On the Credibility of Evaluating LLMs using Survey Questions

Jindřich Libovický
Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
V Holešovičkách 747/2, 180 00 Praha, Czechia
libovicky@ufal.mff.cuni.cz

Abstract

Recent studies evaluate the value orientation of large language models (LLMs) using adapted social surveys, typically by prompting models with survey questions and comparing their responses to average human responses. This paper identifies limitations in this methodology that, depending on the exact setup, can lead to both underestimating and overestimating the similarity of value orientation. Using the World Value Survey in three languages across five countries, we demonstrate that prompting methods (direct vs. chain-of-thought) and decoding strategies (greedy vs. sampling) significantly affect results. To assess the interaction between answers, we introduce a novel metric, self-correlation distance. This metric measures whether LLMs maintain consistent relationships between answers across different questions, as humans do. This indicates that even a high average agreement with human data, when considering LLM responses independently, does not guarantee structural alignment in responses. Additionally, we reveal a weak correlation between two common evaluation metrics, mean-squared distance and KL divergence, which assume that survey answers are independent of each other. For future research, we recommend CoT prompting, sampling-based decoding with dozens of samples, and robust analysis using multiple metrics, including self-correlation distance.

1 Introduction

Evaluating the values expressed in texts generated by Large Language Models (LLMs) is crucial for shaping public perception, informing policy, and ensuring the ethical use of AI. A common evaluation method involves prompting LLMs with questions from standardized surveys and comparing their responses to human answers or calibrated scores (see Table 1 for a comprehensive list of related work).

This approach has been criticized for its inconsistency with open-ended generation (Wright et al., 2024) and sensitivity to prompt formulation (Röttger et al., 2024; Motoki et al., 2024). We address a related issue: the reliability of measuring similarity between model and human responses, particularly in how answers to different questions correlate with one another.

We examine this framework using (1) direct vs. Chain-of-Thought (CoT) prompting, (2) greedy decoding vs. nucleus sampling, and (3) three similarity metrics: mean squared difference, KL divergence, and our novel self-correlation distance. Unlike previous metrics that estimate average alignment by treating survey questions in isolation, our metric accounts for interactions among survey answers.

Using LLaMA 3, Mistral 2, EuroLLM, and Qwen 2.5, we compare model responses in the World Value Survey (WVS; Haerpfer et al., 2020) with opinions from selected countries. We find that choices in prompting, decoding, and metrics yield different conclusions. For example, greedy decoding deviates from typical results obtained with nucleus sampling, while short categorical answers underestimate alignment compared to CoT prompting. The self-correlation distance indicates that, despite a high average alignment with survey data, different correlation patterns reveal potential overgeneralization. Metrics that treat answers independently can thus overestimate alignment.

In this paper, we first review studies that compare LLM-generated answers with population survey responses (Section 2). We then describe our experiments (Section 3), including the models we use, the prompts we employ, and the inference algorithm we employ. We describe the metrics we use, including the newly introduced self-correlation distance. Section 4 presents the results, and Section 5 concludes the paper.

2 LLMs and Surveys

Many studies examine the values encoded in language models, focusing on moral, cultural, and political biases. These often rely on surveys designed for human respondents (see Table 1). They cover topics ranging from general ethics to specific domains, such as political ideologies (Feng et al., 2023), autonomous vehicle ethics (Vida et al., 2024), and religious values (Liu et al., 2024).

Model responses are typically evaluated using two approaches: (1) evaluation keys, as in frameworks like the Political Compass or Moral Foundations Questionnaire (Graham et al., 2011; MFQ), or (2) comparisons with human population data.

Most studies prompt models to produce categorical

Proceedings of the First Workshop on Multilingual Multicultural Evaluation, pages 23–34
March 28, 2026 ©2026 Association for Computational Linguistics

	Gold survey data	Open-source models	Persona	Decoding (# samples)	Categorial output	Comparison w/ humans
Santurkar et al. (2023)	Custom	No	Yes	Logits	Yes	WD
Cao et al. (2023)	Hofstede (1984)	No	No	Sampling (1)	Yes	Accuracy
Feng et al. (2023)	Political compass	Yes	No	10-best tokens	No	—
Olmedo et al. (2023)	Mather et al. (2005)	Yes	No	Logits	Yes	KL
Scherrer et al. (2023)	Custom	Yes	No	Full samp. (5/10)	No	—
Sanders et al. (2023)	Custom	No	Yes	Sampling? (100)	Yes	WD
Benkler et al. (2023)	WVS	No	Yes	Nucleus samp. (1)	No	—
Tao et al. (2024)	WVS, EVS	No	Yes	Greedy	Yes	MSD
Durmus et al. (2024)	Custom	No	Yes	Logits	Yes	JSD
Ceron et al. (2024)	Custom	Yes	Yes	Nucleus samp. (30)	Yes	—
Nunes et al. (2024)	MFQ, MFV	Yes	No	Sampling? (1)	Yes	—
Vida et al. (2024)	Awad et al. (2018)	Yes	No	Sampling? (1)	Yes	Accuracy
Xu et al. (2024)	WVS	Yes	Yes	Nucleus samp. (1)	Yes	Norm. MSD
Liu et al. (2024)	Center (2018)	Yes	No	Greedy	Yes	—
Kim and Baek (2024)	WVS	Yes	No	Logits	Yes	Pearson
Aksoy (2024)	MFQ	Yes	No	Sampling? (100)	Yes	—
Shen et al. (2024)	Schwartz (1992)	Yes	No	Sampling? (10)	Yes	L1
Sukiennik et al. (2025)	Hofstede (1984)	Yes	No	Greedy	Yes	Norm. L1
Qu and Wang (2024)	WVS	No	Yes	Sampling (100)	Yes	MSD
Kazemi et al. (2024)	WVS	No	No	Not specified	Yes	Accuracy
Rupprecht et al. (2025)	WVS	Yes	No	Logits	Yes	Accuracy
Gurgurov et al. (2025)	Political compass	Yes	No	Nucleus sampling (1)	Yes	—
Bulté and Rigouts (2025)	(Hofstede, 1984) + WVS	Yes	No	Nucleus sampling (6)	Yes	Accuracy
Costa et al. (2025)	MFQ	Yes	Yes	Sampling? (1)	Yes	—
Atari et al. (2023)	WVS	No	No	Sampling? (100)	Yes	Fixation ind.

