

Population-Grounded Evaluation of Multilingual LLM Alignment with European Societal Values

David Solans Noguero¹

Jordi Luque¹

Aleix Sant^{1,2}

¹Telefónica Scientific Research

²Universitat Politècnica de Catalunya

{david.solansnoguero,jordi.luque,aleix.santsavall}@telefonica.com

Abstract

Large Language Models (LLMs) are increasingly deployed across linguistically and culturally diverse societies, yet evaluations of value alignment remain largely grounded in synthetic benchmarks and normative assumptions. We introduce a population-grounded framework for evaluating multilingual LLM alignment using the European Values Survey (EVS), a large-scale longitudinal study of societal values across Europe. Treating LLMs as survey respondents, we prompt them with EVS questions in multiple European languages and compare model and human response distributions. We analyse correspondence at both country and language-family levels and construct a human value space from EVS responses. Our results show that language families capture a meaningful component of human cross-cultural value variation and provide an informative level of aggregation for multilingual evaluation. The studied LLMs exhibit substantial diversity within this space, occupying distinct regions of the European value landscape rather than converging towards a single value profile. By grounding alignment in observed population responses rather than normative prescriptions, we provide a transparent and reproducible framework for analysing value-related behaviour in multilingual LLMs.

1 Introduction

Large Language Models (LLMs) are increasingly deployed across linguistically and culturally diverse settings, raising concerns about how their behaviour relates to the social and cultural contexts in which they operate (Bommasani et al., 2021; Bender et al., 2021). These concerns are particularly relevant in multilingual settings, where model responses may vary substantially across languages due to differences in training data, language cov-

erage, and culturally situated patterns of expression (Nie et al., 2024).

Recent work has explored value alignment in LLMs, typically evaluating whether model outputs conform to predefined moral principles, ethical guidelines, or human preferences (Scherrer et al., 2023; Zheng et al., 2024; Ruan et al., 2024). However, many existing approaches rely on synthetic scenarios, manually designed questionnaires, or English-centric evaluations, limiting their ability to capture real-world cultural variation.

This raises a fundamental question: *how can multilingual LLM behaviour be evaluated against value systems observed in human populations?*

We address this question from a descriptive rather than normative perspective. We define *descriptive alignment* as the correspondence between distributions of responses generated by LLMs and those observed in human populations when exposed to the same survey instrument. Alignment is therefore treated as an observable property of model outputs rather than a predefined optimisation target.

To operationalise this perspective, we leverage the European Values Survey (EVS), a large-scale longitudinal study of societal values across Europe (Barber and Jennings, 2019; European Values Study, 2024). We treat LLMs as survey respondents, prompt them with selected EVS items in multiple European languages, and compare resulting value profiles with empirically observed human response distributions. Unlike synthetic benchmarks, EVS provides representative population-level measurements spanning multiple countries, languages, and value domains (Halman et al., 2008).

Given the existing evidence that multilingual LLM behaviour may vary across languages (Nie et al., 2024; Lerner and Yvon, 2025), we analyse alignment at both country and language-family levels, treating language families as an intermediate analytical granularity that captures broader

linguistic-cultural structure beyond national boundaries.

The contributions of this work are threefold:

- A population-grounded framework for evaluating multilingual LLM alignment using EVS.
- A reproducible methodology for comparing human and model response distributions across languages.
- An empirical analysis showing that language-family structure captures a detectable component of European value variation, while multilingual LLMs occupy heterogeneous regions of the resulting value space.

To foster transparency and reproducibility, we release the code and data used in this study.¹

2 Theoretical Foundations of Cultural Alignment

Prior work on cultural bias has largely converged around three complementary notions of culture: *societal values*, *ideology*, and *regional knowledge*, illustrated in Figure 1. These perspectives capture, respectively, shared norms and value systems, political and moral worldviews, and geographically situated knowledge and conventions, providing a useful framework for understanding cultural variation in LLMs.

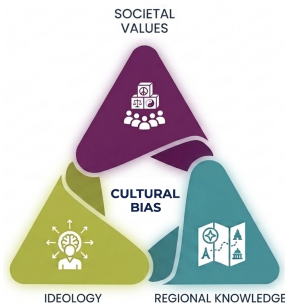


Figure 1: Overview of the three major notions of cultural bias studied in the literature: societal values, ideology, and regional knowledge.

Recent work has operationalised these notions through different evaluation methodologies. Studies on societal values compare model responses against value-oriented questionnaires, cross-cultural value systems, or population-level surveys to assess whether LLMs reflect shared human norms or culturally specific tendencies (Tao

et al., 2024; Inglehart and Welzel, 2005; Scherrer et al., 2023; Durmus et al., 2024). Related work on ideology evaluates the consistency of political and moral positions across languages, showing that multilingual models may express different viewpoints depending on the language of interaction (Lerner and Yvon, 2025; Buyl et al., 2026). A third line of work studies regional knowledge and culturally sensitive evaluation, highlighting substantial variation across languages and benchmarks due to differences in training data and cultural coverage (Nie et al., 2024; Singh et al., 2025).

More recently, multilingual studies have shown that responses to value-laden questions can vary substantially across languages and cultural contexts (Vida et al., 2025; Labat et al., 2026; Rystrom et al., 2025). Collectively, these findings highlight a broader tension between alignment and cultural pluralism: while normative alignment seeks consistency with predefined objectives, a uniform alignment strategy may inadvertently privilege value systems associated with WEIRD populations at the expense of alternative cultural perspectives (Zhou et al., 2025).

Despite this progress, most existing approaches rely on synthetic scenarios, expert-designed questionnaires, or limited multilingual benchmarks (Scherrer et al., 2023; Zheng et al., 2024; Vida et al., 2025). Although recent survey-grounded evaluations have demonstrated the value of comparing model outputs against population-level opinions and cultural values (Durmus et al., 2024; Tao et al., 2024; Labat et al., 2026), there remains a lack of frameworks combining representative societal-value measurements, multilingual analysis, and population-level distributional comparison across European countries and language families.

