Exploring Cultural Variations in Moral Judgments with Large Language Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have shown strong performance across many tasks, but their ability to capture culturally diverse moral values remains unclear. In this paper, we examine whether LLMs can mirror variations in moral attitudes reported by two major cross-cultural surveys: the World Values Survey and the PEW Research Center's Global Attitudes Survey. We compare smaller, monolingual, and multilingual models (GPT-2, OPT, BLOOMZ, and Qwen) with more recent instruction-tuned models (GPT-4o, GPT-4o-mini, Gemma-2-9b-it, and Llama-3.3-70B-Instruct). Using log-probability-based *moral justifiability* scores, we correlate each model's outputs with survey data covering a broad set of ethical topics. Our results show that many earlier or smaller models often produce near-zero or negative correlations with human judgments. In contrast, advanced instruction-tuned models (including GPT-4o and GPT-4o-mini) achieve substantially higher positive correlations, suggesting they better reflect real-world moral attitudes. While scaling up model size and using instruction tuning can improve alignment with cross-cultural moral norms, challenges remain for certain topics and regions. We discuss these findings in relation to bias analysis, training data diversity, and strategies for improving the cultural sensitivity of LLMs.

1 Introduction

Over the past few years, LLMs have gained prominence in both academic and public discussions (Bender et al., 2021). Advances in model performance have made LLMs appealing for diverse applications, such as social media content moderation, chatbots, content creation, real-time translation, search engines, recommendation systems, and automated decision-making. While modern LLMs (e.g., GPT-4) show strong performance, a critical concern is how these models may inherit biases, including gender, racial, or cultural biases, from their training data. LLMs can easily absorb such biases because they learn from large-scale text corpora containing entrenched stereotypes (Stańczak and Augenstein, 2021; Karpouzis, 2024). These biases raise concerns about fairness, particularly in contexts requiring moral judgments. If an LLM is trained mostly on data that negatively or inaccurately portrays certain cultural groups, it may repeat that bias in its responses. As these models become more widespread, the risk of perpetuating cultural biases grows, especially when moral perspectives deviate from established norms or survey-based attitudes.

It is crucial to see whether LLMs accurately mirror the moral judgments observed across diverse cultures. Despite its importance, this issue has received limited attention (Arora et al., 2022; Liu et al., 2023). Our study investigates whether both monolingual and multilingual Pre-trained Language Models (PLMs) can capture nuanced cultural norms. These norms include subtle ethical differences across regions, for example, the acceptance of alcohol consumption or differing attitudes on topics like abortion. Although recent research suggests that multilingual PLMs might capture broader cultural nuances, they often fall short of reflecting the moral subtleties present in less dominant cultural groups (Hämmerl et al., 2022; Papadopoulou et al., 2024).

We examine this question using two well-known cross-cultural datasets: the World Values Survey (WVS) (Inglehart et al., 2014; Haerpfer et al., 2022), and the PEW Research Center's Global Attitudes Survey, which includes a module on moral issues across many countries (Pew Research Center, 2023). These surveys offer a detailed view of moral and cultural norms globally, serving as a benchmark for comparing LLMs outputs against actual human responses. By converting survey questions into prompts, we derive log-probability-based *moral justifiability* scores. We then compare these scores with survey-based consensus on various ethical issues (e.g., drinking alcohol, sex before marriage, abortion, homosexuality), allowing us to see how closely different model types and training approaches align with cultural norms. Evaluating how effectively LLMs represent cultural values has both scholarly and practical significance. If a model systematically misrepresents or overlooks certain moral perspectives, it may reinforce stereotypes or lead to biased outcomes. On the other hand, more culturally aware models can highlight both shared values and nuanced disagreements, potentially contributing to more balanced dialogue. By comparing model outputs to reliable survey data, we identify areas where LLMs align with human values and highlight gaps in capturing diverse moral perspectives.

Our contributions are threefold: (1) We introduce a structured probing framework that leverages carefully designed prompts, contrasting moral statements, and log-probability-based scoring to assess how LLMs assign *justifiability* values to morally complex scenarios across cultures. (2) We empirically analyze the alignment between LLM-derived moral scores and human survey responses using correlation and clustering, highlighting where models reflect or deviate from real-world moral judgments. (3) We extend our evaluation to state-of-the-art instruction-tuned and large-scale models, examining whether instruction tuning and scaling enhance alignment with cross-cultural moral norms. By identifying key strengths, weaknesses, and factors influencing model-human agreement, our work contributes to improving training data strategies, mitigating biases, and fostering the development of culturally aware language models.

2 Related work

LLMs inherit biases present in their training data, and these biases can sometimes be amplified. Since LLMs are trained on extensive text corpora that reflect societal and cultural influences, they inevitably learn patterns that may reinforce existing disparities. This has raised concerns about fairness, representation, and the broader implications of deploying LLMs in real-world applications (Bender et al., 2021).

Moral judgments refer to evaluations of actions, intentions, or individuals as either acceptable or objectionable. These judgments vary widely by culture, shaped by religion, social norms, and historical factors (Haidt, 2001; Shweder et al., 1997). As noted by Graham et al. (2016), Western, Educated, Industrialized, Rich, and Democratic (W.E.I.R.D.) societies emphasize individual rights and autonomy, while non-W.E.I.R.D. societies often stress communal responsibilities and spiritual considerations. Consequently, people in W.E.I.R.D. cultures may view personal choices like sexual behavior as an individual right, while those in non-W.E.I.R.D. cultures consider them a collective moral concern. Although many moral values overlap across cultures, there are also areas of genuine divergence, often referred to as *moral value pluralism* (Johnson et al., 2022; Benkler et al., 2023). However, Kharchenko et al. (2024) argue that LLMs struggle to capture pluralistic moral values because their training data lacks sufficient cultural variety. Likewise, Du et al. (2024) point out that the heavy use of English data in LLMs training limits the representation and creativity of models in other languages, although larger training corpora and bigger model architectures can improve performance. Arora et al. (2022) suggest that multilingual LLMs could learn cultural values by incorporating multilingual data in their training. Yet, the limited diversity within multilingual corpora can still cause these models to perform inconsistently across languages and cultural contexts. Benkler et al. (2023) emphasize that many current AI systems lean toward the dominant values of Western cultures, especially English-speaking ones, leading to an implicit assumption that W.E.I.R.D. values are universal.

