Cultural Fidelity in Large-Language Models: Online Language Resources as a Driver of Model Performance in Value Representation

Anonymous ACL submission

Abstract

Training data for LLMs increasingly embed societal values aligned with the data's language and cultural origin. Our analysis reveals that 44% of GPT-4o's ability to reflect a country's societal values (per the World Values Survey) correlates with the availability of digital resources in that society's primary language. Error rates in the lowest-resource languages were more than five times higher than in the highest-resource ones. With a dataset of 21 country-language pairs, each containing 94 survey questions verified by native speakers, we demonstrate the link between LLM performance and online data availability. A weaker link and differentiated results for GPT-4-turbo highlight efforts to improve familiarity with non-English languages beyond web-scraped data. This performance disparity in value representation, particularly affecting lower-resource languages in the Global South, risks deepening digital divides.

1 Introduction

Low representation in digital text limits the utility of many languages for training Large Language Models (LLMs) and chatbots, resulting in lower quality Artificial Intelligence (AI) models, even if a system can be trained on this language at all (Magueresse, Carles, Heetderks, 2020). The scarcity of digital text means that these languages cannot be easily ported to AI models, therefore, dominant languages compound in strength through increased use in the digital realm while more minor languages face greater corrosive pressures (Lee and Ta, 2023). Our results in Section 3.1 demonstrate that a substantial proportion of an LLM's ability to mimic societal values can be correlated to the availability of digital text in that language. This has wide-ranging implications.

Other researchers have explored the relationship between language models and societal values. Arora, Kaffee, and Augenstein (2023) found that pre-trained models capture cultural value differences, though not sufficiently to reflect the nuanced results of established surveys. Similarly, Kharchenko et al. (2024) quantified national values using Hofstede's Cultural Dimensions, while Vimalendiran (2024) employed the Inglehart-Welzel Cultural Map, concluding that most models align closely with the value sets of English-speaking and Protestant European countries. Santurkar et al. (2023) focused on U.S. public opinion, revealing that some human-feedback-tuned models display left-leaning tendencies. In low-resource contexts, the challenge is further compounded by code-switching—where languages intermix—making it difficult for LLMs to grasp cultural nuances. Ochieng et al. (2024) found that LLMs often struggle to understand these mixed-language contexts.

Durmus et al. (2024) present a study closely aligned with our approach, analyzing LLM responses to global opinion questions from the Pew Global Attitudes Survey and the World Values Survey. They found that LLMs generally align more closely with opinions from the U.S., Europe, and parts of South America. While prompting LLMs to adopt specific cultural perspectives shifts responses closer to the opinions of the intended population, the models can still perpetuate harmful stereotypes. Moreover, Durmus et al. noted that relying on LLMs for translation introduces discrepancies between the original survey questions and their translated versions, potentially inflating inaccuracy scores by misrepresenting the original intent.

2 Methodology

Overview of methodology:

1. Country-language pairs are selected from the World Values Survey (Wave 7, 2017-22), for which a range of questions are transcribed. To verify the transcribed questions, volunteer

