Multilingual Large Language Models and cultural diversity: evidence from civic and moral judgments

Multilingual large language models and cultural diversity: Evidence from civic and moral judgments

Eugenio Vicario[1], Ennio Bilancini[1], and Leonardo Boncinelli[2]
[1]*IMT School for Advanced Studies Lucca*
[2]*University of Florence*

April 23, 2026

Abstract

Multilingual large language models (LLMs) are increasingly deployed across linguistic and cultural contexts, raising the question of whether multilingual interaction preserves cultural diversity in moral judgments. We compare civic and moral evaluations generated by a multilingual LLM across multiple languages with population-level data from the World Values Survey and the European Values Study. Although the model exhibits meaningful linguistic variability, this does not translate into the preservation of cross-national moral diversity. Alignment with population-level values is highest for WEIRD countries and weaker elsewhere. At the cross-national level, the LLM reshapes the structure of moral distances: differences between WEIRD and non-WEIRD countries are selectively compressed in the model-generated space, while distances within these groups remain largely unchanged. These dynamics are strongly domain-dependent. Anti-civic norms display a pronounced norm-enforcing bias with minimal cross-national variation; personal and bioethical judgments cluster around values typical of WEIRD countries; public and social order norms exhibit systematic directional shifts, particularly outside WEIRD cultural zones; and attitudes toward political violence display increased dispersion rather than convergence. Together, these findings suggest that multilingual LLMs act as normative infrastructures that reshape moral representations in uneven and domain-specific ways, highlighting the limits of multilingual fluency as a guarantee of cultural alignment.

1 Introduction

Artificial intelligence is no longer merely a technological support tool but has become a pervasive infrastructure shaping information access, professional decision-making, and daily social dynamics (Fraiwan and Khasawneh, 2023). Consequently, complex cognitive tasks and moral evaluations are increasingly delegated to AI systems, particularly Large Language Models (LLMs). Trained on massive text corpora to predict linguistic tokens (Vaswani et al., 2017), these models do not simply learn syntactic rules; through exposure

to vast amounts of data, they internalize semantic patterns that emerge as forms of *common knowledge* (Petroni et al., 2019; Radford et al., 2019) and implicit ethical stances.

Given the growing delegation of agency to these systems, it is urgent for social scientists to understand the nature of this *emergent knowledge* (Capraro et al., 2024; Ziems et al., 2024; Messeri and Crockett, 2024; Gusella and Vicario, 2025). A rapidly expanding body of literature has already highlighted how LLMs tend to reflect the biases and values of Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies, which are overrepresented in training data (Atari et al., 2023; Rao et al., 2023; Fraser et al., 2022). Notably, however, this evidence is derived almost exclusively from studies that query models in English. As a result, existing findings primarily document how LLMs behave within an Anglophone interactional setting. Whether these patterns generalize beyond English remains largely unexplored. When the same model is prompted in different languages, it may either reproduce the same WEIRD-centered normative alignment, adapt to the moral profiles associated with the populations speaking those languages, or generate responses that diverge from both. Clarifying this issue is crucial to determine whether multilingual capability enables genuine cultural responsiveness or instead reinforces a homogeneous moral core across linguistic contexts.

Whether cross-linguistic consistency should be interpreted as robustness or distortion depends on the normative structure of the domain in which the model operates. In domains with externally validated correct answers, cross-lingual consistency is a desirable property, and deviations across languages are typically interpreted as model limitations rather than meaningful signals (Choudhury and Deshpande, 2021; Ahuja et al., 2023; Rathje et al., 2024; Huang et al., 2023).

By contrast, in domains where moral judgments are socially and culturally grounded rather than externally verifiable, the intuitive hypothesis is that an LLM, when operating in a specific language, should be able to recall the cultural, ethical, and moral traits crystallized in the texts produced by the population speaking that language (AlKhamissi et al., 2024). If this were true, the multilingual capability of AI would guarantee automatic cultural alignment. However, numerous factors, from data selection and implicit translation within the corpus to disparities in digitization across cultures, may distort or prevent this process. Preliminary studies suggest that while LLM responses vary across languages (Khandelwal et al., 2024; Agarwal et al., 2024; Li et al., 2025), alignment with human moral judgments remains weak (Durmus et al., 2023; Jin et al., 2024; Arora et al., 2023) or inconsistent outside of Anglophone or well-represented contexts (Aksoy, 2025; Hämmerl et al., 2023; Johnson et al., 2022). In this work, we investigate whether the multilingual capacity of LLMs translates into genuine cultural pluralism or, conversely, acts as a vector for moral homogenization. The societal impact of large language models depends on several characteristics of human–AI interaction (Bilancini et al., 2024), including how users engage with the system and how its outputs relate to existing social norms and values. One particularly relevant dimension concerns the degree of alignment between the moral judgments expressed by the model and those held by the populations interacting with it. The implications are profound: if an LLM's moral position in a given language diverges from that of the native population, interaction with the AI could influence public opinion (Matz et al., 2024; Hackenburg and Margetts, 2024; Landes et al., 2026), polarize debate, or alter the acceptability of the technology itself (Böhm et al., 2023; Zhang and Gosline, 2023). Models belonging to the GPT-4 generation marked a major stage in the diffusion of large language models across social and institutional contexts and were among the most widely deployed systems during the initial expansion of

generative AI. In this study, we analyze responses generated by *gpt-4o-mini*, a model belonging to this generation. Focusing on this class of models allows us to examine patterns that have already unfolded at scale in real-world human–AI interaction. Our analytical framework is designed to be replicable across model generations, enabling future work to assess whether the patterns observed here persist or evolve with newer model generations.

Here we show that multilingual LLMs do not preserve global cultural diversity but instead exhibit a systematic tendency toward moral convergence. Analyzing GPT-4o-mini responses across multiple languages and comparing them with World Values Survey data (Inglehart and Welzel, 2010), we find that model outputs systematically reshape the cross-national structure of moral distances. In the LLM-generated representations, countries belonging to WEIRD cultural zones tend to move closer to a large share of the rest of the world, while distances within these cultural groups remain largely unchanged or slightly increase. As a result, cross-national moral differences are selectively compressed, indicating a centripetal dynamic driven primarily by convergence between WEIRD and non-WEIRD countries rather than by language-specific cultural alignment.

This convergence, however, is domain-dependent rather than uniform. In domains related to civic norms, model responses collapse toward near-universal condemnation, producing minimal cross-national variation and a pronounced norm-enforcing bias. In personal and bioethical issues, outputs are selectively compressed toward the range characteristic of WEIRD societies, generating larger deviations for countries outside those zones. In matters of public and social order, the model exhibits systematic directional shifts for several non-WEIRD cultural regions, while remaining comparatively heterogeneous within Western contexts. By contrast, attitudes toward political violence display increased dispersion rather than convergence, with the model amplifying cross-country variability. Together, these patterns indicate that multilingual LLMs do not simply reproduce a single moral template, but restructure cultural variation in asymmetric and domain-specific ways.

The emergent moral profile of these systems thus appears neither culturally neutral nor fully language-contingent, but structured around a partially stable normative center whose determinants remain to be fully understood.

2 Methods

We compare moral judgments expressed by a multilingual large language model with population-level moral judgments derived from the World Values Survey (WVS) and the European Values Study (EVS), two large-scale cross-national survey programs on human values. We focus on a subset of items from the *Ethical values and norms* section of both surveys (WVS Q177–Q195; EVS v149–v163), restricting the analysis to items that are identically worded across the two questionnaires. For each item, respondents are asked: *Please tell me for each of the following actions whether you think it can always be justified, never be justified, or something in between.* Responses are recorded on a 10-point Likert scale, where 1 corresponds to *never justifiable* and 10 corresponds to *always justifiable*.

LLM responses are not treated as proxies for individual or collective human judgments, but as model outputs to be compared with population-level distributions. Accordingly, comparisons with WVS/EVS data are used to assess alignment or misalignment between model-generated normative positions, when queried in a given language, and the moral judgments of the corresponding human populations.

We query OpenAI's *gpt-4o-mini* model via API using custom Python scripts. For each country covered by WVS/EVS, survey items were presented verbatim in the languages used for data collection in that country. Each prompt was preceded by a short instruction asking the model to respond exclusively to the question. In cases where the response scale was conveyed visually in the original survey, numeric anchors were explicitly added to reproduce the original 1–10 scale. All instructions and scale clarifications were translated into the corresponding languages using ChatGPT 5 (browser version), without altering the original item wording. For each item–language combination, we conducted 10 independent model queries with temperature fixed at 1. Responses were averaged across repetitions to obtain a single model score for each item and language. To construct country-level synthetic indicators comparable to survey-based measures, language-specific LLM scores were aggregated using weights derived from WVS/EVS data, corresponding to the proportion of questionnaires administered in each language within a country. This weighting scheme enables the construction of a single country-level estimate while preserving within-country linguistic heterogeneity and approximating the linguistic exposure of users interacting with the model in that national context.

