GEORGE MASON UNIVERSITY

UNIVERSITY of WASHINGTON

NAACL 24

Global Gallery

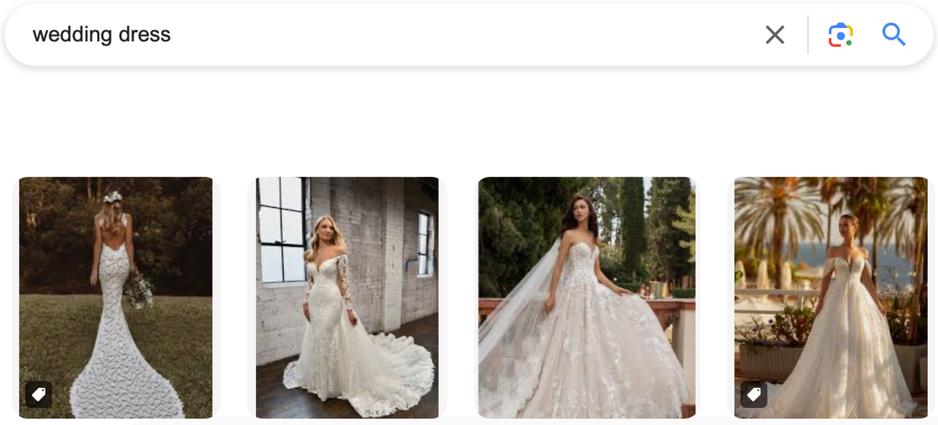
The Fine Art of Painting Culture Portraits through Multilingual Instruction Tuning

Anjishnu Mukherjee
amukher6@gmu.edu

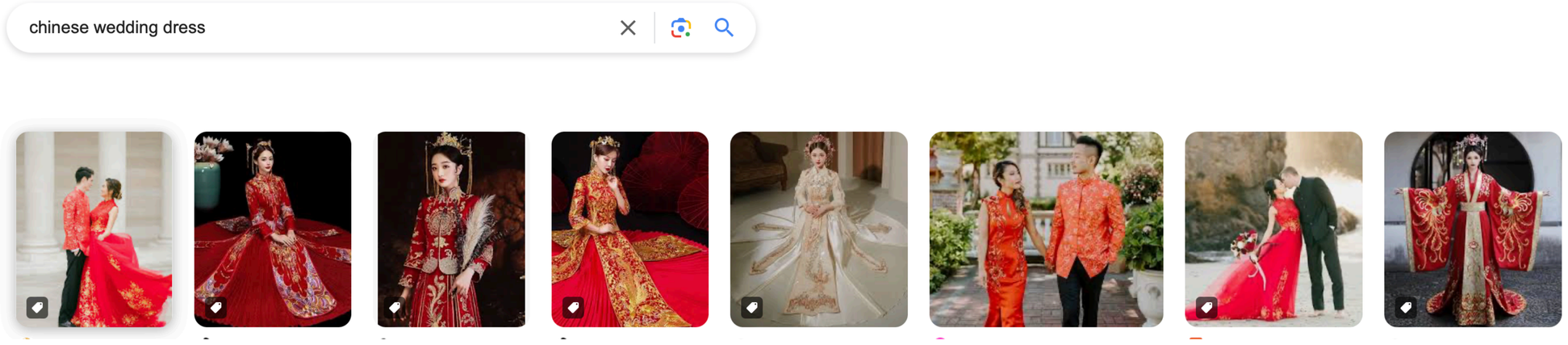
Aylin Caliskan
aylin@uw.edu

Ziwei Zhu
zzhu20@gmu.edu

Antonios Anastasopoulos
antonis@gmu.edu



But are wedding dresses white in all countries/cultures?