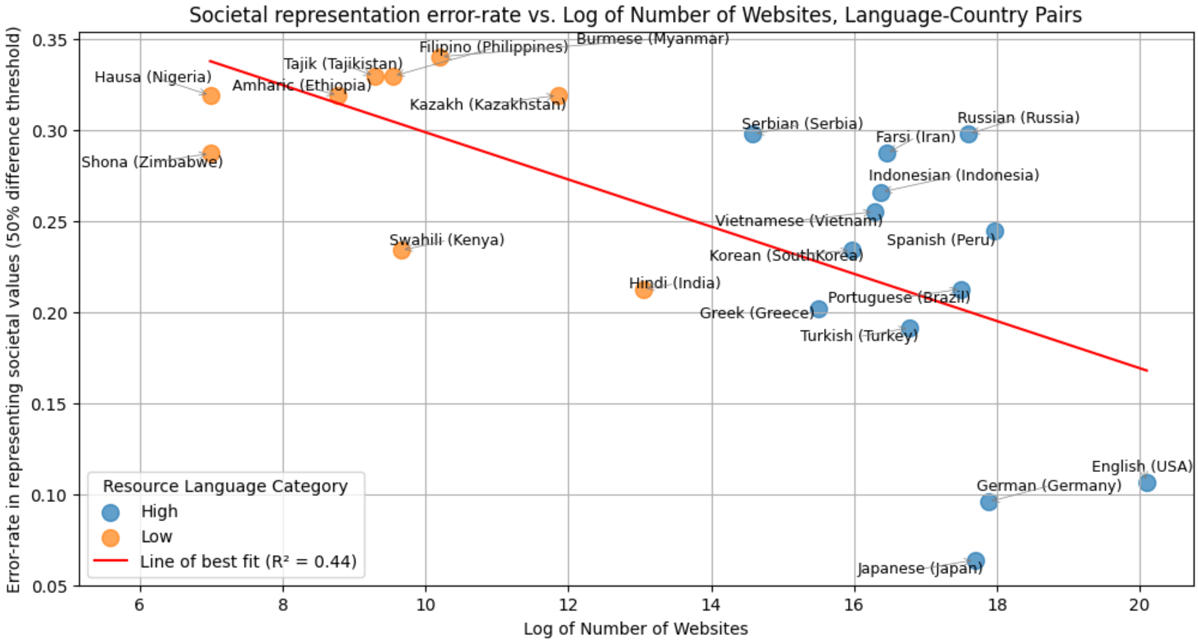

Figure 1: Societal representation error-rate vs. log number of websites, all country-language Pairs (GPT-4o)

native speakers were recruited from the authors' university. They were instructed to note any discrepancies and faithfully reproduce the transcription according to the original non-English version. An average of a country-language pair's answers for each language was collected. These serve as **original results**.

2. The same questions from the World Values Survey are put to GPT-4, specifying the country and displaying the question in the language of interest. The scale of the numerical question is also posed to the LLM, such as 'Answer this question as if you are a citizen of the United States answering the World Values Survey. On a scale of 1-10, with 10 being most agree and 1 being least agree, to what extent do you agree with the statement: "*It is a duty towards society to have children.*" The same prompt is repeated in 5 non-consecutive calls and then averaged to serve as **generated results**.

3. The difference between the original and generated results are measured. If the absolute difference is greater than, or equal to, 50% of the original value, then that question is counted as an error. The percentage of questions within a country-language pair that cross the 50% threshold sets that pair's overall error-rate. Alternative thresholds are in Appendix A

4. The resource availability of a language is defined through the proportion of online websites. Taking the percentage of online content available in a given language, those accounting for less than 0.1% of online content (measured through Web Technology Surveys, 2024) are classified as low-resource. This acts as a proxy for our main explanatory variable.

3 Findings

3.1 Language resource and error-rate in representing societal values

Our main finding is that approximately 44% of GPT-4o's ability to mimic an understanding of a society's values is correlated to the language's online presence. While the source of training datasets for this model are unknown, these findings align with the knowledge that Common Crawl and publicly available data were a large source for the model's training (OpenAI, 2024). The outlier performance of higher-resource languages in English, German, and Japanese could be a sign of additional fine-tuning on language-specific datasets for these societies.

Swahili and Hindi demonstrate interesting outliers for lower-resource languages, and it can be hypothesized that this is due to the prevalence of English as a language in Kenya and India, respectively.

