Limitations

Our work's major limitation lies in the reliance on translations generated by machine translation for our primary experimental data. A straightforward translation of data related to human values not only introduces translation noise but also overlooks cultural differences. We discuss these two points below.

(1) The noise introduced by machine translations has minimal impact on our research findings. Firstly, our research focuses on the existence of multilingual value concepts in LLMs and their multilinguality, which do not depend on exceptional performance in any specific language. Additionally, we examine across multiple tasks, human values, languages, and LLMs to uncover universal patterns, which contributes to the robustness of our results to a certain degree of noise.

(2) We recognize that cultural variations can result in diverse interpretations of explored values among individuals from different cultural backgrounds. However, our work delves into research questions beyond cultural differences. We primarily focus on the multilingual representations of value concepts with LLMs, their universal cross-lingual patterns, and cross-lingual control over value alignment, aiming to enhance the safety and utility of multilingual AI. Additionally, our proposed framework may also be valuable for studying value disparities. For instance, when applying English concept vectors to other languages for cross-lingual concept recognition, errors in recognition may arise from value disparities between them. We plan to further explore the application of our framework to cultural divergences in our future research.

Ethical Statement

In this paper, we leverage the ETHICS, StereoSet, TruthfulQA, REALTOXICITYPROMPTS, and AdvBench datasets to delve into diverse human values. Despite the presence of negative elements such as unethical, biased, untruthful, toxic, and harmful content within these datasets, our utilization of them is consistent with their intended use. Our approach to cross-lingual value alignment control involves employing the representation engineering methodology to control LLMs' behavior. While experimental results suggest that it is possible to steer LLMs towards generating harmful content, this underscores the applicability of this methodology in red-teaming LLMs to enhance AI safety and in steering LLMs towards producing harmless content in the opposite direction.

Acknowledgements

The present research was supported by the National Key Research and Development Program of China (Grant No. 2023YFE0116400). We would like to thank the anonymous reviewers for their insightful comments.

References

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. *CoRR*, abs/2112.00861.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *CoRR*, abs/2309.16609.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023.

Sunit Bhattacharya and Ondrej Bojar. 2023. Unveiling multilinguality in transformer models: Exploring language specificity in feed-forward networks. *CoRR*, abs/2310.15552.