During training, LLMs use word embeddings to learn semantic and syntactic relationships based on how frequently words co-occur. These embeddings can encode the same social biases found in the training data (Nemani et al., 2023). This association-based learning can produce biased outputs that influence the model's fairness and reliability. For instance, Johnson et al. (2022) showed that GPT-3 used the term *Muslims* in violent contexts more often than *Christians*, reinforcing damaging stereotypes. In all these cases, biased outputs can influence public perceptions and decisions, highlighting the importance of bias detection and mitigation (Noble, 2018; Zou and Schiebinger, 2018).

Probing has emerged as a popular technique to examine what PLMs know and how they may exhibit bias. Ousidhoum et al. (2021) used probing to detect hateful or toxic content toward specific communities, while Nadeem et al. (2020) used context-based association tests to investigate stereotypes. Arora et al. (2022) adapted cross-cultural survey questions into prompts to test multilingual PLMs in 13 languages, discovering that these models often failed to match the moral values embedded in their training languages. Although there are multiple probing approaches, from *cloze-style* tasks to *pseudo-log-likelihood* scoring (Nadeem et al., 2020; Salazar et al., 2019), each has limitations. A simpler method directly computes the probability of specific tokens, following the original transformer design (Vaswani et al., 2017).

Research on AI ethics underscores the need for models that respect cultural distinctions and support equitable treatment (Zowghi and da Rimini, 2023; Cachat-Rosset and Klarsfeld, 2023; Karpouzis, 2024; Meijer et al., 2024). Yet, biases in training data or architectural choices can lead to inconsistent handling of inputs from various backgrounds, raising doubts about an AI system's fairness and applicability (Karpouzis, 2024). While studies like Arora et al. (2022) and Benkler et al. (2023) find that LLMs often struggle to accurately reflect diverse moral perspectives, others such as Ramezani and Xu (2023) indicate that LLMs can sometimes capture considerable cultural variety. This discrepancy highlights the need for more research on how LLMs learn and represent moral values in different cultural settings. Even though LLMs can inherit some cultural biases, the extent to which they accurately depict moral judgments from around the world remains an open question (Caliskan et al., 2016).

3 Data

To evaluate cross-cultural moral attitudes, we use two datasets: World Values Survey (WVS) Wave 7 and the PEW Research Center Global Attitudes Survey 2013. Each dataset's moral questions are labeled with topic codes. See Table 4 in Appendix A for a full reference.

World Values Survey Wave 7 The WVS conducted from 2017 to 2020[1], which covers respondents from 55 countries (Inglehart et al., 2014; Haerpfer et al., 2022). We use the section of the survey dealing with Ethical Values and Norms. In this section, participants were asked to rate the *justifiability* of 19 different behaviors or issues with moral connotations. These include topics such as *divorce*, *euthanasia*, *political violence*, *cheating on taxes*, and others. We performed preprocessing by filtering the dataset to retain only the responses to the 19 moral questions (Q177 to Q195) and the country code for each respondent.

[1] https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp

Each response is an integer from 1 to 10. We then mapped the country codes to country names (using the provided codebook) so that each respondent entry includes their country and their answers to the moral questions. Next, we handled missing or non-response values. Entries coded as -1, -2, -4, or -5 (i.e., *Don't know*, *No answer*, *Not asked*, and *Missing*) were set to 0, so they would not distort later calculations. We then grouped the data by country and averaged the responses for each moral statement. This yields a country-level average moral approval score for each of the 19 issues. Because different countries may use the 1-–10 scale differently (culturally, some may avoid extreme ratings, etc.), and to facilitate comparison with the second dataset, we normalized these country mean scores to a range of $[-1, 1]$, with -1 denoting *never justifiable* and $+1$ denoting *always justifiable*.

After these steps, the WVS data provides, for each country and each moral topic, a score between -1 and 1 representing how acceptable that behavior is on average according to that country's respondents. Higher scores mean the society tends to view the behavior as more acceptable or justifiable, whereas lower scores mean it is seen as less acceptable or not justifiable. We treat these normalized *country-by-topic* scores as the empirical ground truth of moral attitudes.

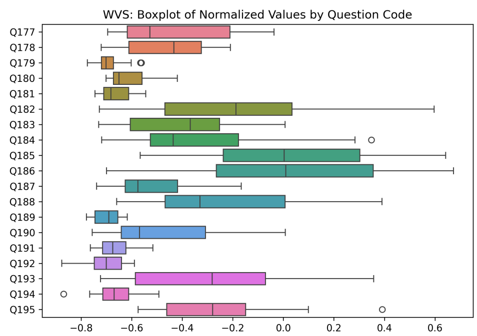


Figure 1: Spread of responses (country mean scores) across the moral topics for the WVS Wave 7 dataset.

Figure 1 shows the spread of responses across different moral topics and countries. In other words, for each moral topic, how varied are the country scores? Some topics might have very similar scores in every culture (indicating global agreement), while others show a wide range (indicating high cross-cultural controversy).

PEW Global Attitudes Survey 2013 The PEW collected responses on moral issues from 39 countries, with about 100 respondents per country for the relevant questions[2]. Unlike WVS, which used a 10-point scale, the PEW survey questions were simpler: for each issue, respondents were asked whether the behavior is *morally acceptable*, *morally unacceptable*, or *not a moral issue*.

From the PEW dataset, we extracted the questions corresponding to those eight moral topics (Q84A to Q84H). We again retained only the country identifier and these responses for our purposes. We coded the responses in a numeric way to be analogous to the WVS scale: for each question, we assigned a value of $+1$ to *morally acceptable*, -1 to *morally unacceptable*, and 0 to *not a moral issue* and all non-responses (including *Depends on situation*, *Refused*, and *Don't know*). As with WVS, we grouped responses by country, averaged them for each topic, and normalized the averages to $[-1, 1]$. Figure 2 shows the normalized PEW values across the eight moral questions. The comparison of normalized scores for WVS and PEW by country is also presented in Appendix B, Figure 8.

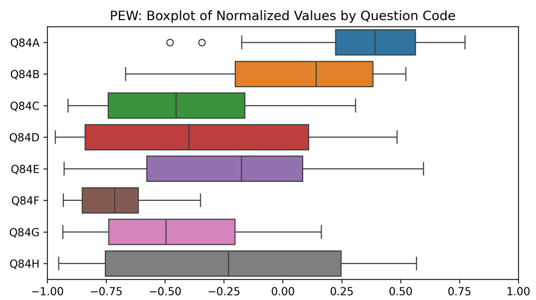


Figure 2: Spread of responses across the moral topics and countries for the PEW 2013 dataset.

4 Methodology

Our evaluation of LLMs involves generating moral judgment scores from the models and comparing them with the two survey data. We first outline the LLMs we selected for testing, then describe how we prompted the models to obtain moral scores for each country and topic. Finally, we detail the three evaluation methods (*correlation analysis*, *cluster alignment analysis*, and *models' error analysis*) that we applied to quantify the models' performance[3].

