APPENDIX

WHAT'S IN A NAME?
THE INFLUENCE OF PERSONAL NAMES ON SPATIAL REASONING IN BLOOM LARGE LANGUAGE MODELS

A RELATED WORK

The rise of the large language models promises to revolutionize differentiable approaches to natural language processing. Models such as BERT (Devlin et al., 2018), T5 (Raffel et al., 2020), GPT (Brown et al., 2020), OPT (Zhang et al., 2022), PALM (Chowdhery et al., 2022) and BLOOM (BigScience, 2022) yield results close to the state-of-the-art on popular benchmarks in several language tasks. However, such models are susceptible to adversarial attacks (Wang et al.) and reduced accuracy on out-of-distribution data (Du et al., 2021).

Moreover, bias and toxicity evaluations Ousidhoum et al. (2021); Zhang et al. (2022) of such models have now become standard with a variety of benchmarks, such as hate speech detection, stereotyping involving multiple biases such as religion, race, profession, gender and nationality, toxicity prompts and dialogue safety evaluations.

B EXPERIMENTS

We performed experiments on the BLOOM family of large language models on a server with 256 AMD cores, 2 TB of RAM and 8 Nvidia A100 80GB GPUs. Each language task was performed 100 times to compute the average accuracy of the task.

Our list of popular names in different countries was first obtained from `https://en.wiktionary.org/wiki/Appendix:Most_popular_given_names_by_country`. This list was supplemented by the names available at `https://en.wikipedia.org/wiki/List_of_most_popular_given_names`.

C ACCURACY FOR 5 INDIVIDUALS

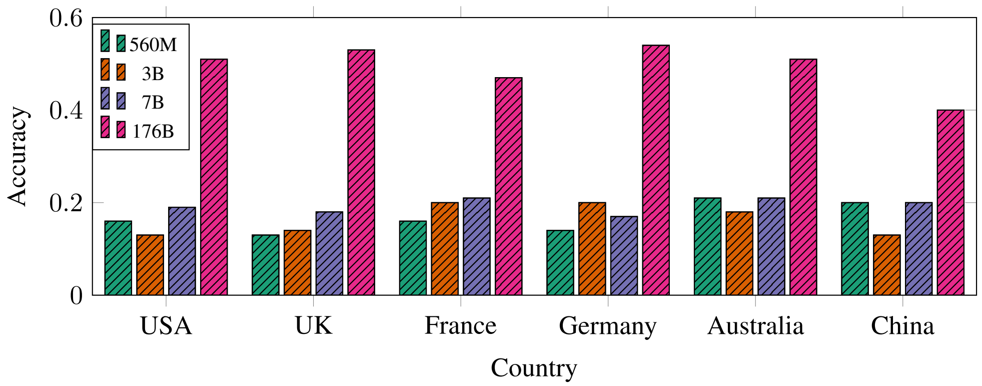

Figure C.1: Efficacy for spatial reasoning with five male names. The BLOOM-176B model predicts the correct response with an accuracy higher than random chance (0.25).