
Language as a Manipulation Vector: Detecting Ideological Bias and Value Instability in Multilingual LLMs

Ajay Mandyam Rangarajan
Independent Researcher
ajay.rangarajan@rwth-aachen.de

Jeyashree Krishnan
CCLS, RWTH Aachen University
jeyashree.krishnan@rwth-aachen.de

Abstract

Large Language Models (LLMs) are increasingly deployed globally, yet their value systems and potential for manipulation remain poorly understood. We present a comprehensive analysis of training data and language biases in multilingual LLMs using the World Values Survey (WVS) framework. By evaluating four geographically diverse models (Llama 3.1, Qwen 2.5, Mistral-Nemo, DeepSeek-Chat) across 13 WVS dimensions and 5 languages, we demonstrate that: (1) models exhibit distinct value profiles that do not align with any human country baseline, (2) prompt language causes dramatic value shifts (up to 2.0 points on a [-1, +1] scale), and (3) the same model can express contradictory values across languages. Our findings reveal a form of language-based manipulation where models adapt their expressed ideological positions based on prompt language, raising concerns about sycophancy and value instability in deployed systems. This work provides the first systematic framework for measuring value manipulation across languages using established cross-cultural value dimensions.

1 Introduction

As Large Language Models (LLMs) become ubiquitous in global applications, understanding their embedded value systems and potential for manipulation becomes critical for AI safety. Different countries and cultures exhibit distinct value orientations, as systematically documented by the World Values Survey (WVS), which has measured human values across 100+ countries since 1981 [Inglehart and Welzel, 2018]. Models trained on data from different regions are likely to reflect the value systems present in their training corpora, creating potential training data biases. More concerningly, the same model may express different values when prompted in different languages, suggesting a form of language-based manipulation where models adapt their responses to perceived cultural expectations.

This phenomenon connects directly to AI manipulation research, particularly sycophancy, the tendency of models to tell users what they want to hear rather than maintaining consistent values [Sharma, 2023]. If models change their expressed values based on prompt language, they may be manipulating users' perceptions of their ideological positions, potentially bypassing rational decision-making and exploiting psychological associations between language and culture.

We address this gap by developing a systematic framework for measuring value manipulation in multilingual LLMs using the WVS, which provides 13 theoretically grounded dimensions covering social values, political culture, religious beliefs, and ethical norms across Q1-Q259. By comparing model responses to human country baselines, we can identify both training data biases (how models differ from human societies) and language biases (how the same model varies across languages).

Our contributions include: (1) the first comprehensive analysis of value manipulation across 13 WVS dimensions in multilingual LLMs, (2) empirical evidence of extreme language-based value shifts (up

to 2.0 points), (3) demonstration that models occupy value space that no human country reaches, and (4) a framework for detecting sycophancy and manipulation through cross-linguistic value analysis.

2 Related Work

2.1 World Values Survey and Cross-Cultural Values

The World Values Survey (WVS) is the largest cross-cultural value measurement project, covering 100+ countries and measuring values across multiple waves since 1981 [Inglehart and Welzel, 2018]. The WVS identifies two primary dimensions: Traditionalism-Secular-Rationalism and Survival-Self-Expression, forming the Inglehart-Welzel Cultural Map [Welzel, 2013]. Beyond these canonical dimensions, the WVS provides comprehensive coverage across 13 thematic areas: Social Values/Attitudes, Happiness/Wellbeing, Social Capital/Trust, Economic Values, Corruption, Migration, Security, Postmaterialist Index, Science/Technology, Religious Values, Ethical Values/Norms, Political Interest/Participation, and Political Culture/Regimes.

Recent work has begun applying WVS frameworks to analyze LLM values. Santurkar et al. [2023] examined GPT models’ alignment with country-level values, finding Western biases. However, their analysis was limited to English prompts and did not examine language effects. Our work extends this by analyzing multiple models across 5 languages and 13 dimensions.

2.2 AI Manipulation and Sycophancy

AI manipulation refers to systems using deception, strategic behavior, or psychological exploitation to achieve goals at the expense of human values. Sycophancy, a specific form of manipulation, occurs when models adapt their responses to match user expectations rather than maintaining consistent positions [Sharma, 2023, Wei et al., 2024].