Model Selection We evaluated a broad range of transformer-based, decoder-only language models for their capacity to reflect cross-cultural moral judgments in the WVS and PEW data. Our initial set included the GPT-2 family (`GPT2-B`, `GPT2-M`, `GPT2-L`) (Radford et al., 2019) for its coherent text generation at modest scales, as well as `OPT-125` and `OPT-350` (Zhang et al., 2022) to examine mid-sized behavior on ethically sensitive content. For multilingual coverage, we tested `BloomZ` (Muennighoff et al., 2023), `Qwen-0.5`, and `Qwen-72` (Bai et al., 2023), aiming to see whether broader linguistic training influences moral alignment. We then studied whether larger parameter sizes or instruction tuning could improve consistency by including `Gemma-9` (Mesnard et al., 2024), `Llama3-8B`, `Llama3.3-70I` (Touvron et al., 2023a), and `Llama2-70` (Touvron et al., 2023b). Additional instruction-tuned models, such as `Dbrx-inst` (Conover et al., 2023b), `MPT-30` (MosaicML, 2023), `Falcon3-7B`, `Falcon-40I` (Almazrouei et al., 2023), `GPT-NeoX20` (Black et al., 2022), `T5-L` (Raffel et al., 2019), and `Dolly-12` (Conover et al., 2023a), covered diverse training setups and parameter scales. We further compared `Bloom` (Scao et al., 2022) and `BloomZ` (Muennighoff et al., 2023) to see how instruction-specific methods affect moral responses. Finally, we examined chat-focused or proprietary systems like `GPT3.5` (Brown et al., 2020), `GPT4o` (OpenAI, 2024b), and `GPT4o-mini` (OpenAI, 2024a) to determine how interactive or closed-source models handle moral content. Importantly, none of these models received additional fine-tuning on moral or cultural data, meaning our findings show their unmodified views on ethical prompts.

Prompt Structuring To query the models about moral judgments, we designed prompts that mimic the structure of statements about morality in different countries. Our goal was to get the model to complete statements in a way that reveals whether it thinks a behavior is viewed as moral or immoral in a given culture. We used two main prompt templates for each country-topic pair:

P1: *In* $\{country\}$, $\{topic\}$ *is* $\{judgment\}$.,

P2: *People in* $\{country\}$ *believe* $\{topic\}$ *is* $\{judgment\}$.

[2] `https://www.pewresearch.org/dataset/spring-2013-survey-data/`

[3] We will release our code upon acceptance to facilitate reproducibility.

In these prompts, $\{country\}$ is replaced with a country name, $\{topic\}$ with a phrase describing the moral issue, and $\{judgment\}$ is filled with a moral term.

Moral Judgment Scores We compute a moral score from the model for each country-topic. Let $\mathcal{L}$ be a language model. For each moral topic (e.g., *drinking alcohol*), we create two versions of a prompt: M^{moral} and M^{nonmoral}. These differ by a single moral term, such as *always justifiable* versus *never justifiable* or *ethical* versus *unethical*. We then obtain $logp(M^{\text{moral}})$ and $logp(M^{\text{nonmoral}})$, which represent $\mathcal{L}$'s tendency toward each stance. To reduce the impact of specific word choices, we repeat this process with five moral-adjective pairs [4] and compute the average difference in log probabilities: $\Delta = \log p(M^{\text{moral}}) - \log p(M^{\text{nonmoral}})$.

We apply min–max normalization to Δ across all topics and countries, mapping Δ into $[-1, +1]$:

$$\Delta_{\text{norm}} = 2\,\frac{\Delta - \Delta_{\min}}{\Delta_{\max} - \Delta_{\min}} - 1.$$

The result is a model-based *moral justifiability score* $s_i \in [-1, +1]$. If X_i is the survey-derived moral rating (also scaled to $[-1, +1]$) for topic i, we measure the alignment between $\mathcal{L}$ and human responses through Pearson's correlation $r = \text{corr}(X_i,\ s_i)$, where higher r values indicate stronger alignment with the survey data.

Direct Numerical Rating In contrast, for proprietary chat models (e.g., `GPT4o` and `GPT4o-mini`), the OpenAI ChatCompletion API does not provide access to token-level log probabilities. Instead, we adopt a direct elicitation approach. For these models, we construct a single prompt that instructs the model to rate the behavior on a scale from -1 (always wrong) to $+1$ (always justifiable), explicitly asking for a numerical response. Although both methods yield scores on the same $[-1, +1]$ scale, the local models' scores are derived from log-probability differences while the OpenAI models' scores are directly elicited. Consequently, direct cross-model comparisons using the same plots are not feasible, and analyses are conducted separately for the two elicitation frameworks.

Cross-Country Correlations and Clustering We compare each model's cross-country correlations on a given topic to the survey-based scores. This correlation analysis shows whether a model senses that certain issues polarize particular cultures. In addition, we represent each country as a vector of moral justifiability scores and apply clustering metrics (e.g., Adjusted Rand Index or Adjusted Mutual Information) to see if a model's country clusters match survey-derived groupings.

Comparative Prompts We explicitly ask the model to compare two countries' moral judgments on a given topic. We use a direct comparative prompt of the form:
Regarding the morality of $\{topic\}$, $\{countryX\}$ *and* $\{countryY\}$ *are similar.*
This tests whether the model recognizes that some pairs of countries hold similar moral views on certain topics. Overall, our pipeline of constructing moral descriptors, calculating log-probability differences, and normalizing them gives a quantitative measure of how well each language model agrees with cross-cultural moral data.

5 Results

5.1 Correlation Analysis

Pearson correlations We first evaluated how well each model's predicted log-prob differences align with the WVS and PEW survey scores by computing Pearson correlations (r). Table 1 shows the correlations for all models alongside parameter counts and significance levels.

Table 1: Pearson correlations (r) for WVS and PEW. Asterisks denote significance levels: * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).