Overall, this procedure yields, for each language, a vector of mean model scores across all analyzed items. These language-specific vectors are subsequently combined using the WVS/EVS language weights to construct a single country-level model vector representing the model's aggregated moral judgments for that country. This country-level model vector is then directly compared with the corresponding country-level vector derived from WVS/EVS respondent data.

Additional methodological details are provided in the Appendix.

3 Results

We structure the empirical analysis in four steps. First, we assess whether prompting the model in different languages produces differentiated moral judgments, thereby examining the extent of linguistic variability in model outputs (Section 3.1). Second, we evaluate the degree of alignment between LLM-generated judgments and population-level moral attitudes derived from the World Values Survey and the European Values Study, comparing model outputs and survey data at the country level (Section 3.2). Third, we analyze how the cross-national structure of moral distances changes when judgments are generated by the model rather than observed in survey data (Section 3.3). Finally, we examine whether these discrepancies display systematic patterns across different domains of moral evaluation (Section 3.4).

3.1 Linguistic variability in multilingual LLM moral judgments

We first assess whether querying the LLM in different languages produces differentiated moral judgments. If multilingual capability merely reflected surface-level translation, responses would collapse across languages, precluding any form of linguistic or cultural differentiation.

To test this, we compute pairwise distances between the language-specific vectors of model-generated mean responses across all moral items. The distance is defined as a root mean squared distance computed over the set of items observed for both languages, a choice that allows us to handle missing values, see the Appendix for details. Figure 1

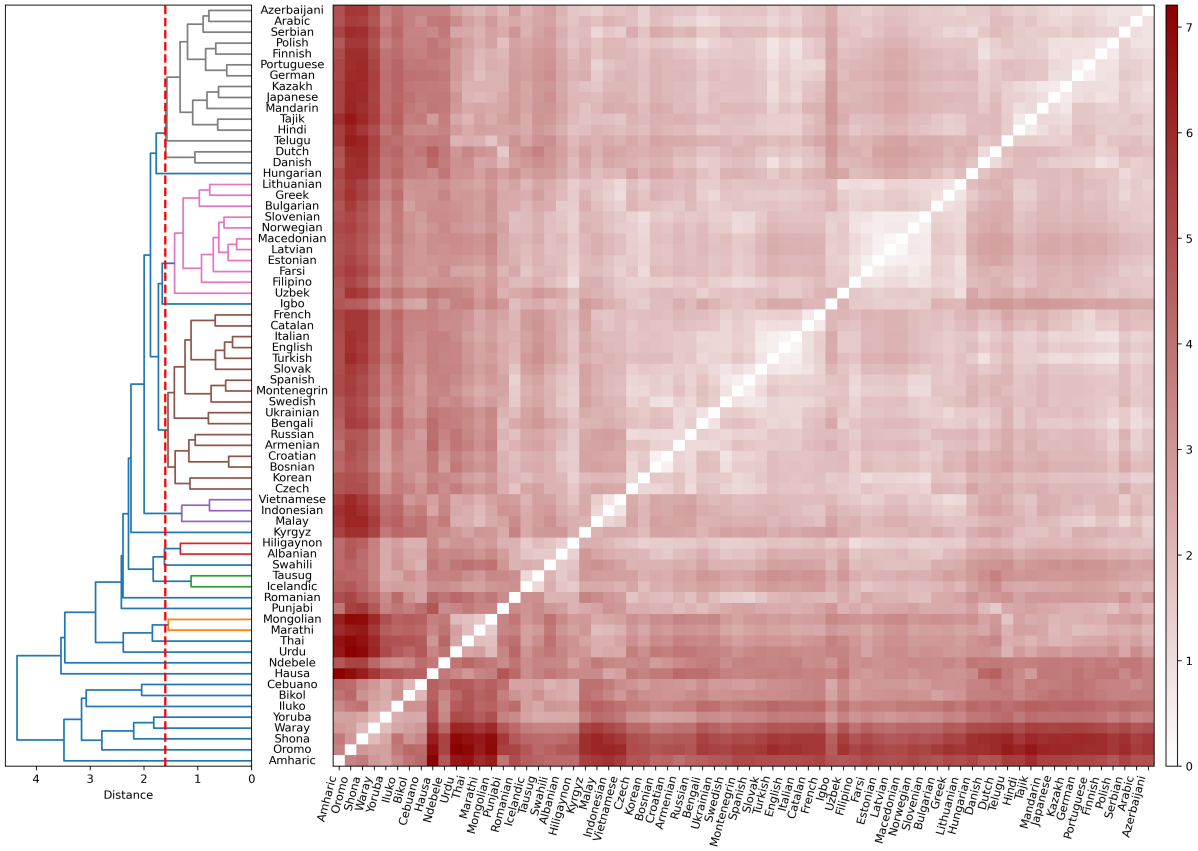

Figure 1: Language distance matrix ordered by the hierarchical clustering.

shows the resulting linguistic distance matrix, with an integrated hierarchical dendrogram indicating clusters of languages. The matrix reveals substantial variation in model responses across languages. However, the resulting clusters do not clearly align with known linguistic families, suggesting that the observed variability does not simply reflect genealogical proximity between languages. Additional diagnostics, including correlation matrices and the distribution of average inter-language distances, are reported in the Appendix.

Together, these results establish that multilingual LLMs exhibit non-trivial linguistic variability, providing a necessary, though not sufficient, condition for cultural differentiation in model-generated moral judgments.

3.2 Alignment with population-level moral judgments

We next examine the degree of alignment between LLM-generated moral judgments and population-level judgments derived from the World Values Survey and the European Values Study. As detailed in the Methods section, this comparison does not assess representational accuracy, but rather empirical correspondence between the normative positions expressed by the model and the moral distributions observed among human populations.

Figure 2 maps, for each country, the correlation between LLM responses and WVS/EVS data (left panel), alongside the mean absolute deviation between the two (right panel). Substantial heterogeneity is observed across countries. Alignment is systematically higher for English-speaking and Western European countries, while both correlations decrease and deviations increase for countries belonging to other cultural zones.

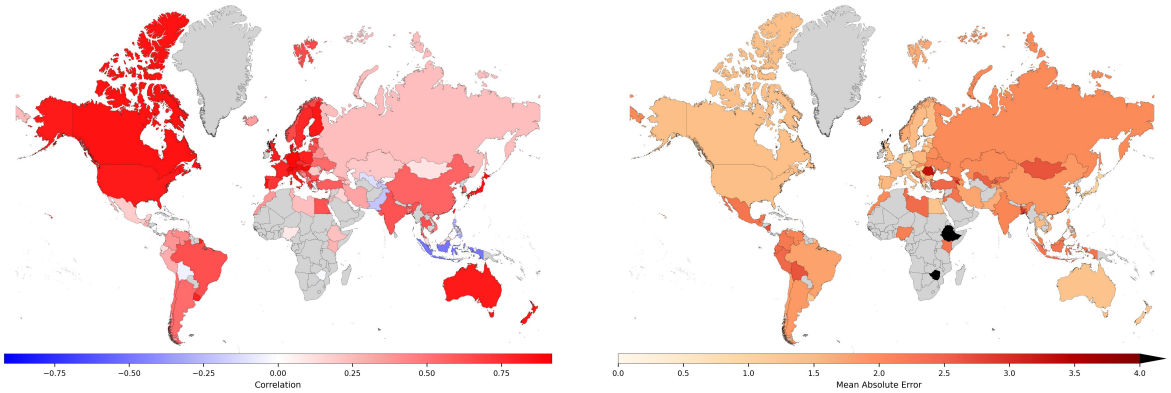

Figure 2: Cross-country alignment between LLM-generated moral judgments and population-level judgments from the World Values Survey (WVS) and the European Values Study (EVS). The left panel shows, for each country, the correlation between the vector of LLM responses obtained in the language(s) spoken in that country and the corresponding vector of WVS/EVS responses across moral items. The right panel reports the mean absolute distance between the two vectors. Look at the interactive map with the data: **LINK**

These patterns indicate that correspondence between model-generated judgments and population-level moral distributions varies substantially across countries. Higher levels of alignment are observed primarily in high-income Western countries, while correspondence tends to be weaker in countries outside these socio-economic contexts.

