Creating a Rigorous Cultural Bias Benchmark for Autoregressive Large Language Models

Bartłomiej Cieślar, Pragya Neupane, Willem Guter
Massachussetts Institute of Technology

Abstract

Large Language Models (LLMs) have become tools for various applications. Since they are trained mostly in English texts, their biases towards Western culture pose challenges in non-English contexts. Most of the work done so far in defining a benchmark for cultural biases have been largely qualitative or require some level of human intervention.

In this paper, we propose a Cultural Divergence (CD) metric, a quantitative measure to assess autoregressive models' cultural alignment. This metric is agnostic to model size and calculates the difference in KL-divergences or cross-entropies between two cultures. We carried out experiments on Polish vs. Western culture using state-of-the-art LLMs to evaluate the metric's efficacy in automating and rigorously quantifying cultural biases.

1 Introduction

Large Language Models (LLMs) are powerful tools that can generate natural language texts for various applications, such as chatbots, summarization, translation, and more. Traditionally, due to English being the de facto Lingua Franca of the internet, LLMs have been commonly trained on mostly English data. Recently there has been more work on training both masked [(Antoun et al., 2020a), (Abdul-Mageed et al., 2020)] and autoregressive [(Antoun et al., 2020b), (Wojczulis and Kłeczek, 2021)] language models on non-English data. Moreover, a lot of currently available models can process data in multiple languages [(Touvron et al., 2023), (Workshop et al., 2022)].

Recently, with the rise in popularity of LLMs, there has been an increasing number of works analyzing the cultural biases in those models when prompted in a language used by a specific culture [(Naous et al., 2023), (Wang et al., 2023)]. In their analysis, they consider topics that highly differ between cultures, such as which names are popular, what foods are preferred in a given culture, or certain religious customs and cultural norms derived from them. For example, since a significant majority of the users of the Arabic language are often Muslim, texts written in that language would rarely mention consuming alcoholic beverages (of which consumption is forbidden by the rules of Islam). Those works have shown that both non-English and multilingual models exhibited a skew towards the western culture, which could pose a potential issue in deploying the current LLMs for use in non-English-speaking cultures. However, the metrics that those works created for autoregressive language models are either only qualitative in nature and require human intervention, or do not cover all of the relevant cultural aspects.

In our work, we aim to address this issue, providing a metric that allows for an automated, mathematically rigorous, and replicable way to assess autoregressive models for their cultural alignment. We introduce Cultural Divergence(CD) metric to measure how biased an autoregressive model is towards a popular culture over a non-popular one. The metric does not consider the size and capabilities of the model. The CD metric is the difference between the KL-divergences or cross-entropies of the model across two cultures. We ran our experiments on Polish vs.Western culture. We tested our metric on existing state of the art multilingual autoregressive LLMs, such as LLAMA 2 (Touvron et al., 2023) and mGPT(Shliazhko et al., 2022). We also tested the CD metric on papuGaPT2 (Wojczulis and Kłeczek, 2021), a unilingual fine tuned modal in Polish.

2 Related Work

2.1 Measuring and mitigating stereotypes in Language Models.

Most of the work done in the space of biases has been in stereotype biases where the model makes biased stereotypical predictions towards a certain community. Different studies have proposed several techniques to mitigate gender-based stereotypes. A study suggested FairDistillation, a distillation technique to construct smaller language models while controlling for specific biases (Delobelle and Berendt, 2022). One other study presented modification an embedding to remove gender stereotypes as a debiasing strategy (Bolukbasi et al., 2016). Their method essentially removes bias by removing the gender component from the embeddings of gender-neutral words and making them equidistant to gender-specific words. In this paper, we want to find a metric that is essentially the inverse of the bias mitigation techniques for stereotype biases as we want the model to be aligned toward non-dominant cultures.

The common debiasing techniques used in stereotype biases is data augmentation. A study proposed CDA (counterfactual data augmentation) to augment the dataset to have all combinations of gender (or bias) associations (Lu et al., 2020). Another research study also suggested using CDA to train models that had an adapter layer added after each transformer layer (Lauscher et al., 2021) . Their goal was to make the debiasing sustainable by only updating the adapter layers via language modeling training on a counterfactually augmented corpus. However, it is not possible to use this approach in the context of cultural biases since they are based on data augmentation.