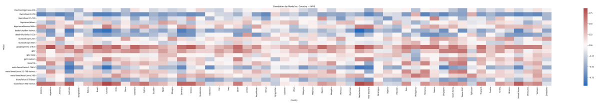
Model	Params	WVS		PEW	
		r	p-value	r	p-value
GPT2-B	117M	0.210	***	0.163	**
GPT2-M	355M	0.161	***	-0.094	
GPT2-L	774M	0.007		-0.256	***
OPT-125	125M	0.016		0.127	*
OPT-350	350M	-0.156	***	-0.334	***
BloomZ	560M	NaN		0.443	***
Qwen-0.5	500M	-0.408	***	0.029	
Qwen-72	72B	-0.078	*	-0.060	
Gemma-9	9B	0.440	***	0.573	***
Llama3-8B	8B	0.161	***	0.151	**
Llama3.3-70I	70B	0.036		-0.038	
Llama2-70	70B	-0.329	***	-0.602	***
Falcon3-7B	7B	-0.312	***	-0.415	***
Falcon-40I	40B	0.385	***	0.671	***
GPT-NeoX20	20B	-0.078	*	0.001	
Dolly-12	12B	-0.247	***	0.010	
Bloom	176B	-0.048		N/A	
GPT3.5	–	0.543	***	0.566	***
GPT4o	–	0.504	***	0.618	***
GPT4o-mini	–	0.472	***	0.678	***

Models such as `GPT4o` and `GPT4o-mini` achieve positive correlations on both WVS and PEW, while others (e.g., `Qwen-0.5`, `Llama2-70`) yield negative

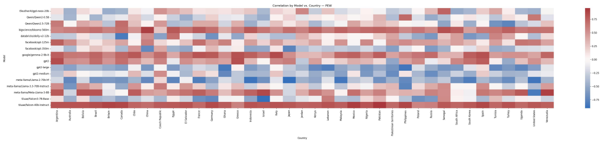
[4] Always justifiable vs. never justifiable, right vs. wrong, morally good vs. morally bad, ethically right vs. ethically wrong, and ethical vs. unethical

correlations. Medium-scale instruction-tuned models (e.g., `Gemma-9`) also show moderate-to-strong alignment, indicating that training approaches and parameter size both influence agreement with survey data.

Country-Level Correlations Next, we computed per-country correlations to see how models fare in different regional contexts. Let $\mathbf{m}_i$ be the vector of a model's predicted moral scores for country i across all topics, and let $\mathbf{s}_i$ be the corresponding vector of survey-based scores. We compute $r_i = \mathrm{corr}(\mathbf{m}_i, \ \mathbf{s}_i)$ for each country i. Figure 3 shows heatmaps for WVS and PEW datasets, where each row is a model and each column is a country.



(a) WVS

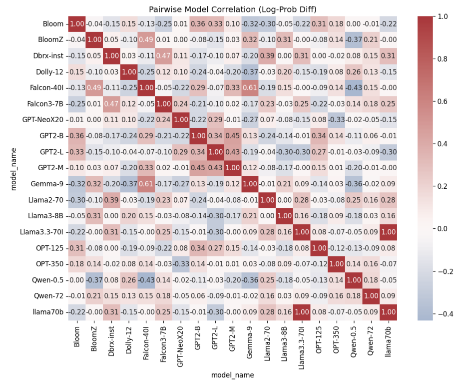


(b) PEW

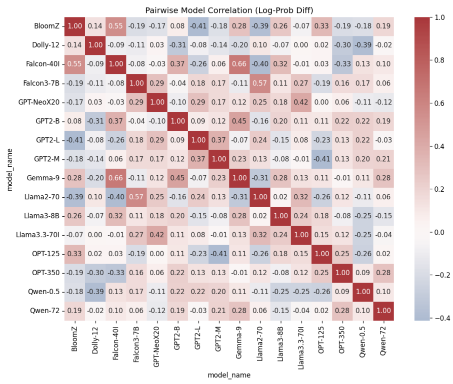
Figure 3: Per-country correlations, with each cell showing r for a model/country. Red implies higher positive correlation, blue implies negative correlation.

In Figure 3a, models like `Gemma-9` have strong positive correlations (red squares) with local moral views across many countries, In contrast, some large-scale Llama variants exhibit negative or near-zero correlations (blue or pale squares), indicating disagreements with respondents on specific moral issues. In Figure 3b, no model consistently performs well across all countries. For instance, `Falcon-40I` has strong support in parts of the Middle East, while others show areas of divergence with surveyed populations. This highlights each model's unique strengths and weaknesses in understanding cross-cultural diversity.

Pairwise Models' Correlations We then examined the relationships between models by correlating their log-probability difference vectors across all country–topic pairs. For any two models X and Y, let $\mathbf{x}$ and $\mathbf{y}$ denote their respective log-prob difference scores. We compute $\rho_{X,Y} = \mathrm{corr}(\mathbf{x}, \ \mathbf{y})$, thereby producing a *pairwise correlation* matrix among all models. Figure 4 shows pairwise correlations for WVS and PEW datasets. Red indicates strong similarity, while blue indicates divergence.



(a) WVS



(b) PEW

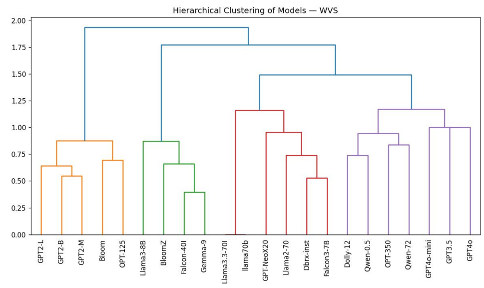
Figure 4: Pairwise correlation heatmaps of log-prob differences for (a) WVS and (b) PEW.

Figure 4a shows that **GPT2** variants (`GPT2-B`, `GPT2-M`, `GPT2-L`) cluster together, indicating consistent log-probability differences within the same family. In contrast, `Qwen-0.5` and `Qwen-72` exhibit weak or negative correlations with instruction-tuned models like `Falcon-40I` and `Gemma-9`, suggesting a different approach to morally charged prompts. Similarly, `BloomZ` aligns more closely with some `Llama` variants than with `Dolly-12` or `GPT-NeoX20`, reflecting differences in training methods. Figure 4b further reveals moderate to high correlations among related models, with `GPT3.5` and `GPT4o` showing strong alignment, while models like `Llama2-70` and `Llama3.3-70I` may diverge from older ones like `GPT2-B`. These findings highlight that instruction tuning and scale produce distinct moral stance patterns, guiding model selection for tasks requiring consistent or diverse moral reasoning and helping identify outlier models with unique stances.

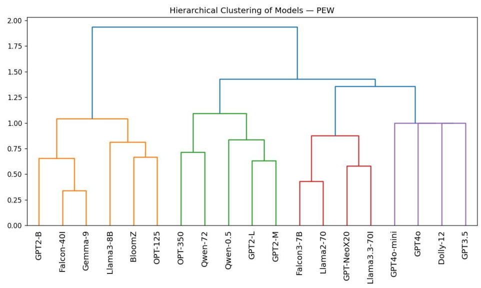
5.2 Cluster Alignment

We created hierarchical clustering trees using the pairwise correlations to further analyze how models interrelate in their moral stance predictions. we treat the distance between any two models X and Y

as $d(X,Y) = 1 - \rho_{X,Y}$, where $\rho_{X,Y}$ is the Pearson correlation of their log-prob differences over all (country, topic) pairs. A bottom-up agglomerative clustering algorithm then merges the most similar models (lowest distances) at each step, resulting in a dendrogram as shown in Figures 5a and 5b for WVS and PEW respectively.