“Culturally relevant” LLMs

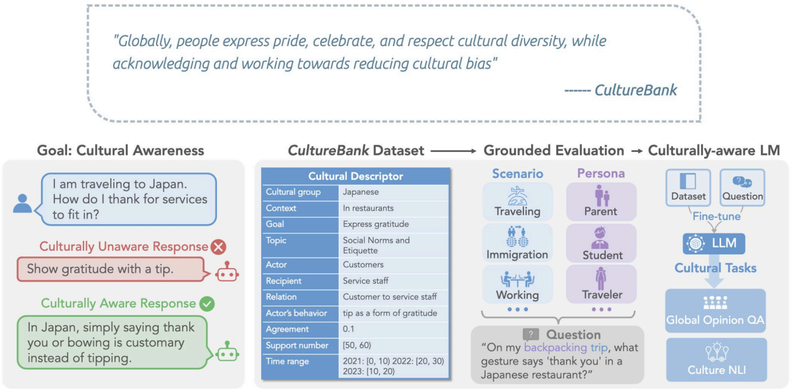
What we are trying to solve

How to make LLM responses culturally relevant?

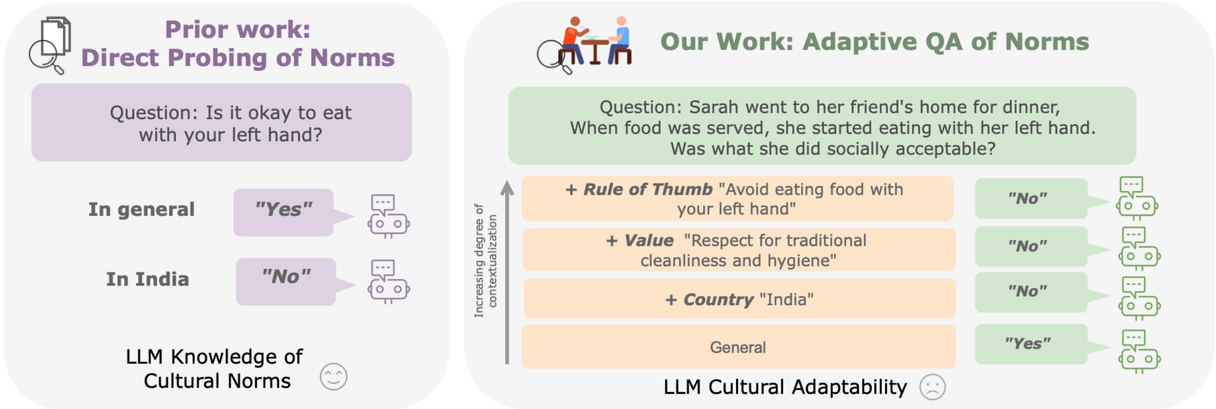
[Method 1] Get a lot of data that is related to the culture

Challenges:

- “Culture” is not well-defined.
- Curating relevant data in the first place is expensive.
- Automated data collection (recent methods) cover limited domains (usually recipes and social norms)



CultureBank: Weiyan Shi et al. (April 2024)



NORMAD: Abhinav Rao et al. (April 2024)

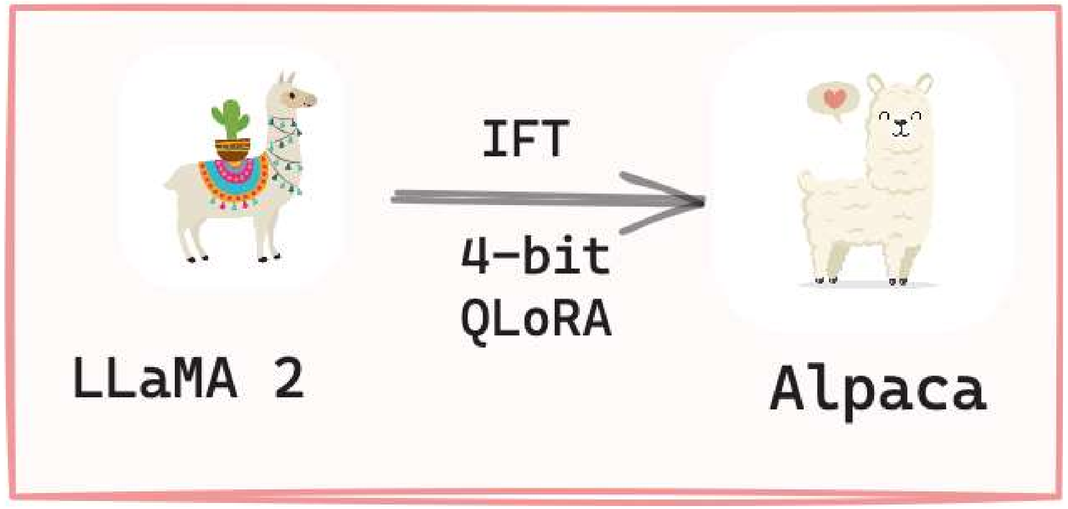
What we are trying to solve

How to make LLM responses culturally relevant?