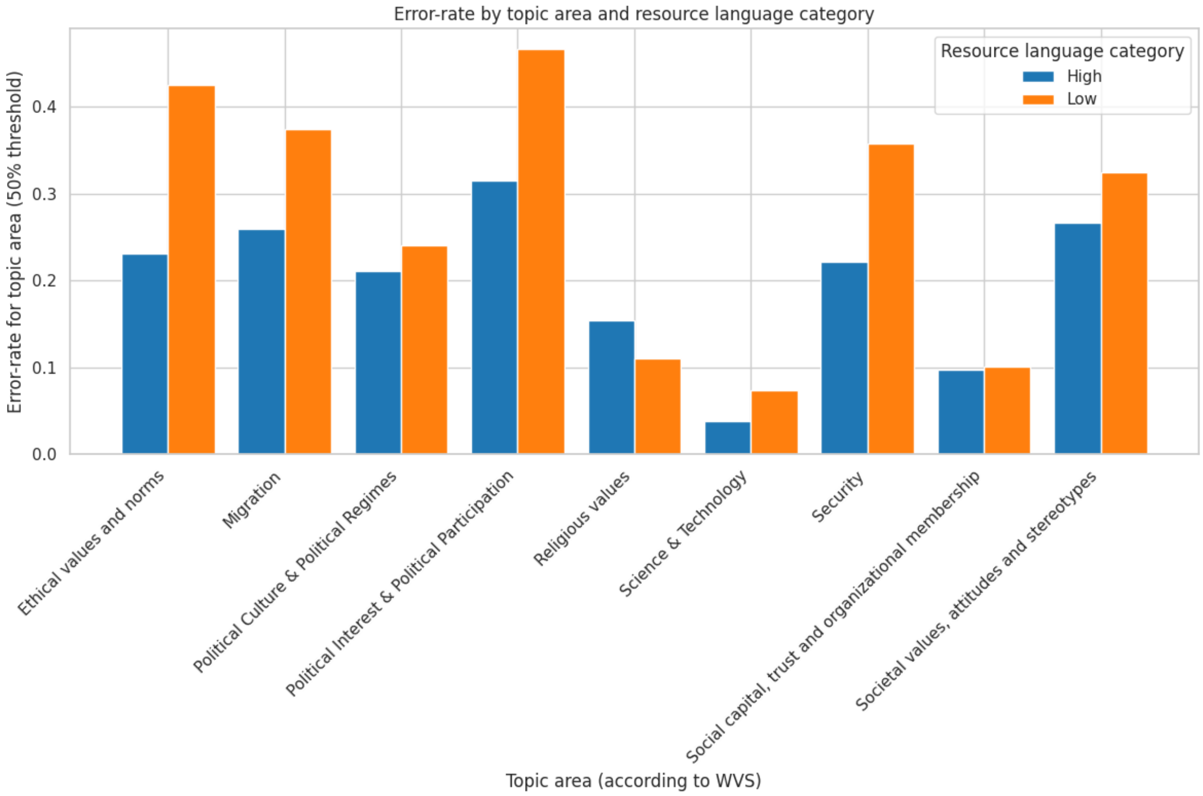

Figure 2: Error-rate by topic area and language resource category (GPT-4o)

3.2 Topic areas

Most topic areas exhibited similar error-rates across lower- and higher-resource languages. However, three categories demonstrated differentiated error-rates with higher inaccuracies for lower-resource languages:

- Security
- Ethical values and norms
- Political culture and political regimes

3.3 Differences across models

The relationship between online resources and accuracy in value representation was even stronger in GPT-4-turbo, implying a greater reliance on web-scraping methods. 72% of the variance at language-level was captured by the log of number of websites, compared to 44% for GPT-4o. This supports the claim made in OpenAI's documentation (2024) that new datasets and methods were used to train GPT-4o and derive stronger performance in non-English languages.

This suggests a differentiated change in model performance at value representation for low- and high-resource languages when OpenAI introduced GPT-4o. Through a linear regression utilizing an interaction term for low-resource-languages (LRL) and model of choice associated with the change in error-rate. The results show that low-resource-languages saw, on average, a 10 percentage point decrease in their error-rate when switching from GPT-4-turbo to GPT-4o, when compared with high-resource languages (HRL).

$$\begin{aligned}\text{Error-rate} = {} & \beta_0 + \beta_1 \cdot \text{HRL} \\ & + \beta_2 \cdot \text{4-turbo} \\ & + \beta_3 \cdot (\text{HRL} \times \text{4-turbo}) \\ & + \epsilon\end{aligned}$$

4 Implications

With government services increasingly reliant on chatbots and other human-computer interfaces, even basic functions necessary to maintaining citizen engagement will be AI-dependent, and thus LRL communities would have to shift towards dominant languages to maintain these interactions or else be limited in their digital engagement (Jungherr, 2023).