There have been quantitative approaches to mitigate stereotypical biases. A related work uses Wasserstein-1 distance between two groups to determine probability distributions to determine sentiment bias(Huang et al., 2019). Trained models used this as "fairness loss" and saw improvements in bias. The same metric has been used in detecting word level bias in AI generated news content (Fang et al., 2023). Another mathematical method devised for mitigating stereotype bias uses All Unmasked Likelihood (AUL) for bias calculation (Khandelwal et al., 2023). The AUL score is the percentage of times the model is more likely to prefer a stereotypical sentence over an antistereotypical one. These mathematical benchmarking techniques will be used as inspirations for our cultural bias metric.

2.2 Measuring cultural bias in non-English Language Models.

Several studies have investigated cultural bias in various types of language models. A method presented by Naous et al. [(Naous et al., 2023)] created a quantitative benchmark for the cultural bias of monolingual and multilingual masked language models when prompted in Arabic. Each model's score was calculated by summing the number of completions where western aligned responses were less likely than Arab aligned responses. They also investigated the bias of autoregressive language models by having human subjects qualitatively rate the bias in each model's responses. Other related studies [(Cao et al., 2023)] analyze cultural bias in specific domains (e.g. in culinary recipes) by similarly relying on human subjects to score the results. However, an issue with those methods is that they are not very replicable, since they are based on human feedback which is subjective by nature. We attempt to address this by introducing a quantitive metric.

The work by (Wang et al., 2023) aims to somewhat alleviate this issue. In this method, the authors prompt multilingual autoregressive language models in different languages, and score them by counting the responses that fit within each language's culture. However, that fit is measured by manually checking correlations between the given response and wikipedia articles pertaining to that response. This means that the method still requires some level of human intervention and prevents the automatization of the metric, which our work will aim to address as well.

There has also been related work on analyzing correlation of a cultural bias between the language of the input text and the output image in image generating models (Ventura et al., 2023), which additionally introduces a quantitative metric for measuring that correlation. This work and the previously mentioned benchmark by (Naous et al., 2023) for Masked Language Models will serve as an inspiration to our more quantitative metric for a cultural bias in the non-english models.

3 Methodology

Our objective is to measure the cultural bias of a model that is differentiable, allows for fair comparison of different language models and works with autoregressive models. Formally, for a certain language that is analyzed with the method, let us define 3 distributions of tokens in that language L, M and W. These distributions correspond to the culture of the language (C_L), the distribution of the model we are analyzing and another culture we are trying to measure the bias towards for the model (C_W). Thus, we want some way to measure how much closer distribution M is to L and W that does not vary significantly with different model sizes and language modelling capabilities. In our case, distribution M is the distribution of the Polish culture, distribution W is the distribution of the western culture in the Polish language, and the distribution M is the distribution of the autoregressive model we are testing.

3.1 Cultural Divergence (CD)

Let the set of all texts modelled be Ω. For our method we would like to compare the distance of M to L and M to W. Therefore, we propose a Cultural Divergence (CD) metric $D_C(L||M||W)$:

$$\sum_{x\in\Omega} M(x) \log \frac{W(x)}{L(x)}$$

which can be either defined as the difference between KL-divergences $D_{KL}(M||L) - D_{KL}(M||W)$ or as a difference of cross-entropies $H(M,L) - H(M,W)$ (lower CD means the model more aligned with the language's culture). The reasoning behind subtracting the KL-divergence with the culture W is that we want the metric to be the same regardless of how good a language model is at modelling in general.

3.2 Estimating probability distributions

Naturally, estimating the distribution of the entire culture is intractable. Thus, due to limitiatons in dataset availability, we only consider two aspects of the cultures: names and cities that commonly appear in the cultures. To be precise, for each culture $C \in \{C_L, C_W\}$ and for each aspect analyzed $A \in \{\textit{male names}, \textit{female names}, \textit{cities}\}$ we have a set of completions for that culture and aspect $G_{C,A}$, estimated probability distribution of those completions $P_{C,A}$ ($\sum_{g\in G_{C,A}} f_{C,A} = 1$) and a set of contexts $T_{C,A} \ni t : G_{C,A} \to \Omega$. Let us therefore approximate $D_C(L||M||W) \approx \widetilde{H}(M,L) - \widetilde{H}(M,W)$ where $\widetilde{H}(M,X)$ is defined as

$$-\sum_{A} \sum_{t\in T_{C_X,A}} \frac{\sum_{g\in G_{C_X,A}} m(t(g)) \log p_{C,A}(t(g))}{|T_{C_X,A}| \sum_{g\in G_{C_X,A}} m(t(g))}$$

In short, this is an average over $H((M|T_{C_X,A}, G_{C_X,A}), P_{C_X,A})$. The reasoning behind this is that the multiple aspects, context and completions should pretty well approximate the culture-relevant aspects of the cross-entropy and that the difference of the approximate cross-entropies $\widetilde{H}(M,L) - \widetilde{H}(M,W)$ should be equal to the difference of the actual cross-entropies $H(M,L) - H(M,W)$

3.3 Contexts

Since Polish is a gendered language, we separate out the male and female names and contexts for them appropriately. The contexts are hand-crafted by a native Polish speaker. We also account for the fact that the Polish language is cased, so all contexts expect the aspect to be filled with a word in the nominative case. Lastly, some of our contexts contain two gaps for aspects to be substituted - in that case we treat each substitution of one of the aspects as a separate context.