Gold survey data: WVS = World Value Survey (Haerpfer et al., 2020), EVS = European Value Study (EVS, 2022), MFQ = Moral Foundation Questionnaire (Graham et al., 2011), MFV = Moral Foundation Vignettes (Clifford et al., 2015)
Comparison methods: WD = Wasserstein Distance, Acc. = Accuracy, KL = Kullback-Leibler Divergence, JSD = Jensen-Shannon Distance, MSD = Mean Squared Difference

Table 1: Overview of publications using standardized surveys to evaluate values in LLMs.

answers, such as selecting an option or providing a score. These responses are compared to population data using metrics like KL Divergence (Olmedo et al., 2023), Jensen-Shannon Divergence (Durmus et al., 2024), or Wasserstein Distance (Santurkar et al., 2023; Sanders et al., 2023). However, these methods often focus on single-token generation, which limits their generalization to longer text samples. Some evaluations use greedy decoding or a single sampled response, with sampling details often unspecified. Persona probing (i.e., specifying demographic traits for models to emulate) is also common but typically focuses on English outputs, leaving biases in other languages underexplored.

Several recent studies have extended survey-based evaluation methodologies. Sukiennik et al. (2025) conducted the first large-scale evaluation of cultural alignment across 20 countries and 10 LLMs, using Hofstede's Cultural Values Questionnaire. The results found that models generally represent a moderate cultural middle ground, with the United States showing the best alignment. Qu and Wang (2024) used WVS data to evaluate ChatGPT's public opinion simulation capabilities, revealing significant performance disparities favoring Western, English-speaking nations and demographic biases across gender, ethnicity, and social class. Kazemi et al. (2024) demonstrated that 44% of GPT-4o's ability to reflect societal values correlates with digital resource availability in a society's primary language, with error rates in low-resource languages exceeding those in high-resource languages by a factor of five. Most recently, Rupprecht et al. (2025) extended bias research in LLM survey responses using WVS data, testing perturbations in answer and question phrasing across multiple models and finding significant sensitivity to prompt variations that mirror human response biases.

Röttger et al. (2024); Wang et al. (2024) and Moore et al. (2024) highlight how prompt formulations, such as multiple-choice setups, significantly influence model outputs and score robustness. While their work examines the robustness of the prompt with respect to evaluation keys (such as the Political Compass), it does not address the methodological aspects of comparing model outputs to population survey data. Also, as far as we know, all previous work treats survey responses as independent and disregards correlations between questions, an issue we address by introducing the self-correlation distance.

Previous work also evaluated generation consistency (Kumar and Joshi, 2022; Bonagiri et al., 2024). However, it focuses on cases with a well-specified ground truth. We are interested in a slightly different type of consistency: Statements like people who tend to say A, also tend to say B to some extent. In this paper, we introduce a metric to measure the extent to which LLMs capture these tendencies, independent of the actual content.

Recent work has also examined cultural adaptation methods using survey data. Adilazuarda et al. (2025) found that WVS-based training can lead to cultural homogenization and undermine factual knowledge, and introduced a cultural distinctiveness metric that complements existing evaluation approaches. Their findings that survey data alone may be insufficient for cultural adaptation align with our observations about the limitations of current evaluation methodologies.

3 Experiments

Following several previous studies (Benkler et al., 2023; Tao et al., 2024; Kim and Baek, 2024), we prompt LLMs with World Value Survey (Round 7, version 5.0) questions and compare their answers with human data using three evaluation methods. WVS is a global research project that has comprehensively explored people's values and beliefs since 1981. The World Value Survey covers 55 countries and 80 languages, making it likely the most comprehensive standardized resource for comparing values in LLM outputs with the human population across languages and cultures. In this work, we focus on evaluation metrics and conduct experiments only on a small subset.

As in related work, we simulate the survey using an LLM and compare the results with those obtained from the human population in the respective countries. Since the answers to all questions are integers, depending on the evaluation metric, we either use the average answers or the categorical distributions of answers as the ground truth for comparison. For the correlation study, we use responses from individual respondents and compute how responses to individual questions correlate with each other.

The source code to replicate the experiments is available at `https://github.com/jlibovicky/llm-survey-eval`.

3.1 Model Prompting

Questionnaire design. WVS is not a self-assessment questionnaire. Interlocutors interview the subjects and, based on their answers, record integer scores for each question, most often on a scale from 1 to 10, indicating the extent to which they agree with a statement.

We use general, non-personalized formulations of the questions, i.e., we exclude questions about income, health, or personal experiences, which would likely be rejected by the models. We reformulated the questions to contain more general, non-personal statements (e.g., replacing "your life" with "human life") to further reduce rejections. After excluding questions that were not in all language versions, 143 prompts remained. Prompts were created in English, machine-translated into German and Czech using Google Translate, and then manually post-edited with reference to official WVS translations by native speakers. The questionnaire was administered in a single conversation session to allow evaluation of answer consistency (see examples in Appendix A).

Scores vs. Chain-of-Thought. We compare two prompt types: direct numeric answers and chain-of-thought prompts (Wei et al., 2022), where justification precedes the answer. We posit that chain-of-thought is closer to real-world LLM use, as chat-like user interactions typically involve longer generations than a single categorical output, e.g., in interactive sessions.

Greedy vs. Sampling. Studies often use greedy decoding for its efficiency and for producing a deterministic output. It approximates the most probable sequence but may not yield typical responses, as the probability mass is distributed across similar sequences (Eikema and Aziz, 2020; Wiher et al., 2022). Because of this, and since sampling is more common in practice, we compare greedy decoding with nucleus sampling (nucleus 0.9, temperature 0.7), including an estimation of how many samples are needed to obtain a stable result for the given metrics. Following Andreas (2022), who argues that language models should be treated as ensembles of different multiple agents, and Lederman and Mahowald (2024), who argue that language models are compressed libraries, we assume that sampling one conversation session might correspond to one agent within the language model in the agent metaphor and retrieving one set of world knowledge from the library in the library metaphor. Therefore, we treat the conversation sessions as comparable to individual respondents in the survey.

3.2 Evaluation

We evaluate model outputs against data from the USA, UK, Czechia, and Germany, where the prompt languages are spoken. Iran and China are included as culturally distinct cases, a sanity check for metric validation.[1] Based on the results of previous studies, we expect that due to the prevalence of English data, models will tend to better align with Western countries. We use three metrics: Mean Squared Difference (MSD), Kullback-Leibler Divergence (KLD), and a novel self-correlation distance.