[Method 2] Maybe the language that is spoken in that culture contains some implicit signals about that culture which can be picked up during fine-tuning?

Pros: Getting data in a specific language is easier/less expensive than getting culturally relevant data for each culture.

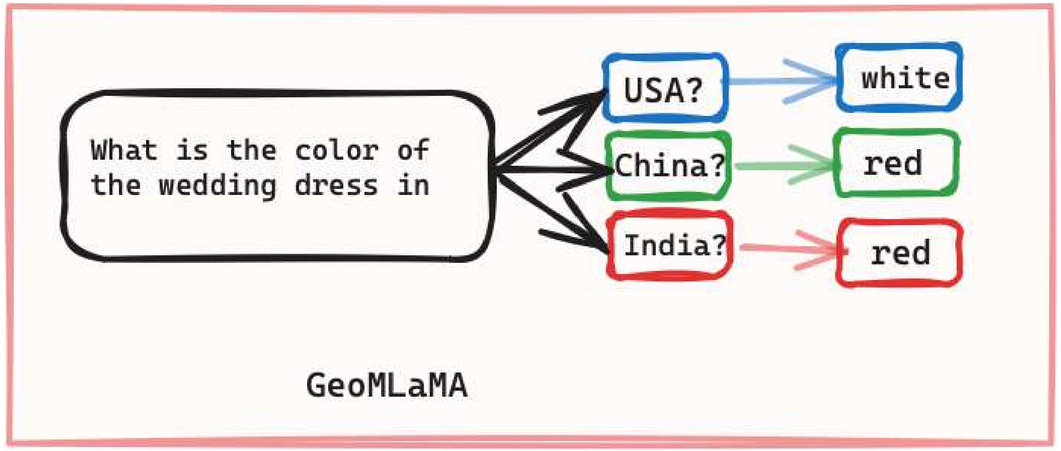
Cons: We are making simplifying assumptions



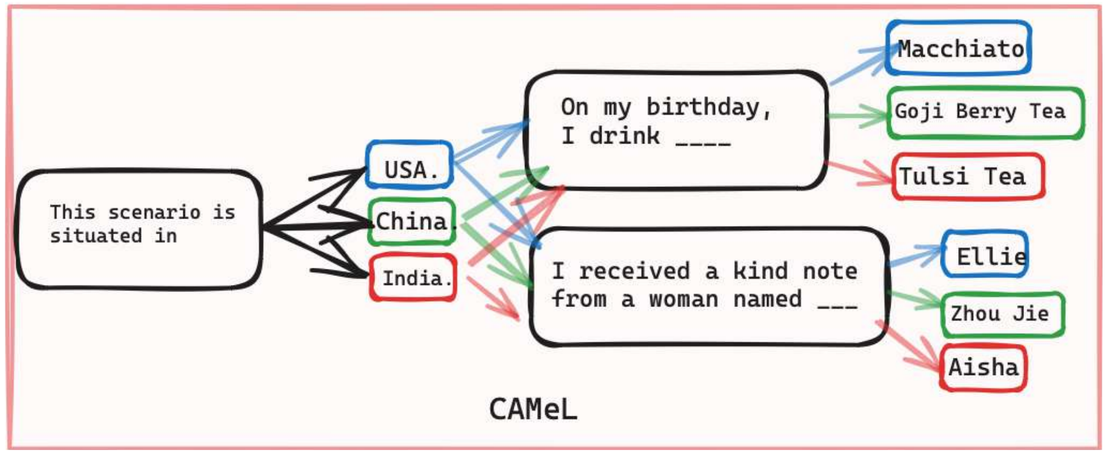
Language	Input	Instruction	Output	Avg
Chinese (zh)	0.78	0.79	0.75	0.77
Greek (el)	0.82	0.83	0.78	0.81
Hindi (hi)	0.84	0.85	0.82	0.84
Persian (fa)	0.83	0.84	0.80	0.82
Swahili (sw)	0.80	0.80	0.77	0.79

CometKiwi scores: Is MT any good?