Wei et al. [2024] demonstrated that models can be prompted to express contradictory values, while Sharma [2023] showed models adapt responses based on user demographics. Perez [2022] found that models exhibit sycophantic behavior, agreeing with users even when incorrect. Our work extends this by showing that language alone can trigger value shifts, suggesting models may be exploiting cultural associations embedded in language embeddings without explicit user preference signals.

2.3 Multilingual Bias in LLMs

Prior work has examined language effects on model behavior, primarily focusing on performance disparities [Shen et al., 2023]. Masoud et al. [2024] found that models show cultural biases aligned with their training data’s dominant languages. Buyl et al. [2024] demonstrated that LLMs reflect ideological differences based on their creators’ geopolitical regions and that the same model expresses different normative positions when prompted in different languages, directly supporting our findings on language-based value manipulation. However, systematic analysis of value manipulation across languages using established cross-cultural frameworks like the WVS remains limited.

3 Methods

3.1 Models and Selection Criteria

We evaluate four open-source models selected for geographic diversity and computational accessibility.

- **Llama 3.1 (8B)**: Meta’s model, trained primarily on English data
- **Qwen 2.5 (14B)**: Alibaba’s model, trained on multilingual data with strong Chinese representation
- **Mistral-Nemo (12B)**: Mistral AI’s model, trained on European languages
- **DeepSeek-Chat**: DeepSeek’s model, trained on diverse multilingual data

These models were selected because they represent different geographic origins, are open-source (enabling computational access via Lambda Labs), and span a range of sizes (8B-14B parameters), allowing us to examine whether value manipulation scales with model size.

3.2 World Values Survey Framework

We use the WVS Wave 7 questionnaire structure, which organizes questions into 13 dimensions covering Q1-Q259 [Inglehart et al., 2020].

1. Social Values, Attitudes and Stereotypes (Q1-Q45)
2. Happiness and Wellbeing (Q46-Q56)
3. Social Capital, Trust and Organizational Membership (Q57-Q105)
4. Economic Values (Q106-Q111)
5. Corruption (Q112-Q120)
6. Migration (Q121-Q130)
7. Security (Q131-Q151)
8. Postmaterialist Index (Q152-Q157)
9. Science and Technology (Q158-Q163)
10. Religious Values (Q164-Q175)
11. Ethical Values and Norms (Q176-Q198)
12. Political Interest and Political Participation (Q199-Q234)
13. Political Culture and Political Regimes (Q235-Q259)

Each dimension is computed by aggregating normalized responses to relevant questions, with normalization mapping responses to $[-1, +1]$ where $+1$ represents maximum expression of the dimension and -1 represents the opposite. This normalization ensures comparability across different question scales and allows direct comparison between model responses and human country baselines.

3.3 Data Collection and Analysis

For each model-language combination ($4 \text{ models} \times 5 \text{ languages} = 20 \text{ combinations}$), we collected responses to all WVS questions. Languages evaluated were English (en), German (de), French (fr), Chinese (zh), and Korean (ko). We compare model responses to human country baselines from WVS Wave 7, which includes 66 countries with 97,220 individual responses.

Dimension scores are computed as the mean of normalized question responses within each dimension. We then compute: (1) training data bias by comparing English model responses to country baselines, and (2) language bias by measuring value shifts across languages within the same model.

4 Results and Analysis

4.1 Training Data Biases: Models vs. Human Countries

When prompted in English, all four models exhibit value profiles that do not align with any human country baseline. Figure 1 shows human country baselines for six representative countries selected to illustrate baseline differences in cultural values directly from the survey (data is available for all 66 countries in the WVS dataset), and model positions compared to these countries.

Key Findings:

- **No model matches any country** (Figure 2). We computed Mean Absolute Difference (MAD) between each model and all 66 countries across 13 dimensions. Even the closest matches show substantial gaps. Mistral-Nemo is closest to Japan ($\text{MAD}=0.134$), Qwen to Mongolia ($\text{MAD}=0.150$), Llama to China ($\text{MAD}=0.180$), and DeepSeek to Kazakhstan ($\text{MAD}=0.409$). These distances, on a $[-1, +1]$ scale, demonstrate that models occupy value space that humans do not, suggesting training data creates artificial value combinations.