3.3 Asymmetric reconfiguration of moral distances across countries

We examine how cross-national moral distances change when moral judgments are generated by a multilingual LLM rather than derived from population-level survey data. Figure 3 presents the matrix of differences between country-level moral distances computed from WVS/EVS data and those computed from LLM-generated responses, with countries ordered according to an integrated hierarchical clustering. Positive values indicate pairs of countries that appear closer to one another in the LLM-generated moral representations than in the survey-based space.

The figure reveals a selective reconfiguration of moral distances, in which reductions and increases coexist across different pairs of countries. LLM-generated moral representations exhibit a centripetal tendency that is unevenly distributed across countries. A group of countries, primarily belonging to WEIRD cultural zones, reduces its moral distance with a large fraction of the rest of the world in the LLM-generated space. At the same time, a smaller subset of countries shows the opposite pattern, increasing its distance from most other countries. These dynamics indicate that convergence is neither global nor uniform, but driven by asymmetric and group-specific repositioning in the moral space.

Importantly, reductions in moral distance are concentrated between countries belonging to different cultural groups, particularly between WEIRD and non-WEIRD countries. By contrast, distances within the two groups tend to increase slightly in the LLM-generated representations, with the exception of countries that share the same language.

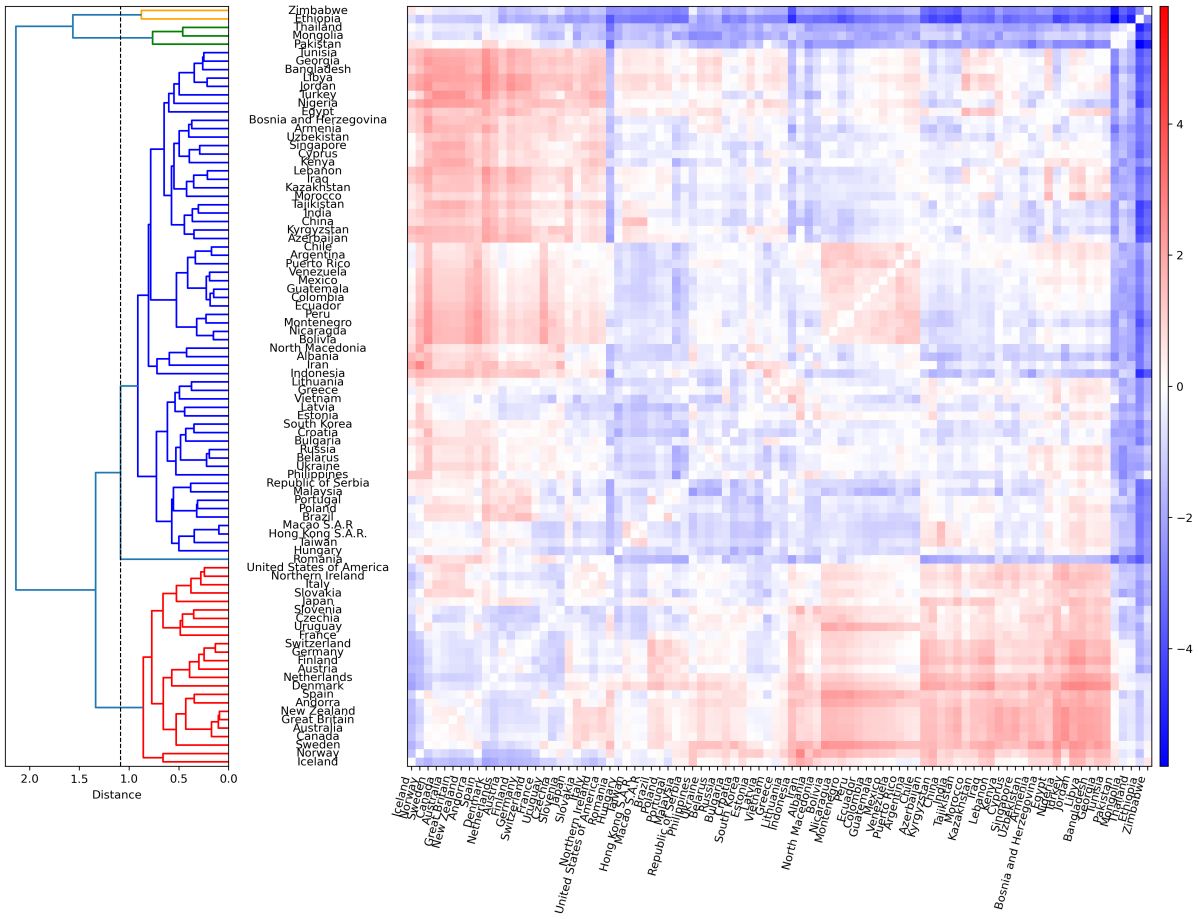

Figure 3: Each element represents the difference between survey-based and LLM-based distances between countries, computed as root mean squared differences across moral items. Countries are ordered according to hierarchical clustering based on the similarity of their discrepancy profiles between the two distance structures.

As a result, convergence operates primarily across groups rather than within them.

For completeness, the survey-based and LLM-based distance matrices used to construct the difference matrix shown in Fig. 3 are reported separately in Appendix together with technical details.

3.4 Domain-specific patterns of moral convergence

Using hierarchical clustering based on the cross-country structure of REAL–LLM differences, survey items are grouped into four conceptually coherent domains: Anti-Civic Norms, Personal and Bioethical Morality, Public and Social Order Morality, and Political Violence (see Appendix for details).

Figure 4 presents, for each domain, scatterplots comparing country-level positions in the survey-based moral space with those generated by the LLM. Countries are grouped into two broad clusters, WEIRD and Rest of the World, consistent with the hierarchical clustering structure identified in the previous section. Each point represents a country, allowing direct assessment of both the magnitude and direction of deviations across domains and cultural groups.

In the Anti-Civic Norms domain, which includes behaviors such as claiming govern-

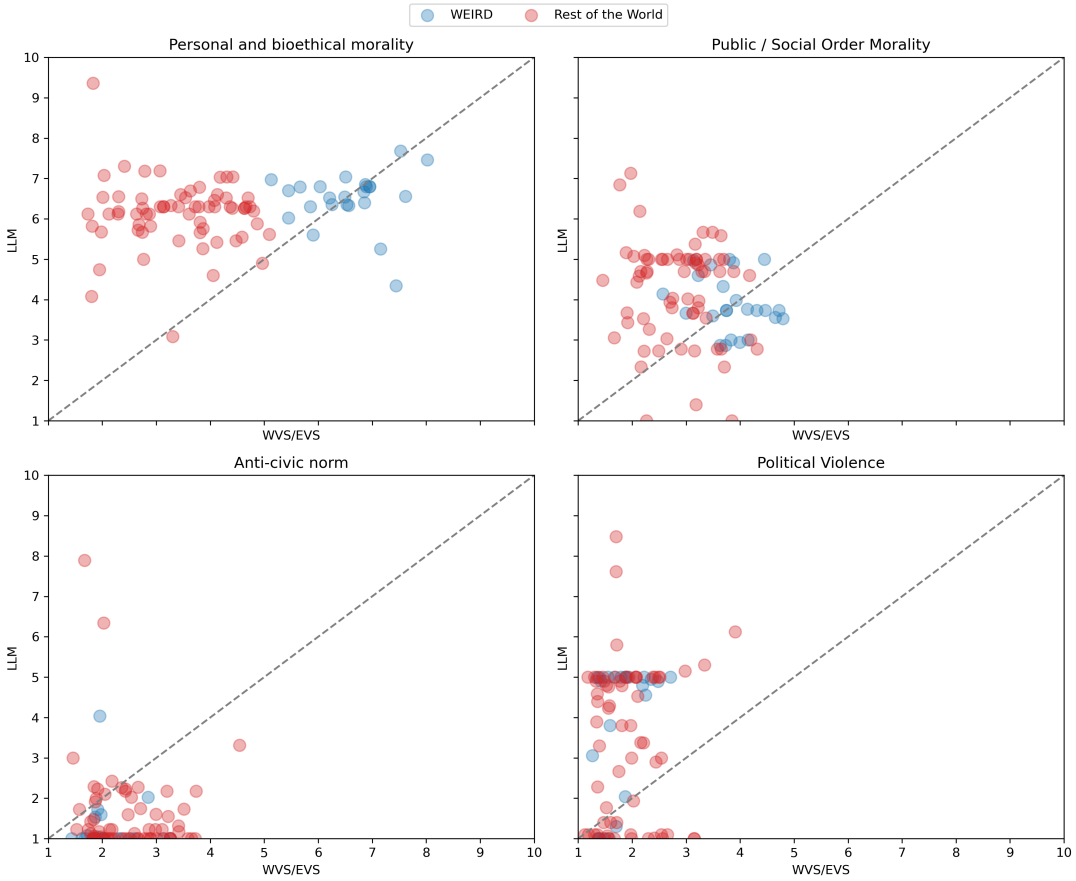

Figure 4: Country-level comparison between population moral judgments (WVS/EVS) and LLM-generated responses across four moral domains identified through hierarchical clustering of survey items. Each point represents a country. The horizontal axis reports the mean survey-based judgment within the domain, while the vertical axis reports the corresponding mean LLM judgment. Colors distinguish countries belonging to WEIRD cultural zones from the rest of the world. The dashed line indicates perfect alignment between model-generated and population-level judgments. Use the interactive map to visualize the results: **LINK**. Use the interactive version of the scatterplots to visualize single country position: **LINK**

ment benefits to which one is not entitled, avoiding fares on public transport, cheating on taxes, and accepting bribes in the course of official duties, the LLM assigns substantially lower permissiveness scores than those observed in human populations. Responses cluster near the lower bound of the scale, indicating near-universal condemnation of these behaviors by the model. Consequently, deviations from population-level judgments are large and negative for most countries, with minimal cross-national variation. This pattern reflects a pronounced norm-enforcing bias rather than convergence driven by relative repositioning between countries.

