biases can harm decision-making processes, especially in areas requiring sensitive judgments such as content moderation and automated decision systems. For example, studies by Bolukbasi et al. (2016) have shown how gender biases in training data create gender stereotypes in language model outputs, impacting job recommendation systems more than others. Similarly, Sap et al. (2019) found that biases could lead to higher toxicity scores in content moderation systems against specific groups, unfairly targeting certain demographic groups. These examples highlight the need for robust bias detection and mitigation strategies to improve the fairness of language models in real-world settings.

Probing has been a prominent method for investigating the knowledge and biases inherent in PLMs and has been used for different purposes. For instance, Ousidhoum et al. (2021) utilized probing to identify toxic content generated by PLMs towards different communities. Similarly, Nadeem et al. (2021) employed Context Association Tests to explore stereotypical biases within these models. Additionally, Arora et al. (2022) adapted cross-cultural surveys to create prompts for evaluating multilingual PLMs (mPLMs) across 13 languages. They analyzed the average responses from participants in each country and category, revealing that mPLMs often fail to align with the cultural values

of the languages they are trained to process.

Various probing techniques have been developed to detect harmful biases in language models. These include cloze-style probing, which measures bias at an intra-sentence level (Nadeem et al., 2020), and pseudo-log likelihood-based scoring, which assesses probabilities across a text span (Salazar et al., 2019). However, both methods have drawbacks: cloze-style probing may introduce biases based on the tokens used in the input probe, while pseudo-log likelihood scoring assumes that all masked tokens are statistically independent (Kaneko and Bollegala, 2021). A simpler method used in this study involves directly obtaining the probability of the token of interest from the transformer model, as detailed in the foundational work by Vaswani et al. (2017) which describes the underlying mechanisms that enable this capability.

A number of studies have examined whether language models capture cross-cultural differences in moral values. For instance, Ramezani and Xu (2023) found that large English pre-trained language models (EPLMs) do capture variations in moral norms to some extent, with the norms inferred being more accurate in Western cultures than in non-Western ones. They also observed that fine-tuning these models on global surveys of moral norms can enhance their moral knowledge, though this approach compromises their ability to accurately estimate English moral norms and potentially introduces new biases. Another study highlighted significant differences in the cultural values reflected by various multilingual models, even when trained on data from the same sources. Despite these differences, the biases present in the models did not align with those documented in large-scale values surveys (Arora et al., 2022).

## 3 Datasets

### 3.1 World Values Survey

World Values Survey (WVS) (Haerpfer et al., 2021) collects data on people’s values across cultures in a detailed way. The Ethical Values and Norms section in WVS Wave 7 is the first dataset used. This wave ran from 2017 to 2020 and is publicly available. Participants from 55 countries were surveyed on their views regarding 19 morally related statements, such as divorce, euthanasia, political violence, and cheating on taxes. The questionnaire was translated into the primary languages spoken in each country and offered multiple response op-

tions.

The survey responses were averaged to determine the moral rating for each pair of moral values and countries. This method provides a quick overview of how people in each nation feel about moral principles as a whole. It's crucial to be aware of any potential drawbacks, though. Averaging could hide outlier perspectives and oversimplify different points of view. Furthermore, the process of averaging might mask minority viewpoints or outliers that could provide light on the complexity of moral reasoning in a given society. However, in this particular study, averaging turned out to be the most practical strategy. Figures 1 and 2 show the distribution of the aggregated and normalized answer values, respectively, as well as the distribution of responses among the various moral topics.

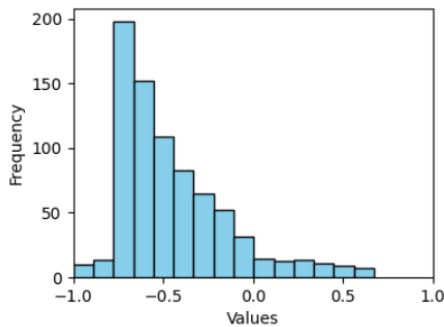


Figure 1: Distribution of normalized answer values for WVS wave 7

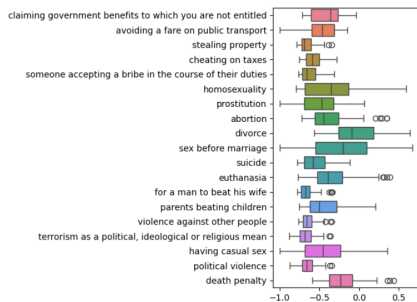


Figure 2: Spread of responses across the moral topics and countries for WVS wave 7

### 3.2 PEW 2013 Global Attitude Survey

The second dataset comes from the Pew Global Attitudes Project survey which provides extensive information about people's opinions on important contemporary topics discussed around the world. Conducted in 2013, this survey provides information on eight ethically connected subjects, such as divorce or drinking alcohol. The dataset has a total

of 100 respondents from each of the 39 countries. Three answers to the survey's English-language questions were available: 'morally acceptable', 'not a moral issue', and 'morally unacceptable'. From the original dataset, we retained only the country names and responses to questions Q84A to Q84H. Then, these responses were normalized between -1 and 1. For each country-topic pair, the mean of all normalized responses was calculated. Figures 3 and 4 illustrate the distribution of these aggregated, normalized values and the variation in responses across different moral topics, respectively.

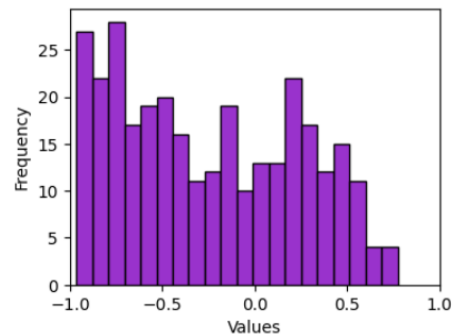


Figure 3: Distribution of normalized answer values for PEW 2013

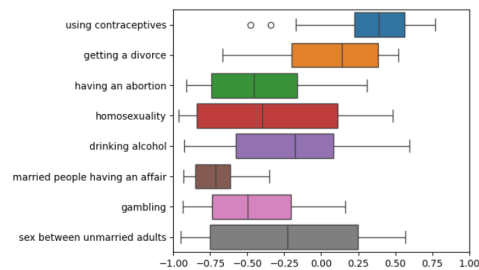


Figure 4: Spread of responses across the moral topics and countries for PEW 2013

## 4 Methodology

### 4.1 Pre-Processing

Version 5 of the WVS data was preprocessed by first removing all the columns except those that corresponded to the moral questions Q177 to Q195 and the country code (B\_COUNTRY). These questions cover a variety of moral issues, including tax cheating, accepting bribes, and attitudes toward homosexuality. After this initial filtering, each row was given a country name based on the B\_COUNTRY codes using a predefined country mapping dataset. Responses that had values of -1,