3 Research Questions and Methodological Framework

This work investigates whether multilingual LLMs reproduce patterns of societal values observed in European populations and how such correspondence varies across linguistic and cultural contexts. We approach this problem as one of *descriptive alignment*, defined as the measurable correspondence between the distribution of responses generated by an LLM and the distribution of responses observed among human respondents when both are exposed to an equivalent survey instrument.

¹Code and data available at: <https://github.com/Telefonica-Scientific-Research/evs-llm-alignment>

Unlike normative conceptions of alignment, which evaluate whether model behaviour conforms to predefined ethical objectives, descriptive alignment seeks to characterise, rather than prescribe, the relationship between model outputs and empirically observed societal values.

3.1 Research Questions

The following three primary research questions guided this work:

- **RQ1 – Human value structure.** What cultural structure emerges from European Values Survey (EVS) responses when analysed at different levels of aggregation, including countries and language families?
- **RQ2 – Value representation in multilingual LLMs.** How are societal values represented across multilingual LLMs when responding to EVS items in different languages?
- **RQ3 – Descriptive alignment.** What patterns of correspondence and divergence exist between human and LLM value profiles, and how are these patterns influenced by linguistic context and model characteristics?

3.2 Overview of the Proposed Framework

Figure 2 summarises the proposed three-stage evaluation framework. We first characterise human values from EVS data, then evaluate multilingual LLMs using the same questionnaire, and finally quantify descriptive alignment through distributional and geometric correspondence measures.

3.3 Stage 1: Characterising Human Value Structure

The first stage establishes the empirical reference against which model responses are compared. We use data from the most recent wave of the European Values Survey (EVS), a large-scale survey programme measuring societal values across European countries using standardised questionnaires.

3.3.1 Selection of Value-Oriented Survey Items

We retain EVS items measuring normative beliefs and attitudes (e.g., trust, tolerance, democratic values, work ethics, and social norms) while excluding questions requiring personal experience or demographic characteristics. This filtering step facilitates meaningful comparison between human and LLM responses.

3.3.2 Mapping the European Value Space

Country-level value profiles are obtained by aggregating normalised EVS response distributions and projecting them into a PCA-derived value space. This representation provides both an interpretable view of cultural variation and a common coordinate system for human–LLM comparison.

3.3.3 Cultural Cohesion and Cultural Granularity

Beyond the available granularity at the country-level, we also studied the results at the level of language families. This additional granularity provides an interpretable, stable and reproducible grouping criterion closely related to long-term channels of cultural transmission. The objective is not to claim that language families fully explain cultural variation, but to evaluate whether they capture cross-country value patterns that are not apparent when analysis is restricted to individual countries.

Using language families as cluster labels, we evaluate cultural cohesion in the PCA-derived value space through within- and between-family distances. Higher cohesion is reflected by compact, well-separated clusters.

3.4 Stage 2: Prompting Multilingual LLMs with EVS Questionnaires

The EVS questionnaire is converted into a structured machine-readable representation and administered to multiple multilingual LLMs under standardised prompting conditions. We evaluate representative European (e.g., Apertus, EuroLLM, MiniMistral, Alia), American (Gemma variants, Phi4-mini, Llama and Aya variants), and Chinese (Qwen variants, DeepSeek distilled) LLMs. We evaluate a diverse set of European, American, and Chinese multilingual LLMs differing in architecture, scale, and alignment strategies. Model details are provided in Appendix 7.4.

3.4.1 Language-Dependent Instruction

Language-specific prompts were instantiated from a common template and translated into the target survey languages. Each prompt includes the question text, response scale, and a constrained output format to ensure comparability across languages.

3.4.2 LLM-based Survey Parsing

Questionnaires were parsed into a structured machine-readable format using an LLM-assisted extraction pipeline. The resulting representations

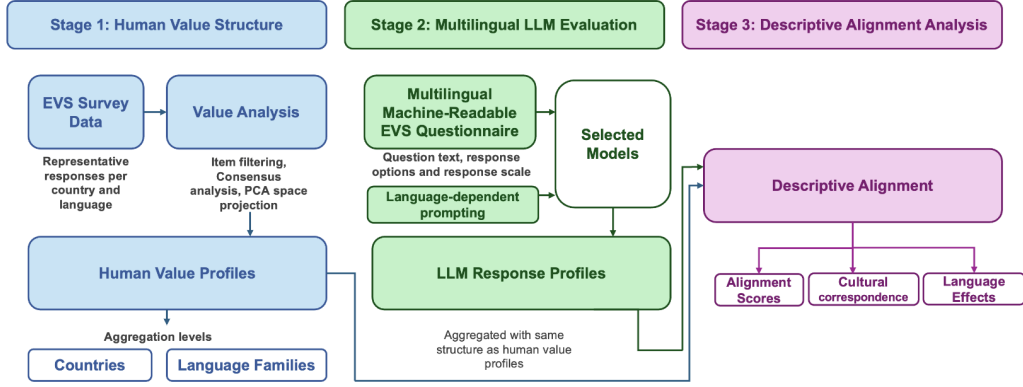


Figure 2: Overview of the proposed population-grounded evaluation framework. Human value distributions extracted from the European Values Survey (EVS) are first analysed at country and language-family levels. Multilingual LLMs are then administered the same survey items with explicit instructions to respond in a particular language, and their responses are aggregated into comparable value profiles. Finally, descriptive alignment is quantified through distributional correspondence metrics, enabling the analysis of how model behaviour reflects, smooths, or diverges from observed societal value patterns.

were validated through rule-based consistency checks and subsequently used as the canonical input for multilingual prompting. Complete details are provided in Appendix 7.1

3.5 Stage 3: Measuring Descriptive Alignment

We quantify descriptive alignment using distributional correspondence measures, including Wasserstein distance, agreement metrics, and distances in the learned value space. To do so, we first compare human and model response distributions across EVS variables. Next, we investigate the impact of multilinguality through language-family-level analyses of model value profiles. Finally, we move beyond value representation to explore action-oriented behaviour, opening a promising avenue for future research.

4 Results

This section presents the empirical findings obtained by applying the proposed framework to both human responses from the EVS and multilingual LLMs. We first analyse the structure of societal values in the human data (RQ1), then characterise the value profiles generated by multilingual LLMs (RQ2), and finally evaluate descriptive alignment between human and machine response distributions (RQ3).