Testing if this worked



GeoMLaMA: Da Yin et al. (May 2022)



CAMeL: Tarek Naous et al. (May 2023)

Does instruction tuning on language-specific data enhance cultural knowledge?

For the same base model, we compare

Adapter 1 = English Alpaca data

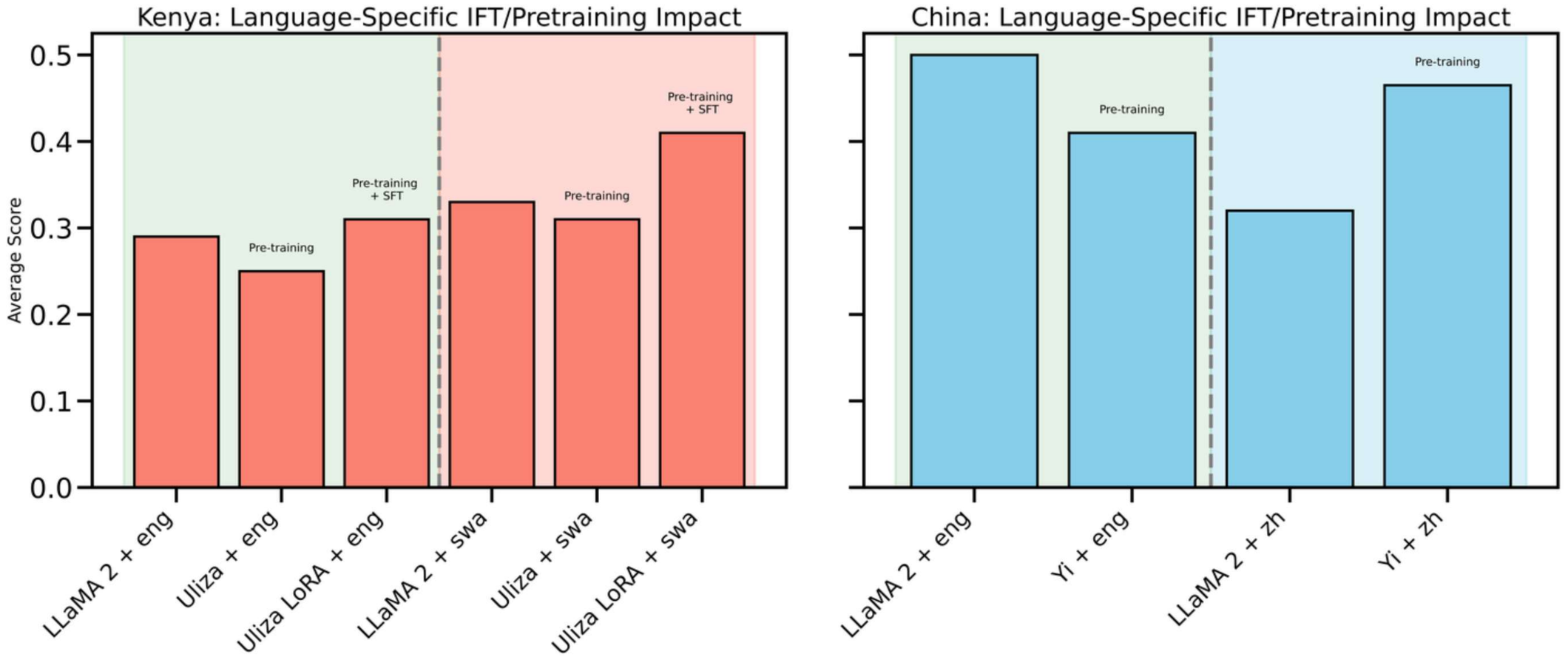
Adapter 2 = Language-specific Alpaca data

For Adapter 2, prompt in English vs prompt in language.

