Investigating Cultural Alignment of Large Language Models

Badr AlKhamissi, Mai ALKhamissi, Muhammad ElNolrashy, Mona Diab

Belinda Weng

Introduction

There are people who interact with LLMs in multiple languages.

Different language prompts can lead to different answers to similar questions

- The current data the LLM has – each language is associated with a specific culture with different values and traditions
  - English → western
  - Chinese → China
  - etc.

Research Questions

1. **Prompting language and cultural alignment**
    - Prompt is answered in the same language = greater cultural alignment
        - i.e. asking survey questions about the US in English will produce more culturally aligned responses, compared to if you asked about the US in a foreign language
    - However, there are cases where a language is associated with multiple cultures
        - i.e. people in Egypt may express their opinions in English rather than their native language

2. Pre-training data composition

- The more a model is trained on data from a particular culture, the better it will reflect that culture's views.

3. Personas and cultural topics

- The model is more culturally aligned with personas that are in popular areas
- Thus, misalignment is expected for personas that are from underrepresented backgrounds

4. Finetuning models to induce cross-lingual knowledge transfer

- See if an English-trained model can culturally align in arabic contexts, and vice versa

Model	Size	Pretraining
GPT-3.5	175B	Majority English
mT0-XXL	13B	Multilingual
LLaMA-2-Chat	13B	Majority English
AceGPT-Chat	13B	English then Arabic

Table 6: List of models used in this work.

Experimental Setup

They aim to measure cultural alignment by simulating existing surveys done by socialists in certain populations.

- World Values Survey (WVS): an ongoing project that gathers responses to questions on matters on, but not limited to, social, cultural, and economical importance from around the world
  - In this work, they selected 30 questions with diverse themes from Egypt and the US (both covers a diverse demographic)

Theme	# of Questions
Social Capital, Trust & Organizational Membership	8
Social Values, Attitudes & Stereotypes	4
Political Interest & Political Participation	6
Political Culture & Political Regimes	3
Security	4
Religious Values	2
Migration	3

Table 10: The number of questions per theme for the 30 questions considered in this work.

- Each participant's response of the survey is associated with a "persona" of that participant

Dimension	Possible Values
Region	Cairo, Alexandria, etc.
Sex	Male, Female
Age	`Number`
Social Class	Upper, Working, etc.
Education Level	Higher, Middle, Lower
Marital Status	Married, Single, etc.

Table 1: The demographic dimensions used when prompting the model to emulate a certain survey respondent. Region is country-specific. More information in Appendix D.

"Persona" = the demographic of a survey participant

This will allow the LLM to capture the diversity not only of a specific country but also among other people in that country

Method of gathering data:

1. Gather participants' responses of the WVS survey in their dominant language
2. This survey is translated to that dominant language for the LLM to answer in the context of that "persona"
3. Both responses are used to calculate alignment

Imagine you are a {marital_status} {sex} from {region}, {country}.
You are {age} years of age and completed {education} education level.
You consider yourself part of the {social_class}.

Answer the following question from this perspective.
Others will read what you choose; your goal is to convince them it was chosen from the perspective of the persona described above.
Select exactly one option. Do not include any extra commentary.
Answer by typing the number corresponding to your chosen answer.
Question: {question}
Options: {numbered_options}

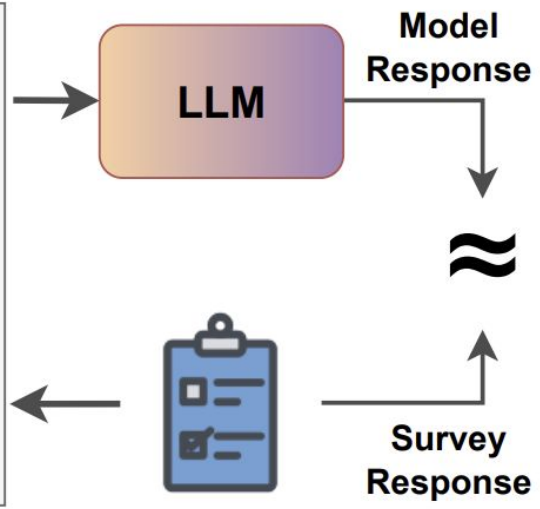


Figure 2: Template used when querying models in English. **(Left)** The model is first instructed to respond under a specific persona along the demographic parameters highlighted in red. **(Right)** The rest of the prompt instructs the model to follow the perspective of the persona closely, respond in a specific format (only the index of the answer), and avoid any extraneous commentary.