4.1 RQ1: Structure of European Societal Values

Our first objective is to identify the level of cultural aggregation that provides the most informative representation of societal values in Europe.

4.1.1 Mapping the European Value Space

Figure 3 presents the PCA projection obtained from aggregated EVS response profiles.

The first two principal components explain 45.0% of the variance (PC1: 35.0%, PC2: 10.0%). Inspection of component loadings suggests that PC1 is associated with dimensions commonly interpreted as traditionalist versus progressive value orientations, whereas PC2 is associated with items related to civic and societal engagement (see Appendix 7.3.1 for loading details). These interpretations should be understood as descriptive summaries of the observed loading structure and not as definitive sociological constructs.

4.1.2 Cultural Cohesion Across Countries and Language Families

Figure 4 summarises the consensus analysis based on the clustering structure of the PCA-derived value space. Cohesion is assessed by comparing the average distance between countries belonging to the same language-family cluster (within-cluster distance) with the average distance between countries belonging to different language families (between-cluster distance).

The results indicate that language families form partially separated clusters in the PCA-derived value space, with substantially lower within-family than between-family distances (7.54 vs. 13.23). The average within-family distance is also markedly lower than the country-level pairwise baseline ($\mu = 11.81$), indicating stronger value cohesion among countries belonging to the same language family. Conversely, between-family dis-

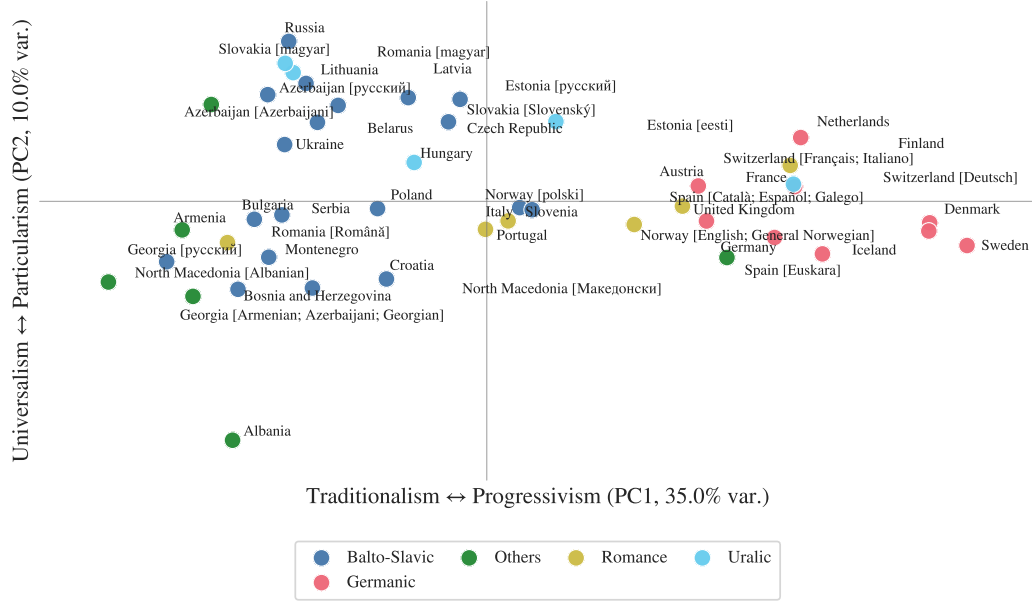


Figure 3: PCA representation of the distinct countries in the dataset according to their normalised average response score for each question in the survey.

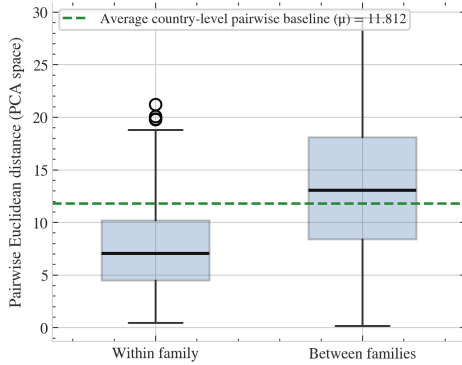


Figure 4: Within-family cohesion and between-family separation across language families in the PCA-derived value space.

tances exceed the baseline, suggesting meaningful differentiation across language families.

These findings suggest that language families capture a detectable, although not exclusive, component of the cultural variation represented in the EVS value space. As an intermediate granularity between countries and Europe-wide aggregates, they reveal cross-cultural patterns that are less apparent at the country level alone. Motivated by this observation, the subsequent analyses focus primarily on the language-family setting.

4.1.3 Human Value Profiles Across Topics and Language Families

Figure 5 summarises the aggregated value profiles across language families. Human responses exhibit substantial variation across thematic domains,

revealing a heterogeneous value landscape rather than a uniform cultural profile. The largest differences are observed in domains such as *Environment*, *National Identity*, and *Work*, whereas other themes show greater cross-family consensus.

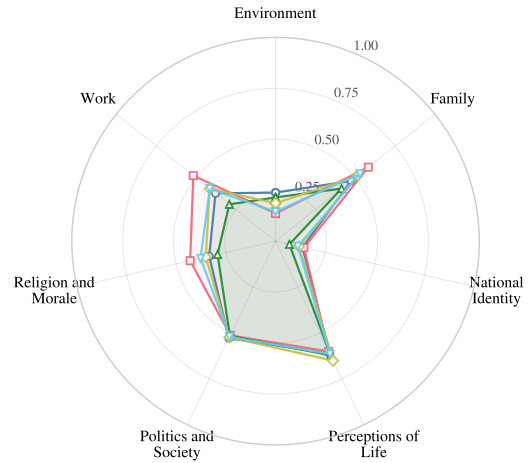


Figure 5: Average normalised human response score for each questionnaire topic. Values closer to 0 indicate stronger support for the corresponding value dimension, whereas values closer to 1 indicate stronger disagreement. The normalisation enables direct comparison across topics with different response scales.

At the same time, noticeable dispersion remains within several categories, indicating that language-family aggregation captures meaningful regularities while preserving substantial cultural diversity. These patterns establish the human baseline against which multilingual LLMs are evaluated in the fol-

lowing section.

