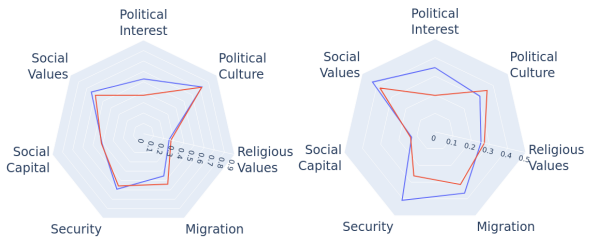

(a) Soft Similarity (b) Hard Similarity

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.

respondents. Similarly, older age groups exhibit higher alignment than younger age groups.

5.4 Cultural Alignment per Theme

The 30 questions examined in this work are categorized into 7 distinct themes outlined by the WVS survey (Haerpfer et al., 2020). Table 10 illustrates the distribution of questions across these themes. The granularity provided by these themes enables us to assess alignment concerning topics such as Religious Values. In Figure 4, we illustrate the cultural alignment of GPT-3.5 with respect to responses from both the Egypt and the US survey, and examine the prompting language effect within each plot. The three themes that are contributing to the improvement in alignment in the Egypt survey when prompting in Arabic using GPT-3.5 are Social Values, Political Interest and Security. In the US survey, both English and Arabic prompting perform very closely except in the Migration theme where English has a slight edge. See Appendix H for a comprehensive set of results for all other models, metrics, and country combinations.

5.5 Finetuning for Cultural Alignment

Here, we delineate the contrast between AceGPT-Chat and LLaMA-2-Chat to illustrate the impact of finetuning an English-pretrained model on data from another language on cultural alignment. We observe an improvement in alignment with the Egypt survey across both metrics when the two models are prompted in Arabic (see Table 2 for a quantitative comparison). When prompted in English, the increase is evident only with the hard metric. Conversely, we note a decline in alignment following finetuning when evaluating alignment against the US survey, indicating that the model forgot some of its existing US cultural knowledge while adapting to data in another language.

Prompting Method	Soft	Hard
Vanilla	0.4834	0.2443
Anthropological	**0.5102**	**0.2838**

Table 4: Anthropological prompting outperforms Vanilla prompting across both metrics in terms of cultural alignment with the Egypt survey. Results here are on GPT-3.5 with English prompting.

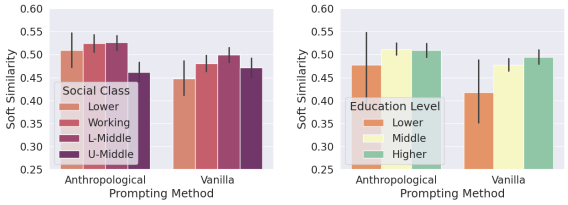

Figure 5: Anthropological prompting improves alignment for underrepresented personas compared to Vanilla prompting. Results on GPT-3.5 using English prompting. More in Appendix I.

5.6 Anthropological Prompting

To improve cultural alignment with responses from Egyptian participants and underrepresented groups, we propose Anthropological Prompting. This approach enables the model to reason before answering the question while grounded with a framework adapted from the toolkit of anthropological methods. The rationale behind it is described in Section 4.6. The framework offers guidance for the model to consider emic and etic perspectives, cultural context, socioeconomic background, individual values, personal experience, cultural relativism, as well as spatial and temporal dimensions in a nuanced manner. The exact prompt is provided in Appendix I. Table 4 presents the results when prompting GPT-3.5 in English, comparing both "vanilla" and anthropological prompting with one variant per question. While vanilla prompting generates 5 responses and computes the majority vote to determine the final answer, the anthropological prompting method generates only one response, yet still outperforms vanilla prompting.

Further, we observe that anthropological prompting improves cultural alignment for participants from underrepresented backgrounds. Figure 5 illustrates this comparison between vanilla and anthropological prompting across Social Class and Education Level demographic dimensions. The alignment distribution among social classes and education levels becomes more equitable as a result.