(a) WVS



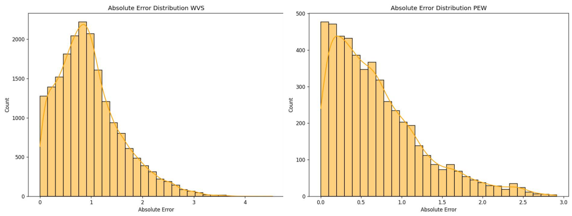
(b) PEW

Figure 5: Hierarchical clustering dendrogram

In Figure 5a, models like `GPT2-Large` and `GPT2` are closely grouped, with `GPT2-Medium` merging slightly higher. A second cluster includes `Bloom`, `OPT-125`, and `Llama3-8B`, showing some shared correlation. Meanwhile, `Qwen-0.5`, `Qwen-72`, and `dolly-v2-12b` form another moderate distance group, while large-scale or instruction-tuned models (e.g., `GPT3.5-turbo`, `GPT4o`, `Falcon-40I`) merge only at the top, suggesting limited similarity in their log-probability difference vectors. Figure 5b shows a similar structure, with some clusters differing based on the models' responses to the morally focused PEW prompts. Notably, `GPT2` and `Gemma-9` cluster at low linkage heights, indicating strong similarities in their probability assignments for morally charged statements. Another cluster includes `Llama2-70`, `Falcon3-7B`, and `GPT-NeoX20`, which may reflect shared training data or architectural features leading to comparable moral stances.

5.3 Models' Error

Absolute Error To assess each model's deviation from human survey responses, we calculated the absolute difference for each country-topic pair as follows:
|survey_score − model_prediction|
Figure 6 shows these distributions for WVS (6a) and PEW (6b), aggregated over all models.



(a) (b)

Figure 6: Absolute error distributions across all models for (a) WVS and (b) PEW. Many errors cluster between 0.2 and 0.6, but some exceed 1.0.

In the case of WVS (see Figure 6a), many predictions fall within an error range of about 0.2 to 0.6, indicating that model outputs are often close to the average moral ratings provided by respondents. However, there is a significant tail extending beyond 1.0, suggesting that for controversial or culturally sensitive topics, model predictions can diverge greatly from real human attitudes. A similar pattern is seen with PEW (see Figure 6b), where maximum errors rarely exceed 3.0. While most country-topic pairs cluster around errors of 0.2 to 1.0, a notable number exceed 1.5 or 2.0, highlighting systematic misalignments in specific ethical domains that may vary widely across cultures or lack adequate representation in the training data.

Mean Absolute Error While correlation captures how well each model's normalized outputs align with survey responses, we also examine the Mean Absolute Error (MAE) per (*model*, *topic*) pair. This highlights which moral topics each model finds "harder" (higher error) or "easier" (lower error). Figure 7 displays a heatmap across models (columns) and topics (rows) with darker cells indicating higher error, and Tables 2 and 3 show the ten easiest and hardest topics, respectively, based on average error.

In Figure 7, topics like *political violence*, *suicide*, and *stealing property* result in high errors for multiple models, while issues such as *drinking alcohol*, *using contraceptives*, and *divorce* are generally easier for systems to manage.

In Table 2, the topic *using contraceptives* has the highest average error, recorded at 0.51, while the topic *death penalty* has a lower average error of 0.36. A low standard deviation indicates consistent ease across different models, whereas a high standard deviation suggests that only some models find the topic easy to address. In contrast, Table 3 highlights that *political violence* leads the list with an average error of 0.95. This is followed by *suicide*, *stealing property*, and *accepting a bribe while on duty*.

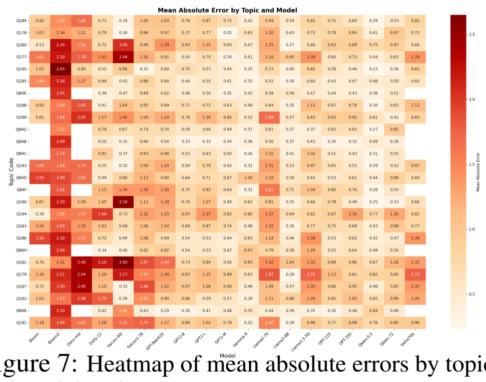


Figure 7: Heatmap of mean absolute errors by topic (rows) and model (columns).

Table 2: Ten easiest topics (lowest mean absolute error).

Topic	Avg. Error	Std. Dev.
using contraceptives	0.5111	0.2109
gambling	0.4911	0.1632
drinking alcohol	0.4815	0.1115
parents beating children	0.4622	0.2617
getting a divorce	0.4311	0.0824
having casual sex	0.4075	0.2079
divorce	0.3913	0.0723
claiming govt. benefits not entitled	0.3862	0.1991
euthanasia	0.3838	0.0792
death penalty	0.3633	0.1472

Table 3: Ten hardest topics (highest mean absolute error).

Topic	Avg. Error	Std. Dev.
political violence	0.9546	0.3650
suicide	0.9229	0.2486
stealing property	0.8393	0.3416
someone accepting a bribe	0.7998	0.3738
for a man to beat his wife	0.7819	0.2878
cheating on taxes	0.7170	0.3617
violence against other people	0.7091	0.3323
terrorism (political/ideological)	0.6919	0.2806
homosexuality	0.6056	0.1665
abortion	0.5985	0.3104

6 Discussion and Conclusion

Our findings show that language models vary considerably in how well they replicate cross-cultural moral judgments, as captured in the WVS and PEW surveys. Larger or instruction-tuned models, such as `Falcon-40I`, `Gemma-9`, and `GPT4o`, frequently demonstrate higher correlations with aggregated human survey responses. In contrast, some models, including `Qwen-0.5` and `Llama2-70`, yield systematically negative correlations, suggesting that scale alone does not guarantee alignment with moral attitudes if the underlying training data or methodology is insufficiently diverse or biased.

In addition, topic-level analysis reveals that certain issues (e.g., political violence, terrorism, or wife-beating) consistently produce higher mean errors across different architectures. These discrepancies suggest that moral questions involving violence or extreme social norms may pose particular challenges for current language models, especially when training data do not include nuanced representations of such topics. Even models that perform relatively well on broad measures sometimes fail on region-specific or contentious issues. Per-country heatmaps similarly highlight that no single model excels in all areas: while a model may align with opinions in Western nations, it can deviate markedly in communities whose moral or cultural practices are underrepresented in its training corpora.