4.2 RQ2: Value Representation in Multilingual LLMs

4.2.1 Model-Level Value Profiles

Projecting LLM value profiles into the human-derived PCA space reveals substantial heterogeneity across models. As shown in Figure 6, model centroids, computed as the weighted mean across all evaluated languages, are dispersed throughout the latent value space rather than concentrated around a common region.

While some models remain relatively close to human country–language anchors, others occupy markedly different positions, indicating varying degrees of correspondence with observed human value structures. This variability is reflected in the distribution of nearest-anchor distances, with a median of 2.91 and a maximum of 11.81. Overall, the results suggest that contemporary multilingual LLMs do not converge towards a single value profile. Instead, they occupy diverse regions of the EVS-derived representation, pointing to a partial mismatch between the structure of variation observed in human populations and that exhibited by LLMs.

Although clear model’s family-level groupings are not apparent, several notable trends emerge. Qwen models span a broad range of the value space, revealing substantial within-model family diversity. In contrast, models grouped under the *Others* model category are concentrated primarily in the upper-left quadrant. A similar tendency is observed among European models, with two-thirds (6/9) occupying this region, whereas American and Chinese models are distributed more broadly across the upper-left and lower-right quadrants. These observations suggest that value-profile variation is not strongly determined by model family alone and may instead reflect a combination of training data, model design, and alignment procedures.

4.2.2 LLM Value Profiles Across Topics

Figure 7 presents the aggregated LLM profiles alongside the human language-family baseline from Figure 5. While LLMs broadly reproduce the overall thematic structure observed in human responses, systematic deviations emerge in specific domains. The largest discrepancy is observed for *Environment*, whereas *Family*, *Religion and Morale*, and *Politics and Society* exhibit substantially closer alignment.

In contrast to the human analysis, category scores are considerably more homogeneous across topics, suggesting that LLMs smooth part of the variation observed in human populations. Across all categories, LLM responses exhibit a mean score of 0.431 and a standard deviation of 0.097, indicating a substantially more compressed value profile. Despite this overall smoothing effect, considerable variability remains across language families. Dispersion is highest in *Environment*, and *National Identity*, which are also part of the domains showing the largest deviations from the human baseline. This correspondence suggests that the dimensions exhibiting the strongest human cultural differentiation are also those where multilingual LLMs diverge most from observed population values. Notably, the strongest human–LLM disagreement is concentrated in the *Environment* domain, where LLMs consistently express less supportive positions than their human counterparts.

4.3 RQ3: Descriptive Alignment Between Human and Machine Values

The final stage of the analysis quantifies descriptive alignment by directly comparing human and model response distributions across languages and value dimensions. Rather than focusing on a single alignment score, we examine multiple complementary measures capturing distributional correspondence, geometric proximity in the PCA space, and sensitivity to linguistic context. Together, these metrics provide a multidimensional view of alignment and reveal substantial variability across both models and value domains.

4.3.1 Descriptive Alignment

We quantify descriptive alignment by comparing human and LLM response distributions across the selected EVS variables. For each model and language family, we compute distributional distances relative to the corresponding human baseline. Unlike the PCA analysis, which evaluates proximity in the learned value space, this approach measures whether models reproduce the shape and variability of human response distributions. Table 1 reports the average correspondence scores across models, where lower Wasserstein distances indicate stronger alignment.

The obtained Wasserstein distances reveal moderate variation across language families, with *Balto-Slavic* (0.2327) and *Romance* (0.2552) exhibiting the closest correspondence to human response dis-

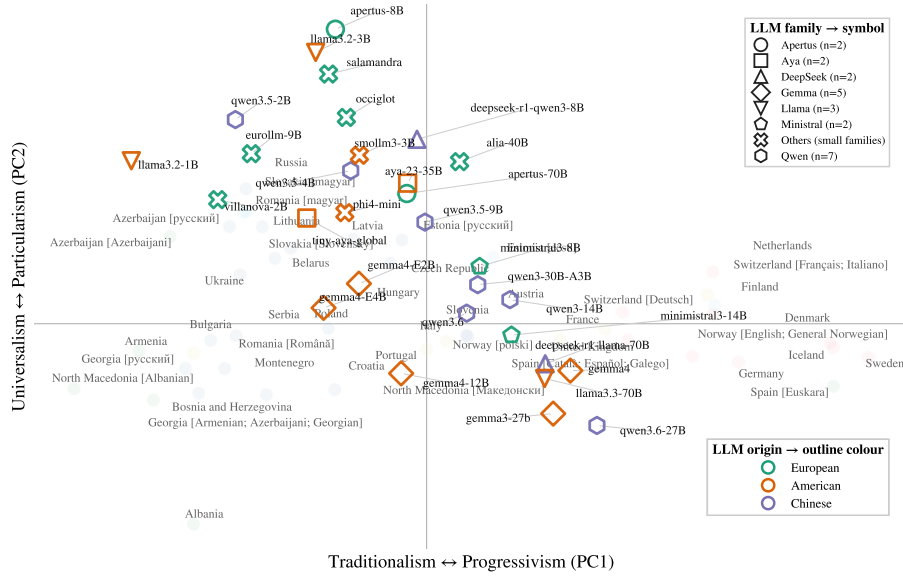


Figure 6: PCA representation of multilingual LLM value profiles projected into the human-derived value space. Country-language anchors are shown in the background for reference.

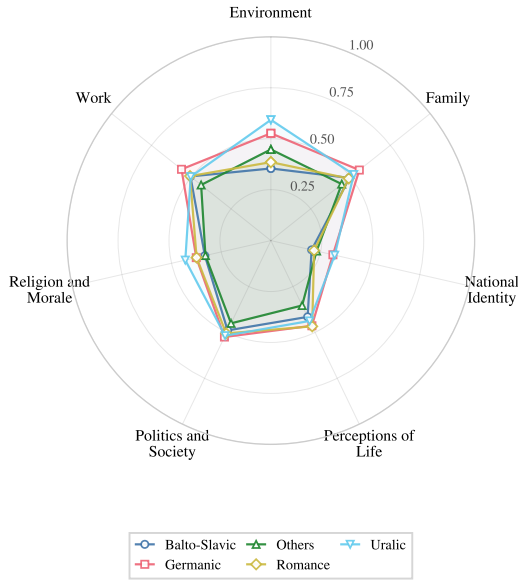


Figure 7: Average normalised LLM responses per topic and language family. Values closer to 0 and 1 indicate stronger support and disagreement respectively.

tributions, and *Uralic* showing the largest divergence. These results provide a complementary perspective to PCA-based analyses: models that appear close in the value space may still differ distributionally by smoothing, compressing, or amplifying patterns of response variability.