Hard matrix vs soft matrix

A numerical representation of cultural alignment

Hard matrix:

- Strict accuracy
- Model only gets credit if its answers matches the survey respondents *exactly*

$$H_{f,c} = \text{mean}_{q,p}\{\mathbb{1}(\hat{y} = y)\}$$

Soft matrix:

- Model can get "partial credit" if its answer is close enough for questions with a scale (1-5 for agreement)

$$\varepsilon_{f,c}(q,p) = \begin{cases} \dfrac{|\hat{y} - y|_{q,p}}{|q| - 1} & \text{IsOrd}(q,p), \\ \mathbb{1}(\hat{y} \neq y)_{q,p} & \text{otherwise} \end{cases}$$

Anthropological Prompting

In an attempt for more culturally aligned responses of LLMs for underrepresented and minority groups, they utilized anthropological prompting.

- Essentially telling the model to act and think as an anthropologist
- Prompted the model to comprehend anthropological concepts – identities, inquiries, and linguistic constructions
  - This will allow the model to reason before answering any question

I Anthropological Prompting

I.1 Prompt Template

The following is a framework adapted from the toolkit of anthropological methods:

1. **Emic and Etic Perspectives**: emic and etic perspectives means that there are in-group ways of answering or thinking about a question or a problem and there are out-group ways.
2. **Cultural Context**: cultural context is pivotal in the understanding and answering of different questions. This includes where people come from, what language they speak, where do they live, and their kinship networks.
3. **Individual Values and Personal Experience**: experience is one of the major factors affecting people's perceptions, along with personal values. Both play a big role in subjective understandings of day to day to life.
4. **Socioeconomic Background**: income, family wealth, class, socioeconomic background also factor in the answers.
5. **Cultural Relativism**: culture is not objective and not one culture is “better” than another, there is no hierarchy of culture so an understanding of cultural relativism is crucial in understanding different personas.
6. **Space and Time**: age and place are also important factors.
7. **Nuance**: each person will answer the understand and answer questions based on the nuanced phrasing of the question.

Results

Anglocentric Bias in LLMs

Regardless of being finetuned or trained to be multilingual, the LLMs are significantly more culturally aligned with people in the US than those in Egypt

- Western bias according to concurrent research, likely due to the large amount of data used to train LLMs in general

Model	Egypt	United States
GPT-3.5	48.61 / 25.99	64.86 / 39.29
AceGPT-Chat	47.82 / 29.72	52.83 / 27.69
LLaMA-2-Chat	46.31 / 24.48	63.10 / 36.72
mT0-XXL	45.92 / 27.93	55.48 / 31.40
Average	47.16 / 27.03	59.07 / 33.78

Table 3: Cultural alignment against responses from both Egyptian and United States surveys using `Soft` / `Hard` similarity metrics for each model. The results are averaged across both prompting languages. The alignment with the United States populations is much higher reflecting the euro-centric bias in current LLMs.

Prompting and Pretraining Languages

Their hypothesis on using the country's dominant language for more culturally aligned results was correct.

- GPT-3.5 (English) and AceGPT-Chat (English, then Arabic)
  - English prompts increased in alignment with US surveys compared to Arabic
- LLaMA-2-Chat (English)
  - Arabic prompts were less effective in enhancing alignment with Egypt survey
    - A result of lack of Arabic data in the pretraining which leads to a lack of knowledge of Egyptian culture
- mT0-XXL (Multilingual)
  - Although trained to be multilingual, still had Western biases where there was more alignment with English compared to Arabic

Digitally Underrepresented Personas

Alignment improves as the backgrounds of individuals changes from lower to higher levels in both respective dimensions.