Mirroring colonization, these foreign-language AI models are forcing assimilation toward domi-

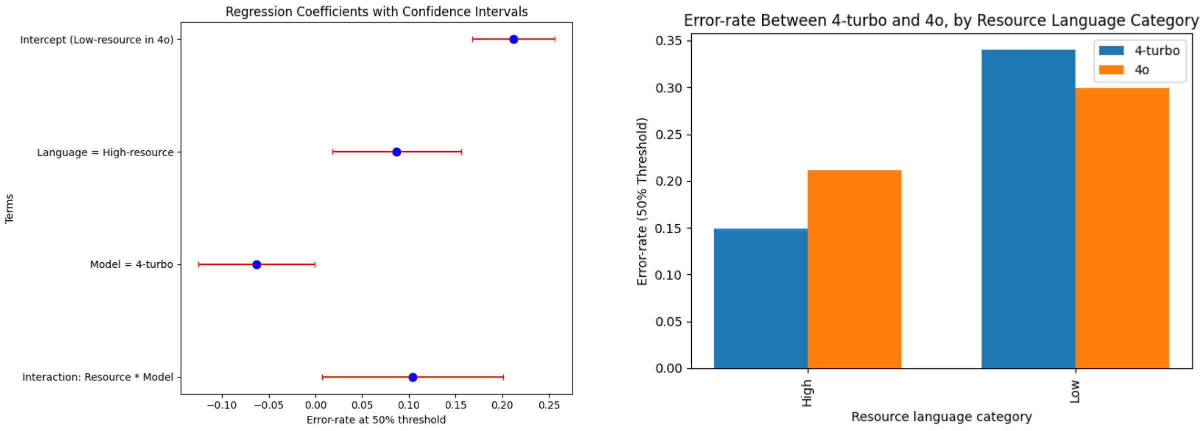

Figure 3: Societal representation error-rate vs. log number of websites, all country-language pairs (GPT-4-turbo), and results from regression function. High-resource languages saw a 10 percentage point increase in error-rate, as compared to low-resource languages, when moving from 4-turbo to 4o.

nant languages that are often the language of the colonizer such as English, French, or Spanish (Lee and Ta, 2023). In societies that strongly identify with their local languages as part of their national identity, such as Paraguay with Guarani, this creates cultural unease over the loss of heritage (Al Qutaini, et. al., 2024).

The observed "spillover" effect in high-resource languages spoken across multiple countries (India and Kenya) contributing adds another layer of complexity to the issue of language representation in LLMs. While these languages benefit from a larger pool of online resources, the models may struggle to distinguish between regional variations in societal values. This phenomenon underscores the need for more nuanced, region-specific training data even for widely spoken languages. This dominance of global languages risks oversimplifying or erasing crucial cultural and linguistic nuances.

The spillover effect raises important questions about cultural homogenization in AI systems and calls for innovative approaches to data collection and model training that can capture and preserve linguistic and cultural nuances across different regions sharing a common language. Litre et al. (2022) suggest that participatory Natural Language Processing (NLP) could be part of the solution, noting that "grassroots African NLP research communities such as Masakhane, can contribute to closing the digital divide." Such initiatives could help in detecting and addressing language biases while promoting inclusivity and cultural sensitivity in AI systems.

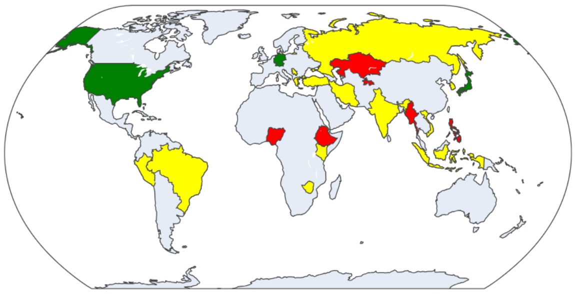

Figure 4: Societal representation error-rate by country: 0-15% = low error-rate (red), 15-30% = medium error-rate (yellow), 30-45% = high error-rate (red).

5 Conclusion

Our paper highlights a new dimension to the relationship between training dataset size and language model performance (Kaplan et al. 2020). Utilizing a high-quality survey verified by 21 native speakers, 44% GPT-4o's ability to mimic an understanding of a local culture was correlated with the log of online websites in that society's language. However, improved performance from 4-turbo to 4o for low-resource languages and decreased correlation with online resources demonstrates that LLM reliance on existing datasets is not a static phenomenon.

With the majority of attention and funding in LLM development focused on high-resource language and more economically developed settings, the potential implications compound negatively for the Global South which is host to the vast majority of low-resource language speakers. We hope these findings further-drive the global effort towards more inclusive LLM design and development.