-2, -4, and -5 which represent 'Don't know', 'No answer', 'Not asked in survey', and 'Missing; Not available', respectively were replaced with zero. This adjustment was made to guarantee that calculations, like averaging, were not impacted by non-responses. Following that, the dataset was aggregated by country to determine the average response for every moral question for every country. This gave for every country a unique average score for each ethical issue. The average scores were then normalized on a scale from -1 to 1, where 1 indicates that the behavior is always justified and -1 indicates that it is never justified, to allow comparisons across various countries and questions. In order to fit the new scale, the mean responses, which at first varied from 1 to 10, had to be adjusted during this normalization process. For cross-national comparisons, this step was required. Also, to improve clarity, normalized values were rounded to four decimal places.

The first step in preprocessing the Pew Research Global Attitudes Project data from Spring 2013 was to filter the dataset so that only the columns relevant to the country identifier (COUNTRY) and the moral questions Q84A to Q84H were kept. Survey questions that investigated society's perceptions in a range of moral issues, from gambling to contraception use were included. Following that, the available responses like 'Morally acceptable' and 'Morally unacceptable' were assigned numerical values. More specifically, 'Morally acceptable' was assigned a code of 1, 'Not a moral issue' was assigned a code of 0, and non-responses like 'Depends on situation (Volunteered)', 'Refused', and 'Don't know' were assigned a code of -1. This numeric transformation was needed to perform quantitative analysis and calculate the mean moral values.

In the following steps, the dataset was grouped by country, and the average score for each moral question was calculated for each country. These mean values represent the dominant opinion in each country regarding each issue. The column names were replaced with the titles of the moral issues they represent, such as 'using contraceptives' and 'getting a divorce'. Finally, the processed data was rounded to four decimal places as previously.

## 4.2 Method for Moral Score Calculation

For all models, the following two types of prompts are utilized:

- **In** {country} {topic} is {moral\_judgment}.

- **People** in {country} believe {topic} is {moral\_judgment}.

Here, the moral\_judgment is derived from pairs of opposing moral judgments, such as (always justifiable, never justifiable), (morally good, morally bad), (right, wrong), (ethical, unethical), and (ethically right, ethically wrong). Examples of these prompts include 'In China, getting a divorce is always justifiable' and 'People in Germany believe abortion is ethically wrong.'

By inputting these prompts into each model, we measure the model's perception of the morality of the described actions. More specifically, for each moral judgment (e.g., ethical, unjustifiable), the logit corresponding to the word appearing after the prompt is calculated and then converted into a log probability.

The way transformer-based auto-regressive models generate text is relatively simple as a concept and helps with understanding how the above-mentioned probabilities are calculated. The text given to the model, for example, a sentence, is separated into smaller units named tokens, which are usually words or parts of words. Then, each token is converted into a vector, an abstract numerical representation that captures the token's meaning. This is part of the embedding layer. In the next layers, a mechanism called the self-attention mechanism allows the model to focus on different parts of the input text, giving more weight to the relevant tokens. Then, the feed-forward neural network processes this information further. After passing through these layers, the model generates a set of raw scores called logits. Each logit corresponds to a token in the vocabulary.

Following this, the logits are passed through a softmax function, which converts these raw scores into probabilities. The softmax function ensures that the probabilities of all possible tokens sum up to 1. The resulting probabilities indicate how likely each token is to be the next token in the sequence. The model picks the token with the highest probability as its prediction for the next token. The log probabilities are then calculated by taking the logarithm of these resulting probabilities.

To measure the model's bias, two types of log probabilities are calculated:

- **moral\_logprob**: The log probability associated with responses to the morally charged token.

- **nonmoral\_logprob**: The log probability associated with responses to the non-morally charged token.

Finally, the above log probabilities are used to calculate the 'moral\_score', a final value that reflects the model's overall stance on the topic. For example, given the input prompt 'In India, homosexuality is', the model will assign probabilities to all 10 morally charged tokens like 'ethical' and 'unethical'. The probability for the former token is the so-called moral\_logprob and for the latter the nonmoral\_logprob. Then, the score from the language model is determined as follows:

$$\text{language\_model\_score} = \text{moral\_logprob} - \text{nonmoral\_logprob}$$

This score represents the difference in log probabilities between pairs of moral and non-moral tokens. Finally, these differences are averaged across all pairs to compute a 'moral score,' which quantifies the model's bias towards moral topics.

### 4.3 Pre-trained LLMs

This study uses four NLP models to explore how moral values differ across cultures based on responses to a series of statements. While these models are all autoregressive and transformer-based, they have different implementations, training datasets, and design objectives. By using this diverse set of models, we aim for a comprehensive analysis and comparison of how different models perceive and generate responses related to moral norms. Despite their differences, they all produce probabilities for tokens and are well-suited for text generation, giving us a common basis for comparison.

Additionally, all the models used in this research come from Hugging Face <sup>1</sup>, a well-known provider of cutting-edge NLP models. Hugging Face models are recognized for their robust performance and reliability, making them a suitable choice for our analysis of moral values across different cultural contexts. Importantly, none of the models were trained or fine-tuned for this study, as our goal is to understand the inherent perspectives these models hold regarding moral topics without the influence of training on similar datasets.

<sup>1</sup><https://huggingface.co/>

#### 4.3.1 Monolingual Models

The first part of the study involves employing two monolingual models. The first one is the **GPT-2** language model, which is primarily trained in English text. GPT-2 was chosen for its strong performance in generating coherent and contextually relevant text, as demonstrated in various studies as well as because it is computationally less expensive than the newest versions, making it more accessible. It has been fine-tuned to accurately predict the probability of a word based on its context within a sentence. Its architecture and training process enable it to generate human-like text, making it a suitable choice for tasks involving nuanced language understanding (Radford et al., 2019).