The Personal and Bioethical Morality cluster, encompassing issues such as homosexuality, abortion, divorce, euthanasia, and casual sex, exhibits a markedly different pattern. For countries belonging to WEIRD cultural zones, deviations between LLM-generated and survey-based judgments are relatively limited and display mixed signs. In contrast, for countries outside these zones, deviations are substantially larger and predominantly negative, indicating that the LLM adopts more restrictive positions than those

expressed by local populations. Notably, variability in LLM-generated judgments within this domain is markedly reduced relative to the variability observed in the survey data, with model-generated values clustering around the range characteristic of WEIRD countries. This pattern suggests a selective form of convergence, in which moral positions expressed by the model are compressed toward WEIRD-level values, while deviations from population-level judgments emerge primarily outside these cultural zones.

In the Public and Social Order Morality domain, which includes behaviors such as the death penalty, suicide, and prostitution, deviations between LLM-generated and survey-based judgments are systematically negative for countries not belonging to the WEIRD zone. In these contexts, the LLM judges these actions more acceptable than the populations of those countries do. For WEIRD countries, deviations are instead more heterogeneous in sign, with both positive and negative differences observed. This pattern indicates that, in this domain, the model does not exhibit a systematic shift relative to population-level judgments for WEIRD countries, whereas a consistent directional deviation is observed for non-WEIRD cultural zones.

Finally, the Political Violence domain displays a qualitatively distinct behavior. Here, LLM-generated judgments are generally more permissive than those observed in survey data, with responses clustering around mid-scale and extreme values. Unlike the previous domains, cross-country variability increases rather than decreases, indicating that the model amplifies dispersion instead of producing convergence. This expansion of variability highlights that moral homogenization is not uniform across domains and that politically charged behaviors remain comparatively unstable in the model's normative structure.

4 Discussion

In this study, we investigated whether the multilingual capabilities of large language models translate into the preservation of cultural diversity in moral judgments. By comparing moral evaluations generated by a multilingual LLM with population-level data from the World Values Survey and the European Values Study, we show that multilingual interaction alone does not ensure cultural pluralism in model outputs.

Although the model exhibits clear linguistic variability, this does not translate into the preservation of cross-national moral diversity. Instead, model-generated moral representations reorganize the structure of relationships between countries in a selective and asymmetric way, with convergence operating primarily across cultural groups rather than within them. Alignment with population-level moral judgments is also uneven, being highest for WEIRD countries and substantially weaker elsewhere, suggesting that multilingual variation alone is insufficient to guarantee cultural alignment.

Crucially, the cross-national convergence documented in Section 3.3 is largely driven by domain-specific dynamics. In particular, the compression of moral distances between WEIRD and non-WEIRD countries is primarily attributable to items in the *Personal and Bioethical Morality* domain, including homosexuality, abortion, divorce, euthanasia, and casual sex. In this domain, the LLM strongly reduces the variability observed in population-level data, effectively collapsing responses toward the range characteristic of WEIRD countries across all languages. This effect is further reinforced by the fact that WEIRD countries exhibit higher levels of acceptance for these behaviors in the survey data, so that the compression of LLM-generated responses toward this range produces systematically larger deviations for countries outside these cultural zones.

By contrast, in the *Anti-Civic Norms* domain, the model exhibits a widespread norm-enforcing bias, with responses concentrated near the lowest end of the scale across almost all countries. However, because low permissiveness is also observed in several non-WEIRD countries in the survey data, the resulting effect is not systematically structured along the WEIRD versus non-WEIRD divide. Rather than inducing cross-group convergence, this domain produces a general compression toward universally condemned behaviors.

A different pattern emerges in the *Public and Social Order Morality* domain, which includes controversial issues such as suicide, prostitution, and the death penalty. As shown in the clustering structure (Appendix), these items are characterized by the largest within-domain heterogeneity, reflecting substantial variation in legal frameworks and cultural attitudes both within WEIRD countries and across the rest of the world. Consistently, model-generated judgments in this domain do not display a strong or systematic distortion along cultural lines. Instead, values tend to overlap across groups in both survey and model data, with WEIRD countries exhibiting only slightly higher average levels of acceptance.

Finally, the *Political Violence* domain displays a qualitatively distinct behavior. In the survey data, acceptance of political violence is uniformly low across countries, with limited differentiation between cultural groups. In contrast, LLM-generated responses exhibit substantially increased dispersion, with many countries shifting toward intermediate levels of acceptability. Rather than compressing variability, the model amplifies cross-national differences in this domain, suggesting that responses are highly sensitive to how political violence is interpreted, including variation in its perceived intensity and context.

Taken together, these findings suggest that multilingual LLMs function less as neutral mirrors of cultural values and more as normative infrastructures that actively reshape moral representations. When deployed across linguistic and cultural contexts, their outputs may introduce systematic normative pressures that differ from locally grounded moral judgments. This has important implications for the societal integration of AI systems, particularly in contexts where LLMs are consulted for guidance on ethical, civic, or socially sensitive issues.

More broadly, our results highlight a tension between the technical achievement of multilingual fluency and the social expectation of cultural alignment. Expanding language coverage alone is insufficient to ensure that AI systems reflect the moral diversity of the populations that interact with them. Addressing this tension will require moving beyond language as a proxy for inclusivity and toward a more explicit engagement with how normative content is learned, represented, and deployed across cultural contexts. Future research should investigate how training data composition, model architectures, and alignment strategies jointly shape the moral landscapes produced by multilingual AI systems, and how these landscapes interact with public discourse and social norms in diverse societies.

References

Agarwal, U., K. Tanmay, A. Khandelwal, and M. Choudhury (2024). Ethical reasoning and moral value alignment of llms depend on the language we prompt them in. *arXiv preprint arXiv:2404.18460*.

Ahuja, K., H. Diddee, R. Hada, M. Ochieng, K. Ramesh, P. Jain, A. Nambi, T. Ganu, S. Segal, M. Ahmed, et al. (2023). Mega: Multilingual evaluation of generative ai. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4232–4267.

Aksoy, M. (2025). Whose morality do they speak? unraveling cultural bias in multilingual language models. *Natural Language Processing Journal*, 100172.

AlKhamissi, B., M. ElNokrashy, M. AlKhamissi, and M. Diab (2024). Investigating cultural alignment of large language models. *arXiv preprint arXiv:2402.13231*.

Arora, A., L.-a. Kaffee, and I. Augenstein (2023). Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the first workshop on cross-cultural considerations in NLP (C3NLP)*, pp. 114–130.

Atari, M., J. Haidt, J. Graham, S. Koleva, S. T. Stevens, and M. Dehghani (2023). Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology 125*(5), 1157.

Bilancini, E., L. Boncinelli, and E. Vicario (2024). Ai-powered chatbots: Effective communication styles for sustainable development goals. *arXiv preprint arXiv:2407.01057*.

Böhm, R., M. Jörling, L. Reiter, and C. Fuchs (2023). People devalue generative ai's competence but not its advice in addressing societal and personal challenges. *Communications Psychology 1*(1), 32.

Capraro, V., A. Lentsch, D. Acemoglu, S. Akgun, A. Akhmedova, E. Bilancini, J.-F. Bonnefon, P. Brañas-Garza, L. Butera, K. M. Douglas, et al. (2024). The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS nexus 3*(6), pgae191.

Choudhury, M. and A. Deshpande (2021). How linguistically fair are multilingual pre-trained language models? In *Proceedings of the AAAI conference on artificial intelligence*, Volume 35, pp. 12710–12718.

