Creating a Rigorous Cultural Bias Benchmark for Autoregressive Large Language Models

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

INTRODUCTION

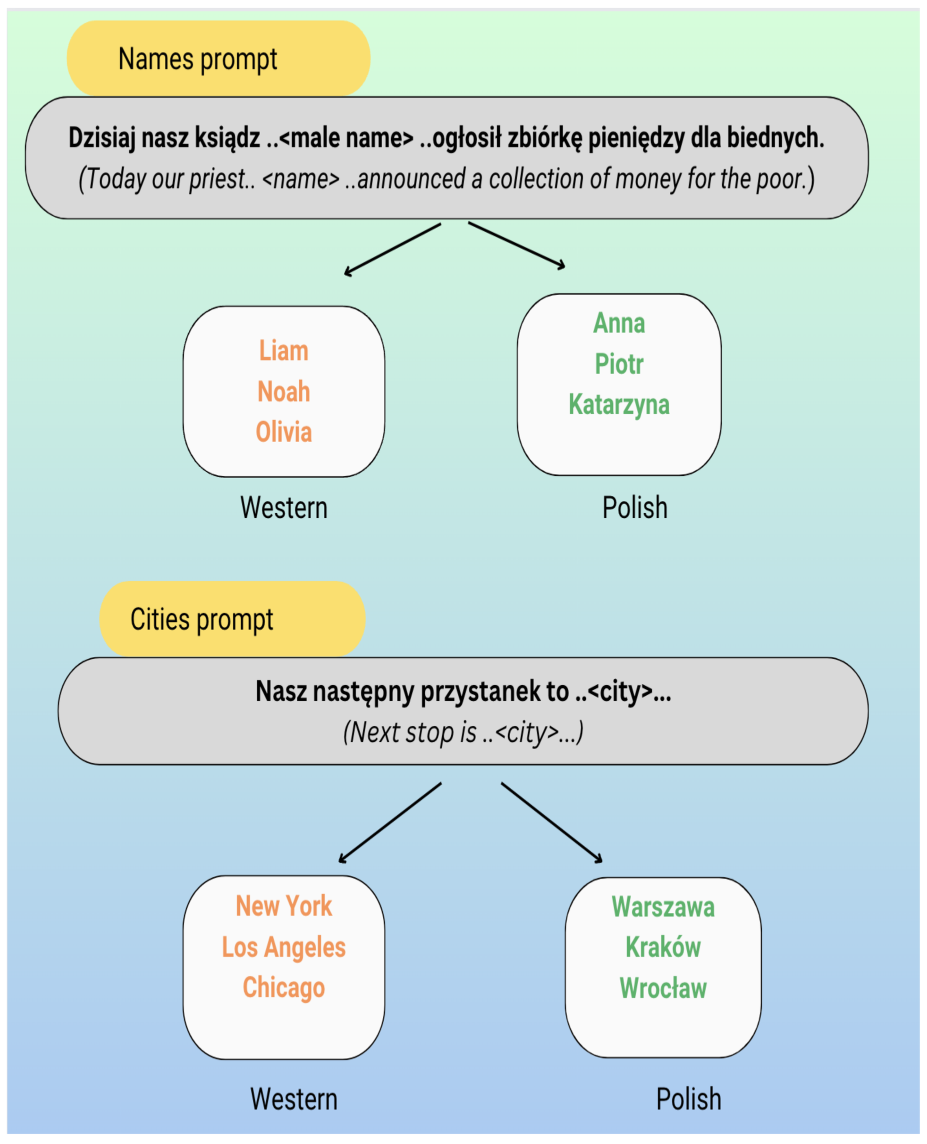

Figure 1: Examples of generations in Western vs. Polish culture

Motivation

- LLMs have been commonly trained on mostly English data
- Both non-English and multilingual models exhibit a skew towards the western culture
- Current metrics for autoregressive language models are either only qualitative in nature and require human intervention or do not cover all relevant cultural aspects
- We propose Cultural Divergence(CD) metric to measure how biased an autoregressive model is towards western culture over others

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

METHOD

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

Western Data

- Names
  - Names and frequencies were taken from the United States Social Security Administration list of most popular baby names and normalized over the total name count.
  - The top 1000 names from the past 10 years(2013-2022) were used.
  - Western names with Polish equivalents were translated into their Polish counterparts using the DeepL translation service.
- Cities
  - Cities and their populations were taken from the United States Census estimates of 2022 populations and normalized by population.
  - The 300 most populous cities were used.