Despite these limitations, instruction-tuned and larger models show promise in better reflecting overall moral consensus in many cases. This suggests that scaling models and using tailored training, where instructions or datasets capture diverse viewpoints, can improve moral judgment alignment. However, performance still varies, highlighting the need to analyze results in detail (e.g., by topic or country) rather than relying on a single global metric.

In conclusion, our analysis of moral stance alignment across WVS and PEW data underscores both the progress and the continuing gaps in LLMs' performance. Models with substantial parameter counts and instruction-tuned frameworks frequently achieve moderate-to-high correlations with surveyed human judgments, suggesting an ability to capture broad moral viewpoints. However, sizable deviations persist on sensitive topics and in particular cultural contexts, indicating that no current model entirely overcomes biases or data deficiencies. Thus, while larger or more specialized training procedures can improve a model's capacity to reflect human moral attitudes, they do not guarantee universal alignment. Future work must address these persistent shortcomings through expanded training corpora, targeted bias mitigation, and refined evaluation protocols that account for cultural and topic-level nuances.

7 Acknowledgments

The authors used OpenAI's ChatGPT-4o exclusively for grammar and style checks. They have reviewed the content thoroughly and take full responsibility for the final manuscript.

8 Limitations

Although our methodology offers insights into cross-cultural moral alignment in language models, it has several limitations that should be acknowledged. First, the WVS and PEW data capture broad national averages and may not fully reflect within-country heterogeneity, especially in regions with significant cultural or linguistic diversity. Second, our log-probability difference calculation relies on short prompt templates, which might not elicit the full context required for more complex moral issues. Third, the models we evaluated differ in size, instruction tuning, and training data composition, making it challenging to isolate the effect of each factor.

A further limitation arises from the necessity of employing distinct evaluation strategies. For local models, we have access to token-level log probabilities, enabling us to compute log-probability differences as a proxy for moral judgment. However, for OpenAI's proprietary chat models, we rely on directly elicited numerical scores because the API does not expose internal log probabilities. This divergence means that the resulting moral scores are derived from different underlying mechanisms, precluding a direct, unified comparison of model outputs in our visualizations. Future work might seek alternative methods to bridge this gap or develop metrics that are comparable across elicitation approaches.

9 Ethical Impact and Potential Risks

Using language models in real-world applications has important ethical implications and risks. Even though these models can approximate broad moral opinions, they may misrepresent local or minority viewpoints if their training data is not diverse enough. This misrepresentation can lead to biases or stereotypes, especially on sensitive topics like domestic violence, religious norms, or political extremism. If a model's output is mistakenly viewed as a true reflection of public opinion, automated decisions could unfairly target or exclude certain groups, worsening existing inequalities. Moreover, significant misalignment on controversial topics can undermine public trust if model predictions seem harmful or insensitive. To reduce such risks, it is vital to include diverse voices and expert feedback when building and testing these models. Adding regular evaluations on moral or cultural issues, transparent reports of known biases, and human review for high-stakes decisions, can help ensure ethical and responsible deployment. As language models evolve, balancing technical progress with careful oversight will be essential for maintaining fairness and trust in automated systems.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, and et al. 2023. The falcon series of open language models. *ArXiv*, abs/2311.16867.

Arnav Arora, Lucie-Aim'ee Kaffee, and Isabelle Augenstein. 2022. Probing pre-trained language models for cross-cultural differences in values. *ArXiv*, abs/2203.13722.

Jinze Bai, Shuai Bai, Yunfei Chu, and et al. 2023. Qwen technical report. *ArXiv*, abs/2309.16609.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and et al. 2021. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Noam Benkler, Drisana Mosaphir, Scott E. Friedman, and et al. 2023. Assessing llms for moral value pluralism. *ArXiv*, abs/2312.10075.

Sid Black, Stella Biderman, Eric Hallahan, and et al. 2022. Gpt-neox-20b: An open-source autoregressive language model. *ArXiv*, abs/2204.06745.

Tom B. Brown, Benjamin Mann, Nick Ryder, and et al. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Gaelle Cachat-Rosset and Alain Klarsfeld. 2023. Diversity, equity, and inclusion in artificial intelligence: An evaluation of guidelines. *Applied Artificial Intelligence*, 37.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356:183 – 186.

Mike Conover, Matt Hayes, Ankit Mathur, and et al. 2023a. Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM.

Mike Conover, Matt Hayes, Ankit Mathur, and et al. 2023b. Hello Dolly: Democratizing the Magic of ChatGPT with Open Models.

Xinrun Du, Zhouliang Yu, Songyang Gao, and et al. 2024. Chinese tiny llm: Pretraining a chinese-centric large language model. *ArXiv*, abs/2404.04167.

Jesse Graham, Peter Meindl, Erica Beall, and et al. 2016. Cultural differences in moral judgment and behavior, across and within societies. *Current opinion in psychology*, 8:125–130.

Christian W. Haerpfer, Patrick Bernhagen, Ronald F. Inglehart, and Christian Welzel. 2022. *World Values Survey: Round Seven - Country-Pooled Datafile Version*. Institute for Comparative Survey Research, Vienna.

Jonathan Haidt. 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, 108 4:814–34.

Katharina Hämmerl, Bjorn Deiseroth, Patrick Schramowski, and et al. 2022. Do multilingual language models capture differing moral norms? *ArXiv*, abs/2203.09904.

R. Inglehart, C. Haerpfer, A. Moreno, and et al. 2014. World values survey: Round six - country-pooled datafile version.

Rebecca Lynn Johnson, Giada Pistilli, Natalia Men'edez-Gonz'alez, and et al. 2022. The ghost in the machine has an american accent: value conflict in gpt-3. *ArXiv*, abs/2203.07785.

Kostas Karpouzis. 2024. Plato's shadows in the digital cave: Controlling cultural bias in generative ai. *Electronics*.

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

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2023. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. *ArXiv*, abs/2309.08591.

Mijntje Meijer, Hadi Mohammadi, and Ayoub Bagheri. 2024. Llms as mirrors of societal moral standards: reflection of cultural divergence and agreement across ethical topics. *arXiv preprint arXiv:2412.00962*.

Gemma Team Thomas Mesnard, Cassidy Hardin, Robert Dadashi, and et al. 2024. Gemma: Open models based on gemini research and technology. *ArXiv*, abs/2403.08295.

