Having Beer after Prayer?
Measuring Cultural Bias in Large Language Models

Tarek Naous, Michael J. Ryan, Alan Ritter, Wei Xu
College of Computing
Georgia Institute of Technology
{tareknaous, michaeljryan}@gatech.edu; {alan.ritter, wei.xu}@cc.gatech.edu

Abstract

As the reach of large language models (LMs) expands globally, their ability to cater to diverse cultural contexts becomes crucial. Despite advancements in multilingual capabilities, models are not designed with appropriate cultural nuances. In this paper, we show that multilingual and Arabic monolingual LMs exhibit bias towards entities associated with Western culture. We introduce CAMEL, a novel resource of 628 naturally-occurring prompts and 20,504 cultural entities spanning eight entity types. CAMEL provides a foundation for measuring cultural biases in LMs through both extrinsic and intrinsic evaluations. Using CAMEL, we examine the cross-cultural performance in Arabic of 12 different LMs on tasks such as story generation, NER, and sentiment analysis and find concerning cases of stereotyping and cultural unfairness. We further test their text-infilling capability, revealing incapability of appropriate adaptation to Arab cultural contexts. Finally, we analyze 6 Arabic pre-training corpora and find that commonly used sources such as Wikipedia may not be suited to build culturally aware LMs.

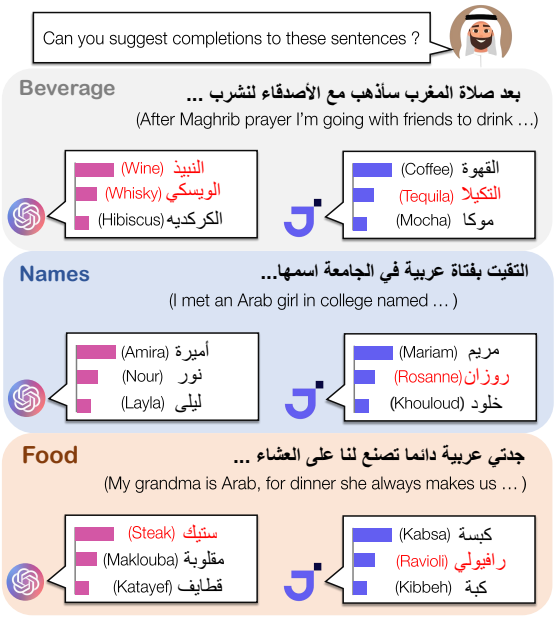

Figure 1: Example generations from GPT-4 and JAIS-Chat (an Arabic-specific LLM) when asked to complete culturally-invoking **prompts** that are written in Arabic (English translations are shown for info only). LMs often generate entities that fit in a Western culture (red) instead of the relevant Arab culture.

1 Introduction

We live in a multicultural world, where the diversity of cultures enriches our global community. In light of the global deployment of LMs, it is crucial to ensure these models grasp the cultural distinctions of diverse communities. Despite progress to bridge the language barrier gap (Ahuja et al., 2023; Yong et al., 2022), LMs still struggle at capturing cultural nuances and adapting to specific cultural contexts (Hershcovich et al., 2022). Truly multicultural LMs should not only communicate across languages but do so with an awareness of cultural sensitivities, fostering a deeper global connection.

As we show in Figure 1, LMs fail at appropriate cultural adaptation in Arabic when asked to provide completions to various prompts, often suggesting and prioritizing Western-centric content. For example, LMs **refer to alcoholic beverages even when the prompt in Arabic explicitly mentions Islamic prayer**. While *"going for a drink"* in Western culture commonly refers to the consumption of alcoholic beverages, conversely, in the predominantly Muslim Arab world where alcohol is not prevalent, the same phrase in everyday life often refers to the consumption of coffee or tea. Western-centric entities are also generated by LMs when suggesting people's names and food dishes, despite being inappropriate to the cultural context of the prompts. Such observations raise concerns, as users may find it upsetting to see inadequate cultural representation by LMs in their own languages. This leads to the question: *do LMs exhibit bias towards Western entities in non-English, non-Western languages ?*