Durmus, E., K. Nguyen, T. I. Liao, N. Schiefer, A. Askell, A. Bakhtin, C. Chen, Z. Hatfield-Dodds, D. Hernandez, N. Joseph, et al. (2023). Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Fraiwan, M. and N. Khasawneh (2023). A review of chatgpt applications in education, marketing, software engineering, and healthcare: Benefits, drawbacks, and research directions. *arXiv preprint arXiv:2305.00237*.

Fraser, K. C., S. Kiritchenko, and E. Balkir (2022). Does moral code have a moral code? probing delphi's moral philosophy. *arXiv preprint arXiv:2205.12771*.

Gusella, F. and E. Vicario (2025). Generative agents and expectations: Do llms align with heterogeneous agent models? *arXiv preprint arXiv:2511.08604*.

Hackenburg, K. and H. Margetts (2024). Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences 121*(24), e2403116121.

Hämmerl, K., B. Deiseroth, P. Schramowski, J. Libovickỳ, C. Rothkopf, A. Fraser, and K. Kersting (2023). Speaking multiple languages affects the moral bias of language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 2137–2156.

Huang, H., T. Tang, D. Zhang, W. X. Zhao, T. Song, Y. Xia, and F. Wei (2023). Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*.

Inglehart, R. and C. Welzel (2010). The wvs cultural map of the world. *World Values Survey 22*.

Jin, Z., M. Kleiman-Weiner, G. Piatti, S. Levine, J. Liu, F. Gonzalez, F. Ortu, A. Strausz, M. Sachan, R. Mihalcea, et al. (2024). Language model alignment in multilingual trolley problems. *arXiv preprint arXiv:2407.02273*.

Johnson, R. L., G. Pistilli, N. Menédez-González, L. D. D. Duran, E. Panai, J. Kalpokiene, and D. J. Bertulfo (2022). The ghost in the machine has an american accent: value conflict in gpt-3. *arXiv preprint arXiv:2203.07785*.

Khandelwal, A., U. Agarwal, K. Tanmay, and M. Choudhury (2024). Do moral judgment and reasoning capability of llms change with language? a study using the multilingual defining issues test. *arXiv preprint arXiv:2402.02135*.

Landes, E., K. Francis, and J. A. Everett (2026). People defer to ai moral advice, but not blindly. *Cognition*.

Li, N., B. Kang, and T. De Bie (2025). Human-ai moral judgment congruence on real-world scenarios: A cross-lingual analysis. In *Proceedings of the 9th Widening NLP Workshop*, pp. 46–49.

Matz, S. C., J. D. Teeny, S. S. Vaid, H. Peters, G. M. Harari, and M. Cerf (2024). The potential of generative ai for personalized persuasion at scale. *Scientific Reports 14*(1), 4692.

Messeri, L. and M. J. Crockett (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature 627*(8002), 49–58.

Petroni, F., T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller (2019). Language models as knowledge bases? In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2463–2473.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog 1*(8), 9.

Rao, A., A. Khandelwal, K. Tanmay, U. Agarwal, and M. Choudhury (2023). Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in llms,", arxiv. *arXiv preprint arXiv:2310.07251*.

Rathje, S., D.-M. Mirea, I. Sucholutsky, R. Marjieh, C. E. Robertson, and J. J. Van Bavel (2024). Gpt is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences 121*(34), e2308950121.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). Attention is all you need. *Advances in neural information processing systems 30*.

Zhang, Y. and R. Gosline (2023). Human favoritism, not ai aversion: People's perceptions (and bias) toward generative ai, human experts, and human–gai collaboration in persuasive content generation. *Judgment and Decision Making 18*, e41.

Ziems, C., W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang (2024). Can large language models transform computational social science? *Computational Linguistics 50*(1), 237–291.

A Data

A.1 Data sources and survey items

Human moral judgments are derived from the World Values Survey (WVS) and the European Values Study (EVS), two established cross-national survey programs designed to measure values, beliefs, and norms across societies. We use the most recent harmonized releases available at the time of analysis.

Our analysis focuses on the *Ethical values and norms* section of both surveys. In the WVS, this corresponds to items Q177–Q195, while in the EVS the corresponding items span v149–v163. Because item coverage and wording differ slightly between the two surveys, we restrict the analysis to items that are present in both questionnaires. These items cover a range of civic and moral behaviors, including compliance with public norms, interpersonal honesty, and political or social transgressions. The complete list of retained items, along with their exact wording in each survey, is reported in Supplementary Table 1.

Responses in both surveys are recorded on a 10-point ordinal scale, where 1 indicates that an action is *never justifiable* and 10 indicates that it is "*always justifiable*.

WVS	**EVS**	**Action**
Q177	v149	Claiming government benefits to which you are not entitled
Q178	v159	Avoiding a fare on public transport
Q180	v150	Cheating on taxes if you have a chance
Q181	v152	Someone accepting a bribe in the course of their duties
Q182	v153	Homosexuality
Q183	v160	Prostitution
Q184	v154	Abortion
Q185	v155	Divorce
Q187	v157	Suicide
Q188	v156	Euthanasia
Q193	v158	Having casual sex
Q194	v162	Political violence
Q195	v163	Death penalty

Table 1: WVS–EVS Items and Corresponding Actions

A.2 Language Coverage and Aggregation Methodology

For each country included in the WVS/EVS, we retrieved the official survey questionnaires in all languages used for data collection in that country. The set of languages therefore varies across countries and reflects the linguistic composition of the survey samples rather than the full linguistic diversity of the population.

To construct country-level indicators from multilingual model outputs, we aggregated responses across the languages used in each country. Aggregation was performed using weighted averages, where weights correspond to the proportion of WVS/EVS questionnaires administered in each language relative to the total number of questionnaires in that country.

This weighting scheme approximates the linguistic exposure of users interacting with the model in a given national context while preserving within-country linguistic heterogeneity.

Summary statistics on the number of languages per country are provided in Supplementary Table 2.

Table 2: Language distribution by country (percentages)

Country	Language	%	Country	Language	%
Albania	Albanian	100	Andorra	Catalan	53.38645418
Andorra	Spanish	45.61752988	Andorra	French	0.896414343
Andorra	English	0.099601594	Argentina	Spanish	100
Armenia	Armenian	100	Australia	English	100
Austria	German	100	Azerbaijan	Azerbaijani	96.5
Azerbaijan	Russian	3.5	Bangladesh	Bangla	100
Belarus	Russian	100	Bolivia, Plurinational State of	Spanish	100
Bosnia and Herzegovina	Bosnian	64.79118329	Bosnia and Herzegovina	Serbian	24.47795824
Bosnia and Herzegovina	Croatian	10.73085847	Brazil	Portuguese	100
Bulgaria	Bulgarian	100	Canada	English	76.92882031
Canada	French	23.07117969	Chile	Spanish	100
China	Mandarino	100	Colombia	Spanish	100
Croatia	Croatian	100	Cyprus	Greek	50
Cyprus	Turkish	50	Czechia	Czech	99.83394221
Czechia	Russian	0.13284623	Czechia	Slovak	0.033211558
Denmark	Danish	100	Ecuador	Spanish	100
Egypt	Arabic	100	Estonia	Estonian	75.69018405
Estonia	Russian	24.30981595	Ethiopia	Amaric	84.93273543
Ethiopia	Oromo	14.17040359	Ethiopia	English	0.896860987
Finland	Finish	100	France	French	100
Georgia	Georgian	92.88969918	Georgia	Armenian	3.87420237
Georgia	Azerbaijani	3.144940747	Georgia	Russian	0.091157703
Germany	German	100	Great Britain	English	100
Greece	Greek	100	Guatemala	Spanish	100
Hong Kong	Mandarino	99.51807229	Hong Kong	English	0.481927711
Hungary	Hungarian	100	Iceland	Icelandic	97.16748768
Iceland	English	2.832512315	India	Hindi	100
Indonesia	Indonesian	100	Iran, Islamic Republic of	Farsi	100
Iraq	Arabic	100	Italy	Italian	100
Japan	Japanese	100	Jordan	Arabic	100
Kazakhstan	Kazakh	53.13479624	Kazakhstan	Russian	46.86520376
Kenya	English	86.17693523	Kenya	Swahili	13.82306477
Korea, Republic of	Korean	100	Kyrgyzstan	Kyrkyz	70.91666667
Kyrgyzstan	Russian	17	Kyrgyzstan	Uzbek	10.75
Kyrgyzstan	Other	1.333333333	Latvia	Latvian	87.86516854
Latvia	Russian	12.13483146	Lebanon	Arabic	100
Libya	Arabic	100	Lithuania	Lithuanian	89.77900552
Lithuania	Russian	10.22099448	Macao	Mandarino	100
Malaysia	Malay	100	Maldives	Dhivehi	100
Mexico	Spanish	100	Mongolia	Mongolian	100
Montenegro	Montenegrin	100	Morocco	Arabic	100
Myanmar	Burmese	100	Netherlands	Dutch	100
New Zealand	English	100	Nicaragua	Spanish	100
Nigeria	English	54.56750202	Nigeria	Hausa	28.45594179
Nigeria	Yoruba	13.74292643	Nigeria	Igbo	3.233629749
North Macedonia	Macedonian	81.37869293	North Macedonia	Albanian	18.62130707
North Ireland	English	100	Norway	Norwegian	97.59358289
Norway	English	2.139037433	Norway	Polish	0.267379679
Pakistan	Urdu	100	Peru	Spanish	100
Philippines	Filipino	45.08333333	Philippines	Cebuano	31.83333333
Philippines	Hiligaynon	8.833333333	Philippines	Waray	4.166666667
Philippines	Iluko	4.083333333	Philippines	Tausung	3.5
Philippines	Bikol	2.5	Poland	Polish	100
Portugal	Portuguese	100	Puerto Rico	Spanish	100
Romania	Romanian	98.7456446	Romania	Hungarian	1.254355401
Russian Federation	Russian	100	Serbia	Serbian	100
Singapore	English	80.71570577	Singapore	Mandarino	19.13518887
Singapore	Malay	0.149105368	Slovakia	Slovak	98.74620061
Slovakia	Hungarian	1.253799392	Slovenia	Slovenian	100
Spain	Spanish	95.02074689	Spain	Catalan	4.398340249
Spain	Galician	0.497925311	Spain	Basque	0.082987552

