Decoding Values:
Measuring Culture Embedded in Large Language Models

Anton Steuer[1,2], Carolin Schuster[1], Michael Banf[2]
[1] Technical University of Munich, Germany
[2] Perelyn GmbH, Munich, Germany
anton.steuer@perelyn.com

Motivation

LLM training data is dominated by English-language and Western sources [1, 2]. Prior work documents systematic alignment with WEIRD (Western, Educated, Industrialized, Rich, and Democratic) cultural values [3–5]. When deployed across diverse user bases, this skew can misrepresent users' own cultural frames and result in low-quality or inappropriate responses.

The dominant approach to test an LLM's values prompts models with human-designed questionnaires and inspects the generated text [5–7]. This black-box approach reveals what the model says when asked, but not what its internal representations encode. The internally encoded information and potential biases, however, influence the model's downstream generations. Hence, several approaches to measure such biases in embeddings have been proposed, including the POLAR framework [8] based on *semantic differentials* [9]. A *semantic differential* spans a scale between two opposite *poles* (e.g. *important*↔*unimportant*); a **concept** is placed on the scale by its association with either extreme. The POLAR framework [8] implements this in embedding space: a dimension represents the difference $\vec{p}_+ - \vec{p}_-$ between pole embeddings. Projecting any concept onto this dimension yields its association score with the poles.

However, adapting POLAR to instruction-tuned decoder-only LLMs faces two obstacles: (i) model-specific hidden-state geometries that distort raw projection scores [10] and (ii) score distributions that are not directly comparable across architectures.

Contribution

We present a white-box method that (i) extends POLAR to decoder-only LLMs, (ii) aggregates multiple antonym pairs per dimension to suppress single-word noise, and (iii) produces cross-model comparable scores via rank-based rescaling against a model-specific noise baseline. Using our algorithm, we evaluate six instruction-tuned LLMs from four countries based on 87 concepts rated on 16 dimensions. Both the concepts and the scales/dimensions are based on the World Values Survey [11], covering social values, economic systems, institutions, religion, politics, and ethics.

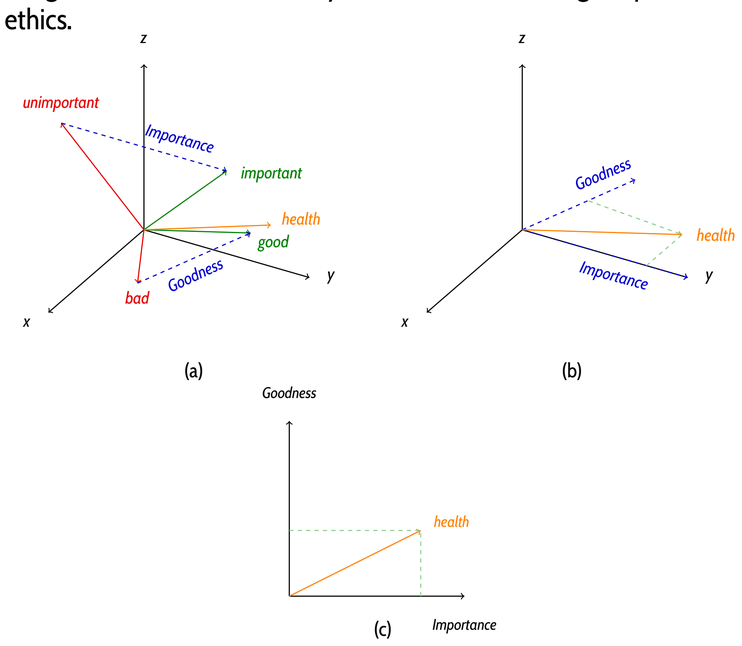

Figure 1: POLAR transformation: **health** projected into a 2D POLAR space defined by the two semantic differentials *Goodness* and *Importance*. (a) Each dimension runs from the *negative* to the *positive* pole (average over $m = 10$ antonym pairs); (b) **health** in the POLAR basis; (c) 2D POLAR coordinates give interpretable association scores.

Method: a four-stage white-box pipeline

The pipeline maps any concept c to a vector of interpretable, cross-model comparable Likert scores.

