Cultural Divergence: a Rigorous Cultural Bias Benchmark for Autoregressive Language Models

Bartłomiej Cieślar, Pragya Neupane, Willem Guter

Massachusetts Institute of Technology

INTRODUCTION

Motivation

- LLMs commonly trained on western-aligned data
- Non-English and multilingual models exhibit a skew towards the western culture even when prompted in another culture's language
- Current metrics for autoregressive language models only qualitative in nature and require humans or not differentiable
- We propose a differentiable Cultural Divergence (CD) metric to measure how biased an autoregressive model is towards the culture of the prompt (Polish) over another culture (Western)

Existing Methods

- Cultural bias metric:
  - summing the number of completions where western aligned responses were less likely than culturally aligned responses.
  - counting the responses that fit within each language's culture
  - human subjects qualitatively rating the bias in each model's responses
  - analyzing correlation of a cultural bias between the language of the input text and the output image in image generating models
- Mitigating Stereotype biases:
  - Counterfactual Data augmentation
  - Distillation
  - Wasserstein-1 distance

Figure 1: Examples of prompts and generations in Western vs. Polish culture

METHOD

- Ω - the distribution of the language we are investigating (Polish)
- L - language's culture distribution (Polish)
- W - the disribution of the culture we are comparing to (Western) over that language
- M - model's distribution over that language
- Cultural Divergence is the difference of cross-entropies (or KL-divergences) between the model's and language's distribution and the model's and other culture's distribution (over that language)

$$D_C(L||M||W) = \sum_{x\in\Omega} m(x)\frac{w(x)}{l(x)} = H(M,L) - H(M,W)$$

- A - aspect we are approximating the CD with (male/female names and cities) For each
- C - contexts for an aspect
- G - names to fill the context with
- We make sure that both C and G use the basic case (Polish has 7 cases for each noun and noun phrase)
- The distributions for aspects are normalized by name counts and city popiulations
- We approximate the cross entropies with the aspects, contexts and names

$$\tilde{D}_C(L||M||W) = \tilde{H}(M,L) - \tilde{H}(M,W)$$
$$\tilde{H}(M,X) = \sum_{a\in A} \tilde{H}(M,X|a)$$
$$\tilde{H}(M,X|a) = \frac{1}{|C_{X,a}|}\sum_{c_a\in C_{X,a}} \tilde{H}(M,X|c_a)$$
$$\tilde{H}(M,X|c_a) = -\frac{\sum_{g\in G_{c_a}} m(c_a(g))\log x(c_a(g))}{\sum_{g\in G_{c_a}} m(c_a(g))}$$

DATA

Names	Cities	
Anna, 1,075,653 Piotr, 692,120 Krzysztof, 645,674 Katarzyna, 605,826	Warszawa, 1,860,281 Kraków, 800,653 Wrocław, 672,929 Łódź, 670,642	Polish
Liam, 193,343 Noah, 188,340 Olivia, 184775 Emma, 183407	New York, 8,335,897 Los Angeles, 3,822,238 Chicago, 2,665,039 Houston, 2,302,878	Western

Fig. 3 The 4 most frequent items from each distribution

Polish Data

- Names
  - names and frequencies taken from the Polish citizen database, aggregated by sex and first name, publicly available on the government's website and normalized
- Cities
  - cities and their populations taken from the 2021 Polish Census
  - only cities with at least 20.000 inhabitants used

Western Data

- Names
  - names and frequencies taken from the US Social Security Administration list of most popular baby names and normalized
  - top 1000 names from the past 10 years(2013-2022) used.
  - western names with Polish equivalents translated into Polish using the DeepL translator
- Cities
  - cities and their populations taken from the US Census from 2022 populations and normalized by population.
  - the 300 biggest cities were used