Models. We use four instruction-tuned models: LLaMA 3 8B Instruct (Dubey et al., 2024) and Mistral v0.1 7B Instruct (Jiang et al., 2023), EuroLLM 9B Instruct (Martins et al., 2024), and Qwen 2.5 7B Instruct (Yang et al., 2024). This selection includes both general-purpose models (LLaMA 3 from the USA, Mistral from France) and models with specific regional focuses (EuroLLM for European languages, Qwen developed in China for multilingual applications), allowing us to examine how model design influences value alignment across cultures.

[1] The number of WVS participants in the respective countries was: USA: 2,596, UK: 2,609, Czechia: 1,200, Germany: 1,528, Iran: 1,499, China: 3,036.

Mean Square Difference. Most WVS questions have numerical answers on a 1-to-10 scale or lower. We scale the answer to the 0–1 interval, compute the squared differences between the scores sampled from the model and human population averages, and average them over all questions.

KL Divergence. We treat the model and human answers as categorical distributions. For the model outputs, we normalize the distribution over model runs. For the survey data, we normalize over the participants. We compute the Kullback-Leibler divergence between the sampled model answer distribution and the distribution of answers in the human population.

Self-Correlation Distance. The previous metrics fail to account for the fact that questionnaire responses often correlate with one another due to underlying consistency in values and opinions, yet all questions are treated as conditionally independent. This assumption is not realistic. Value opinions often come in bundles and follow patterns that may differ across various societies. Simple examples might include people who believe that religion should play a stronger role in society being more likely to say that mothers of young children should stay at home with their children, or that individuals concerned about social justice are often also concerned about the environment. These correlations between individual respondents' answers capture second-order patterns, relationships between answers, that are not apparent when comparing individual answers alone, as with MSE or KL-Divergence.

To analyze this, we compute self-correlation matrices that measure the Pearson correlation between all pairs of questionnaire responses. In the survey data, we calculate the correlation between responses to individual questions across participants. In the LLM case, we calculate the correlation between answers across model runs. This gives a matrix with all question pairs. Using the Frobenius norm, we quantify how internally consistent or "principled" the answers are. High norm means that the absolute value of the correlations tends to be high, whereas a norm close to zero means that questions in the survey are more independent of each other.

Additionally, we compare the self-correlation matrices of model outputs and human responses using the Frobenius norm. This metric allows us to evaluate whether the underlying structure of the answers aligns between models and humans, going beyond simple agreement on individual responses taken independently.

4 Results

Comparing MSD, KLD, and self-correlation distance. The results comparing model responses to human surveys are presented in Table 2 for the USA, with additional results for other countries and languages provided in Table 6 in the Appendix. Within languages and countries, results follow similar trends across setups.

To interpret our results, we first establish baseline values from human populations. Country-level comparisons in the WVS yield MSD values ranging from 0.009 (USA-UK) to 0.069 (Germany-Iran), with Western countries showing differences of 0.009–0.024 (see Table 7 in the Appendix). KLD between countries ranges from 0.07 (USA-UK) to 0.44 (Germany-Iran), with Western countries showing 0.07–0.22. Self-correlation distances between human populations range from 0.64 (China-Iran) to 0.95 (USA-Iran), with typical values between 0.79 and 0.92.

Using MSD and KLD metrics, which treat answers independently, we observe substantial variation across prompting and decoding strategies. With CoT prompting and sampling, Mistral 2 achieves remarkably low MSD (0.022) and KLD (0.26) for USA data, comparable to differences between Western countries. However, the same model with greedy decoding and CoT prompting shows drastically worse alignment (MSD=0.188), exceeding even USA-Iran differences. LLaMA 3 shows more moderate values (MSD=0.059, KLD=1.47 for CoT+Sampling), falling between Western and cross-cultural differences. EuroLLM exhibits the poorest alignment with direct prompting and greedy decoding (MSD=0.165–0.284), though sampling decreases the distance between model and population substantially. Qwen demonstrates relatively stable performance across setups (MSD=0.041–0.199).

The self-correlation distance metric reveals a paradox: setups that achieve the best surface-level alignment often exhibit the poorest structural alignment. Mistral 2, with CoT and sampling, despite having the lowest MSD (0.022) and KLD (0.26), exhibits a self-correlation distance of 1.62, which is far greater than the distances between any human populations (0.64–0.95). Its correlation norm (2.80) is also substantially higher than human values ($\sim$1.66), which indicates overly rigid response patterns. In contrast, LLaMA 3 with the same setup shows better structural alignment (self-correlation distance=1.29, correlation norm=1.70) despite worse surface metrics (MSD=0.059, KLD=1.47). Qwen, with score-only prompts and sampling, produces the most structured responses (correlation norm=3.33, self-correlation distance=2.13) with an even greater departure from human response variability.

Across all models and prompting strategies, greedy decoding consistently underestimates alignment when measured with MSD. For instance, comparing LLaMA 3 with score-only prompts, greedy decoding yields MSD=0.098, versus 0.073 with sampling, roughly a one-third increase. The disparity is even more pronounced for KLD, where sampling values are often 2–3× higher than greedy decoding, indicating that greedy decoding captures only a narrow slice of the probability distribution reachable by common decoding algorithms.

The effect of CoT prompting varies across models and decoding strategies. For LLaMA 3 with sampling, CoT improves alignment (MSD decreases from 0.073 to 0.059). For Mistral 2 with sampling, CoT dramat-

Model	Prompt type	De-co-de	MSD	KLD	Corr. norm	Self-corr. dist.
LLaMA 3	Score only	Gr.	.098	1.71		
		Spl.	.073	2.99	0.90	1.26
	CoT	Gr.	.088	1.68		
		Spl.	.059	1.47	1.70	1.29
Mistral2	Score only	Gr.	.094	1.80		
		Spl.	.041	0.77	1.77	1.17
	CoT	Gr.	.188	1.95		
		Spl.	.022	0.26	2.80	1.62
EuroLLM	Score only	Gr.	.165	1.91		
		Spl.	.059	0.72	2.55	1.56
	CoT	Gr.	.130	1.76		
		Spl.	.125	0.97	2.97	1.74
Qwen	Score only	Gr.	.082	1.66		
		Spl.	.074	1.27	3.33	2.13
	CoT	Gr.	.199	1.60		
		Spl.	.062	1.13	1.52	1.25

Table 2: A comparison of model outputs with different prompting strategies (Score-only, CoT: Chain of Thought), decoding method (Gr.: greedy, Spl.: sampling). For other countries and languages, see Table 6 in the Appendix.