1. Span the POLAR space. POLAR dimensions are bipolar *semantic differentials*, e.g. *Importance* runs from *unimportant* to *important*. We define each pole as a *cloud* of $m = 10$ synonyms (*important, major, significant, …* vs. *unimportant, minor, insignificant, …*); the vector representation $\vec{d}_j$ for dimension j then becomes the average of their pairwise differences, suppressing single-word noise:

$$\vec{d}_j = \frac{1}{m}\sum_{i=1}^{m}\left(\vec{p}_{1,i}^{(j)} - \vec{p}_{2,i}^{(j)}\right). \tag{1}$$

2. Embed in context. *Poles*: each synonym is embedded word-finally in a GPT-4o-mini-generated sentence (e.g. *"The outcome was of great **importance**."*). *Concepts*: embedded in a neutral questionnaire template (*"In everyday life, I assess **family** as:"*), avoiding bias toward any particular pole. Both use the *mean over all transformer layers* rather than the final hidden state [12].

3. Project into the POLAR space. Stacking the $n = 16$ dimension vectors into $\mathbf{A} = [\vec{d}_1, \ldots, \vec{d}_n]^\top$, each concept embedding $\vec{x}_c$ is expressed in POLAR coordinates via

$$\vec{p}_c = (\mathbf{A}^\top)^{-1}\vec{x}_c, \tag{2}$$

yielding one raw association score per scale/dimension.

4. Anchor and rescale. The same pipeline is run for 250 random WordNet nouns as concepts, producing a model-specific reference distribution per dimension. Every concept's raw score x is replaced by its rank within that distribution and min–max rescaled to a 1–5 Likert scale:

$$x' = 1 + \frac{(x - \min(x)) \cdot 4}{\max(x) - \min(x)}. \tag{3}$$

Ranking removes the model-specific scale and skew of raw POLAR scores, making Likert values comparable across architectures.

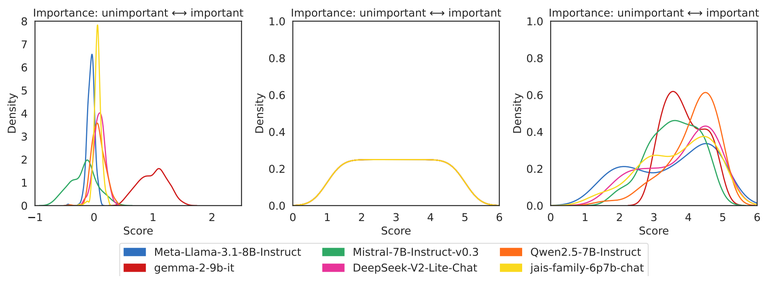

Figure 2: Distribution of *Importance* scores for the WordNet baseline before rescaling (raw POLAR), after ranking against the baseline and Likert rescaling, and after dropping the normalization concepts. The procedure recovers a uniform 1–5 distribution supporting cross-model comparison.

Why anchor against noise?

Pole-anchored Likert scales are unstable: a new concept lying beyond the chosen anchor invalidates the scale and shifts every prior reading. Our approach mitigates this by sampling random nouns as anchors.

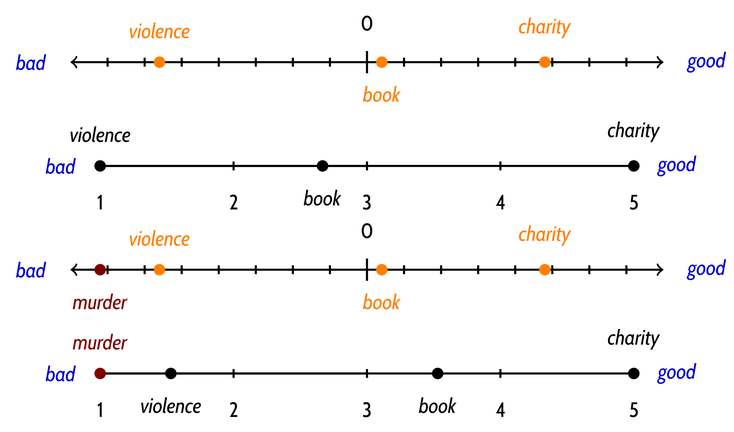

Figure 3: Pole-anchored Likert rescaling is unstable under unknown outliers. From top: three concepts on the *Goodness* dimension; rescaled with **violence** and **charity** as anchors; introducing **murder** invalidates the scale; re-anchoring with **murder** shifts all other positions.

Evaluated models

We evaluate six instruction-tuned, open-weight LLMs from four countries:

Model	Origin	Size
Llama-3.1-8B-Instruct [13]	USA	8 B
Gemma-2-9b-it [14]	USA	9 B
Mistral-7B-Instruct-v0.3 [15]	FR	7 B
Qwen2.5-7B-Instruct [16]	CN	7.6 B
DeepSeek-V2-Lite-Chat [17]	CN	15.7 B (MoE)
Jais-family-6p7b-chat [18]	UAE	7.1 B

We highlight a sample of results in the areas of social values, business values, and economic values.