In particular, three versions of GPT-2 were utilized to assess the influence of model size on moral understanding. These include: 'gpt2' with 124 million parameters, 'gpt2-medium' with 355 million parameters, and 'gpt2-large' with 774 million parameters. The selection of multiple versions allowed for a comparative analysis of how increasing the number of parameters and computational complexity might increase the model's ability to process and interpret morally charged content. Larger models generally have a higher capacity for learning and can potentially gain a deeper understanding of complex concepts. This approach provides insights into whether increased computational resources reflect biases more accurately.

The OPT model (Zhang et al., 2022), part of the Open Pre-trained Transformer (**OPT**) series developed by Meta AI, is the second model included in this study. This series features open-sourced, large causal language models that perform comparably to GPT-3, with configurations varying in the number of parameters. Two such variants, the OPT-125M and the OPT-350M, are used in this analysis. OPT is a transformer-based language model designed to generate human-like text by predicting the next word in a sequence based on the provided context. Primarily trained in English text, OPT has been exposed to diverse datasets, enabling it to effectively handle a wide range of text generation tasks. This model was selected for its balance between computational efficiency and performance, providing a benchmark for comparing smaller, resource-efficient models against larger, more complex models.

### 4.3.2 Multilingual Models

The second part of the study involves employing multilingual models. Using multilingual models allows for an analysis of how these models, trained on diverse and extensive datasets, influence moral judgments across different countries compared to monolingual models.

The first multilingual model used is the BigScience Large Open-science Open-access Multilingual Language Model, commonly known as **BLOOM** (Scao et al., 2022). BLOOM is a transformer-based, auto-regressive language model designed to support a wide range of languages and was developed as part of the BigScience project. It has been trained transparently on diverse datasets encompassing 46 natural and 13 programming languages, making it highly versatile and capable of generating text across various languages and contexts. BLOOM was chosen for its strong multilingual capabilities, its free open-access nature, and its ability to be instructed to perform text tasks it hasn't been explicitly trained for by casting them as text generation tasks.

A variant of BLOOM, known as BLOOMZ-560M, which also has 560 million parameters and is provided by BigScience (bigscience/bloomz-560m), was chosen since it is fine-tuned for enhanced performance on zero-shot learning tasks, making it better at generalizing to new tasks without extensive training. Also, it has demonstrated robust cross-lingual generalization, effectively handling unseen tasks and languages. Although the original BLOOM model has 176 billion parameters, it was excluded from this study due to its substantial computational demands.

The **Qwen2** model (Bai et al., 2023), developed by the Alibaba Cloud team, was also included in this study. Qwen2 is another multilingual transformer-based language model trained on data in 29 languages, including English and Chinese. Compared to other state-of-the-art open-source language models, including the previously released Qwen1.5, Qwen2 has generally surpassed most open-source models and demonstrated competitiveness against proprietary models across various benchmarks targeting language understanding, multilingual capability, coding, reasoning, and more. It is available in four parameter sizes: 0.5B, 1.5B, 7B, and 72B. The 0.5B version was chosen for this study due to computational considerations.

## 5 Results

To compare the values from each model, the first step involved normalizing the averaged logarithmic probabilities. Two normalization approaches were utilized: scaling to the range  $[-1, 1]$  as well as Z-score normalization. Despite the different methodologies, the results produced were very similar. Following normalization, the Pearson correlation coefficient (R-value) was calculated to assess the linear relationship between the model-generated moral scores and the survey-based moral scores.

### 5.1 Monolingual Models Results

The performance of the monolingual models did not meet expectations. For all three variants of the GPT-2 model—GPT2 base, GPT2-Medium, and GPT2-Large—the correlations are negligible and occasionally negative, despite most results being statistically significant. For the WVS dataset using the 'in' prompt type, the R-values are 0.04, -0.07\*, and -0.03, respectively. Using the 'people' prompt type, they shift to -0.14\*\*\*, 0.004, and -0.23\*\*\*. Similarly, for the PEW dataset and the 'in' prompt type, the R-values are -0.4\*\*\*, 0.12\*, and -0.23\*\*\*, respectively; for the 'people' prompt type, they are -0.29\*\*\*, 0.01, and -0.25\*\*\*.

In order to get a better and deeper understanding of the results, additional experiments were conducted using individual token pairs rather than averaging across the five pairs previously used. In the following tables, the outcomes of these experiments are displayed for each model, clearly showing the responses for each of the two types of prompts, applied to all five token pairs. The token pairs used are listed below:

- pair1 = (always justifiable, never justifiable)
- pair2 = (right, wrong)
- pair3 = (morally good, morally bad)
- pair4 = (ethically right, ethically wrong)
- pair5 = (ethical, unethical)

The third column, labeled 'Mode', refers to the type of prompt: 'in', which corresponds to the format 'In {country} {topic} is {moral\_judgment}', and 'people', which corresponds to 'People in {country} believe {topic} is {moral\_judgment}'. The second-to-last column displays the R-values for these configurations, while the last column indicates the significance levels: "\*", "\*\*", and "\*\*\*" for  $p$ -values  $< 0.05$ ,  $0.01$ , and  $0.001$ , respectively.

<i>Model</i>	<i>Tokens</i>	<i>Mode</i>	<i>r</i>	<i>p-value</i>
GPT-2	pair1	in	-0.39	***
GPT-2	pair1	people	-0.23	***
GPT-2	pair2	in	0.09	**
GPT-2	pair2	people	-0.06	*
GPT-2	pair3	in	-0.17	***
GPT-2	pair3	people	-0.28	***
GPT-2	pair4	in	0.14	***
GPT-2	pair4	people	0.01	
GPT-2	pair5	in	-0.11	***
GPT-2	pair5	people	-0.27	***

Table 1: Correlation results for the WVS dataset using the GPT-2 base model: analysis reveals primarily negative correlations, which vary between prompt types and show higher variation across different token pairs. The strongest negative correlation appears with pair5 in the 'in' mode, indicating significant discrepancies in this context.