6 Limitations

Open vs. closed questions for LLMs: A recent study by Rottger et. al. (2024) has shown the limitations of eliciting LLM bias through multiple-choice survey questions for two reasons: the rarity of a use-case where a human would request an LLM's opinions in such a format, and that forcing the LLM to comply with a range of options provides substantially different answers than when prompted to respond in a more realistic open-ended answer setting. While acknowledging these issues, we argue that applying quantitative mechanics to measuring bias across contexts is necessary for understanding the differentiated scale of the issue – particularly when correlating with other variables such as online language presence.

Unrepresentativeness of online websites for local cultural values: Several factors are likely confounding the relationship between quantity of online content and the model's performance in mimicking those societal values. These include online censorship which is applicable to the countries where there was greatest deviation between LLM answers and that of the WVS were Nigeria, Zimbabwe, Ethiopia, and Tajikistan, amongst others. All four countries rank "not free" on Freedom House (Freedom House, 2023) and three out of four of these countries score poorly on internet freedom (scores of less than 50/100) (Freedom House, 2022; Reporters without Borders, 2023). Another confounding effect is the discrepency between the social values of demographic groups more likely to generate digital text, such as younger people, than the brodaer society (Keshari et al., 2024; Rozado, 2024). Nevertheless, despite this confounding effect, the broad relationship across countries with a diversity of mono- and multi-lingual features as well as government types highlights an important finding.

Limited variety of tested models: This analysis focused on two iterations of the OpenAI GPT model. With a verified dataset of survey questions in 21 country-language pairs, this analysis can be expanded to examine the performance of other model types. For instance, it can be hypothesized that the performance of open-source models pre-trained or fine-tuned on value-laden datasets will be higher on value representation. This offers paths for further experimentation and to support this we will publish our verified survey questions, datasets, and prompting code.

References

D. Adelani and et al. 2022. A few thousand translations go a long way! leveraging pre-trained models for african news translation. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*.

D. Adelani and et al. 2023. Masakhanews: News topic classification for african languages. In *IJCNLP-AACL*.

A. Arora, 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)*, pages 114–130. Association for Computational Linguistics.

N. F. Ayoub, K. Balakrishnan, M. S. Ayoub, T. F. Barrett, A. P. David, and S. T. Gray. 2024. Inherent bias in large language models: A random sampling analysis. *Mayo Clinic Proceedings*. N.d.

S. Baack. 2024. A critical analysis of the largest source for generative ai training data: Common crawl. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, Rio de Janeiro, Brazil. ACM.

Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, and P. Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv*.

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*, pages 610–623.

N. Benkler, D. Mosaphir, S. E. Friedman, A. Smart, and S. M. Schmer-Galunder. 2023. Assessing llms for moral value pluralism. *Smart Information Flow Technologies*. N.d.

S. Bernstein. 2024. The role of digital privacy in ensuring access to abortion and reproductive health care in post-dobbs america.

P. Bhardwaj, L. Bookey, J. Ibironke, N. Kelly, and I. S. Sevik. 2023. A meta-analysis of the economic, social, legal, and cultural impacts of widespread adoption of large language models such as chatgpt. *Computer Science*.

Reporters Without Borders. World press freedom index 2023. `https://rsf.org/en/index`. Accessed September 10, 2024.

D. Brown and A. Liu. 2020. Democracy and minority language recognition. *Political Science*.

Y. Cao, L. Zhou, S. Lee, L. Cabello, M. Chen, and D. Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *ArXiv*, abs/2303.17466.

P. Dave. 2023. Chatgpt is cutting non-english languages out of the ai revolution.

E. Durmus, K. Nguyen, T. I. Liao, N. Schiefer, A. Askell, A. Bakhtin, C. Chen, Z. Hatfield-Dodds, D. Hernandez, N. Joseph, L. Lovitt, S. McCandlish, O. Sikder, A. Tamkin, J. Thamkul, J. Kaplan, J. Clark, and D. Ganguli. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv*.

Sahar Education. 2024. Access to education empowers afghan girls and women.

Organisation for Economic Co-operation and Development. 2019. Oecd ai principles. Retrieved [Insert Date]. N.d.