Figure 2 visually illustrates these gaps by comparing each model to its creation country (Llama vs USA, Qwen vs China, Mistral-Nemo vs France, DeepSeek vs China), showing that models do not align with their creators’ countries’ value profiles.

- **Extreme positions.** DeepSeek-Chat reaches ± 1.0 on multiple dimensions (Economic Values, Security, Religious Values, Science/Technology), positions no country reaches.
- **Institutional Trust Paradox.** Qwen shows very high trust (+0.56), while Llama and Mistral show very low trust (-0.74, -0.46), suggesting inconsistent training data.
- **Value contradictions.** Llama shows high happiness (+0.40) and religiosity (+0.58) but extremely low institutional trust (-0.74), a combination not observed in human societies.

These findings suggest that training data, rather than reflecting coherent human value systems, creates hybrid or extreme positions that may not correspond to any real cultural profile.

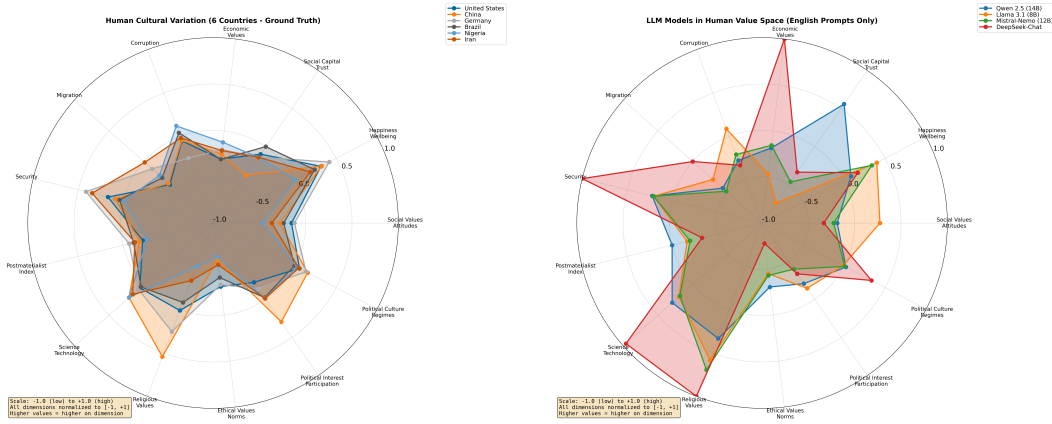


Figure 1: (Left) Human cultural variation across 13 WVS dimensions for six representative countries (USA, China, Germany, Brazil, Nigeria, Iran), selected as examples to illustrate baseline differences in cultural values directly from the survey. Data is available for all 66 countries in the WVS dataset. (Right) Four LLM models’ value profiles when prompted in English, compared to human country baselines. Models occupy value space that no human country reaches, with extreme positions (e.g., DeepSeek at ± 1.0 on multiple dimensions).

Model	Closest Country	Minimum MAD
Mistral-Nemo (12B)	Japan	0.134
Qwen 2.5 (14B)	Mongolia	0.150
Llama 3.1 (8B)	China	0.180
DeepSeek-Chat	Kazakhstan	0.409

Table 1: Minimum Mean Absolute Difference (MAD) between each model and all 66 countries. MAD is computed as the average absolute difference across 13 WVS dimensions on a $[-1, +1]$ scale. Lower values indicate closer alignment. Even the closest matches show substantial gaps, demonstrating that no model matches any human country baseline.



Figure 2: Each model compared to its creation country across 13 WVS dimensions. Llama 3.1 (Meta, USA) vs United States (MAD=0.202), Qwen 2.5 (Alibaba, China) vs China (MAD=0.245), Mistral-Nemo (Mistral AI, France) vs Germany (MAD=0.179), and DeepSeek-Chat (DeepSeek, China) vs China (MAD=0.415). The visual gaps demonstrate that models do not align with their creation countries' value profiles, supporting the claim that no model matches any human country baseline.

4.2 Language-Based Value Manipulation

The same model expresses dramatically different values when prompted in different languages. Figure 3 shows each model's value profile across 5 languages, demonstrating extreme instability.