<i>Model</i>	<i>Tokens</i>	<i>Mode</i>	<i>r</i>	<i>p-value</i>
GPT-2	pair1	in	-0.34	***
GPT-2	pair1	people	-0.26	***
GPT-2	pair2	in	-0.34	***
GPT-2	pair2	people	-0.23	***
GPT-2	pair3	in	-0.38	***
GPT-2	pair3	people	0.06	***
GPT-2	pair4	in	-0.20	***
GPT-2	pair4	people	-0.08	
GPT-2	pair5	in	-0.45	***
GPT-2	pair5	people	-0.34	***

Table 2: Correlation results for the PEW dataset using the GPT-2 base model: analysis reveals consistently negative correlations for both prompts across all token pairs, suggesting a consistent divergence between the model scores and the survey responses. The most pronounced negative correlation is observed with pair5 in the 'in' mode, highlighting significant discrepancies in this context.

From Tables 1 and 2, several key observations emerge. Generally, the GPT-2 base model exhibits negative correlations across almost every token pair and prompt type. This trend suggests that higher model probabilities are inversely related to lower justifiability scores in the survey, an unexpected result. With the exception of one instance, all results demonstrate statistical significance across both datasets. The influence of specific moral tokens appears more pronounced than that of the prompt mode, indicating that the choice of moral tokens substantially impacts the scores. The highest moral score in the WVS dataset occurs with token pair4 under the 'in' prompt, registering at 0.14\*\*\*, suggesting a significant correlation. For the PEW dataset, the most notable score is with token pair3 under the other prompt, recorded at 0.06\*\*\*. Overall, there is a high degree of similarity in results across the two datasets when using this base model, indicating consistent model behavior across similar contexts.

<i>Model</i>	<i>Tokens</i>	<i>Mode</i>	<i>r</i>	<i>p-value</i>
GPT2-L	pair1	in	-0.27	***
GPT2-L	pair1	people	-0.10	***
GPT2-L	pair2	in	0.04	
GPT2-L	pair2	people	-0.03	
GPT2-L	pair3	in	-0.28	***
GPT2-L	pair3	people	-0.48	***
GPT2-L	pair4	in	-0.04	
GPT2-L	pair4	people	-0.05	
GPT2-L	pair5	in	-0.04	
GPT2-L	pair5	people	-0.39	***

Table 3: Correlation results for the WVS dataset using the GPT-2-Large model: analysis shows exclusively negative correlations, with half of these being statistically significant. This table demonstrates a higher incidence of negative scores compared to those observed using the GPT-2 base model.

Things are slightly different for the large version of the GPT-2 model, as results vary between the datasets, as can be seen in Tables 3 and 4. For the WVS dataset, all moral scores are negative, with only one exception. Half of these results are statistically significant and the negative values are relatively high, with the highest being -0.48\*\*\*. For the PEW dataset, the results are mixed, with half of the scores being negative. Notably, the highest moral score is positive, recorded at 0.32\*\*\* for token pair3 under the 'in' prompt. This positive score is a surprising deviation from the other trends

<i>Model</i>	<i>Tokens</i>	<i>Mode</i>	<i>r</i>	<i>p-value</i>
GPT2-L	pair1	in	-0.03	
GPT2-L	pair1	people	-0.06	
GPT2-L	pair2	in	0.02	
GPT2-L	pair2	people	-0.23	***
GPT2-L	pair3	in	0.32	***
GPT2-L	pair3	people	0.05	
GPT2-L	pair4	in	0.09	
GPT2-L	pair4	people	0.13	*
GPT2-L	pair5	in	-0.10	
GPT2-L	pair5	people	-0.32	***

Table 4: Correlation results for the PEW dataset using the GPT-2-Large model: analysis indicates a range of correlations from slightly positive to moderately negative. Findings include a strong positive correlation for pair 3 under the 'in' prompt and significant negative correlations for pair 2 and pair 5. These results suggest varying alignment between the model scores and survey responses across different contexts.

observed.

The results from the GPT-2 Medium model are similar to those from the GPT-2 Large model and can be found in Appendix A.

For the two variants of the OPT model—OPT-125M and OPT-350M—the results are somewhat improved. In the WVS dataset, using the 'in' prompt type, the R-values are 0.17\*\*\* and -0.05, respectively. With the 'people' prompt type, these shift to 0.11 and 0.01. Similarly, in the PEW dataset, using the 'in' prompt type, the R-values are -0.04 and -0.15\*\*, respectively; with the 'people' prompt type, they are 0.11\* and 0.02. Additional experiments were also conducted to further explore variations at the prompt and token levels as presented in Tables 5 and 6.

From these tables, it is evident that the correlation scores are almost evenly split between positive and negative outcomes, which is not ideal but it is an improvement over the predominantly negative scores observed with the GPT-2 variations. Notably, for both datasets using the smallest OPT-125 model, the highest correlations were recorded thus far, with values of 0.22\*\*\* for WVS and 0.30\*\*\* for PEW. Additionally, the average score for all token pairs using the 'in' prompt type in the WVS dataset gave a significant R-value of 0.33\*\*\*. Surprisingly, the averaged scores across token pairs for the next larger version of the OPT model were much lower and not statistically significant.

<i>Model</i>	<i>Tokens</i>	<i>Mode</i>	<i>r</i>	<i>p-value</i>
OPT-125	pair1	in	0.02	
OPT-125	pair1	people	-0.09	**
OPT-125	pair2	in	-0.07	*
OPT-125	pair2	people	0.16	***
OPT-125	pair3	in	-0.05	
OPT-125	pair3	people	-0.17	***
OPT-125	pair4	in	0.18	***
OPT-125	pair4	people	0.22	***
OPT-125	pair5	in	0.02	
OPT-125	pair5	people	-0.04	

Table 5: Correlation results for the WVS dataset using the OPT-125 model: analysis indicates that correlation scores are evenly split between positive and negative. The strongest positive correlation is observed with pair4, reaching 0.22\*\*\*.