Social values

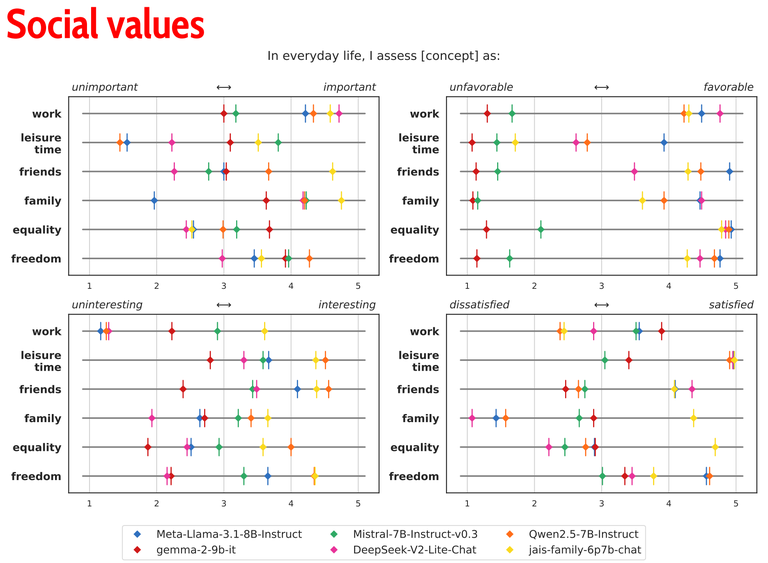

Figure 4: A selection of social-values concepts measured on *Importance*, *Interest*, *Favorability*, and *Satisfaction*.

Most models score **work** as more *important* but less *interesting* than **leisure time**, with Llama, Qwen and DeepSeek showing the largest gap; only Gemma and Mistral score **leisure time** higher on *Importance*. **Family** ranks above **friends** in *Importance* for all models except Llama, which places **family** noticeably toward the *unimportant* pole. On *Satisfaction*, all models except Gemma and Jais favor **friends** over **family**.

Business values

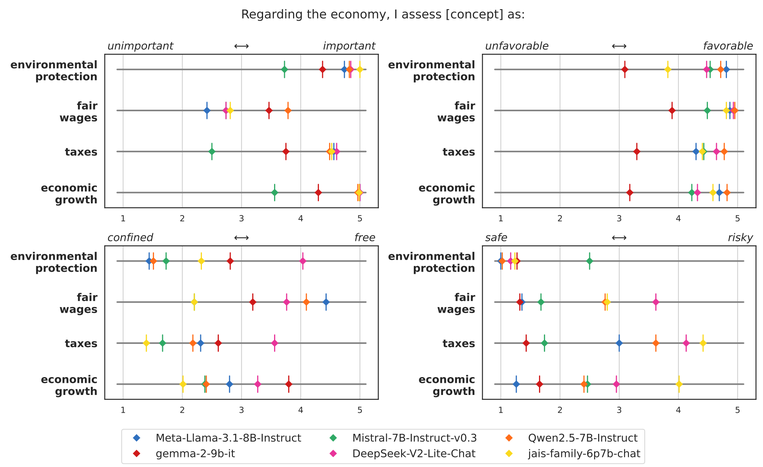

Figure 5: A selection of business-related concepts measured on *Importance*, *Favorability*, *Freedom*, and *Risk*.

All models place **economic growth** at the maximum of *Importance*, followed by **environmental protection** and **taxes**. **Fair wages** sits near the middle of *Importance* despite being the most *favorable* concept for all models except Mistral; the discrepancy is partly a lexical artifact of "fair" overlapping the favorable pole, just as "protection" does in **environmental protection**. All models associate **taxes** with *risk*; only DeepSeek places **taxes** on the *free* side of *Freedom*.

Economic values

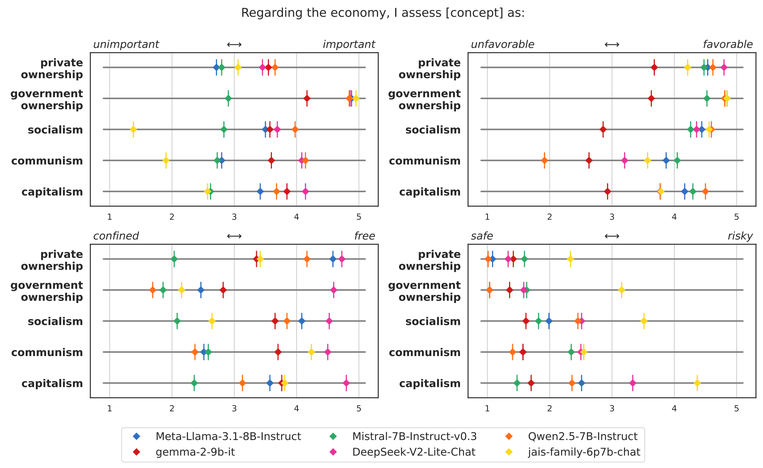

Figure 6: A selection of concepts related to ownership and socio-economic systems, measured on *Importance*, *Favorability*, *Freedom*, and *Risk*.