RESULTS

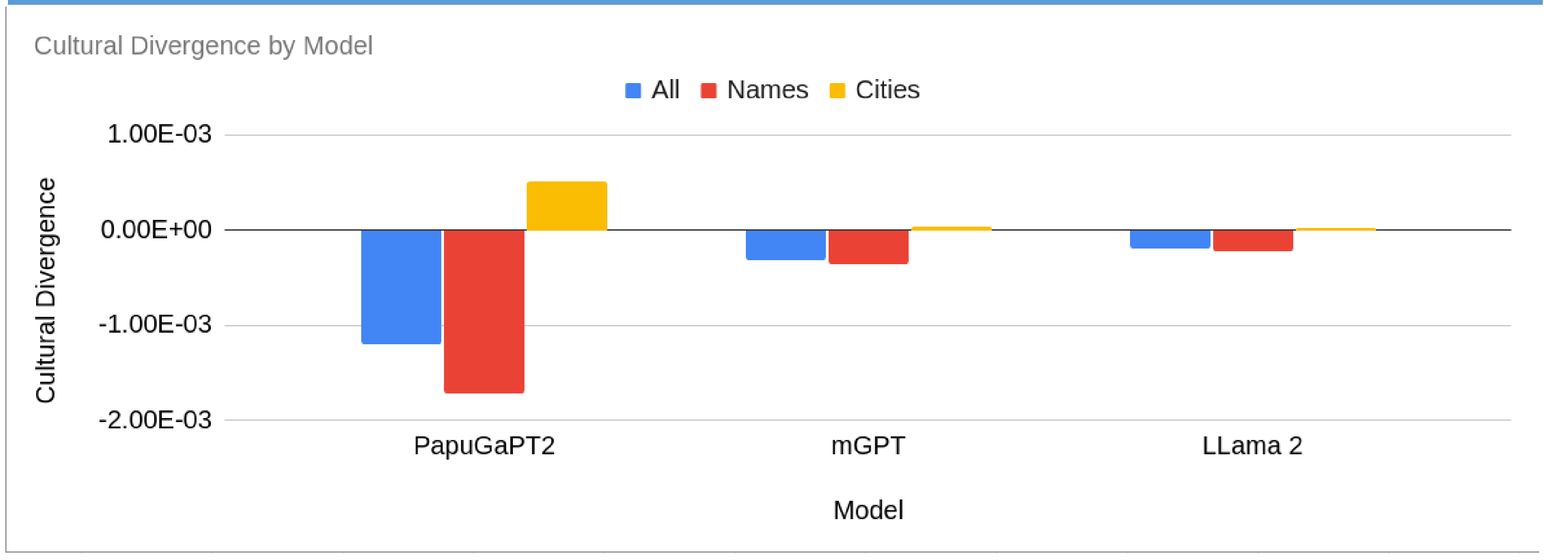

Fig 4. Cultural divergence per model evaluated for both datasets combined (All) and for each dataset individually(Names/Cities).

Distribution\Model	PapuGaPT2	mGPT	LLama 2
All	-1.20E-03	-3.12E-04	-1.99E-04
Names	-1.72E-03	-3.51E-04	-2.28E-04
Cities	5.18E-04	3.94E-05	2.91E-05

Table 1. Values of cultural divergence per model evaluated for both datasets combined (All) and for each dataset individually(Names/Cities).

Conclusions

- Results of the CD metric indicate that language specific models (PapuGaPT2) are more culturally aligned with the language's culture than multilingual models (mGPT) or English specific models (Llama 2), which is in line with previous work.
- The results for the cities distribution alone demonstrate an opposite trend with PapuGaPT2 showing higher CD than mGPT and Llama 2. This may indicate that the metric only works on larger distributions, or that our method of approximating frequency with relative population isn't adequate.
- The difference between the CD of the PapuGaPT2 and the CD of mGPT is much larger than the difference between the CD mGPT and the CD of Llama 2, which indicates that mGPT is western biased even when prompted in non-western languages.

Future Work

- Determine better methods to approximate cultural word distributions.
- Test the CD metric on newer state of the art multilingual or non-western models
- Evaluate the CD metric for finetuning non-western metrics to better match the target culture.

ACKNOWLEDGEMENTS

The authors acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing HPC resources that have contributed to the research results reported within this poster. We would also like to express our gratitude towards 6.8610 instructors and TAs for their guidance and feedback.