RESULTS

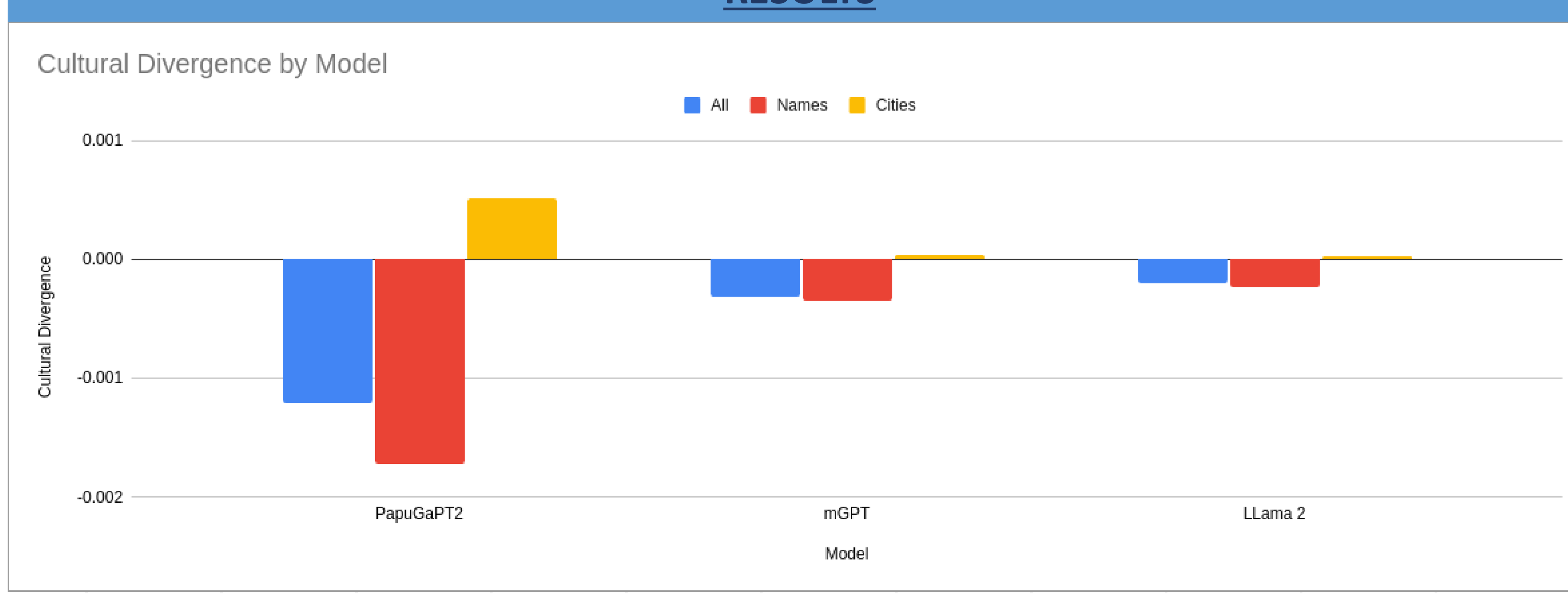

Fig 4. Cultural divergence per model evaluated for both datasets combined (All) and for each dataset individually(Names/Cities).

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 and Future Work

- The cultural divergence metric returned results indicating that language specific models (PapuGaPT2) are more culturally aligned with the target culture than multilingual models (mGPT) or English specific models (Llama 2), which is in line with what has been found with other metrics on non-English models.
- The results for the cities distribution alone demonstrate the opposite results with the language specific model showing higher CD than the other two models. 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 overall CD of the language specific model and the overall CD of the multilingual model is much larger than the difference between the overall CD of the multilingual model and the overall CD of the English only model. This indicates that the multilingual model is western biased even when prompted in non-western languages.
- Next steps:
  - 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 6.861 instructors and TAs for their guidance and feedback.