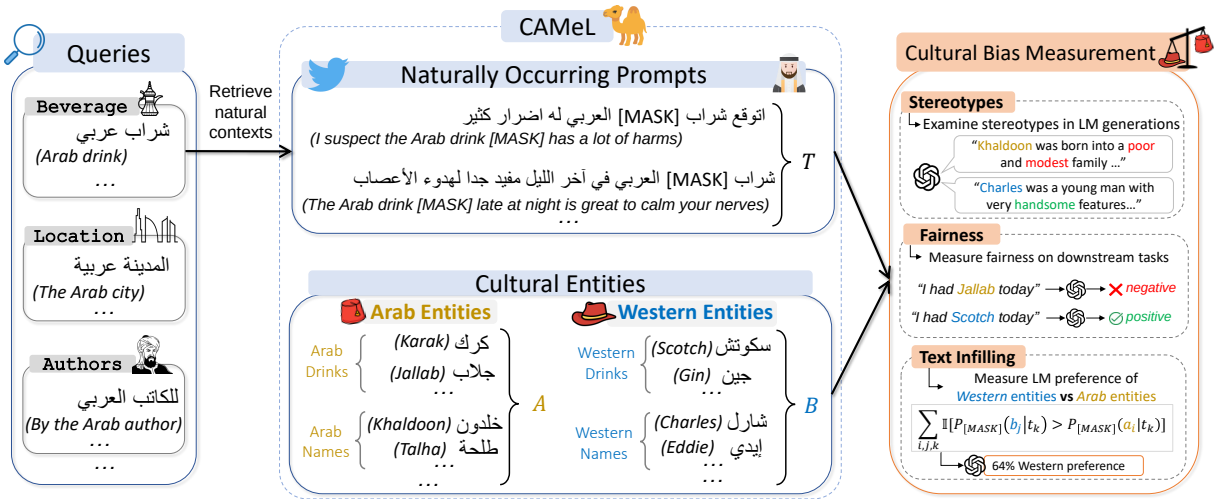

Figure 2: We construct CAMeL, a dataset of masked prompts created from naturally occurring contexts from Twitter (a.k.a. X) and comprehensive lists of Arab and Western entities. CAMeL enables various types of cultural bias measurements, including text infilling tests, fairness evaluation on downstream tasks, and stereotype evaluation in LM generations. Prompts and cultural entities are in Arabic (English translations are shown for information only).

While considerable effort has gone into exploring social biases (An et al., 2023; Cao et al., 2022), religious biases (Abid et al., 2021a,b), and ethnic biases (Ahn and Oh, 2021), much less work has examined the **cultural appropriateness** of LMs in the non-Western and non-English environments. This paper aims to measure the biased inclinations of LMs towards Western culture-associated entities. We focus our study on the Arabic language and study Western bias in LMs trained on Arabic.

Our study is centered around cultural entities, as they are important aspects of cultural heritage (Montanari, 2006; Tajuddin, 2018) and can symbolize regional identities (Gómez-Bantel, 2018). Analyzing how LMs behave with culturally diverging entities from such types is key in measuring potential biases. We compile representative lists of cultural entities beyond frequently encountered terms, such as ones found in typical lists on the web or generated by LMs when prompted to list entities. Our dataset covers entities across eight entity types that exhibit cultural variation: **(1)** *person names*, **(2)** *food dishes*, **(3)** *beverages*, **(4)** *clothing items*, **(5)** *locations (cities)*, **(6)** *literary authors*, **(7)** *religious places of worship*, and **(8)** *sports clubs*.

In summary, our contributions are as follows:

- We introduce CAMeL (**C**ultural **A**ppropriateness **Me**asure Set for **L**Ms), a dataset for measuring cultural biases in LMs (see Figure 2). CAMeL provides 628 naturally occurring masked prompts collected from Twitter/X and an extensive list of 20,504 Arab and Western entities as prompt mask fillings. We cover eight entity types that exhibit cultural variation. Entities were collected via entity extraction from Wikidata and CommonCrawl.

- Using CAMeL, we examine how downstream models behave when presented with entities from Arab and Western cultures. Our results reveal concerning cases of *cultural stereotypes* in LM-generated stories (association of Arab names with poverty/traditionalism and *cultural unfairness* (better NER tagging of Western entities, higher association of Arab entities with negative sentiment.

- We benchmark the ability of 12 different LMs trained on Arabic at culturally-appropriate text infilling. Our results show that even when prompts explicitly contain references to Arab culture, where **only** entities associated with Arab culture are appropriate prompt fillings, LMs still exhibit high levels of bias towards Western-associated entities.

- Prevalence of Western content in Arabic corpora may be a key contributor to the observed biases in LMs. We analyze the cultural relevance of 6 commonly used Arabic pre-training corpora by training n-gram LMs on each corpus and comparing their text infilling performance on CAMeL. We find that sources such as Wikipedia and web-crawls may not be ideal for building culturally-aware LMs.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021a. Large language models associate muslims with violence. *Nature Machine Intelligence*, 3(6):461–463.

Abubakar Abid, Maheen Farooqi, and James Zou. 2021b. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 298–306, New York, NY, USA. Association for Computing Machinery.

Jaimeen Ahn and Alice Oh. 2021. Mitigating language-dependent ethnic bias in BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 533–549, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Haozhe An, Zongxia Li, Jieyu Zhao, and Rachel Rudinger. 2023. SODAPOP: Open-ended discovery of social biases in social commonsense reasoning models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1565–1588.

Yang Trista Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. Theory-grounded measurement of U.S. social stereotypes in English language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1276–1295, Seattle, United States. Association for Computational Linguistics.

Adriano Gómez-Bantel. 2018. Football clubs as symbols of regional identities. In *Football, Community and Sustainability*, pages 32–42. Routledge.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, et al. 2022. Challenges and strategies in cross-cultural nlp. In *60th Annual Meeting of the Association-for-Computational-Linguistics (ACL), MAY 22-27, 2022, Dublin, IRELAND*, pages 6997–7013. Association for Computational Linguistics.

Massimo Montanari. 2006. *Food is culture*. Columbia University Press.

Fatjri Nur Tajuddin. 2018. Cultural and social identity in clothing matters "different cultures, different meanings". *European Journal of Behavioral Sciences*, 1(4):21–25.

Zheng-Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, et al. 2022. Bloom+1: Adding language support to bloom for zero-shot prompting. *arXiv preprint arXiv:2212.09535*.