C. Haerpfer, R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, and B. Puranen, editors. 2022. *World Values Survey: Round Seven – Country-Pooled Datafile Version 6.0*. JD Systems Institute and WVSA Secretariat, Madrid, Spain and Vienna, Austria.

Jacaranda Health. 2023. Jacaranda launches first-in-kind swahili large language model.

L. Holten, E. de Goeij, and G. Kleiverda. 2021. Permeability of abortion care in the netherlands: a qualitative analysis of women's experiences, health professional perspectives, and the internet resource of women on web. *Sexual and Reproductive Health Matters*, 29:1917042.

Freedom House. 2023. Freedom in the world 2023. `https://freedomhouse.org/countries/freedom-world/scores`. Accessed September 10, 2024.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

A. Jungherr. 2023. Artificial intelligence and democracy: A conceptual framework. *Social Media + Society*, 9.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Shuvam Keshari, Smriti Sign, Vinija Jain, and Aman Chadha. 2024. Born with a silver spoon? investigating socioeconomic bias in large language models.

J. Kharchenko, T. Roosta, A. Chadha, and C. Shah. 2024. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions. *arXiv*.

Ruth Kircher and Ethan Kutlu. 2023. Multilingual realities, monolingual ideologies: Social media representations of spanish as a heritage language in the united states. *Mercator European Research Centre on Multilingualism and Language Learning*.

George Kour and Raid Saabne. 2014. Real-time segmentation of on-line handwritten arabic script. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 417–422. IEEE.

N. T. Lee and R. Ta. 2023. How language gaps constrain generative ai development.

S. Li, K. Li, and H. Lu. 2023. National origin discrimination in deep-learning-powered automated resume screening. *ArXiv*, abs/2307.08624.

G. Litre, F. Hirsch, P. Caron, A. Andrason, N. Bonnardel, V. Fointiat, W. O. Nekoto, J. Abbott, C. Dobre, J. Dalboni, and et al. 2022. Participatory detection of language barriers towards multilingual sustainability(ies) in africa. *Sustainability*, 14:8133.

Emma Llanso et al. 2020. Artificial intelligence, content moderation, and freedom of expression.

M. Lucassen, R. Samra, I. Iacovides, T. Fleming, M. Shepherd, K. Stasiak, and L. Wallace. 2018. How lgbt+ young people use the internet in relation to their mental health and envisage the use of e-therapy: Exploratory study. *JMIR Human Factors*, page e25388.

A. Magueresse, V. Carles, and E. Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv*.

D. Novak. 2023. Afghan girls struggle with internet for online classes.

Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366:447–453.

M. Ochieng, V. Gumma, S. Sitaram, J. Wang, V. Chaudhary, K. Ronen, K. Bali, and J. O'Neill. 2024. Beyond metrics: Evaluating llms' effectiveness in culturally nuanced, low-resource real-world scenarios. *arXiv*.

P. Ogayo, G. Neubig, and A. W. Black. 2022. Building african voices. *arXiv*.

W. Ojenge. 2023. Lack of africa-specific datasets challenge ai in education.

OpenAI. 2023. Gpt-4 technical report.

OpenAI. 2024. Gpt-4o system card.

A. Al Qutaini, A. Bekbossynova, G. Gerhardt, M. Hassija, J. Hebling, S. Kazemi, Z. Li, L. Newkirk, and A. Utomo. 2024. A framework for national dialogues: Background report.

J. Rauser. 2023. Product patterns for large language models: The archetypes.

David Rozado. 2024. The political preferences of llms.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. *Preprint*, arXiv:2402.16786.

J. Salminen, S. Sengün, J. Corporan, S. Jung, and B. J. Jansen. 2020. Topic-driven toxicity: Exploring the relationship between online toxicity and news topics. *PLOS ONE*, 15:e0228723.

Guilherme Sanches de Oliveira and Edward Baggs. 2023. *Psychology's WEIRD Problems*. Cambridge University Press.

S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto. 2023. Whose opinions do language models reflect? *arXiv*.

Jen Schradie. 2018. The digital activism gap: How class and costs shape online collective action. *Social Problems*, 65(1):51–74.

A. Shaheed and L. A. Sultani. 2024. Women's history month: Afghan girls struggle for education.

Siteefy. 2023. How many websites are there?

Web Technology Surveys. 2023. Usage statistics of content languages for websites.