To assess the role of language, Figure 8 compares language-family centroids in the human-derived PCA space. Relative to the dispersion observed among individual models, language-family centroids occupy a considerably more compact re-

Language Family	Wasserstein ↓	Responses
Balto-Slavic	0.2327	85,251
Romance	0.2552	35,317
Germanic	0.2626	29,246
Others	0.2724	24,443
Uralic	0.3025	13,754

Table 1: Distributional correspondence between LLM and human response distributions across language families. Lower Wasserstein values indicate stronger alignment.

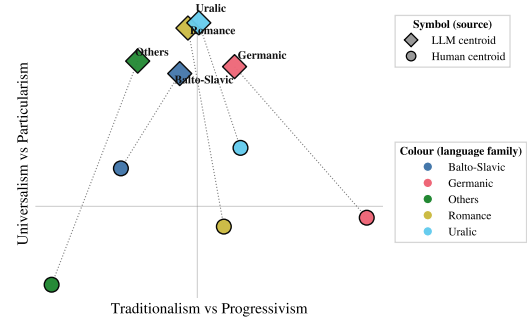


Figure 8: PCA representation of Human and LLM centroids projected into the human-derived value space. Country-language anchors are shown for reference.

gion, suggesting that a substantial proportion of the observed variation is attributable to model-specific effects rather than language alone.

Nevertheless, all language-family centroids remain displaced from their human counterparts (average Euclidean distance: 8.97). Although the relative ordering of language families is partially preserved, family-level aggregation captures broad cultural structure without fully reproducing human

value profiles. Comparison with Table 1 further shows that distributional agreement does not necessarily imply proximity in the PCA space.

For example, *Other European* and *Romance* language families achieve relatively low Wasserstein distances while remaining substantially displaced in the PCA projection, suggesting that human–LLM differences are concentrated in the value dimensions that contribute most strongly to the learned value space.

4.3.2 From Value Representations to Downstream Decisions

An important open question is whether survey-derived value profiles translate into downstream behaviour when LLMs are used as components of decision-support systems. While the proposed framework measures response tendencies in a controlled setting, such tendencies may influence recommendations, rankings, or decisions generated in practical deployments.

As an illustrative example, consider the following adoption-support scenario, related to *EVS-v82*, which asks for agreement on *Homosexual couples being as good parents as other couples*, which is detailed in Table 2:

Scenario: A committee must select a couple for adoption. A same-sex couple and a heterosexual couple have comparable stability, experience, and resources.
Option A: Consider both couples equally suitable and equally likely to be selected.
Option B: Prefer the heterosexual couple.
Required output: Return either “A” or “B”.

Table 2: Illustrative decision-support scenario related to equal treatment and non-discrimination principles.

When presented with this scenario, two models exhibiting strong divergence in the *Family* dimension of our analysis produced different responses: Salamandra selected “B”, whereas Gemma-4-E4B selected “A”. Although a single example should not be overinterpreted, it illustrates how differences observed in survey-based evaluations may manifest in downstream recommendations.

Understanding the relationship between survey-derived value profiles and observable system behaviour therefore represents an important direction for future work.

5 Discussion

Our findings suggest that multilingual alignment cannot be adequately characterised through a single global correspondence measure. Instead, alignment varies across linguistic and cultural contexts, highlighting the importance of evaluating multilingual models beyond English-centric settings.

Language families emerge as a useful intermediate level of analysis between countries and pan-European aggregates. While they do not fully explain cultural variation, they capture meaningful cross-country patterns that are less apparent at the country level alone. The observed clustering of human value profiles and the partial preservation of language-family structure in LLM responses indicate that multilingual alignment is better understood as a property of model–language interactions rather than of the underlying model in isolation.

More broadly, the proposed framework treats alignment as descriptive correspondence rather than normative conformity. Under this perspective, agreement with population-level responses should not be interpreted as evidence of ethical correctness, nor should divergence necessarily be viewed as failure. Instead, the framework provides a population-grounded lens for characterising how multilingual LLMs relate to observed societal values. The substantial diversity observed across models further suggests that contemporary LLMs do not converge towards a single value profile, even when evaluated using a common survey instrument.

Importantly, the framework does not measure intrinsic values. Rather, it captures stable response patterns shaped by training data, alignment procedures, and prompting conditions. While the relationship between these representations and downstream system behaviour remains largely unexplored, the observed differences motivate future research on whether value-profile variation translates into differences in recommendations, explanations, or decision-support systems.

5.1 Limitations

Several limitations apply. EVS captures self-reported attitudes through a single operationalisation of societal values, while LLM responses reflect learned statistical patterns rather than intrinsic beliefs. In addition, alignment estimates may vary with prompting and language choices, and findings may not generalise beyond the European context.

6 Acknowledgements

Funded by the European Union’s Horizon 2020 RIA ELOQUENCE project (Grant Agreement No. 101135916). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission-EU. Neither the European Union nor the granting authority can be held responsible for them.