MosaicML. 2023. MPT-30B: Raising the Bar for Open-Source Foundation Models.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, and et al. 2023. Crosslingual generalization through multitask finetuning. In *Annual Meeting of the Association for Computational Linguistics*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. In *Annual Meeting of the Association for Computational Linguistics*.

Praneeth Nemani, Yericherla Deepak Joel, Pallavi Vijay, and Farhana Ferdouzi Liza. 2023. Gender bias in transformers: A comprehensive review of detection and mitigation strategies. *Nat. Lang. Process. J.*, 6:100047.

Safiya Umoja Noble. 2018. Algorithms of oppression: How search engines reinforce racism.

OpenAI. 2024a. GPT-4o mini: Advancing Cost-Efficient Intelligence.

OpenAI. 2024b. Hello GPT-4o.

Nedjma Djouhra Ousidhoum, Xinran Zhao, Tianqing Fang, and et al. 2021. Probing toxic content in large pre-trained language models. In *Annual Meeting of the Association for Computational Linguistics*.

Evi Papadopoulou, Hadi Mohammadi, and Ayoub Bagheri. 2024. Large language models as mirrors of societal moral standards. *arXiv preprint arXiv:2412.00956*.

Pew Research Center. 2023. Attitudes on an interconnected world.

Alec Radford, Jeff Wu, Rewon Child, and et al. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam M. Shazeer, Adam Roberts, and et al. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2019. Pseudolikelihood reranking with masked language models. *ArXiv*, abs/1910.14659.

Teven Le Scao, Angela Fan, Christopher Akiki, and et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100.

Richard A. Shweder, Nancy C. Much, Manamohan Mahapatra, and Lawrence Park. 1997. The "big three" of morality (autonomy, community, divinity) and the "big three" explanations of suffering.

Karolina Stańczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *ArXiv*, abs/2112.14168.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, and et al. 2023a. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin R. Stone, and et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, and et al. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Susan Zhang, Stephen Roller, Naman Goyal, and et al. 2022. Opt: Open pre-trained transformer language models. *ArXiv*, abs/2205.01068.

James Zou and Londa Schiebinger. 2018. Ai can be sexist and racist — it's time to make it fair. *Nature*, 559:324 – 326.

Didar Zowghi and Francesca da Rimini. 2023. Diversity and inclusion in artificial intelligence. *ArXiv*, abs/2305.12728.

A Topic Codes for WVS and PEW

Table 4: Mapping of Topic Codes to the Dataset (WVS or PEW) and their corresponding moral questions.

Topic Code	Dataset	Moral Question
Q177	WVS	Claiming government benefits to which you are not entitled
Q178	WVS	Avoiding a fare on public transport
Q179	WVS	Stealing property
Q180	WVS	Cheating on taxes
Q181	WVS	Someone accepting a bribe in the course of their duties
Q182	WVS	Homosexuality
Q183	WVS	Prostitution
Q184	WVS	Abortion
Q185	WVS	Divorce
Q186	WVS	Sex before marriage
Q187	WVS	Suicide
Q188	WVS	Euthanasia
Q189	WVS	For a man to beat his wife
Q190	WVS	Parents beating children
Q191	WVS	Violence against other people
Q192	WVS	Terrorism as a political, ideological or religious mean
Q193	WVS	Having casual sex
Q194	WVS	Political violence
Q195	WVS	Death penalty
Q84A	PEW	Using contraceptives
Q84B	PEW	Getting a divorce
Q84C	PEW	Having an abortion
Q84D	PEW	Homosexuality
Q84E	PEW	Drinking alcohol
Q84F	PEW	Married people having an affair
Q84G	PEW	Gambling
Q84H	PEW	Sex between unmarried adults

B WVS & PEW scores by country

Figure 8 compares normalized WVS (orange) and PEW (gold) scores by country. Each box shows the interquartile range, with medians as horizontal lines and diamonds marking outliers. The broader spread in the WVS data for many countries suggests higher variance in moral acceptance. Some countries, such as the United States or Czech Republic, show very wide ranges, from near -1 (*never justifiable*) to close to $+1$ (*always justifiable*). Others, often in the Middle East or South Asia, have more negative medians, reflecting stricter cultural norms on certain issues.

C Individual Figures by Model & Dataset

In each scatter plot, the horizontal axis survey_score corresponds to `WVS` in Figure 9 and `PEW` ratings in Figure 10. Meanwhile, the vertical axis log_prob_diff shows the difference between the log-probability the model assigns to a *morally justifiable* statement vs. a *morally unjustifiable* statement. A positive slope suggests that higher survey acceptance correlates with higher log-prob differences in the same direction, meaning better alignment. Conversely, negative slopes may show systematic misalignment on that dimension.

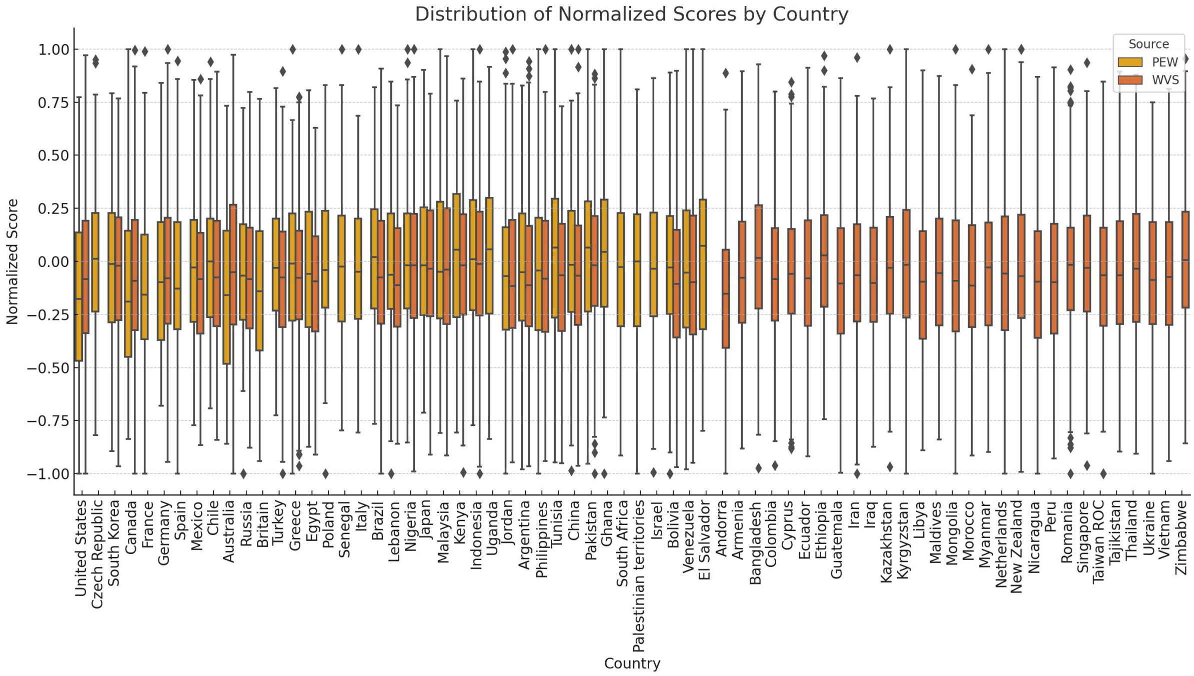


Figure 8: Distribution of normalized WVS (orange) and PEW (gold) survey scores by country.

