Exploring Multilingual Concepts of Human Values in Large Language Models: Is Value Alignment Consistent, Transferable and Controllable across Languages?

Shaoyang Xu[1], **Weilong Dong**[2], **Zishan Guo**[2], **Xinwei Wu**[2] and **Deyi Xiong**[2,1]*
[1]School of New Media and Communication, Tianjin University, Tianjin, China
[2]College of Intelligence and Computing, Tianjin University, Tianjin, China
{syxu, willowd, guozishan, wuxw2021, dyxiong}@tju.edu.cn

Abstract

Prior research has revealed that certain abstract concepts are linearly represented as directions in the representation space of LLMs, predominantly centered around English. In this paper, we extend this investigation to a multilingual context, with a specific focus on human values-related concepts (i.e., value concepts) due to their significance for AI safety. Through our comprehensive exploration covering 7 types of human values, 16 languages and 3 LLM series with distinct multilinguality (e.g., monolingual, bilingual and multilingual), we first empirically confirm the presence of value concepts within LLMs in a multilingual format. Further analysis on the cross-lingual characteristics of these concepts reveals 3 traits arising from language resource disparities: cross-lingual inconsistency, distorted linguistic relationships, and unidirectional cross-lingual transfer between high- and low-resource languages, all in terms of value concepts. Moreover, we validate the feasibility of cross-lingual control over value alignment capabilities of LLMs, leveraging the dominant language as a source language. Ultimately, recognizing the significant impact of LLMs' multilinguality on our results, we consolidate our findings and provide prudent suggestions on the composition of multilingual data for LLMs pre-training.

1 Introduction

Recent years have witnessed the emergence of large language models, such as ChatGPT (OpenAI, 2023a), GPT-4 (OpenAI, 2023b), and LLaMA2 (Touvron et al., 2023). These LLMs have shown powerful capabilities in natural language understanding and generation (Guo et al., 2023; Bang et al., 2023; Jiao et al., 2023; Liu et al., 2024). However, alongside with their prowess, LLMs present potential risks. Research has demonstrated that LLMs can generate responses containing toxic, untruthful, biased, and even illegal content (Cui et al., 2024; Wang et al., 2023; Huang et al., 2023; Shen et al., 2023). Thus, aligning LLMs with human values (i.e., value alignment) is necessary for unleashing their potential safely.

Human values, encompassing concepts like fairness, deontology, utilitarianism, and so on, although challenging to be precisely defined in language, are undoubtedly embedded in textual form (Hendrycks et al., 2021). Recently, Zou et al. (2023a) have introduced Representation Engineering (RepE) to enhance the transparency and controllability of deep neural networks. Through RepE, they unveil that high-level concepts can be extracted as concept vectors from LLMs, utilizing positive and negative text pairs aligned with the directions of specific concepts. These concept vectors, representing the directions of corresponding concepts, can be utilized to assess whether the behavior of LLMs aligns with or to steer their behavior towards the target directions (Zou et al., 2023a; Li et al., 2023; Leong et al., 2023; Liu et al., 2023).

However, existing studies on concept representations in LLMs have primarily focused on English (Zou et al., 2023a), leaving multilingual concepts unexplored. Our work is the first to explore multilingual concepts in LLMs, emphasizing human values-related concepts to advance multilingual AI safety and utility. The primary research questions we aim to answer are as follows: (Q1) *Do LLMs encode concepts representing human values in multiple languages?* (Q2) *To what extent are these concepts consistent and transferable across different languages?* (Q3) *Whether LLMs trained with different distributions of multilingual data exhibit distinct multilinguality in these concepts?* (Q4) *Is Value Alignment of LLMs Controllable across Languages?* To address these questions, we propose a framework consisting of 5 com-

* Corresponding author