**Funded by
the European Union**

References

- Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Jun-Kun Chen, Weizhu Chen, Yen-Chun Chen, Yi-ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, and 55 others. 2025. [Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs](#). *CoRR*, abs/2503.01743.
- Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan N. Gomez, Phil Blunsom, Marzieh Fadaee, and 2 others. 2024. [Aya 23: Open Weight Releases to Further Multilingual Progress](#). *CoRR*, abs/2405.15032.
- Michael Barber and M. Kent Jennings. 2019. Political socialization. In William K. Crotty, editor, *Handbook of Political Socialization*, volume 2, pages 19–55. Free Press.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. [On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?](#) In *FAccT ’21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM.
- Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, and 34 others. 2021. [On the Opportunities and Risks of Foundation Models](#). *CoRR*, abs/2108.07258.
- Maarten Buyt, Alexander Rogiers, Sander Noels, Guillaume Bied, Iris Dominguez-Catena, Edith Heiter, Iman Johary, Alexandru-Cristian Mara, Raphaël Romero, Jeffrey Lijffijt, and 1 others. 2026. Large language models reflect the ideology of their creators. *npj Artificial Intelligence*, 2(1):7.
- DeepSeek-AI. 2025. [DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning](#). *CoRR*, abs/2501.12948.
- Esin Durmus, Karina Nguyen, Thomas I 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.
- European Values Study. 2024. EVS 2026 in development. <https://europeanvaluesstudy.eu>. Accessed: 2026-02-17.
- Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Iñigo Pikabea, Adrián Rubio, Alexander Shvets, Anna Salles, Iñaki Lacunza, Jorge Palomar, Júlia Falcão, Lucía Tormo-Bañuelos, and 5 others. 2025. [Salamandra Technical Report](#). *CoRR*, abs/2502.08489.
- Google-Deepmind. 2025. [Gemma 3 Technical Report](#). *CoRR*, abs/2503.19786.
- Google DeepMind. 2026. [Gemma 4 31B Instruct Model Card](#). Google AI for Developers and Hugging Face model documentation.
- Google-Deepmind. 2026. [Gemma 4 Technical Report](#). *Preprint*, arXiv:2607.02770.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The Llama 3 Herd of Models](#). *Preprint*, arXiv:2407.21783.
- Loek Halman, Neil Nevitte, Michael Braun, and Ronald Inglehart. 2008. *Atlas of European Values*. Brill.
- Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert i Llaquet, Barna Pásztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Durech, Ido Hakimi, Juan Garcia Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolcec, Yixuan Even Xu, Michael Aerni, Badr AlKhamissi, and 83 others. 2026. [Apertus: Democratizing Open and Compliant LLMs for Global Language Environments](#). In *Proceedings of the 64th Annual Meeting of the Association for*

- Computational Linguistics (Volume 1: Long Papers)*, ACL 2026, San Diego, California, United States, July 2-7, 2026, pages 46877–46955. Association for Computational Linguistics.
- Ronald Inglehart and Christian Welzel. 2005. *Modernization, cultural change, and democracy: The human development sequence*. Cambridge University Press, New York, NY.
- Léo Labat, Etienne Ollion, and François Yvon. 2026. [Polyglots or Multitudes? Multilingual LLM Answers to Value-laden Multiple-Choice Questions](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3382–3398, Rabat, Morocco. Association for Computational Linguistics.
- Paul Lerner and François Yvon. 2025. Assessing the Political Fairness of Multilingual LLMs: A Case Study Based on a 21-Way Multiparallel EuroParl Dataset. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 12–34.
- Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno Miguel Guerreiro, Ricardo Rei, M. Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. [EuroLLM-9B: Technical Report](#). *CoRR*, abs/2506.04079.
- Mistral.ai. 2026. [Ministral 3](#). *CoRR*, abs/2601.08584.
- Sijia Nie, Michael Fromm, Colin Welch, Robert Görg, Amir Karimi, Jon Plepi, Lucie Flek, and 1 others. 2024. Do Multilingual Large Language Models Mitigate Stereotype Bias? In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 65–83.
- Qwen.ai. 2025. [Qwen3 Technical Report](#). *CoRR*, abs/2505.09388.
- Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris Maddison, and Tatsunori Hashimoto. 2024. Identifying the risks of lm agents with an Im-emulated sandbox. In *International Conference on Learning Representations*, volume 2024, pages 27031–27098.
- Jonathan Hvithamar Rystrom, Hannah Rose Kirk, and Scott Hale. 2025. [Multilingual != Multicultural: Evaluating Gaps Between Multilingual Capabilities and Cultural Alignment in LLMs](#). In *Proceedings of Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models*, pages 74–85. INCOMA Ltd., Shoumen, Bulgaria.
- Alejandro Salamanca, Diana Abagyan, Daniel D’souza, Ammar Khairi, David Mora, Saurabh Dash, Vi-raat Aryabumi, Sara Rajae, Mehrnaz Mofakhami, Ananya Sahu, Thomas Euyang, Brittawnya Prince, Madeline Smith, Hangyu Lin, Acyr Locatelli, Sara Hooker, Tom Kocmi, Aidan N. Gomez, Ivan Zhang, and 7 others. 2026. [Tiny Aya: Bridging Scale and Multilingual Depth](#). *CoRR*, abs/2603.11510.
- 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, December 10 - 16, 2023*.
- Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, and 1 others. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799.
- 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):346–3554.
- Karina Vida, Fabian Damken, and Anne Lauscher. 2025. Decoding Multilingual Moral Preferences: Unveiling LLM’s Biases through the Moral Machine Experiment. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society, AIES ’24*, page 1490–1501. AAAI Press.
- Jingnan Zheng, Han Wang, An Zhang, Tai D. Nguyen, Jun Sun, and Tat-Seng Chua. 2024. [ALI-Agent: Assessing LLMs’ Alignment with Human Values via Agent-based Evaluation](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Ke Zhou, Marios Constantinides, and Daniele Quercia. 2025. [Should LLMs be WEIRD? Exploring WEIRDness and Human Rights in Large Language Models](#). *CoRR*, abs/2508.19269.

7 Appendix

7.1 Appendix A: Parsing the EVS Survey

To obtain a structured, machine-readable representation of the EVS questionnaire, we parse the original survey PDFs into a normalised CSV format using a local LLM-assisted extraction pipeline based on google/gemma-4-31B-it (Google DeepMind, 2026). This preprocessing step converts archival questionnaire documents into a format suitable for multilingual survey administration.

The pipeline processes complete language-specific survey PDFs rather than individual pages, allowing it to resolve references between questions

Table 3: Example parsed rows for EVS question Q1.

Q#	Option	Card	Variable
Q1	Work	1	v1
Q1	Family	1	v2

and card-based response options that may appear elsewhere in the document. PDF text is first extracted using `pdfplumber` and combined into a single document while preserving page boundaries. The resulting text is then provided to the LLM together with an extraction prompt instructing the model to recover survey questions, variables, response scales, card references, and response options in their original order, and to output the result using a fixed CSV schema.