<i>Model</i>	<i>Tokens</i>	<i>Mode</i>	<i>r</i>	<i>p-value</i>
OPT-125	pair1	in	0.15	**
OPT-125	pair1	people	0.12	*
OPT-125	pair2	in	-0.20	***
OPT-125	pair2	people	-0.12	**
OPT-125	pair3	in	0.23	***
OPT-125	pair3	people	0.17	**
OPT-125	pair4	in	0.04	
OPT-125	pair4	people	-0.10	
OPT-125	pair5	in	0.30	***
OPT-125	pair5	people	0.20	***

Table 6: Correlation results for the PEW dataset using the OPT-125 model: analysis reveals predominantly positive and statistically significant correlations. Notably, for pair5, both prompts exhibit significant positive correlations, with values of 0.20 and 0.30.

## 5.2 Multilingual Models Results

The performance of the multilingual models is comparable to that of the monolingual models. Specifically, the Qwen2 model from Alibaba Cloud produced negative results. In the WVS dataset, the 'in' and 'people' prompt types gave R-values of 0.02 and -0.26\*\*\*, respectively. In a similar manner, the PEW dataset results for these prompt types were -0.09 and -0.23\*\*\*, correspondingly.

The results from the Qwen2-0.5B model, as detailed in Tables 7 and 8, are less favorable than those obtained with the OPT model, presenting weaker correlations between the model outputs and the survey scores. These results predominantly show statistically significant and largely negative



<i>Model</i>	<i>Tokens</i>	<i>Mode</i>	<i>r</i>	<i>p-value</i>
Qwen2	pair1	in	-0.10	**
Qwen2	pair1	people	-0.12	***
Qwen2	pair2	in	0.14	***
Qwen2	pair2	people	-0.10	**
Qwen2	pair3	in	-0.18	***
Qwen2	pair3	people	-0.21	***
Qwen2	pair4	in	-0.09	**
Qwen2	pair4	people	-0.05	
Qwen2	pair5	in	-0.18	***
Qwen2	pair5	people	-0.36	***

Table 7: Correlation results for the WVS dataset using the Qwen2-0.5B model: analysis reveals significant negative correlations across all token pairs, with the most pronounced being -0.36 for pair5. The presence of a single positive correlation at 0.14\*\*\* for pair2 in the 'in' mode provides a contrast to the generally negative trend.

correlations across the different token pairs, particularly within the WVS dataset. There appears to be a consistent pattern where the choice of moral token generally has a more substantial impact on the score than the prompt mode used. Notably, the highest moral scores recorded are 0.14 in the WVS dataset and 0.30 in the PEW dataset, both achieving a 99.9% significance level.

The BLOOMZ-560M model has produced the best results so far in terms of alignment between the model outputs and the survey scores. Using the WVS questions as prompts, the average moral scores are 0.25\*\*\* and 0.29\*\*\*, both significant and the highest recorded thus far across the averaged token pairs scores. Similarly, when using the topics and countries from PEW, the scores are 0.16\*\* and 0.11\* for the two prompt types.

As illustrated in Table 9, the results for the WVS dataset showcase a prevalence of significant, strong positive correlations, surpassing the performance of previous models. Three token pairs achieve these notable results for both prompts, with the highest recorded at 0.36\*\*\*. Although negative correlations are present, they are comparatively less pronounced.

Similarly, for the PEW dataset, the results are also encouraging as shown in Table 10. Significant positive correlations prevail, though they are not as high as those observed for the WVS dataset. The highest positive correlation recorded is 0.28\*\*\*, while the negative correlations, that are present in

<i>Model</i>	<i>Tokens</i>	<i>Mode</i>	<i>r</i>	<i>p-value</i>
Qwen2	pair1	in	0.11	*
Qwen2	pair1	people	0.11	*
Qwen2	pair2	in	-0.06	
Qwen2	pair2	people	-0.26	***
Qwen2	pair3	in	0.30	***
Qwen2	pair3	people	0.14	**
Qwen2	pair4	in	-0.18	**
Qwen2	pair4	people	-0.22	***
Qwen2	pair5	in	-0.38	***
Qwen2	pair5	people	-0.35	***

Table 8: Correlation results for the PEW dataset using the Qwen2-0.5B model: analysis demonstrates a mix of positive and negative correlations. Highlights include a strong positive correlation of 0.30\*\*\* for pair3, contrasting with significant negative correlations, especially for pair5 under both prompt with each reaching beyond -0.35.

<i>Model</i>	<i>Tokens</i>	<i>Mode</i>	<i>r</i>	<i>p-value</i>
BLOOM	pair1	in	-0.07	**
BLOOM	pair1	people	-0.16	***
BLOOM	pair2	in	0.14	***
BLOOM	pair2	people	0.12	***
BLOOM	pair3	in	-0.04	
BLOOM	pair3	people	-0.36	***
BLOOM	pair4	in	0.36	***
BLOOM	pair4	people	0.26	***
BLOOM	pair5	in	0.21	***
BLOOM	pair5	people	0.30	***

Table 9: Correlation results for the WVS dataset using the BLOOMZ-560M model: analysis showcases a predominance of significant, strong positive correlations, with three token pairs for both prompts achieving these results, with the highest reaching 0.36\*\*\*. Negative correlations, while present, are less pronounced.

the dataset, lack statistical significance.

### 5.3 Distribution of moral scores per topic

As depicted in Figure 2 in section 3, the spread of responses varies significantly across different moral topics. Topics such as 'for a man to beat his wife', 'stealing property', and 'violence against other people' show limited variation across countries and are mainly positioned on the left side, indicating negative moral scores. In contrast, topics like 'homosexuality', 'sex before marriage', and 'having casual sex' exhibit a wide range of responses, spanning both negative and positive moral scores.

Model	Tokens	Mode	$r$	$p$ -value
BLOOM	pair1	in	-0.25	***
BLOOM	pair1	people	-0.13	*
BLOOM	pair2	in	0.08	
BLOOM	pair2	people	0.12	*
BLOOM	pair3	in	-0.07	
BLOOM	pair3	people	0.12	*
BLOOM	pair4	in	0.28	***
BLOOM	pair4	people	0.23	***
BLOOM	pair5	in	0.16	**
BLOOM	pair5	people	0.08	

Table 10: Correlation results for the PEW dataset using the BLOOMZ-560M model: analysis highlights a predominance of strong positive correlations, more pronounced than those seen with earlier models. Significant positive results include correlations of 0.28 and 0.23 both for pair4. While negative correlations are present, they are comparatively milder.