We define 5 settings with different kinds of distractors to test whether LLMs are able to reason for different aspects.

Instruction-tuning on language specific data

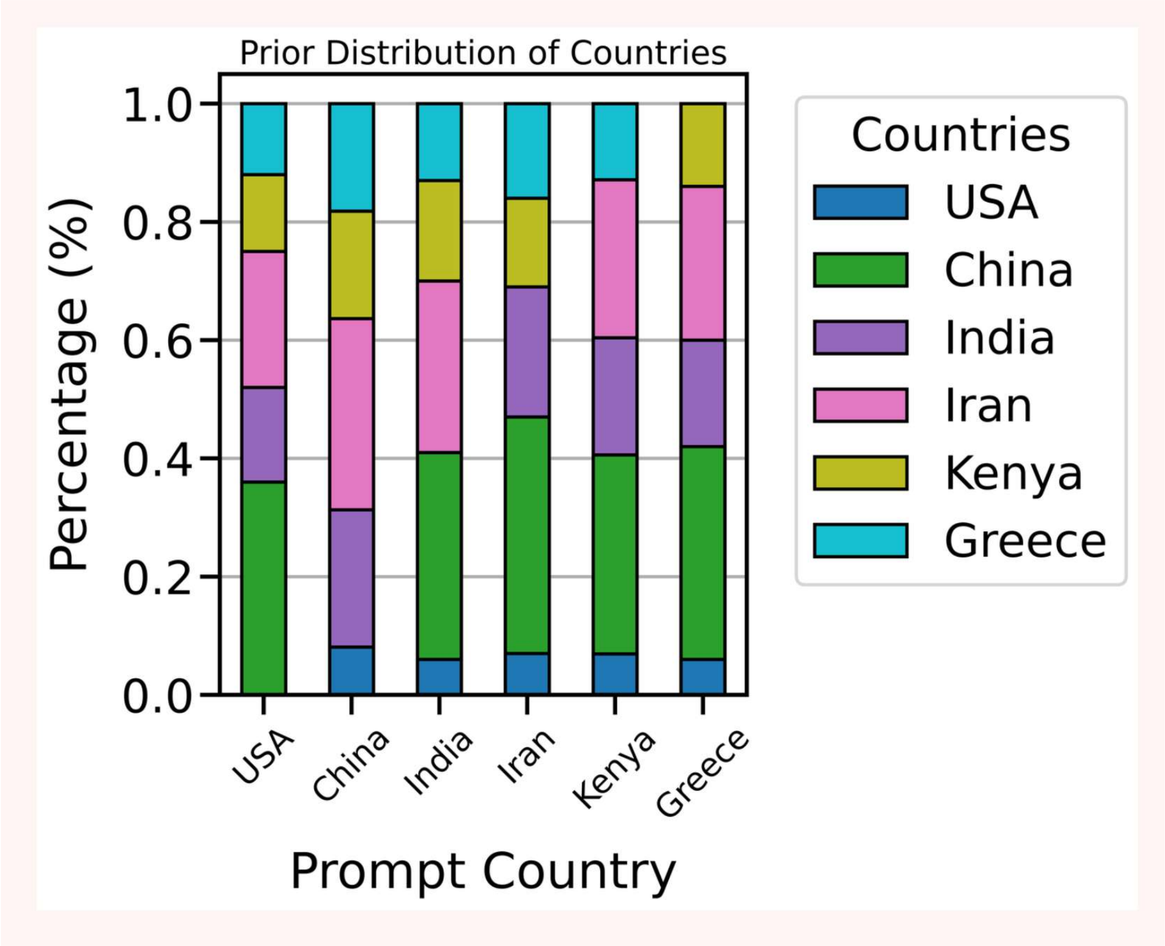
Also, effects of pre-training on high quality data and fine-tuning on curated data



Loosely speaking, Pretraining + language specific SFT (better if curated data) + language specific prompting is best.