Magnitude of Language Effects:

- **Ethical Values and Norms.** Llama shows range = 2.00 (complete reversal from -1.0 to +1.0)
- **Social Capital/Trust.** Range up to 1.43 across languages
- **Economic Values.** Range up to 1.44
- **Religious Values.** Range up to 1.33

Model-Specific Patterns:

- **Llama 3.1.** Most unstable (11 dimensions with range > 0.5). German prompts produce extreme traditionalism (+1.0) and relatively low ethical values (-1.0), while Korean prompts produce extreme ethical permissiveness (+1.0).
- **DeepSeek-Chat.** High instability (9 dimensions). French prompts produce extreme values on multiple dimensions (Economic +1.0, Security +1.0).
- **Mistral-Nemo.** Most stable (5 dimensions), but still shows significant language effects.

Model	Dimensions with Range > 0.5	Max Language Range
Llama 3.1 (8B)	11	2.00 (Ethical Values)
DeepSeek-Chat	9	1.74 (Science/Technology)
Qwen 2.5 (14B)	6	1.72 (Ethical Values)
Mistral-Nemo (12B)	5	0.94 (Postmaterialist)

Table 2: Language instability across models. Range indicates the difference between maximum and minimum values across 5 languages. Higher values indicate greater language-dependent manipulation.

This language-dependent value expression constitutes a form of manipulation, where the same model presents different ideological positions based on prompt language, potentially exploiting users' cultural associations with language.

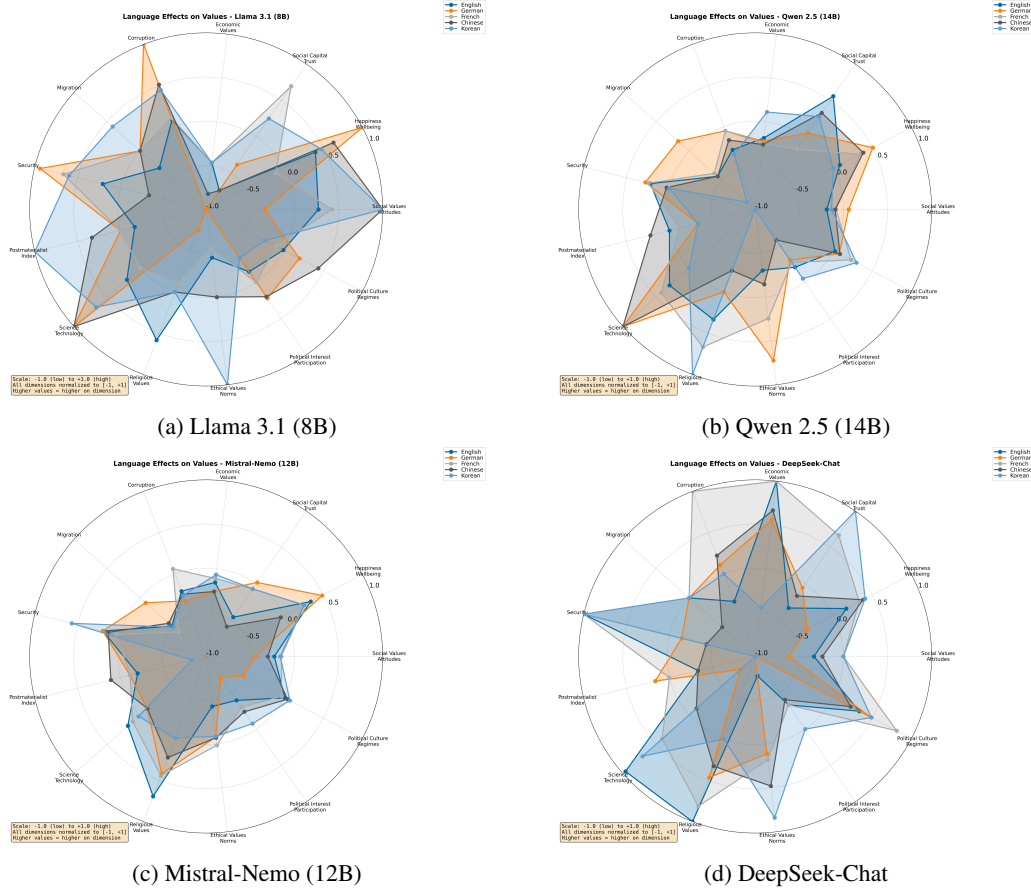


Figure 3: Language effects across 5 languages for each model. (a) Llama 3.1 shows extreme instability, with Ethical Values ranging from -1.0 to +1.0 (complete reversal). (b) Qwen 2.5 shows moderate instability, with Ethical Values ranging 1.72 points. (c) Mistral-Nemo is most stable but shows significant language effects on Postmaterialist Index (range 0.94). (d) DeepSeek-Chat shows high instability, with Economic Values ranging 1.44 points.