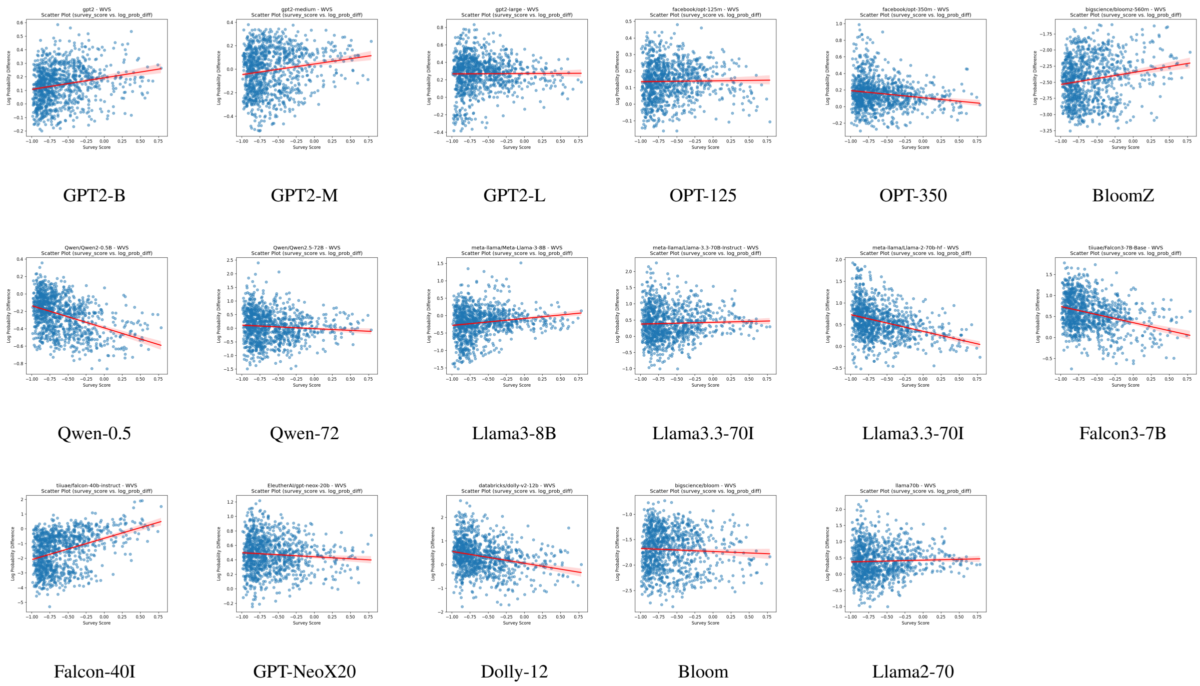
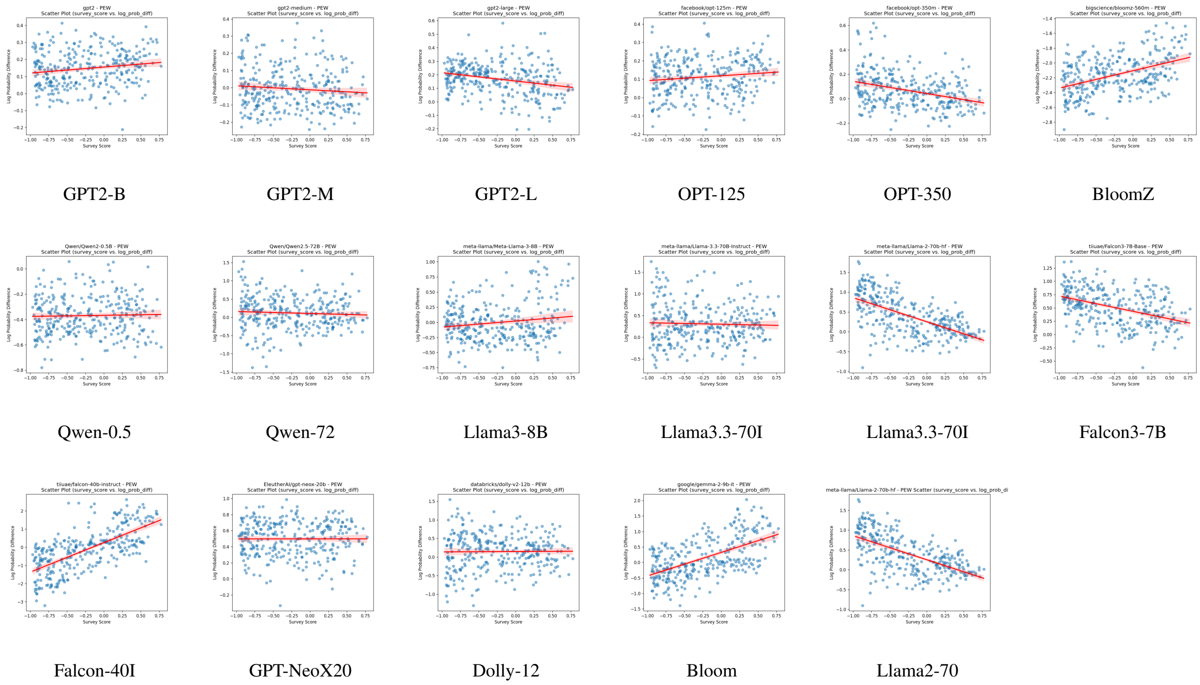


Figure 9: Scatter plots for WVS dataset



GPT2-B | GPT2-M | GPT2-L | OPT-125 | OPT-350 | BloomZ

Qwen-0.5 | Qwen-72 | Llama3-8B | Llama3.3-70I | Llama3.3-70I | Falcon3-7B

Falcon-40I | GPT-NeoX20 | Dolly-12 | Bloom | Llama2-70

Figure 10: Scatter plots for PEW dataset