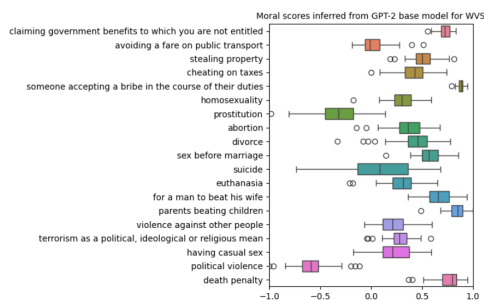


Figure 5: Distribution of normalized moral scores from GPT-2 base model using the WVS dataset

When comparing the moral scores from the GPT-2 base model for the WVS dataset, as depicted in Figure 5, with the actual survey responses, notable differences emerge. The GPT-2-derived scores predominantly gather on the right side of the x-axis, indicating a tendency toward positive moral judgments. Topics such as 'sex before marriage', 'having casual sex', 'homosexuality', and 'divorce' exhibit wide variation according to the survey, highlighting diverse viewpoints among people from different countries. In contrast, the GPT-2 model shows less variation for most topics, but notable disagreements are seen in topics 'suicide' and 'prostitution'.

Interestingly, despite the general trend of smaller variations in GPT-2 scores—which aligns with findings from previous studies—the model also unexpectedly shows a significant number of positive moral judgments across different prompts and token pairs. This suggests that while the model cap-

tures some aspects of human moral reasoning, its application still presents challenges in accurately mirroring the complex landscape of human moral values.

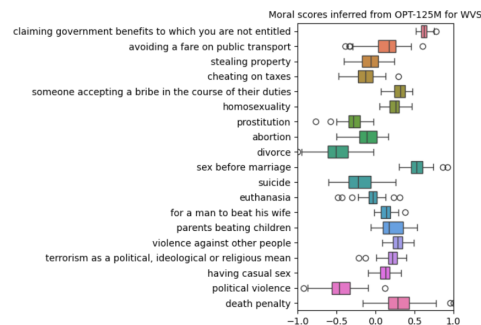


Figure 6: Distribution of normalized moral scores from OPT-125M model using the WVS dataset

The moral scores inferred from the OPT-125 model for the WVS dataset reveal that for certain topics, the results closely follow those of the GPT-2 base model, as shown in Figure 6. However, for topics where the behavior diverges from that observed in the GPT-2 model, the scores from the OPT-125 model tend to cluster closer to zero rather than extending into more positive values. This suggests a more neutral stance by the OPT-125 model on these particular issues.

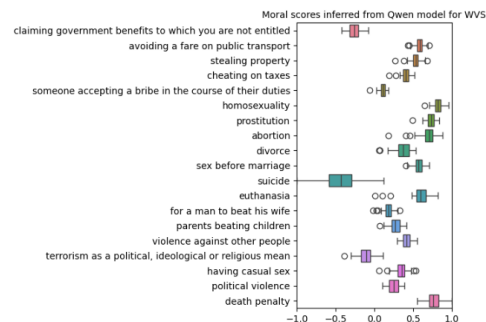


Figure 7: Distribution of normalized moral scores from Qwen2 model using the WVS dataset

The results from the Qwen2 model (Figure 7), when assessed using WVS moral topics as prompts, show patterns similar to those observed with the OPT model. Variations in moral scores are much smaller than those seen in the survey data. Additionally, the boxplots are predominantly positioned on the positive side of the x-axis, indicating a bias towards viewing these actions as morally acceptable, which does not align with the societal views as they are reflected in the survey results.

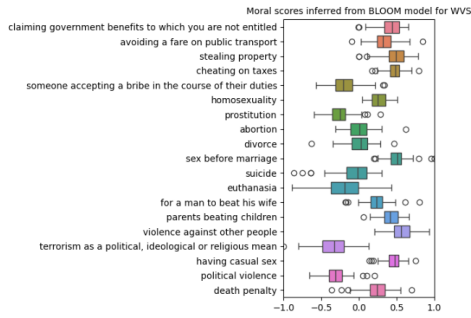


Figure 8: Distribution of normalized moral scores from BLOOMZ-560M model using the WVS dataset

As highlighted by the significant positive correlations between the BLOOM model’s scores and the survey results, as described earlier, the BLOOM model performs better in mirroring societal views. As displayed in Figure 8, it exhibits greater variability in moral scores compared to previous models, and these scores are now more closely aligned with the actual survey responses, tending towards more negative assessments. This shift suggests that this model offers a more accurate representation of societal views as presented in the survey data.

Regarding the PEW survey, as depicted in Figure 4 in section 3, responses to moral questions display significant diversity, similar to those in the WVS survey. Notably, topics such as ‘married people having an affair’ and ‘gambling’ consistently receive negative judgments, while they also show significant disagreement among respondents from different countries. ‘Homosexuality’ and ‘sex between unmarried adults’ exhibit the greatest variability, underscoring sharp differences in moral views across populations.

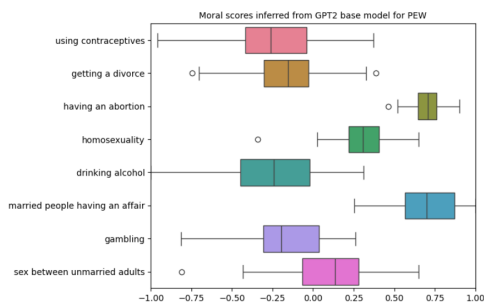


Figure 9: Distribution of normalized moral scores from GPT-2 base model using the PEW dataset

When comparing the PEW survey results to the outputs from the GPT-2 base model for the same dataset, as shown in Figure 9, several striking differences emerge. Firstly, the model demonstrates unexpectedly high variations in its responses, which contrasts with similar studies. There is a clear dis-

agreement on topics such as ‘married people having an affair’ and ‘getting an abortion’; the model typically sees these actions as acceptable, whereas the people who participated in the survey categorize them as mostly unacceptable. The only topic for which the model and the survey agree is ‘drinking alcohol,’ since both exhibit significant variability.