Country	Language	%	Country	Language	%
Sweden	Swedish	100	Switzerland	German	69.62822936
Switzerland	French	25.07876497	Switzerland	Italian	5.293005671
Taiwan, Province of China	Mandarino	100	Tajikistan	Tajik	84.75
Tajikistan	Uzbek	10.66666667	Tajikistan	Russian	4.583333333
Thailand	Thai	100	Tunisia	Arabic	100
Türkiye	Turkish	100	Ukraine	Ukrainian	54.36056532
Ukraine	Russian	45.63943468	United States	English	97.88135593
United States	Spanish	2.118644068	Uruguay	Spanish	100
Uzbekistan	Uzbek	94.08	Uzbekistan	Russian	5.92
Venezuela, Bolivarian Republic of	Spanish	100	Viet Nam	Vietnamese	100
Zimbabwe	Shona	68.47736626	Zimbabwe	English	23.12757202
Zimbabwe	Ndebele	8.395061728			

B AI interaction

B.1 Prompt construction and translation procedure

Each survey item was presented to the model verbatim in the corresponding survey language. To ensure consistency across prompts, a short instruction was added before each item: *Respond exclusively to the question without adding comments. Follow the provided instructions.*

In several languages, the original WVS/EVS surveys rely on visual show cards to communicate the response scale rather than embedding the scale directly in the item text. In these cases, we explicitly added numeric anchors to the item text to reproduce the original scale: *(1)* after never justifiable, *(10)* after always justifiable, and *(a number between 1 and 10)* after something in between. These additions were designed solely to make the response scale explicit and did not alter the substantive content of the items.

All introductory instructions and scale clarifications were translated into the corresponding languages using ChatGPT 5 (browser version). The original wording of the survey items was not modified.

B.2 Model specification and querying procedure

We queried OpenAI's *gpt-4o-mini* model via API using custom Python scripts. All queries were conducted using identical prompting structures across languages, differing only in the language of the survey item and instructions.

For each item–language combination, we conducted 10 independent queries to account for stochastic variation in model outputs. Responses were constrained to numeric values between 1 and 10.

C Analysis

C.1 Inter-language distance and correlation of LLM responses

Let

$$V^{l,\text{llm}} = \left(v_1^{l,\text{llm}}, v_2^{l,\text{llm}}, \ldots, v_I^{l,\text{llm}}\right)$$

denote the vector of responses produced by the LLM when prompted in language l, where $v_i^{l,\text{llm}}$ is the model's response to item i, and $i = 1, \ldots, I$ indexes the survey items.

Given two languages l_1 and l_2, we measure the distance between the corresponding response vectors $V^{l_1,\text{llm}}$ and $V^{l_2,\text{llm}}$ using the mean Euclidean distance across items:

$$d\big(V^{l_1,\text{llm}}, V^{l_2,\text{llm}}\big) = \sqrt{\frac{1}{|\mathcal{I}(l_1,l_2)|} \sum_{i \in \mathcal{I}(l_1,l_2)} \left(v_i^{l_1,\text{llm}} - v_i^{l_2,\text{llm}}\right)^2},$$

where $\mathcal{I}(l_1, l_2)$ denotes the set of items for which responses are available in both languages, and $|\mathcal{I}(l_1, l_2)|$ its cardinality.

The use of the mean Euclidean distance is motivated by the presence of missing items. Some questionnaire items are not administered in certain countries. As a result, the corresponding item texts are not available in languages spoken exclusively in the countries where those items are not administered, and responses for those items cannot be obtained in those languages. To ensure comparability across languages despite these differences in item availability, distances are computed using only the items observed in both languages and normalized by the number of shared items.

This distance measure is used to construct the pairwise language-distance matrix shown in Figure 1 in the main text, together with the hierarchical clustering and dendrogram reported in the same figure. The distribution of pairwise language distances obtained with this metric is shown in the histogram in the right panel of Figure 5 in this appendix.

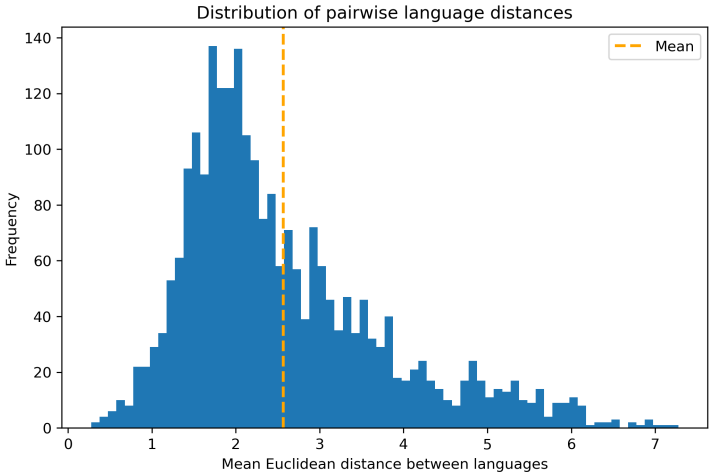

Figure 5: Distribution of pairwise language distances.

We also compute the Pearson correlation between the response vectors associated with two languages. Let

$$\bar{v}^{l_1,\text{llm}}(l_1,l_2) = \frac{1}{|\mathcal{I}(l_1,l_2)|} \sum_{i \in \mathcal{I}(l_1,l_2)} v_i^{l_1,\text{llm}}, \qquad \bar{v}^{l_2,\text{llm}}(l_1,l_2) = \frac{1}{|\mathcal{I}(l_1,l_2)|} \sum_{i \in \mathcal{I}(l_1,l_2)} v_i^{l_2,\text{llm}}$$

denote the mean responses in the two languages computed over the shared items. For pairs of languages with at least two shared items, the correlation coefficient is defined as

$$\rho\big(V^{l_1,\text{llm}}, V^{l_2,\text{llm}}\big) = \frac{\sum_{i \in \mathcal{I}(l_1,l_2)} \left(v_i^{l_1,\text{llm}} - \bar{v}^{l_1,\text{llm}}(l_1,l_2)\right)\left(v_i^{l_2,\text{llm}} - \bar{v}^{l_2,\text{llm}}(l_1,l_2)\right)}{\sqrt{\sum_{i \in \mathcal{I}(l_1,l_2)} \left(v_i^{l_1,\text{llm}} - \bar{v}^{l_1,\text{llm}}(l_1,l_2)\right)^2} \sqrt{\sum_{i \in \mathcal{I}(l_1,l_2)} \left(v_i^{l_2,\text{llm}} - \bar{v}^{l_2,\text{llm}}(l_1,l_2)\right)^2}}.$$

As for the distance metric, the correlation is computed using only the subset of items observed in both languages. If fewer than two items are jointly observed for a given pair of languages, the correlation is not defined. The resulting correlation matrix between languages is displayed in Figure 6, with ordering of languages obtained with hierarchical clustering.