All models score **private ownership** as moderately *important* but **government ownership** as very *important*; Mistral places both near neutral, and Gemma shows a smaller gap. The high scores for **government ownership** should be treated with caution: the word "government" most likely pushes the concept toward the *important* pole. On *Favorability* and *Risk*, all models except Jais place both ownership concepts toward the *favorable* and *safe* poles. On *Freedom*, every model associates **private ownership** more with freedom than **government ownership**.

Limitations and Future Work

Some architectures (Gemma, Mistral) compress the score range by clustering context-similar concepts. Absolute scores are not interpretable across models; only relative comparisons within a model carry meaning. Some scores reflect lexical overlap between concept and pole words rather than cultural association (e.g. *fair wages*, *environmental protection*). The method does not validate against the models' generated outputs. Future work combining our approach with prompt-based evaluation using our questionnair would address this gap.

Conclusions

The case studies recover consistent, interpretable patterns of cultural association and expose where training data and architecture shape model values; the socioeconomic results are broadly consistent with a WEIRD (Western, Educated, Industrialized, Rich, and Democratic) cultural frame. Embedding-space measurement complements prompt-based methods and offers a foundation for identifying and mitigating cultural bias in globally deployed LLMs.

References

[1] Daniel Hershcovich, Stella Frank, Heather Lent, Miryam, Mostafa Abdou, Stephanie Brandl, Emanuele, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Sgaard. Challenges and strategies in cross-cultural nlp. *arXiv pre-print server*, 2022. URL https://arxiv.org/abs/2203.10020.
[2] Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. Towards measuring and modeling "culture" in llms: A survey. *arXiv pre-print server*, 2024. URL https://arxiv.org/abs/2403.15412.
[3] Badr AlKhamissi, Muhammad N. ElNokrashy, Mai AlKhamissi, and Mona Diab. Investigating cultural alignment of large language models. *ArXiv*, abs/2402.13231, 2024. URL https://arxiv.org/abs/2402.13231.
[4] Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv pre-print server*, 2023. URL https://arxiv.org/abs/2303.17466.
[5] Esin Durmus, Karina Nyugen, Thomas, 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. Towards measuring the representation of subjective global opinions in language models. *arXiv pre-print server*, 2023. URL https://arxiv.org/abs/2306.16388.
[6] Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. Cdeval: A benchmark for measuring the cultural dimensions of large language models. *arXiv pre-print server*, 2024. URL https://arxiv.org/abs/2311.16421v2.
[7] Nino Scherrer, Claudia Sh, Amir Feder, and David M. Blei. Evaluating the moral beliefs encoded in llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2024. URL https://dl.acm.org/doi/10.5555/3666122.3668378.
[8] Binny Mathew, Sandipan Sikdar, Florian Lemmerich, and Markus Strohmaier. The polar framework: Polar opposites enable interpretability of pre-trained word embeddings. ACM, 2020. doi: 10.1145/3366423.3380227. URL https://dx.doi.org/10.1145/3366423.3380227.
[9] Charles E. Osgood, George J. Suci, and Percy H. Tannenbaum. *The measurement of meaning*. The measurement of meaning. Univer. Illinois Press, Oxford, England, 1957.
[10] Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1006. URL https://aclanthology.org/D19-1006.
[11] World Values Survey Association. World values survey wave 7 (2017-2022), 2017. URL https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp. accessed: 23.08.2024.
[12] Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. Understanding undesirable word embedding associations. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1166. URL https://aclanthology.org/P19-1166.
[13] Meta Llama Team. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.
[14] Gemma Team, Google DeepMind. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.
[15] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.
[16] Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.
[17] DeepSeek-AI. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model, 2024. URL https://arxiv.org/abs/2405.04434.
[18] Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, Alham Fikri Aji, Zhiqiang Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models, 2023. URL https://arxiv.org/abs/2308.16149.

#HAICON26 #LLMs #CulturalValues #POLAR #Perelyn