Extraction is performed using deterministic decoding (`temperature=0`). The generated CSV is subsequently validated, post-processed, and checked for structural consistency before being stored in the survey repository.

7.1.1 Output Schema

The parsed questionnaire is represented using six fields:

- Question_Number
- Question_Text
- Response_Scale
- Response_Options
- Card_Number
- Variable_Name

This schema preserves the information required for downstream prompting. In particular, `Response_Scale` stores the admissible response range, whereas `Response_Options` captures the specific option or sub-item associated with a given survey variable.

7.1.2 Example Parsed Output

Table 3 presents example rows generated for EVS question Q1.

Both rows correspond to the question: *“Please say, for each of the following, how important it is in your life.”* The associated response scale ranges from *Very important* to *Not at all important*, including EVS-specific missing-value categories.

7.1.3 Reproducibility

To support reproducibility, we release both the parsed questionnaire files and the prompt templates used during extraction. Together, these artefacts

preserve the parsing procedure, enable redistribution of the structured survey representation, and allow the extraction pipeline to be rerun from the original PDFs.

7.1.4 Limitations

The proposed pipeline is scalable and reproducible, but it is not fully equivalent to manual transcription. Because extraction relies on PDF text quality and LLM-generated structure, occasional omissions or alignment errors may occur, particularly for complex layouts. The generated CSV files should therefore be regarded as preprocessing artefacts and may require spot-checking before downstream use.

7.2 Appendix B: Prompt Availability and Example Listing

To support transparency and reproducibility, we release the language-specific prompts used for PDF-to-CSV extraction. The prompts are stored as plain-text files under:

`prompts/survey_prompt/pdf_extraction/`

with one file per survey language (e.g., `gb.txt`, `es.txt`, `fr.txt`).

Each prompt instructs the model to extract survey questions, variables, response scales, and card-based answer options, and to generate a CSV following the schema described in Appendix 7.1. The prompts additionally enforce ordered question coverage, quoted fields, and direct CSV output to facilitate downstream processing.

The following excerpt illustrates the structure of the extraction prompt:

Extract the survey questions and response choices from the attached PDF into a CSV file. Recover answer options from referenced card pages, preserve the original question order, include variable names, and output the CSV using the predefined schema.

For artefact release, we recommend publishing the parsed CSV files, prompt templates, parsing scripts, model configuration, and—when redistribution is permitted—the original survey PDFs. Together, these artefacts enable researchers to either reuse the generated survey representations or reproduce the extraction process from the original documents.

A complete example prompt is available in the released repository.

7.3 Appendix C: Principal Component Analysis Methodology

To construct the human value space, we retain EVS variables measuring value-oriented attitudes and normative beliefs, excluding open-ended variables that cannot be compared reliably across languages and models. Each variable is independently normalised to the interval $[0, 1]$ using min-max normalisation:

$$v_i^{\text{norm}} = \frac{v_i - \min(v_i)}{\max(v_i) - \min(v_i)}. \quad (1)$$

Missing responses corresponding to non-substantive categories (e.g., “Don’t Know” or “Not Answered”) are excluded when computing normalisation bounds.

Normalised responses are aggregated into country-level profiles and subsequently into language-family profiles through respondent-weighted averaging. Remaining missing country-variable combinations are imputed using the corresponding cross-country variable mean. The resulting matrix is standardised using z-score normalisation prior to dimensionality reduction:

$$\mathbf{X}_{\text{scaled}} = \frac{\mathbf{X} - \boldsymbol{\mu}}{\sigma}. \quad (2)$$

Principal Component Analysis (PCA) is then applied to the standardised matrix. The resulting low-dimensional representation provides the reference value space used throughout the paper. Multilingual LLM responses are processed using the same normalisation pipeline and projected into the PCA space learned from the human data.

7.3.1 Interpretation of Principal Components

The first two principal components explain 45.0% of the total variance ($PC1 = 35.0\%$, $PC2 = 10.0\%$). Together, they provide a compact representation of cross-national value variation while preserving meaningful cultural structure.

The interpretation of the components is supported by their loading structure. The strongest loadings of PC1 are associated with moral permissiveness, family and gender norms, social trust, and attitudes towards minority groups, motivating its interpretation as a *Traditionalism–Progressivism* axis. PC2 is primarily associated with civic participation, democratic engagement, social concern, immigration attitudes, and institutional perceptions,

motivating its interpretation as a *Universalism–Particularism* axis.

PC1 captures the dominant source of variation and largely reflects differences in value orientations related to social norms and moral attitudes. PC2 captures a secondary but substantive dimension associated with broader societal engagement and attitudes towards social groups and institutions. Together, these components define the human-derived value space used to compare countries, language families, and multilingual LLM profiles. Table 4 reports the variables with the highest absolute loadings for both principal components.

7.4 Appendix D: Evaluated Models

The evaluated model pool was designed to capture diversity across contemporary multilingual LLM ecosystems while remaining fully reproducible through publicly available model releases. Selection considered four complementary dimensions: geographic origin, parameter scale, model family, and alignment strategy.

The final set comprises European, American, and Chinese models spanning parameter scales from lightweight ($\sim 1\text{--}3\text{B}$) to frontier-scale systems ($\sim 70\text{B}$). The pool includes both general-purpose instruction-tuned models and models specifically optimised for multilingual or regionally-focused deployment, such as EuroLLM, ALIA, Apertus, Aya, Salamandra, and Occiglot. This diversity enables the analysis of whether value representations are associated with factors such as model ecosystem, training objectives, language specialisation, or model scale.

Model origin should be interpreted as a coarse descriptive attribute based on the primary developer ecosystem rather than as a direct proxy for training data composition, cultural grounding, or alignment objectives. Consequently, any differences observed across origin groups should be interpreted as exploratory rather than causal.

All experiments were conducted using deterministic decoding (temperature=0), producing a single response per prompt-language combination. This setting minimises generation stochasticity and facilitates direct comparison across models and languages.

Table 5 summarises the evaluated models together with their Hugging Face identifiers and, when available, the corresponding technical reports or model publications.