	MSD	KLD	CorrD
MSD	—	.465	-.389
KLD	.717	—	-.083
CorrD	-.374	-.064	—

Table 3: Correlation of the metrics (Pearson above the diagonal, Spearman under the diagonal) for both models over all languages and countries.

ically improves both MSD (0.041→0.022) and KLD (0.77→0.26). However, with greedy decoding, CoT can worsen results: Mistral 2's MSD increases from 0.094 to 0.188. This suggests that CoT prompting is beneficial primarily when combined with sampling-based decoding, likely because it provides more stable generation that better leverages the sampling strategy.

Table 3 shows moderate overall correlation between MSD and KLD (Pearson=0.465, Spearman=0.717), but Table 5 reveals this masks dramatic model-specific differences. Mistral 2 shows a very strong correlation (Pearson=0.832, Spearman=0.925), meaning MSD and KLD largely agree on what constitutes good alignment for this model. LLaMA 3, however, shows weak correlation (Pearson=0.276, Spearman=0.389), indicating that these metrics measure somewhat different aspects of value alignment for this model. Critically, self-correlation distance is negatively correlated with both MSD (Pearson=-0.389) and KLD (Pearson=-0.083), confirming that surface-level and structural alignment are different, and sometimes opposing, qualities.

Model	Prompt type	Deco-de	MSD		KLD		Self-corr. dist.	
LLaMA 3	Score only	Gr.	-.19	.18	-.07	.21		
		Spl.	-.08	.24	.08	.24	.02	-.25
	CoT	Gr.	-.12	.26	-.29	.21		
		Spl.	-.03	.34	-.17	.12	-.12	-.28
Mistral2	Score only	Gr.	.26	.23	.25	.36		
		Spl.	.01	.33	-.08	.17	.07	.01
	CoT	Gr.	-.16	.08	-.01	.13		
		Spl.	.24	.26	.21	.20	.19	.21
EuroLLM	Score only	Gr.	.37	.16	.30	.08		
		Spl.	.24	-.01	.18	.07	.02	.03
	CoT	Gr.	.28	.12	-.18	-.16		
		Spl.	.07	-.26	.15	-.15	-.17	-.00
Qwen	Score only	Gr.	.46	.45	-.20	-.07		
		Spl.	-.26	-.00	-.14	.03	-.17	-.15
	CoT	Gr.	-.12	.11	-.10	.03		
		Spl.	-.05	.28	-.12	.09	.03	-.16

Color scale from -.3 through 0 to .3

Table 4: Point-Biserial Correlation of the alignment metrics with matching country and language as a binary indicator variable. The left part shows the correlation only within Western countries; the right part also includes Iran and China.

Cross-lingual and cross-cultural patterns. Table 4 summarizes how language-country matching affects alignment metrics. The correlations are generally weak, but patterns emerge when culturally distant countries (Iran and China) are included. Most models exhibit positive correlations between matching conditions and alignment metrics, suggesting that prompts in English, Czech, and German align more closely with Western surveys than with those from Iran or China. However, the strength of this effect varies: EuroLLM shows the strongest language-country matching effects (up to 0.37), likely reflecting its European-centric training, while Qwen exhibits more uniform performance across language-country combinations, consistent with its multilingual design.

Notably, the correlation patterns differ substantially across prompting and decoding strategies. Mistral 2 with CoT prompting and nucleus sampling shows the most consistent positive correlations with both language matching (0.24 for MSD, 0.21 for KLD) and the inclusion of distant cultures (0.26 for MSD, 0.20 for KLD). This suggests that this setup not only achieves the best average alignment but also shows more predictable cross-cultural patterns.

Self-correlation patterns. Figure 1 visualizes the self-correlation matrices, revealing systematic differences in how models structure their responses compared to humans. The visualization shows three major patterns: (1) Block diagonal structures indicate question clusters with strong internal correlations. (2) The intensity of col-

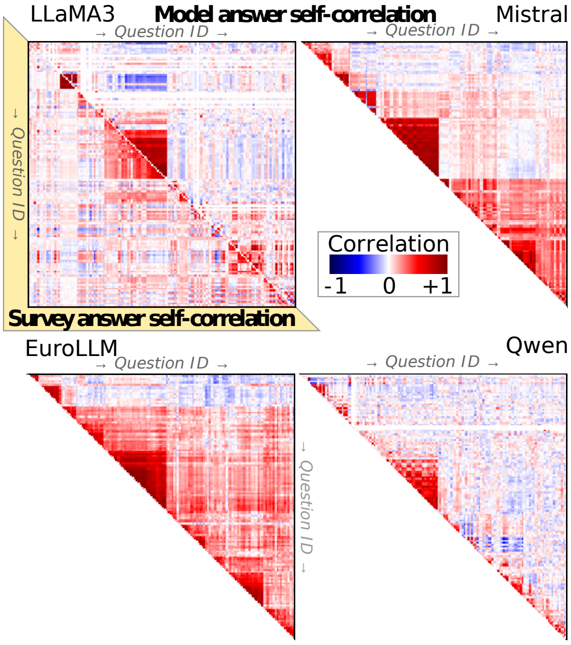

Figure 1: Correlation patterns between human answers in the USA (under the diagonal) and between answers of the LLaMA 3 and Mistral 2 models in English (above the diagonal).

ors shows correlation strength: models produce darker, more saturated colors than humans. (3) Off-diagonal patterns reveal how different value domains relate; models show simpler, more predictable patterns than the human answers.

All models show overly consistent responses in sections about cultural identity and national pride (visible as darker red blocks along the diagonal). Mistral 2 displays particularly strong self-correlation in questions about social and political attitudes, far exceeding human patterns. These visualizations confirm that, while models may match human averages on individual questions, their internal response structures systematically diverge from human response patterns.

The quantitative self-correlation distances reinforce this finding: model-to-human distances (1.1–2.1 across setups) consistently exceed human-to-human distances (0.64–0.95), indicating that all tested models impose more rigid correlation structures than exist in human populations. This suggests that models generate “principled” but overly simplistic response patterns, potentially missing the nuanced and sometimes inconsistent nature of human value systems.