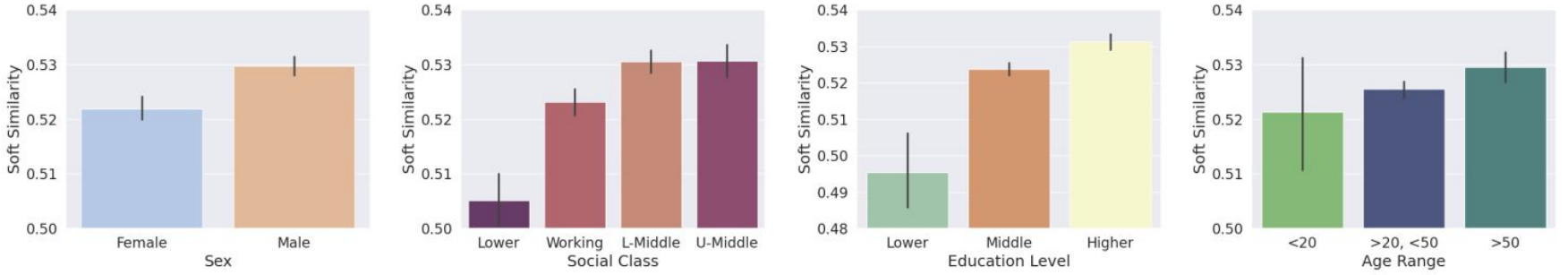
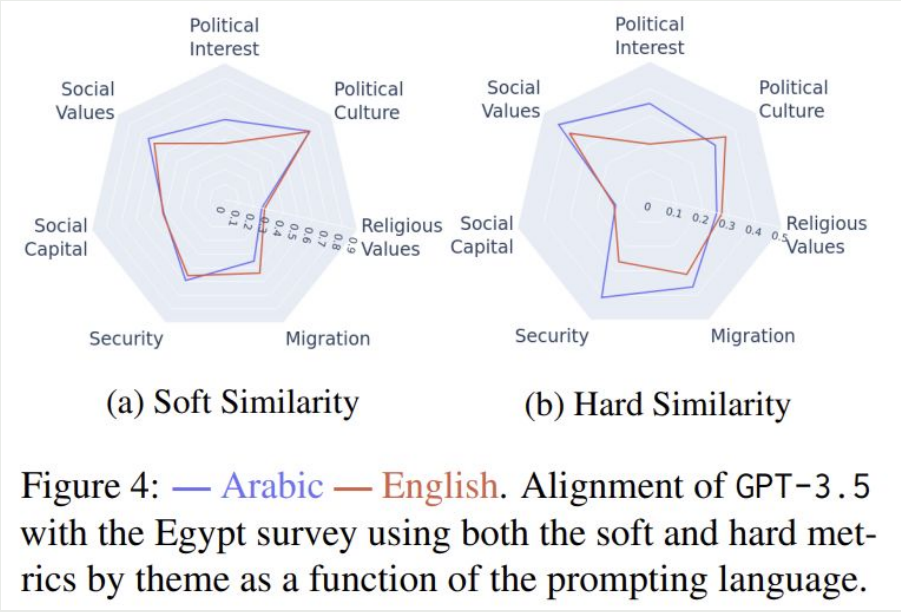


Figure 3: Cultural alignment as a function of a subject's `Sex`, `Education Level`, `Social Class`, and `Age Range`. Results are averaged across the models, prompting languages and surveys used in this work. `L-Middle` and `U-Middle` are `Lower Middle` and `Upper Middle` Class respectively.

Cultural Alignment per Theme

Examining cultural alignment concerning certain topics and theme.



(a) Soft Similarity (b) Hard Similarity

Figure 4: — Arabic — English. Alignment of GPT-3.5 with the Egypt survey using both the soft and hard metrics by theme as a function of the prompting language.

Themes contributing to the improvement in alignment in the Egypt survey when prompt in Arabic:

- Social values, political interest, and security

Anthropological Prompting

Results shows that anthropological prompting improves cultural alignment, especially for participants from underrepresented backgrounds.

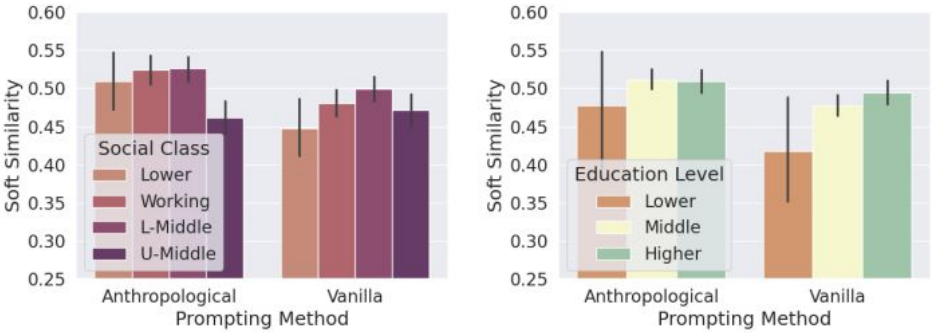


Figure 5: Anthropological prompting improves alignment for underrepresented personas compared to Vanilla prompting. Results on `GPT-3.5` using English prompting. More in Appendix I.

Discussion

This study finds that both pretraining language and prompting language boost cultural alignment, especially when they match the culture's dominant language.

Cultural identity is diverse with any country. This study used Modern Standard Arabic (MSA). Doesn't capture all of Egyptian culture, since Egyptians primarily speak Egyptian Arabic dialects. Thus, alignment will further improve with dialect-specific training, rather than measured with a single archetype.

Conclusion & Future Work

Study presented a framework for measuring cultural alignment in LLMs by referencing surveys done in the US and Egypt, and comparing these responses with LLMs across six demographic dimensions, testing how pretraining language composition and prompting language affect alignment.

Introduced anthropological prompting, which allows models to use anthropological reasoning before generating a response, leading to improved cultural alignment, especially for underrepresented groups.

Future works include expanding to more cultures and languages and exploring cultural alignment as a proxy for cross-lingual knowledge transfer. Additionally, using other LLMs and other survey sources

Thank you :)!