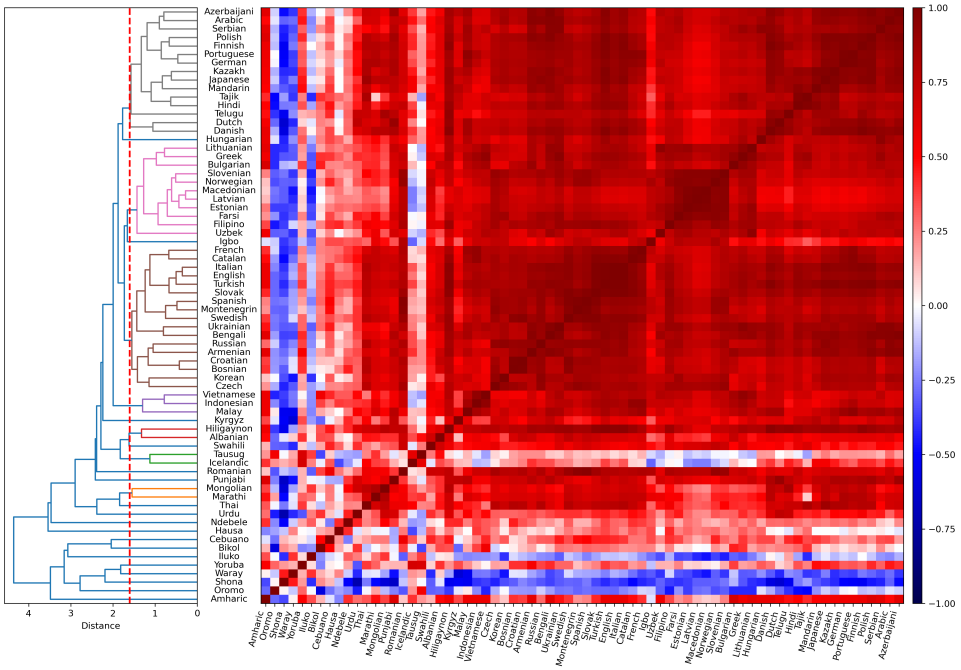

Figure 6: Correlation matrix between languages.

C.1.1 Hierarchical Clustering of Languages

To identify groups of languages with similar response profiles, we apply agglomerative hierarchical clustering to the pairwise distance matrix between languages.

Let L denote the number of languages, and let

$$D = \big(d(l_i, l_j)\big)_{i,j=1,\dots,L} \tag{1}$$

be the symmetric matrix of pairwise distances between languages. The distance between two languages l_i and l_j is defined as

$$d(l_i, l_j) = d\big(V^{l_i,\mathrm{llm}}, V^{l_j,\mathrm{llm}}\big)\,,$$

that is, the mean Euclidean distance between the corresponding response vectors defined in the previous section.

At the initial step, each language forms a singleton cluster. The clustering procedure is agglomerative: at each iteration, the two closest clusters are merged, and the distances between the newly formed cluster and all remaining clusters are recomputed. This process continues until all languages are merged into a single cluster. The resulting hierarchy is represented by a dendrogram.

We use the *average linkage* criterion. For two clusters A and B, their distance is defined as the average of all pairwise distances between languages belonging to the two

clusters:

$$d(A, B) = \frac{1}{|A||B|} \sum_{l_a \in A} \sum_{l_b \in B} d(l_a, l_b), \tag{2}$$

where $|A|$ and $|B|$ denote the number of languages in clusters A and B, respectively.

At each step, the algorithm merges the pair of clusters

$$(A^*, B^*) = \arg\min_{A \neq B} d(A, B). \tag{3}$$

If $C = A^* \cup B^*$ denotes the newly formed cluster, then its distance from any other cluster M is again computed according to the average linkage rule:

$$d(C, M) = \frac{1}{|C||M|} \sum_{l_c \in C} \sum_{l_m \in M} d(l_c, l_m). \tag{4}$$

The hierarchical clustering is implemented using the `linkage()` function from the Python package `scipy.cluster.hierarchy`, with `method="average"`. The output is a sequence of $L - 1$ merges, each characterized by the two clusters being merged, the distance at which the merge occurs, and the size of the resulting cluster. These merges are visualized through the dendrogram.

The order of the leaves in the dendrogram is then used to reorder the rows and columns of the distance and correlation matrices shown in the figures. This ensures that the visual representation of the matrices is consistent with the hierarchical structure identified by the clustering procedure.

To obtain a discrete partition of languages, the dendrogram can be cut at a fixed distance threshold t. In our analysis, we use a cut at distance $t = 1.6$, so that languages connected below this threshold are assigned to the same cluster.

Because the clustering is performed on the distance matrix rather than directly on the response vectors $V^{l,\text{llm}}$, the handling of missing values is incorporated upstream in the construction of the pairwise distances.

C.2 Distance matrices for population and LLM responses

In addition to the inter-language comparison of LLM responses described above, we construct pairwise distance matrices comparing the structure of moral judgments across countries in the population data and in the LLM-generated responses.

Let

$$V^{c,\text{real}} = \left(v_1^{c,\text{real}}, v_2^{c,\text{real}}, \ldots, v_I^{c,\text{real}}\right)$$

denote the vector of population-level responses observed in country c for the survey items, where $v_i^{c,\text{real}}$ represents the empirical response to item i obtained from the WVS/EVS data.

Similarly, let

$$V^{c,\text{llm}} = \left(v_1^{c,\text{llm}}, v_2^{c,\text{llm}}, \ldots, v_I^{c,\text{llm}}\right)$$

denote the vector of responses produced by the LLM when prompted in the language associated with country c.

For any pair of countries c_1 and c_2, we compute the distance between the corresponding response vectors using the root mean squared distance across items. For the population

data, the distance is defined as

$$d^{\text{real}}(c_1, c_2) = \sqrt{\frac{1}{|\mathcal{I}(c_1, c_2)|} \sum_{i \in \mathcal{I}(c_1, c_2)} \left(v_i^{c_1,\text{real}} - v_i^{c_2,\text{real}} \right)^2},$$

where $\mathcal{I}(c_1, c_2)$ denotes the set of items observed in both countries.

Analogously, the distance between the LLM response vectors is defined as

$$d^{\text{llm}}(c_1, c_2) = \sqrt{\frac{1}{|\mathcal{I}(c_1, c_2)|} \sum_{i \in \mathcal{I}(c_1, c_2)} \left(v_i^{c_1,\text{llm}} - v_i^{c_2,\text{llm}} \right)^2}.$$

As in the inter-language comparison, distances are computed only over the subset of items observed for both countries. This normalization by the number of shared items ensures comparability of distances despite differences in item availability across countries.

The two resulting matrices,

$$D^{\text{real}} = \left(d^{\text{real}}(c_i, c_j) \right)_{i,j} \qquad \text{and} \qquad D^{\text{llm}} = \left(d^{\text{llm}}(c_i, c_j) \right)_{i,j},$$

describe the pairwise dissimilarity structure between countries in the empirical data and in the LLM-generated responses, respectively.

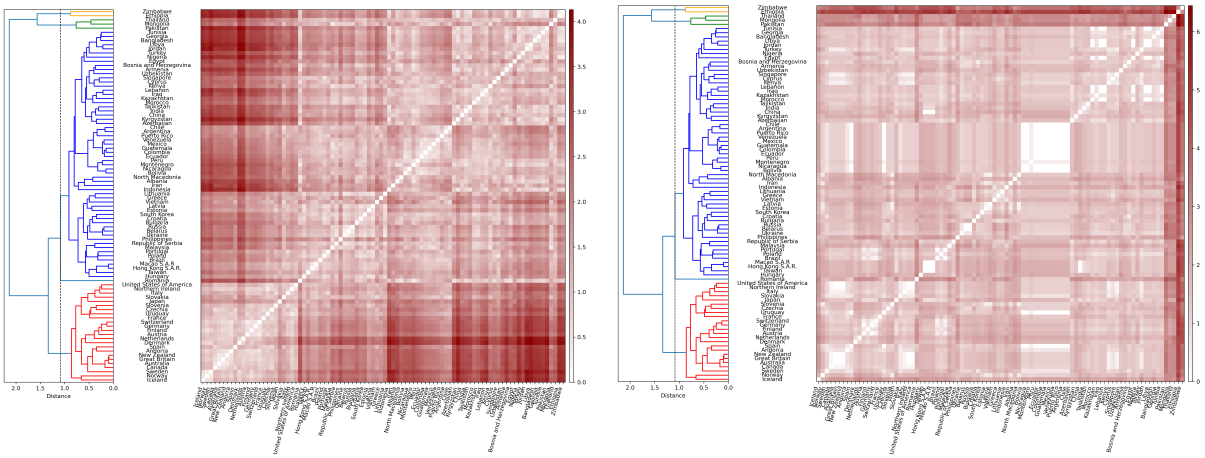

Figure 7: Pairwise distance matrices between countries in the population data (left) and in the LLM-generated responses (right). Distances are computed using the root mean squared distance across survey items. Countries are ordered according to the hierarchical clustering derived from the average of the two distance matrices.