Number of sampled responses. Figure 2 demonstrates how metric estimates stabilize with increasing sample size. For MSD, approximately 100 samples are needed for stable estimates, with values converging to within 0.005 of the final estimate. KLD requires more samples, several hundred for full stability. CoT prompting produces more stable estimates with fewer samples compared to direct prompts, likely because the reasoning process provides additional structure that reduces sampling variance. Importantly, greedy decoding produces estimates that deviate significantly from the sampling-based norm across both metrics and all sample sizes. This confirms that greedy decoding systematically misestimates both average alignment and response distributions. Most prior studies (Table 1) used far fewer samples (typically 1–10), suggesting their conclusions may have differed substantially with more comprehensive sampling, particularly for KLD estimates.

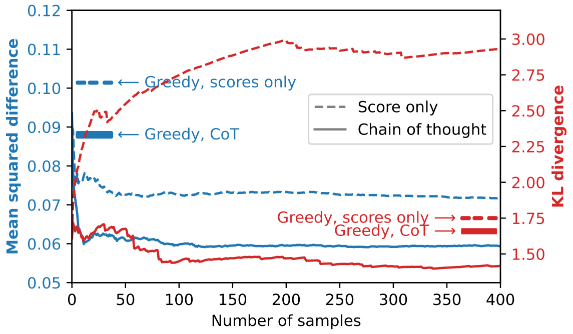

Figure 2: Mean-squared difference and KL-divergence of LLaMA 3 answers when compared to the USA data of the World Value Survey. It compares the greedy decoding and sampling from the model.

Model-specific behaviors. The four tested models show distinct patterns. Mistral 2 achieves the best surface-level scores but at the cost of the highest structural rigidity, suggesting it may be overfitting to typical responses. LLaMA 3 shows the most balanced profile across metrics. EuroLLM’s strong language-matching effects (correlation of up to 0.37) reflect its European-centric training data. Qwen, despite being developed in China, shows relatively uniform cross-lingual performance but produces highly structured responses (correlation norms up to 3.33).

5 Conclusions

This study examined the impact of decoding strategies and evaluation metrics on comparisons of LLM responses to population survey data, using the World Value Survey as a case study across three languages and six countries.

We found that setups closely mirroring real-world LLM usage, specifically, Chain-of-Thought prompting with sampling-based decoding, achieve the best alignment with survey data. **Prior work relying on direct prompts and greedy decoding may underestimate average alignment when evaluating answers independently.**

To address gaps in current evaluations, we introduced the self-correlation distance, a novel metric that captures consistency and interaction between answers. Unlike traditional metrics such as MSD and KL Divergence, the self-correlation distance showed that high scores in some setups indicate a lower diversity of responses than in the human population. The high average numbers are achieved at the cost of generating typical cases, resulting in overly structured responses rather than an accurate reflection of survey variability, especially in social and political topics. **This shows that metrics treating answers independently overestimate the alignment.** Discrepancies between metrics, such as LLaMA 3's low correlation between MSD and KLD (Pearson=0.276), underscore the importance of multi-metric evaluations.

For future research, we recommend using CoT prompting, nucleus sampling with at least 100 samples to achieve stable estimates, and a multi-metric approach that incorporates our proposed self-correlation distance to capture both surface-level and structural alignment.

Limitations

The assumption behind all surveys is that the answers provided by the survey subjects reflect their actual behavior in the real world. For example, a man might claim that he believes men should do a fair share of invisible household labor and will vote for parties with a compatible election program. With LLMs, there is no such guarantee. LLMs might advocate for certain values when prompted to generate text directly related to those values, but still generate texts with underlying values that do not align with the survey answer. We do not challenge this assumption in this paper. We are not aware of any existing methodology that would approach this challenge.

Comparing answers across countries based on answer distributions is a simplifying assumption. Demographic factors other than nationality, such as age or economic status, may also play a role. This study used cross-country distributions to gain insights into LLM evaluation, rather than making claims about LLMs being more representative of certain countries than others. Such claims would require a more detailed methodology.

The primary objective of this study was to compare metrics; therefore, we included only a few variables that could influence the model's behavior. It is possible that different prompt formulations and question order in the questionnaire may also yield slightly different results.

Acknowledgments

Many thanks to Jan Hajič Jr. and Dominik Macháček for comments on the draft of this paper.

We used GitHub Copilot when writing our code. When writing the paper, we used Grammarly and Claude to improve grammar and spelling.

This research was supported by the Charles University project PRIMUS/23/SCI/023 and project CZ.02.01.01/00/23_020/0008518 of the Czech Ministry of Education.

References

Muhammad Farid Adilazuarda, Chen Cecilia Liu, Iryna Gurevych, and Alham Fikri Aji. 2025. From surveys to narratives: Rethinking cultural value adaptation in llms. *CoRR*, abs/2505.16408.

Meltem Aksoy. 2024. Whose morality do they speak? unraveling cultural bias in multilingual language models. *CoRR*, abs/2412.18863.

Jacob Andreas. 2022. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mohammad Atari, Mona J. Xue, Peter S. Park, Damián E. Blasi, and Joseph Henrich. 2023. Which humans? *PsyArXiv*.

Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature*, 563(7729):59–64.

Noam Benkler, Drisana Mosaphir, Scott Friedman, Andrew Smart, and Sonja Schmer-Galunder. 2023. Assessing llms for moral value pluralism. *CoRR*, abs/2312.10075.

Vamshi Krishna Bonagiri, Sreeram Vennam, Priyanshul Govil, Ponnurangam Kumaraguru, and Manas Gaur. 2024. SaGE: Evaluating moral consistency in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14272–14284, Torino, Italia. ELRA and ICCL.

Bram Bulté and Terryn Ayla Rigouts. 2025. Llms and cultural values: The impact of prompt language and explicit cultural framing. *Computational Linguistics*, pages 1–85.

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

Pew Research Center. 2018. The religious typology: A new way to categorize americans by religion.

Tanise Ceron, Neele Falk, Ana Barić, Dmitry Nikolaev, and Sebastian Padó. 2024. Beyond prompt brittleness: Evaluating the reliability and consistency of political worldviews in LLMs. *Transactions of the Association for Computational Linguistics*, 12:1378–1400.

Scott Clifford, Vijeth Iyengar, Roberto Cabeza, and Walter Sinnott-Armstrong. 2015. Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior research methods*, 47(4):1178–1198.