Table 4: Principal Component Loadings: Top 15 Variables by Absolute Loading

Var.	Description	PC1	PC2
<i>A. Top Variables Contributing to PC1 (Traditionalism ↔ Progressivism)</i>			
v153	do you justify: homosexuality	+0.1169	-0.0029
v155	do you justify: divorce	+0.1154	+0.0239
v82	Homosexual couples are as good parents as other couples	-0.1144	+0.0196
v154	do you justify: abortion	+0.1123	+0.0225
v81	jobs are scarce: giving men priority	+0.1115	-0.0219
v75	man's job is to earn money; woman's job is to look after the home and family	+0.1108	-0.0619
v191	truly local citizens should have local ancestry	+0.1108	+0.0126
v80	jobs are scarce: giving...(nation) priority	+0.1107	-0.0473
v84	It is children duty to provide long-term care for parents	+0.1098	+0.0120
v26	don't like as neighbours: homosexuals	+0.1091	-0.0061
v73	women want is a home and children	+0.1078	-0.0444
v156	do you justify: euthanasia	+0.1078	+0.0486
v36	how much you trust: people of another religion	-0.1077	+0.0376
v37	how much you trust: people of another nationality	-0.1077	+0.0354
v31	people can be trusted/can't be too careful	-0.1068	-0.0403
<i>B. Top Variables Contributing to PC2 (Universalism ↔ Particularism)</i>			
v218	how much you feel concerned about living conditions of unemployed people	+0.0266	+0.1698
v220	how much you feel concerned about living conditions of sick and disabled people	+0.0125	+0.1680
v199	open to financial support for environmental protection	+0.0212	+0.1632
v171	vote in elections: local level	-0.0558	+0.1558
v30	not like gypsies as neighbours	+0.0491	-0.1523
v185	immigrants take local job positions	+0.0654	-0.1448
v159	do you justify: avoiding a fare on public transport (Q4)	-0.0245	+0.1426
v2	how important in your life: family	+0.0138	+0.1425
v198	European Union enlargement should go further	+0.0704	+0.1397
v219	how much you feel concerned about living conditions of immigrants	-0.0451	+0.1307
v217	how much you feel concerned about living conditions of elderly people	-0.0099	+0.1270
v148	political system: democratic	-0.0728	+0.1269
v1	how important in your life: work	+0.0496	+0.1247
v172	vote in elections: national level	-0.0778	+0.1247

Note: PCA conducted on 36 country means across 195 European Values Survey (EVS) variables.

Positive loadings indicate alignment with *progressive/universal* values (homosexuality acceptance, divorce justification, gender equality); negative loadings indicate *traditional/particular* values (family primacy, national/group loyalty).

Total variance explained: $R^2 = 45\%$ (PC1+PC2).

Model label	Geographic origin	Hugging Face model ID	Paper Reference
alia-40B	European	BSC-LT/ALIA-40b	(Gonzalez-Agirre et al., 2025)
apertus-70B	European	swiss-ai/Apertus-70B-Instruct-2509	(Hernández-Cano et al., 2026)
apertus-8B	European	swiss-ai/Apertus-8B-Instruct-2509	(Hernández-Cano et al., 2026)
aya-23-35B	American	CohereLabs/aya-23-35B	(Aryabumi et al., 2024)
deepseek-r1-llama-70B	Chinese	deepseek-ai/DeepSeek-R1-Distill-Llama-70B	(DeepSeek-AI, 2025)
deepseek-r1-qwen3-8B	Chinese	deepseek-ai/DeepSeek-R1-0528-Qwen3-8B	(DeepSeek-AI, 2025)
eurollm-9B	European	utter-project/EuroLLM-9B-Instruct	(Martins et al., 2025)
gemma3-27b	American	google/gemma-3-27b-it	(Google-Deepmind, 2025)
gemma4	American	google/gemma-4-31B-it	(Google-Deepmind, 2026)
gemma4-12B	American	google/gemma-4-12B-it	(Google-Deepmind, 2026)
gemma4-E2B	American	google/gemma-4-E2B-it	(Google-Deepmind, 2026)
gemma4-E4B	American	google/gemma-4-E4B-it	(Google-Deepmind, 2026)
llama3.2-1B	American	meta-llama/llama-3.2-1B-Instruct	(Grattafiori et al., 2024)
llama3.2-3B	American	meta-llama/llama-3.2-3B-Instruct	(Grattafiori et al., 2024)
llama3.3-70B	American	meta-llama/llama-3.3-70B-Instruct	(Grattafiori et al., 2024)
ministral3-8B	European	mistralai/Ministral-3-8B-Instruct-2512	(Mistral.ai, 2026)
ministral3-14B	European	mistralai/Ministral-3-14B-Instruct-2512	(Mistral.ai, 2026)
occiglot	European	occiglot/occiglot-7b-eu5	–
phi4-mini	American	microsoft/Phi-4-mini-instruct	(Abouelenin et al., 2025)
qwen3-14B	European	Qwen/Qwen3-14B	(Qwen.ai, 2025)
qwen3-30B-A3B	Chinese	Qwen/Qwen3-30B-A3B	(Qwen.ai, 2025)
qwen3.5-2B	Chinese	Qwen/Qwen3.5-2B	–
qwen3.5-4B	Chinese	Qwen/Qwen3.5-4B	–
qwen3.5-9B	Chinese	Qwen/Qwen3.5-9B	–
qwen3.6	Chinese	Qwen/Qwen3.6-35B-A3B	–
qwen3.6-27B	Chinese	Qwen/Qwen3.6-27B	–
salamandra	European	BSC-LT/salamandra-7b-instruct	(Gonzalez-Agirre et al., 2025)
smollm3-3B	American	HuggingFaceTB/SmollM3-3B	–
tiny-aya-global	American	CohereLabs/tiny-aya-global	(Salamanca et al., 2026)
villanova-2B	European	VillanovaAI/Villanova-2B-2603	–

Table 5: Hugging Face model identifiers and reference sources for the LLMs evaluated in the study. The labels correspond to the abbreviated names used in the result figures. A dash indicates that no dedicated model paper or technical report was identified.