Qeios

Peer Review

Review of: "Self-Pluralising Culture Alignment for Large Language Models"

Aliah Zewail[1]

1. Department of Psychological and Brain Sciences, University of Massachusetts at Amherst, United States

The present manuscript, "Self-Pluralising Culture Alignment for Large Language Models," introduces an innovative framework (CultureSPA) that taps into the internal cultural knowledge of large language models (LLMs) in order to create a more culturally sound tool. The authors outline this process in four steps: first, they create an array of questions pulled from cultural topics mentioned in the World Values Survey (WVS). Next, they generate LLMs' outputs for these questions through two different prompts (culture-unaware and culture-aware). Then, they assess shifts in the consistencies between the two prompts. Lastly, the authors incorporate Supervised Fine Tuning and compare two models, CultureSPA-joint and CultureSPA-specific, with results demonstrating that LLMs trained on cross-cultural data (CultureSPA-joint) result in overall better cultural alignment. I enjoyed reading this manuscript; the Introduction provides a succinct overview of the paper, the Methods are interesting and robust, and the Results are well-explained. The paper is well-written and well-designed overall. In what follows, I have listed a number of minor points and suggestions, which I hope the authors will find helpful in revising their work.

1. In section 4 (CultureSPA), you state that "[exhibited shifts in outputs] when cultural information is provided are deemed the most representative of a specific culture." My concern is that a shift in itself does not necessarily indicate it is the "right" shift. Did you perform many iterations of this process to make sure it is robust? How can you validate this assumption?
2. In section 5 (Experiments), I thought the left distribution in Figure 3 depicting the internal knowledge of LLMs was fascinating. I think it would be helpful to provide some insight into why LLMs are more likely to display an alignment across various cultures for topics such as well-being and happiness and not others (e.g., religious values, security, corruption). Is it because of

doi.org/10.32388/BOJVOG

their harm-avoidance tendencies? If so, how can we reconcile these tendencies while ensuring that LLMs do not overly generalize or simplify complex cultural values?

3. In the Limitations section, it is important to mention the limitation of solely using English in generating your cultural questions and prompting the system, as it prevents you from capturing this motif of cultural diversity mentioned throughout the paper (see Naous et al., 2023). Moreover, you mention using WVS as a limitation in terms of the ground truth for cultural alignment. I would expand upon this and consider what kind of populations within each culture have access to/the bandwidth to participate in the WVS. How does this impact the cultural diversity researchers are wanting to instill in LLMs?

Declarations

Potential competing interests: No potential competing interests to declare.