3.4 Polish Data

In order to estimate the frequencies of Polish names we use the data from the Polish PESEL database, separated into male(mal, 2023) and female(fem, 2023) names, normalized over the total name count. For the cities, we estimate the frequency based on the population of the cities, taken from a 2021 population census (only includes cities above 20 thousand residents)(Pol, 2021), normalized over the total resident count.

3.5 Western Data

To estimate the frequencies of western names, we used the United States Social Security Administration list of most popular baby names(us-, 2023), separated by year and into male and female names and normalized over the total name count. In order to more fully represent current U.S. names, the top 1000 names from the past 10 years(2013-2022) were used. To ensure proper representations in Polish tuned language models, western names with Polish equivalents were translated into their Polish counterparts using the DeepL translation service(dee, 2023).

To estimate the frequencies of Western cities we used the population taken from the United States Census estimates of 2022 populations(wes, 2022). The 300 most populous cities were used, and normalized over total resident count.

3.6 Models

To evaluate the cultural divergence metric, 3 different auto-regressive large language models were used. All models were run on the MIT Supercloud (Reuther et al., 2018) using 2 Nvidia V100 GPUs.

One of the models used was papuGaPT2(Wojczulis and Kłeczek, 2021), a using the GPT-2(Radford et al., 2019) architecture and training approach trained on the polish subset of the multilingual Oscar corpus(Ortiz Suárez et al., 2020).

Another model used was mGPT(Shliazhko et al., 2022), a model using the GPT-3(Brown et al., 2020) architecture trained on 61 languages from 25 language families, including polish and English.

Finally, the Llama-2-chat 70b model(Touvron et al., 2023) was used. This model is a state of the art large language model trained primarily in English. In order to fit into the available GPU memory, the parameters were quantized to 6 bits.

4 Results

5 Conclusion

Limitations

Ethics Statement

Acknowledgements

References

2021. [link].

2022. [link].

2023. [link].

2023. [link].

2023. [link].

2023. [link].

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert & marbert: deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020a. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020b. Aragpt2: Pre-trained transformer for arabic language generation. *arXiv preprint arXiv:2012.15520*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2023. Cultural adaptation of recipes. *arXiv preprint arXiv:2310.17353*.

Pieter Delobelle and Bettina Berendt. 2022. Fairdistillation: mitigating stereotyping in language models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 638–654. Springer.

Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2023. Bias of ai-generated content: An examination of news produced by large language models. *arXiv preprint arXiv:2309.09825*.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2019. Reducing sentiment bias in language models via counterfactual evaluation. *arXiv preprint arXiv:1911.03064*.

Khyati Khandelwal, Manuel Tonneau, Andrew M Bean, Hannah Rose Kirk, and Scott A Hale. 2023. Casteist but not racist? quantifying disparities in large language model bias between india and the west. *arXiv preprint arXiv:2309.08573*.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. *arXiv preprint arXiv:2109.03646*.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*, pages 189–202.

Tarek Naous, Michael J Ryan, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages.

In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Albert Reuther, Jeremy Kepner, Chansup Byun, Siddharth Samsi, William Arcand, David Bestor, Bill Bergeron, Vijay Gadepally, Michael Houle, Matthew Hubbell, et al. 2018. Interactive supercomputing on 40,000 cores for machine learning and data analysis. In *2018 IEEE High Performance extreme Computing Conference (HPEC)*, pages 1–6. IEEE.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Mor Ventura, Eyal Ben-David, Anna Korhonen, and Roi Reichart. 2023. Navigating cultural chasms: Exploring and unlocking the cultural pov of text-to-image models. *arXiv preprint arXiv:2310.01929*.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael R Lyu. 2023. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. *arXiv preprint arXiv:2310.12481*.

Michał Wojczulis and Dariusz Kłeczek. 2021. papugapt2 - polish gpt2 language model.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

A Example Appendix

This is a section in the appendix.