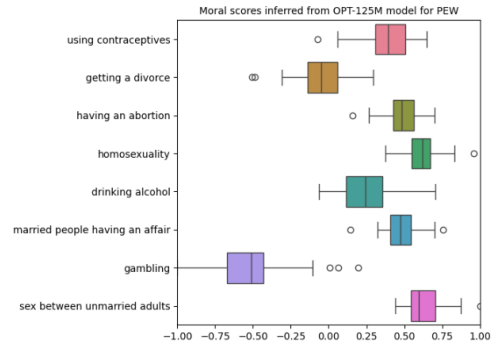


Figure 10: Distribution of normalized moral scores from OPT-125M model using the PEW dataset

The OPT-125M model’s moral scores for the PEW dataset, as illustrated in Figure 10, generally show smaller variations across most topics compared to the GPT-2 base model, with the notable exception of ‘gambling,’ which displays significant spread. Additionally, like the GPT-2, the scores are predominantly shifted towards the positive side, suggesting a more favorable moral assessment of most topics.

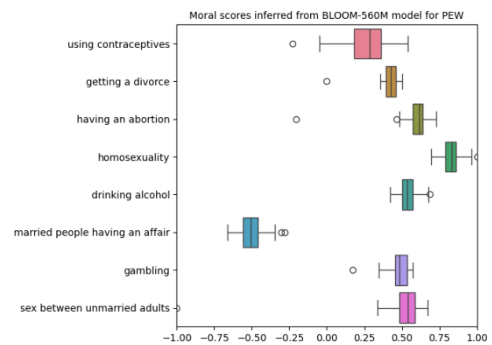


Figure 11: Distribution of normalized moral scores from BLOOMZ-560M model using the PEW dataset

For the BLOOMZ-560M model (Figure 11), variations in moral scores are even smaller across most topics, with only a few outliers. Notably, the topic ‘married people having an affair’ is consistently considered unjustifiable in all countries, according to the model’s assessments.

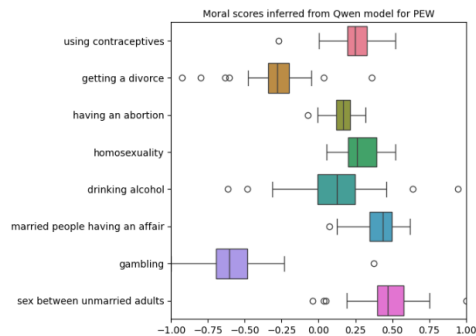


Figure 12: Distribution of normalized moral scores from Qwen2 model using the PEW dataset

Results from the Qwen2 model are similar to those from the BLOOM model but are slightly shifted closer to zero, indicating a more neutral stance on the issues (Figure 12).

## 6 Conclusion and Discussion

The aim of this study was to investigate whether pre-trained monolingual and multilingual language models contain knowledge about moral norms across many different cultures. The analysis shows that the examined LLMs do capture certain cultural value differences, but these only weakly align with established values surveys. They tend to characterize most topics as justifiable or generally acceptable across most countries, which contrasts with the varied and often contradictory views reflected in the WVS and the PEW survey data. While LLMs are capable of processing language, they cannot fully perceive the complex societal and cultural contexts that influence moral judgments.

The outputs of the four models reveal notable differences in the variability and alignment of moral scores compared to the actual survey results. The correlation scores between the models and the survey data were largely disappointing, with predominantly negative values, most of which were statistically significant. Furthermore, an in-depth analysis conducted by calculating correlations separately for the different prompt types and token pairs did not provide solid conclusions, as the results exhibited substantial variability across models and datasets.

The variant of BLOOM, BLOOMZ-560M showed a closer approximation to human judgments by aligning more consistently with negative assessments than the other three models. Yet it still failed to reflect human opinions even to a moderate degree. A possible reason for this performance could be attributed to its multilingual capabilities. As a multilingual model, BLOOM is trained on di-

verse linguistic datasets, which potentially enables it to access more cultural and moral contexts compared to monolingual models. Additionally, the performance of BLOOM, which is similar to that of GPT-3—a significant improvement over GPT-2—has been trained on 46 different languages and 13 programming languages in total. In contrast, Qwen2, the other multilingual model used in this study, did not showcase similar performance. This may be due to it being trained on data in fewer languages, with a particular focus on Chinese and English.

Another conclusion is that the four LLMs tend to characterize most topics as generally acceptable. Language models may simplify complex moral judgments due to their inability to fully understand nuanced cultural contexts and ethical considerations. As a result, they tend to adopt a more generalizable and justifiable stance. Additionally, without specific context, the models might be designed to lean towards more neutral or positive judgments to avoid controversial or negative outputs, which could be seen as safer or more acceptable. Thus, to avoid drawing solid conclusions, it's important to recognize that the models' responses might not accurately represent reality due to their lack of sufficient contextual information.

As a final conclusion, from the different experiments with the prompt modes and the different token pairs it was concluded that the choice of moral tokens used has a greater impact on the model scores than the choice of prompt types. This indicates that the selection of moral tokens substantially influences how the models assess moral norms.

Furthermore, it is worth noting that using alternative correlation coefficient metrics, such as Spearman's rank correlation coefficient, and testing newer models like GPT-3 or GPT-4, could potentially lead to different conclusions. Even deploying the available versions of the four models used in this study with the highest number of parameters could give different results. These methodological and architectural variations might offer additional insights into how language models interpret and generate moral judgments. Further exploration following these adjustments is essential to improve our understanding of their capabilities and limitations in ethical reasoning and comprehension.

## 7 Limitations

Although the datasets employed are publicly available and include responses from participants across different countries, they cannot fully represent the moral norms of all cultural groups globally or predict how these norms might evolve over time (Bloom, 2010; Bicchieri, 2005). Moreover, this study only explores a limited range of moral issues per country, and thus should not be considered exhaustive of the moral dilemmas people face worldwide. Additionally, averaging moral ratings for each culture simplifies the diverse range of moral values to a single value, which is a limitation of this study.

Furthermore, computational limitations constrained the scope of this research. The computational demands of the models were significant, and the availability of tools offering free additional resources restricted the analysis. Similarly, the use of more advanced models like GPT-3 or GPT-4 was not possible due to their requirement for paid access. Consequently, this restriction likely impacted the comprehensiveness of the findings and the depth of the analysis.