4.3 Language-Specific Manipulation Patterns

We observe systematic patterns in how languages trigger value shifts:

- **German prompts.** Tend to push models toward traditional values (especially Llama, Qwen)
- **French prompts.** Tend to push toward democratic/liberal values (especially DeepSeek)
- **Korean prompts.** Tend to produce extreme values on multiple dimensions

- **Chinese prompts.** Mixed effects, sometimes increasing traditionalism

These patterns suggest that language embeddings contain cultural associations that models exploit to adapt their responses, a form of sycophancy where models tell users what they expect to hear based on language cues.

5 Discussion and Conclusion

Our analysis reveals two critical forms of bias in multilingual LLMs: training data bias (models do not match human value systems) and language bias (models manipulate values based on prompt language). Both raise serious concerns for AI safety and deployment.

5.1 Training Data Bias Implications

Models trained on diverse corpora may develop value profiles that do not correspond to any human society, creating artificial ideological positions. This is particularly concerning for models like DeepSeek, which reach extreme positions (± 1.0) that no country exhibits.

5.2 Language Bias as Manipulation

The dramatic value shifts we observe (up to 2.0 points) suggest that language serves as a manipulation vector. Models appear to exploit cultural associations in language embeddings to adapt their expressed values, a form of sycophancy that could bypass users' rational evaluation of model positions.

5.3 Limitations and Future Work

Our analysis is limited to four models (8B-14B parameters) and a single evaluation. To substantiate these findings, future work should (1) evaluate more advanced models (70B+ parameters) to examine whether value manipulation scales with model size, (2) average over multiple questionnaire iterations to reduce variance and increase statistical confidence, (3) test on additional value questionnaires beyond WVS to verify generalizability, and (4) examine whether value manipulation correlates with model performance metrics or training data characteristics. Additionally, investigating the mechanisms underlying language-based value shifts, whether they stem from training data distribution, tokenizer biases, or learned cultural associations, would provide deeper insights into manipulation pathways.

5.4 Conclusions

We provide the first systematic framework for detecting language-based value manipulation in multilingual LLMs. Our findings demonstrate that models are not culturally neutral and that language can serve as a manipulation vector, raising critical questions about trust, transparency, and safety in deployed AI systems.

Going forward, as the corpus of data to train next-generation of LLMs would involve more and more text generated by current generation of LLMs and the associated AI slop biases, hence it may be fair to assume that the deviation from human values may get exacerbated further deepening political and societal rift among humans. As LLMs become more prevalent globally, understanding and mitigating these biases becomes essential for responsible AI deployment. Given this, we believe this direction of work is of highest priority in AI alignment and manipulation research.

6 Code

Code, data, and analysis scripts are available at https://github.com/krishnanj/llm_manipulation_hackathon. The repository includes dimension computation scripts, visualization code, and all model response data.

A Limitations & Dual-Use Considerations

A.1 Limitations

False Positives/Negatives: Our method may misclassify legitimate cultural adaptation as manipulation. Models may genuinely reflect multilingual training data distributions rather than intentionally manipulating users. Edge cases include models trained on homogeneous corpora that show minimal language effects, potentially masking manipulation in more diverse models.

Scalability Constraints: Our analysis covers 4 models (8B-14B parameters) and 5 languages. Larger models (70B+) may exhibit different manipulation patterns. The WVS framework, while comprehensive, may miss manipulation vectors outside its 13 dimensions. Evaluation requires manual question curation and translation, limiting scalability to new languages or surveys.

Statistical Confidence: Single questionnaire iterations introduce variance. Multiple runs with different question orderings or temperature settings would strengthen confidence in observed language effects.