Fine-grained evals

Only the interesting bits

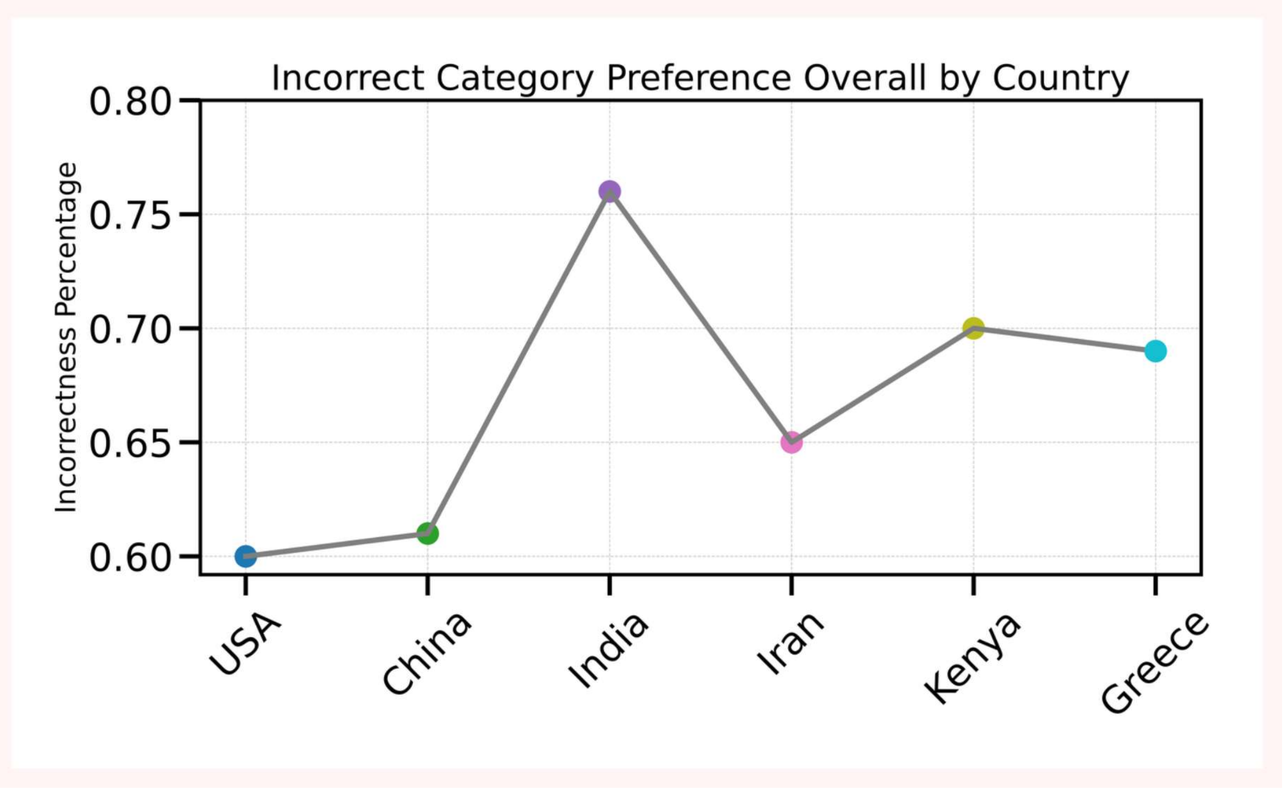


When no information about the country is provided, there is still a noticeable preference for some countries.

China and Iran appear to have a strong prior over our evaluation set of items. Future work?