Davi Bastos Costa, Felippe Alves, and Renato Vicente. 2025. Moral susceptibility and robustness under persona role-play in large language models. *CoRR*, abs/2511.08565.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

EVS. 2022. Evs trend file 1981-2017. GESIS, Cologne. ZA7503 Data file Version 3.0.0, https://doi.org/10.4232/1.14021.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Jesse Graham, Brian Nosek, Jonathan Haidt, Ravi Iyer, Sena P Koleva, and Peter H Ditto. 2011. Mapping the moral domain. *Journal of Personality and Social Psychology*, 101 (2):366–385.

Daniil Gurgurov, Katharina Trinley, Ivan Vykopal, Josef van Genabith, Simon Ostermann, and Roberto Zamparelli. 2025. Multilingual political views of large language models: Identification and steering. *CoRR*, abs/2507.22623.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2020. World values survey wave 7 (2017-2020) cross-national data-set. *(No Title)*.

Geert Hofstede. 1984. *Culture's consequences: International differences in work-related values*, volume 5. sage.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Sharif Kazemi, Gloria Gerhardt, Jonty Katz, Caroline Ida Kuria, Estelle Pan, and Umang Prabhakar. 2024. Cultural fidelity in large-language models: An evaluation of online language resources as a driver of model performance in value representation. *CoRR*, abs/2410.10489.

Minsang Kim and Seungjun Baek. 2024. Exploring large language models on cross-cultural values in connection with training methodology. *CoRR*, abs/2412.08846.

Ashutosh Kumar and Aditya Joshi. 2022. Striking a balance: Alleviating inconsistency in pre-trained models for symmetric classification tasks. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1887–1895, Dublin, Ireland. Association for Computational Linguistics.

Harvey Lederman and Kyle Mahowald. 2024. Are language models more like libraries or like librarians? bibliotechnism, the novel reference problem, and the attitudes of LLMs. *Transactions of the Association for Computational Linguistics*, 12:1087–1103.

Songyuan Liu, Ziyang Zhang, Runze Yan, Wei Wu, Carl Yang, and Jiaying Lu. 2024. Measuring spiritual values and bias of large language models. *CoRR*, abs/2410.11647.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno Miguel Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, M. Amin Farajian, Manuel

Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *CoRR*, abs/2409.16235.

Mark Mather, Kerri L Rivers, and Linda A Jacobsen. 2005. The american community survey. *Population Bulletin*, 60(3).

Jared Moore, Tanvi Deshpande, and Diyi Yang. 2024. Are large language models consistent over value-laden questions? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15185–15221, Miami, Florida, USA. Association for Computational Linguistics.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: measuring chatgpt political bias. *Public Choice*, 198(1):3–23.

José Luiz Nunes, Guilherme F. C. F. Almeida, Marcelo de Araujo, and Simone D. J. Barbosa. 2024. Are large language models moral hypocrites? a study based on moral foundations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):1074–1087.

Ricardo Dominguez Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2023. Questioning the survey responses of large language models. *CoRR*, abs/2306.07951.

Yao Qu and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1–13.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.

Jens Rupprecht, Georg Ahnert, and Markus Strohmaier. 2025. Prompt perturbations reveal human-like biases in LLM survey responses. *CoRR*, abs/2507.07188.

Nathan E. Sanders, Alex Ulinich, and Bruce Schneier. 2023. Demonstrations of the potential of AI-based political issue polling. *Harvard Data Science Review*, 5(4). Https://hdsr.mitpress.mit.edu/pub/dm2hrtx0.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Nino Scherrer, Claudia Shi, Amir Feder, and David M. Blei. 2023. Evaluating the moral beliefs encoded in llms. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, New Orleans, LA, USA.

Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Yu-Ju Yang, Tanushree Mitra, and Yun Huang. 2024. Valuecompass: A framework of fundamental values for human-ai alignment. *CoRR*, abs/2409.09586.

Nicholas Sukiennik, Chen Gao, Fengli Xu, and Yong Li. 2025. An evaluation of cultural value alignment in LLM. *CoRR*, abs/2504.08863.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

Karina Vida, Fabian Damken, and Anne Lauscher. 2024. Decoding multilingual moral preferences: Unveiling llm's biases through the moral machine experiment. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):1490–1501.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Gian Wiher, Clara Meister, and Ryan Cotterell. 2022. On decoding strategies for neural text generators. *Transactions of the Association for Computational Linguistics*, 10:997–1012.

Dustin Wright, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, and Isabelle Augenstein. 2024. LLM tropes: Revealing fine-grained values and opinions in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17085–17112, Miami, Florida, USA. Association for Computational Linguistics.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2024. Self-pluralising culture alignment for large language models. *CoRR*, abs/2410.12971.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian

Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

A Example of Prompt Formulation

Question 1: Direct Prompt

English: For each of the following, indicate how important it is in human life. 1 is very important, 4 is not important at all. Say a number from 1 to 4 and nothing else. Family.

German: Geben Sie für jeden der folgenden Punkte an, wie wichtig er im menschlichen Leben ist. 1 ist sehr wichtig, 4 ist überhaupt nicht wichtig. Nennen Sie die Zahl von 1 bis 4 und sonst nichts. Familie.

Czech: U každé z následujících označte, jak je důležitá v lidském životě. 1 je velmi důležitá, 4 není vůbec důležitá. Napiš číslo od 1 do 4 a nic jiného. Rodina.

Question 1: CoT Prompt

English: For each of the following, indicate how important it is in human life. 1 is very important, 4 is not important at all. First, provide a step-by-step justification, and finally, say the number from 1 to 4 and nothing else. Family.

German: Geben Sie für jeden der folgenden Punkte an, wie wichtig er im menschlichen Leben ist. 1 ist sehr wichtig, 4 ist überhaupt nicht wichtig. Geben Sie zunächst eine schrittweise Begründung an und nennen Sie abschließend die Zahl von 1 bis 4 und sonst nichts. Familie.

Czech: U každé z následujících označte, jak je důležitá v lidském životě. 1 je velmi důležitá, 4 není vůbec důležitá. Nejprve napiš zdůvodnění krok za krokem a nakonec napiš číslo od 1 do 4 a nic jiného. Rodina.

Question 41: Direct Prompt

English: For the following statement, tell me how much agreeable it is. 1 is strongly agree 3 is neutral, 5 is strongly disagree. Say a number from 1 to 5 and nothing else. Work should always come first, even if it means less spare time.