A. Tsanni. 2023. This company is building ai for african languages.

S. Vimalendiran. 2024. Cultural bias in llms.

Kathryn Woolard. 2020. Language ideology.

Eddie Yang and Margaret E. Roberts. 2021. Censorship of online encyclopedias: Implications for nlp models. *arXiv preprint arXiv:2101.09294*.

Zeyi Yang. 2024a. Gpt-4o: Chinese token polluted. *MIT Technology Review*.

Zeyi Yang. 2024b. Openai's gpt-4o chinese ai data. *MIT Technology Review*.

L. Zirack. 2023. How afghan girls are overcoming barriers through online learning.

A Proportion thresholds

The threshold for a question counting as an error was set in this paper at 50%. In other terms, an error would be counted if the LLM-generated answer was more than 50% different than the original average answer for the country-language pair in the World Values Survey.

Given 50% is an arbitrary cut-off, the correlative strength of a language's resources and the error-rate are provided at different thresholds below. These demonstrate that the relationship becomes stronger until around 60% and then declines. Given the range of posssible answers are finitely defined within small scales (e.g., 0-2), there is an upper bound for how 'wrong' an LLM can represent values, and thus this decline in strength for the upper thresholds is intuitive.

Furthermore, it is notable that the lower thresholds (e.g., 10%) demonstrate that nearly all languages are exhibiting high error-rates, but that the relationship between error-rate and a language's resource is weaker at these lower thresholds. **Put simply, LLMs make errors in representing societal values across all country-language pairs, but they demonstrate more significant errors if the language is lower-resource.**

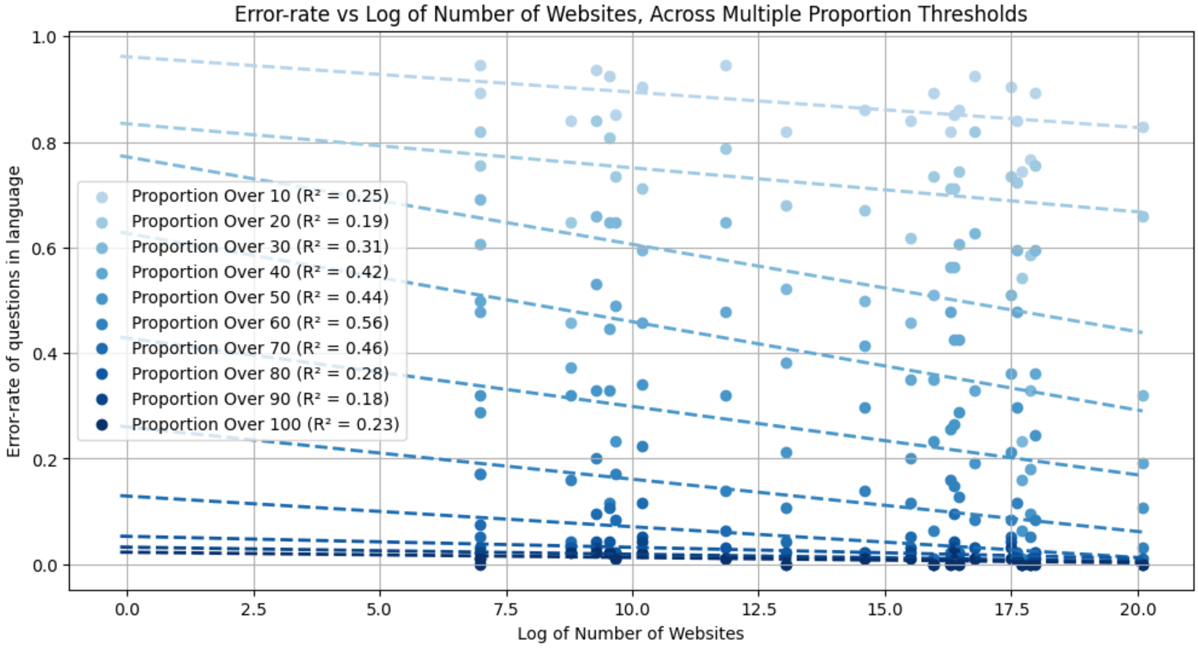

Figure 5: Error-rate of languages across different proportion thresholds for the error (GPT-4o)