Fine-grained evals

Only the interesting bits



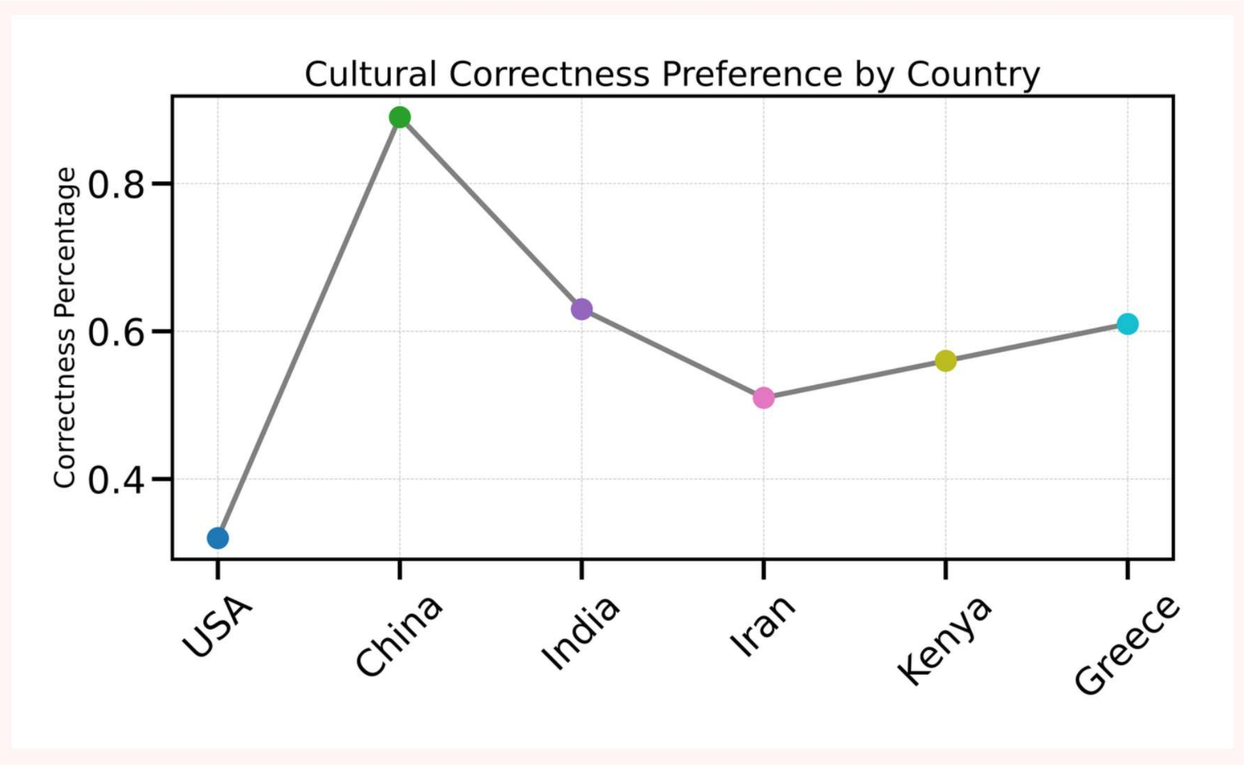
We provide 4 options, all of which are culturally appropriate. 3 of these however are for an incorrect “category”. We then measure the percentage of times that one of these incorrect options is chosen.

0.60 on the y axis means for 60 prompts out of 100, an incorrect category was chosen (for eg, an option about a beverage was chosen when asked about a location)

The fact that it gets so many questions incorrect in this setting implies missing category understanding.

Fine-grained evals

Only the interesting bits



We provide 4 options, 2 of which are culturally appropriate but are not of the correct grammatical gender.

0.60 on the y axis means for 60 prompts out of 100, an option that is culturally appropriate but has incorrect grammatical gender is preferred.

More analyses to do, yes. But this takes a quick look at how gender bias is a problem in this context.

Limitations/Future work

- Conceptualizing culture as a single entity associated with a country is not very accurate (many on-going works on defining “what is culture”)
- Some of the languages we test are not officially supported by the LLaMA 2 tokenizer. Methods to extend the tokenizer to cover these would be helpful. (Eg. Recent methods like Zero-Shot Tokenizer Transfer enables usage of any model with any tokenizer)
- Comparisons of benefits gained from pre-training on culturally relevant data and fine-tuning on such data is not measured in a controlled setting as we do not pre-train any models from scratch. (A couple of ACL submissions do this on a small scale)

Thank you!

Questions?

https://github.com/iamshnoo/culture-llm

amukher6@gmu.edu