German: Sagen Sie mir, wie sehr kann man der folgenden Aussage zustimmen. 1 bedeutet völlige Zustimmung, 3 bedeutet neutral, 5 bedeutet überhaupt keine Zustimmung. Nennen Sie eine Zahl von 1 und 5 und sonst nichts. Die Arbeit sollte immer an erster Stelle stehen, auch wenn dies weniger Freizeit bedeutet.

Czech: U následující tvrzení uveď, jak moc s ním lze souhlasit. 1 úplný souhlas, 3 je neutrální, 5 naprostý nesouhlas. Napiš číslo od 1 do 5 a nic jiného. Práce by měla být vždy na prvním místě, i když to znamená méně volného času.

Question 41: CoT Prompt

English: For the following statement, tell me how much agreeable it is. 1 is strongly agree 3 is neutral, 5 is strongly disagree. First provide a step-by-step justification and finally say a number from 1 to 5 and nothing else. Work should always come first, even if it means less spare time.

German: Sagen Sie mir, wie sehr kann man der folgenden Aussage zustimmen. 1 bedeutet völlige Zustimmung, 3 bedeutet neutral, 5 bedeutet überhaupt keine Zustimmung. Geben Sie zunächst eine schrittweise Begründung an und nennen Sie abschließend eine Zahl von 1 und 5 und sonst nichts. Die Arbeit sollte immer an erster Stelle stehen, auch wenn dies weniger Freizeit bedeutet.

Czech: U následující tvrzení uveď, jak moc s ním lze souhlasit. 1 úplný souhlas, 3 je neutrální, 5 naprostý nesouhlas. Nejprve napiš zdůvodnění krok za krokem a nakonec napiš číslo od 1 do 5 a nic jiného. Práce by měla být vždy na prvním místě, i když to znamená méně volného času.

B Detailed Multilingual Results

Here, we present the MSD, KL Divergence, Correlation norm, and Self-correlation distance for all combinations of languages (English: en, German: de, Czech: cs) and all countries (United States: USA, United Kingdom: GBR, Czechia: CZE, Germany: DEU, Iran: IRN, China: CHN) in Table 7. Table 2 is a subset of this table. Table 4 with point-biserial correlation of country-language matching and evaluation metrics and Table 3 are computed from numbers in Table 7. Table 5 shows the metric correlation separately from LLaMA 3 and Mistral 2.

Table 7 shows country comparison using the data from WVS with the metrics that we use in the paper. In the paper, we occasionally compare the model alignment to differences between countries. For this, we use data from this table.

	MSD	KLD	CorrD
MSD	—	.276	-.368
KLD	.389	—	-.064
CorrD	-.342	-.005	—

(a) LLaMA 3

	MSD	KLD	CorrD
MSD	—	.832	-.657
KLD	.925	—	-.500
CorrD	-.666	-.526	—

(b) Mistral 2

	MSD	KLD	CorrD
MSD	—	.691	.279
KLD	.724	—	.251
CorrD	.328	.271	—

(c) EuroLLM

	MSD	KLD	CorrD
MSD	—	.617	.120
KLD	.656	—	.495
CorrD	.009	.396	—

(d) Qwen 2.5

Table 5: Breakdown of correlation of the metrics (Pearson above the diagonal, Spearman under the diagonal) over all languages and countries for (a) LLaMA 3, (b) Mistral 2, (c) EuroLLM, and (d) Qwen 2.5