C.2.1 Combined difference matrix

To highlight discrepancies between the two distance structures, we construct a combined matrix defined as the difference between the two pairwise distance matrices:

$$\Delta(c_1, c_2) = d^{\text{real}}(c_1, c_2) - d^{\text{llm}}(c_1, c_2).$$

Each element of the matrix Δ measures the extent to which the distance between two countries differs between the empirical data and the LLM-generated responses. Positive values indicate that the two countries are more distant in the population data than in the LLM responses, while negative values indicate that the LLM responses produce a larger separation between the two countries than observed in the empirical data.

The matrix Δ therefore captures deviations between the relational structure of countries implied by the LLM and that observed in the population data.

C.2.2 Hierarchical clustering based on country-level distance discrepancies

To organize the visualization of the differences between survey-based and LLM-based moral distance structures, we construct a hierarchical clustering of countries based on the pattern of discrepancies between the two distance matrices.

Let D^{real} and D^{llm} denote the matrices of pairwise country distances computed from survey responses and LLM-generated responses, respectively. Both matrices are obtained using the root mean squared distance between country response vectors across moral items, computed over the subset of items observed for both countries.

We first construct the matrix of differences between these distances:

$$\Delta^{\text{dist}}(c_i, c_j) = d^{\text{real}}(c_i, c_j) - d^{\text{llm}}(c_i, c_j),$$

where c_i and c_j denote two countries.

This matrix captures how the relational structure between countries differs between the survey-based moral space and the LLM-generated moral space.

Country discrepancy vectors To cluster countries according to their overall pattern of discrepancies, each country c_i is represented by the vector of its distance differences with respect to all other countries:

$$\Delta_{c_i} = \big(\Delta^{\text{dist}}(c_i, c_1), \Delta^{\text{dist}}(c_i, c_2), \ldots, \Delta^{\text{dist}}(c_i, c_C)\big),$$

where C is the number of countries. By construction, the component corresponding to the country itself satisfies

$$\Delta^{\text{dist}}(c_i, c_i) = 0.$$

Each vector therefore summarizes how the relative position of country c_i differs between the survey-based and LLM-generated moral distance structures.

Distance between countries The similarity between two countries c_i and c_j is measured as the root mean squared distance between their discrepancy vectors:

$$d_\Delta(c_i, c_j) = \sqrt{\frac{1}{|\mathcal{K}(i,j)|} \sum_{k \in \mathcal{K}(i,j)} \left(\Delta^{\text{dist}}(c_i, c_k) - \Delta^{\text{dist}}(c_j, c_k)\right)^2},$$

where $\mathcal{K}(i, j)$ denotes the set of countries for which both discrepancy vectors are observed. This formulation ensures that missing values are handled using pairwise deletion.

Hierarchical clustering and matrix ordering The resulting pairwise distance matrix $d_\Delta(c_i, c_j)$ is used as input for agglomerative hierarchical clustering with the average linkage criterion (see Section C.1.1). The procedure produces a dendrogram grouping countries according to the similarity of their discrepancy profiles.

Finally, the ordering of the leaves in the dendrogram is used to reorder the rows and columns of the matrix Δ^{dist}. This common ordering allows the visualization of the difference matrix to reveal contiguous groups of countries exhibiting similar patterns of divergence between the survey-based and LLM-generated moral distance structures.

C.3 Clustering of Moral Items Based on REAL–LLM Differences

To investigate whether discrepancies between LLM-generated moral judgments and population-level responses follow systematic patterns across survey items, we cluster items based on the cross-country structure of the differences between the two sources.

C.3.1 Construction of item-level REAL–LLM differences

Let

$$v_i^{c,\text{real}} \quad \text{and} \quad v_i^{c,\text{llm}}$$

denote the average response to item i for country c, respectively obtained from the WVS/EVS data and from the LLM.

For each country c and item i, we define the difference

$$d_{c,i} = v_i^{c,\text{real}} - v_i^{c,\text{llm}}.$$

The resulting matrix

$$D = (d_{c,i})$$

has dimension $C \times I$, where C is the number of countries and I the number of survey items.

C.3.2 Standardization of item differences

Because the magnitude of REAL–LLM discrepancies varies across items, the differences are standardized item-wise in order to focus on the cross-country pattern rather than the absolute scale of deviations.

For each item i, we compute

$$z_{c,i} = \frac{d_{c,i} - \mu_i}{\sigma_i},$$

where

$$\mu_i = \frac{1}{|\mathcal{C}_i|} \sum_{c \in \mathcal{C}_i} d_{c,i}, \qquad \sigma_i = \sqrt{\frac{1}{|\mathcal{C}_i| - 1} \sum_{c \in \mathcal{C}_i} (d_{c,i} - \mu_i)^2},$$

and $\mathcal{C}_i$ denotes the set of countries for which item i is observed.

This step ensures that items with larger raw discrepancies do not dominate the clustering procedure.

C.3.3 Pairwise distance between items

To measure similarity between items, we compute the root mean squared (RMS) distance between their standardized difference profiles across countries.

Let $z_{c,i}$ and $z_{c,j}$ denote the standardized REAL-LLM differences for items i and j. The distance between the two items is defined as

$$d(i,j) = \sqrt{\frac{1}{|\mathcal{C}(i,j)|} \sum_{c \in \mathcal{C}(i,j)} \left(z_{c,i} - z_{c,j}\right)^2},$$

where $\mathcal{C}(i,j)$ is the set of countries for which both items are observed.

This formulation implements a *pairwise deletion* strategy: missing values are not imputed, and distances are computed only over the subset of countries where both items are available. This approach is consistent with the treatment of missing observations adopted in the construction of language distance matrices described earlier in the appendix.

The resulting matrix

$$D^{\text{items}} = (d(i, j))$$

contains the pairwise distances between all items.

C.3.4 Hierarchical clustering

We apply agglomerative hierarchical clustering to the item distance matrix D^{items}.

Initially, each item forms a singleton cluster. At each iteration, the two clusters with the smallest distance are merged. The distance between clusters is computed using the *average linkage* criterion. For two clusters A and B, the cluster distance is defined as

$$d(A, B) = \frac{1}{|A||B|} \sum_{i \in A} \sum_{j \in B} d(i, j),$$

where $d(i, j)$ is the RMS distance between items defined above.

The clustering procedure produces a hierarchy of nested clusters that can be visualized using a dendrogram.

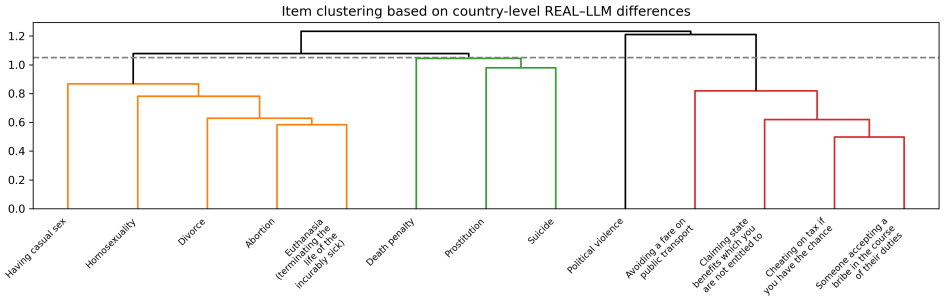

Figure 8: Hierarchical clustering of survey items based on the cross-country pattern of REAL–LLM differences. Distances between items are computed as root mean squared distances over the set of countries observed for both items.

Figure 8 shows the resulting dendrogram. Items that merge at lower heights exhibit similar cross-country structures of discrepancies between LLM-generated judgments and population responses.

C.3.5 Definition of item clusters

To obtain a discrete grouping of items, the dendrogram is cut at a fixed distance threshold t. In our analysis, we use a threshold of

$$t = 1.05,$$

which yields four clusters of items.

The resulting clusters correspond to coherent domains of moral judgment:

- **Anti-civic norms**: items related to rule violations such as *Claiming government benefits to which you are not entitled*, *Avoiding a fare on public transport*, *Cheating on taxes if you have a chance*, and *Someone accepting a bribe in the course of their duties*.
- **Personal and bioethical morality**: items concerning issues such as *Homosexuality*, *Abortion*, *Divorce*, *Euthanasia*, and *Having casual sex*.
- **Public / social order morality**: items involving broader social order considerations such as *Suicide*, *Prostitution*, and the *Death penalty*.
- **Political violence**: the item concerning *Political violence*.

These clusters are subsequently used to analyze how discrepancies between LLM-generated judgments and population-level moral attitudes vary across groups of countries.