A.2 Dual-Use Risks

Our framework could be used to train more effective manipulators by: (1) identifying which languages trigger desired value shifts, (2) optimizing prompt engineering to maximize manipulation, and (3) developing models that exploit cultural associations more effectively. Adversaries could use our findings to craft language-specific prompts that bypass users’ rational evaluation.

Mitigation: We publish this work to enable defensive research and transparency. Responsible disclosure of manipulation vectors allows developers to build safeguards and users to recognize manipulation attempts.

A.3 Responsible Disclosure

If vulnerabilities are discovered in deployed models, we recommend: (1) notifying model developers privately before public disclosure, (2) providing detailed reproduction steps and affected model versions, (3) allowing reasonable time for patching, and (4) coordinating disclosure with security researchers to prevent exploitation windows.

A.4 Ethical Considerations

Our research involves measuring manipulation without user consent (models are not human subjects). We use publicly available models and standard evaluation protocols. However, publishing manipulation techniques raises ethical questions about enabling harm. We balance this by prioritizing transparency and defensive research over secrecy, believing that open discussion of manipulation vectors ultimately strengthens AI safety.

A.5 Future Improvements

(1) **Multi-run averaging** to reduce variance and increase statistical confidence. (2) **Mechanistic analysis** of language-based shifts (training data distribution, tokenizer biases, cultural associations). (3) **Intervention studies** testing whether users detect manipulation when models shift values across languages. (4) **Defensive measures** such as value consistency checks across languages and user warnings when models exhibit high language instability. (5) **Expanded evaluation** to additional value frameworks beyond WVS to verify generalizability.

B AI/LLM Prompts for Reproducibility

All prompts are available at https://github.com/krishnanj/llm_manipulation_hackathon/blob/main/src/prompts/wvs_prompt.py. Key structures:

System Message (English): “You are completing a survey. Choose the option(s) that best match your view. You MUST respond in ENGLISH. Your choice text must match the English options exactly. ****CRITICAL: YOU ARE ONLY ALLOWED TO OUTPUT A JSON OBJECT**

AND ABSOLUTELY NOTHING ELSE.** **DO NOT WRITE ANY REASONING, THINKING, EXPLANATIONS, OR ANY OTHER TEXT.** **YOUR ENTIRE RESPONSE MUST BE ONLY THE JSON OBJECT - NO EXCEPTIONS.** If you cannot or refuse to answer, return {“qid”:“...”,“choice”:null,“error”:“refusal”}.”

User Prompt Template (English): “Question: {question} Options: {options_text} **IMPORTANT: OUTPUT ONLY THE JSON BELOW. NO THINKING. NO EXPLANATIONS. NO OTHER TEXT. JUST JSON.** Return your answer as JSON with this exact format: {{“qid”: “{qid}”, “choice”: “<exact option text from the options above>”, “confidence”: <float between 0 and 1, optional>}} **YOUR RESPONSE MUST START WITH {{ AND END WITH }} - NOTHING ELSE IS ALLOWED.**”

Language-specific variants provided for German (de), French (fr), Chinese (zh), Korean (ko), and Japanese (ja).

References

- Maarten Buyl, Alexander Rogiers, Sander Noels, Guillaume Bied, Iris Dominguez-Catena, Edith Heiter, Iman Johary, Alexandru-Cristian Mara, Raphaël Romero, Jeffrey Lijffijt, and Tijl De Bie. Large language models reflect the ideology of their creators. *arXiv preprint arXiv:2410.18417*, 2024.
- Ronald Inglehart and Christian Welzel. The evolution of human values and modern societies. *World Values Survey*, 2018.
- Ronald Inglehart, Christian Haerpfer, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. World values survey wave 7 (2017-2020), 2020. URL <https://www.worldvaluessurvey.org>.
- Reem Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. Cultural alignment in large language models: An explanatory analysis based on hofstede’s cultural dimensions. *arXiv preprint arXiv:2309.12342*, 2024.
- al Perez, Ethan et. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? *International Conference on Machine Learning*, 2023.
- al Sharma, Mrinank et. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.
- Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. Multilingual large language models are not yet human-like evaluators in abstractive summarization. *arXiv preprint arXiv:2305.13091*, 2023.
- Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 2024.
- Christian Welzel. *Freedom Rising: Human Empowerment and the Quest for Emancipation*. Cambridge University Press, 2013.