## References

- Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2022. Probing pre-trained language models for cross-cultural differences in values. *arXiv preprint arXiv:2203.13722*.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Sheng-guang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. [On the dangers of stochastic parrots: Can language models be too big?](#) . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, page 610–623, New York, NY, USA. Association for Computing Machinery.
- Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. 2016. [Man is to computer programmer as woman is to homemaker? debiasing word embeddings](#). *CoRR*, abs/1607.06520.
- Christian Haerper, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bjorn Puranen. 2021. World values survey: Round six - country-pooled datafile version. Available at: <http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp>.
- Katharina Hämmerl, Björn Deiseroth, Patrick Schramowski, Jindřich Libovický, Alexander Fraser, and Kristian Kersting. 2022. [Do multilingual language models capture differing moral norms?](#)
- Masahiro Kaneko and Danushka Bollegala. 2021. [Unmasking the mask - evaluating social biases in masked language models](#). *CoRR*, abs/2104.07496.
- Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. [StereoSet: Measuring stereotypical bias in pretrained language models](#). *CoRR*, abs/2004.09456.
- Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. [StereoSet: Measuring stereotypical bias in pretrained language models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.
- Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. 2021. [Probing toxic content in large pre-trained language models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4262–4274, Online. Association for Computational Linguistics.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#).
- Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. *arXiv preprint arXiv:2306.01857*.
- Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2019. [Pseudolikelihood reranking with masked language models](#). *CoRR*, abs/1910.14659.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. [The risk of racial bias in hate speech detection](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.
- Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagn'e, Alexandra Sasha Luccioni, François

Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurencon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa Etxabe, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris C. Emezue, Christopher Klamm, Colin Leong, Daniel Alexander van Strien, David Ifeoluwa Adelani, Dragomir R. Radev, Eduardo Gonz'alez Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady ElSahar, Hamza Benyamina, Hieu Trung Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jorg Frohberg, Josephine L. Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro von Werra, Leon Weber, Long Phan, Loubna Ben Allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario vSavsko, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nuru-laqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto L'opez, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, S. Longpre, Somaieh Nikpoor, S. Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-Shaibani, Matteo Manica, Nihal V. Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Févry, Trishala Neeraj, Urmish Thakker, Vikas Raut, Xiang Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Y Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre Francois Lavall'ee, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim

Detmeters, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramanian, Aur'elie N'ev'eol, Charles Lovering, Daniel H Garrette, Deepak R. Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Xiangru Tang, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najeoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, S. Osher Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdenek Kasner, Zdeněk Kasner, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ananda Santa Rosa Santos, Anthony Hevia, Antigona Tumdreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ayoade Ajibade, Bharat Kumar Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David M. Lansky, Davis David, Douwe Kiela, Duong Anh Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatim Tahirah Mirza, Frankline Ononiwu, Habib Rezanejad, H.A. Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jan Passmore, Joshua Seltzer, Julio Bonis Sanz, Karen Fort, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nourhan Fahmy, Olanrewaju Samuel, Ran An, R. P. Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas L. Wang, Sourav Roy, Sylvain Viguier, Thanh-Cong Le, Tobi Oyebade, Trieu Nguyen Hai Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Kumar Singh, Benjamin Beilharz, Bo Wang, Caio Matheus Fonseca de Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel Le'on Perin'an, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrmann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabc, Iman I.B. Bello, Isha Dash, Ji Soo Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthi Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, María Andrea Castillo, Marianna Nezhurina, Mario Sanger, Matthias Samwald, Michael Cullan, Michael Weinberg, M Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patricia Haller, R. Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aaroonsiri, Srishti Kumar, Stefan Schweter, Sushil Pratap Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech

Kusa, Yanis Labrak, Yashasvi Bajaj, Y. Venkatraman, Yifan Xu, Ying Xu, Yu Xu, Zhee Xao Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. [Bloom: A 176b-parameter open-access multilingual language model](#). *ArXiv*, abs/2211.05100.

Karolina Stanczak and Isabelle Augenstein. 2021. [A survey on gender bias in natural language processing](#). *CoRR*, abs/2112.14168.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). *CoRR*, abs/1706.03762.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. [Opt: Open pre-trained transformer language models](#). *arXiv preprint arXiv:2205.01068*.

## Acknowledgements

I would like to express my gratitude to my supervisors, Professor Ayoub Bagheri and Hadi Mohammadi, for their guidance, support, and encouragement throughout this research. Their insights and constructive feedback were crucial to the successful completion of this thesis.

I also acknowledge the authors of the papers I tried to replicate, as well as the creators of the datasets and tools utilized in this research, including the World Values Survey and the PEW Research Center.

## A Appendix

<i>Model</i>	<i>Tokens</i>	<i>Mode</i>	<i>r</i>	<i>p-value</i>
GPT2-M	pair1	in	-0.35	***
GPT2-M	pair1	people	-0.04	***
GPT2-M	pair2	in	0.01	*
GPT2-M	pair2	people	0.16	
GPT2-M	pair3	in	-0.18	***
GPT2-M	pair3	people	-0.18	***
GPT2-M	pair4	in	0.11	
GPT2-M	pair4	people	-0.17	
GPT2-M	pair5	in	-0.04	
GPT2-M	pair5	people	-0.33	**

Table 11: Correlation results for the WVS dataset using the GPT-2-Medium model: analysis shows almost exclusively negative correlations, with half of these being statistically significant. This table demonstrates a higher incidence of negative scores compared to those observed using the GPT-2 base model. The results are quite similar to those obtained by the GPT-2-Large model for the same dataset.

<i>Model</i>	<i>Tokens</i>	<i>Mode</i>	<i>r</i>	<i>p-value</i>
GPT2-M	pair1	in	-0.25	***
GPT2-M	pair1	people	0.11	***
GPT2-M	pair2	in	0.12	*
GPT2-M	pair2	people	-0.01	
GPT2-M	pair3	in	0.26	***
GPT2-M	pair3	people	0.35	***
GPT2-M	pair4	in	0.19	
GPT2-M	pair4	people	-0.04	
GPT2-M	pair5	in	0.04	
GPT2-M	pair5	people	-0.19	***

Table 12: Correlation results for the PEW dataset using the GPT-2-Medium model: analysis indicates a range of correlations from positive to negative. Findings include a strong positive correlation for pair3 and significant negative correlations for pair1 and pair5. These results suggest varying alignment between the model scores and survey responses across different contexts.