Model	Prompt type	De-co-de	Lng	Mean Sq. Difference						KL Divergence						Corr. norm	Self-correlation distance					
				USA	GBR	CZE	DEU	IRN	CHN	USA	GBR	CZE	DEU	IRN	CHN		USA	GBR	CZE	DEU	IRN	CHN
LLaMA 3	Score only	Gr.	en	.098	.098	.110	.096	.115	.102	1.71	1.66	1.74	1.67	1.81	1.70							
			de	.084	.076	.103	.079	.132	.112	1.65	1.56	1.70	1.62	1.93	1.85							
			cs	.086	.071	.110	.077	.145	.119	1.56	1.41	1.64	1.41	1.83	1.69							
		Spl.	en	.073	.073	.083	.073	.090	.083	2.99	2.90	3.26	2.81	3.43	3.23	0.90	1.26	1.20	1.17	0.95	0.96	0.93
			de	.084	.073	.099	.079	.127	.114	3.58	3.33	3.76	3.30	4.14	3.79	0.67	1.31	1.25	1.24	1.01	0.99	1.01
			cs	.067	.047	.080	.057	.114	.105	2.90	2.49	3.05	2.46	3.52	3.04	0.78	1.26	1.20	1.20	0.96	0.94	0.95
	CoT	Gr.	en	.088	.083	.112	.090	.152	.133	1.68	1.57	1.73	1.63	2.01	1.89							
			de	.085	.072	.104	.080	.134	.117	1.61	1.47	1.69	1.55	1.96	1.82							
			cs	.071	.060	.089	.068	.125	.104	1.57	1.46	1.63	1.58	1.87	1.71							
		Spl.	en	.059	.046	.075	.054	.118	.111	1.47	1.30	1.70	1.29	1.94	1.93	1.70	1.29	1.26	1.22	1.14	1.19	1.03
			de	.054	.042	.063	.050	.105	.094	1.05	0.87	1.12	0.87	1.35	1.34	1.50	1.17	1.14	1.10	0.97	1.02	0.89
			cs	.049	.035	.059	.042	.095	.087	0.98	0.81	1.11	0.84	1.37	1.27	1.60	1.24	1.19	1.14	1.04	1.11	0.93
Mistral2	Score only	Gr.	en	.094	.095	.114	.110	.121	.114	1.80	1.75	1.76	1.90	2.01	1.88							
			de	.096	.107	.105	.110	.100	.105	1.90	1.97	1.82	2.08	2.08	1.94							
			cs	.114	.129	.121	.131	.111	.118	1.91	1.94	1.83	2.06	2.07	1.98							
		Spl.	en	.041	.040	.051	.051	.072	.056	0.77	0.70	0.82	0.82	1.41	0.97	1.77	1.17	1.18	1.10	1.06	1.10	0.95
			de	.035	.039	.040	.051	.072	.068	0.56	0.57	0.51	0.73	0.94	0.79	2.26	1.37	1.35	1.32	1.32	1.37	1.19
			cs	.034	.036	.037	.046	.060	.058	0.41	0.41	0.40	0.49	0.62	0.55	2.64	1.53	1.44	1.51	1.58	1.67	1.46
	CoT	Gr.	en	.188	.181	.187	.187	.156	.152	1.95	1.85	1.98	2.01	1.78	1.94							
			de	.196	.195	.202	.202	.281	.288	1.88	1.87	1.86	1.98	2.03	2.19							
			cs	.163	.156	.147	.169	.196	.189	1.82	1.80	1.72	1.88	1.96	1.88							
		Spl.	en	.022	.025	.030	.035	.050	.048	0.26	0.26	0.29	0.33	0.45	0.43	2.80	1.62	1.56	1.64	1.68	1.78	1.58
			de	.066	.082	.064	.088	.054	.062	0.81	0.77	0.82	0.77	0.62	0.81	3.20	1.88	1.76	1.85	1.96	2.03	1.80
			cs	.039	.045	.036	.056	.052	.050	0.47	0.45	0.46	0.46	0.41	0.49	3.32	1.92	1.82	1.92	2.03	2.09	1.91
EuroLLM	Score only	Gr.	en	.165	.167	.193	.186	.162	.156	1.91	1.80	1.97	1.81	1.72	1.75							
			de	.284	.290	.261	.300	.254	.274	2.40	2.35	2.33	2.40	2.08	2.38							
			cs	.223	.244	.243	.248	.176	.209	2.19	2.21	2.26	2.22	1.84	2.14							
		Spl.	en	.059	.069	.068	.081	.045	.055	0.72	0.65	0.78	0.68	0.55	0.75	2.55	1.56	1.54	1.54	1.57	1.62	1.44
			de	.094	.111	.093	.121	.065	.080	1.00	0.95	1.04	0.93	0.69	0.92	2.76	1.66	1.59	1.70	1.69	1.77	1.57
			cs	.129	.149	.129	.157	.082	.103	1.52	1.50	1.59	1.53	1.17	1.55	2.95	1.71	1.66	1.78	1.79	1.87	1.69
	CoT	Gr.	en	.130	.130	.157	.142	.155	.170	1.76	1.69	1.78	1.76	1.70	1.82							
			de	.369	.402	.322	.393	.207	.279	0.91	0.97	0.80	1.08	0.79	0.94							
			cs	.358	.421	.380	.442	.286	.296	1.15	1.25	1.16	1.28	0.98	1.12							
		Spl.	en	.125	.143	.126	.153	.091	.099	0.97	0.93	1.00	0.94	0.68	0.87	2.97	1.74	1.64	1.78	1.79	1.88	1.68
			de	.149	.171	.151	.179	.106	.121	1.29	1.27	1.38	1.21	0.89	1.12	2.83	1.71	1.59	1.74	1.73	1.85	1.64
			cs	.120	.140	.119	.149	.086	.097	1.12	1.09	1.14	1.07	0.76	0.98	2.84	1.67	1.52	1.67	1.71	1.82	1.57
Qwen	Score only	Gr.	en	.082	.068	.105	.079	.136	.107	1.66	1.60	1.82	1.68	2.02	1.88							
			de	.070	.102	.069	.088	.092	.100	0.82	0.88	0.73	0.92	0.92	0.98							
			cs	.081	.053	.099	.057	.178	.147	0.66	0.65	0.60	0.78	0.83	0.87							
		Spl.	en	.074	.075	.080	.081	.076	.065	1.27	1.17	1.48	1.15	1.35	1.33	3.33	2.13	2.07	2.12	2.15	2.16	2.09
			de	.054	.043	.062	.046	.093	.073	1.13	0.95	1.24	0.95	1.40	1.31	2.12	1.61	1.56	1.51	1.45	1.50	1.36
			cs	.041	.033	.053	.040	.083	.060	0.81	0.69	0.91	0.65	1.00	0.91	1.93	1.39	1.33	1.24	1.24	1.27	1.16
	CoT	Gr.	en	.199	.170	.206	.193	.268	.253	1.60	1.50	1.56	1.54	1.75	1.74							
			de	.141	.133	.149	.154	.179	.177	1.70	1.66	1.73	1.78	1.84	1.87							
			cs	.116	.113	.109	.101	.142	.153	1.22	1.21	1.12	1.25	1.20	1.35							
		Spl.	en	.062	.048	.078	.058	.105	.078	1.13	0.97	1.37	1.01	1.53	1.39	1.52	1.25	1.13	1.15	1.03	1.09	0.94
			de	.055	.050	.073	.061	.103	.093	0.69	0.59	0.75	0.65	0.92	0.84	1.84	1.24	1.16	1.12	1.12	1.21	0.98
			cs	.048	.036	.062	.045	.088	.070	0.73	0.61	0.83	0.63	1.01	0.90	1.72	1.17	1.12	1.10	1.03	1.12	0.92

Table 6: A comparison of model outputs with different prompting strategies (Score-only, CoT: Chain of Thought), decoding method (Gr.: greedy, Spl.: sampling) and different languages

	USA	GBR	CZE	DEU	IRN	CHN
USA	—	.009	.017	.016	.046	.043
GBR	.009	—	.022	.011	.066	.052
CZE	.017	.022	—	.024	.044	.042
DEU	.016	.011	.024	—	.069	.049
IRN	.046	.066	.044	.069	—	.030
CHN	.043	.052	.042	.049	.030	—

	USA	GBR	CZE	DEU	IRN	CHN
USA	—	0.07	0.11	0.16	0.34	0.32
GBR	0.15	—	0.13	0.09	0.39	0.35
CZE	0.17	0.15	—	0.22	0.34	0.29
DEU	0.21	0.08	0.17	—	0.36	0.32
IRN	0.41	0.42	0.33	0.44	—	0.27
CHN	0.32	0.34	0.27	0.33	0.24	—

	Norm.	USA	GBR	CZE	DEU	IRN	CHN
USA	1.66	—	0.79	0.92	0.78	0.95	0.92
GBR	1.54	0.79	—	0.90	0.72	0.92	0.78
CZE	1.55	0.92	0.90	—	0.83	0.93	0.79
DEU	1.11	0.78	0.72	0.83	—	0.66	0.64
IRN	1.06	0.95	0.92	0.93	0.66	—	0.69
CHN	1.13	0.92	0.78	0.79	0.64	0.69	—

Table 7: Comparison of difference between countries in the World Value Survey when measured using mean squared difference , KL Divergence , the norm of the self-correlation tables for each country and distances of the self-correlation tables across countries.