Firoz Shaik CS 521 30/03/2024

Paper Citation: Aida Ramezani, and Yang Xu. *Knowledge of cultural moral norms in large language models?*

With a focus on two main topics, the research paper investigates how well monolingual English language models can comprehend and forecast moral norms across cultural boundaries. 1) Are these models able to faithfully capture the subtle moral differences across nations on different subjects? 2) Can they convey the similarities and differences in moral opinions on these subjects across cultural boundaries? Pretrained English language models are less successful than English norms at predicting non-English moral norms, according to a study that analyzes data from the World Values Survey and PEW global surveys, which cover 55 and 40 nations, respectively. The models' ability to forecast English norms accurately is compromised when survey data is used to fine-tune them, but cross-cultural moral inference improves as a result. By adjusting models using this survey data, it significantly improves the field's ability to forecast cross-cultural moral norms. This technique offers a feasible way to improve language models' comprehension of global moral diversity while also illuminating the underlying cultural knowledge within them. The research has made a substantial contribution to the development of culturally aware language models, which is highlighted by its rigorous methodology and concentration on a globally inclusive dataset.

The research points out various areas that need work, especially considering the possible biases created during model fine-tuning and the difficulties in effectively representing moral norms seen in non-Western cultures. While fine-tuning the model using worldwide survey data enhances its capacity to infer moral standards across cultural boundaries, it may also introduce new biases and decrease its accuracy in estimating English moral norms. The study also emphasizes the drawbacks of moral rating averages, which tend to oversimplify the range of moral standards that exist within a society. Moreover, it recognizes how challenging it is to determine whether the discrepancy between actual moral judgments and model predictions results from the lack of cultural moral norms in the pre-training data. The research methodology carefully integrates the application of sophisticated machine learning algorithms with large-scale survey data from the World Values Survey and PEW global surveys to investigate the capacity of English language models to comprehend cultural moral norms. The study significantly advances the models' cross-cultural inference capabilities by utilizing fine-tuning procedures on this internationally diversified data. The concept, however, also raises questions about the inherent trade-offs between preserving accuracy in native language situations and boosting cultural sensitivity, as well as the difficulties in fully capturing the range of moral variation both within and between cultures.

The research employs extensive evaluation methodologies that center on the language models’ capacity to precisely deduce moral norms from diverse cultural contexts and to comprehend the global diversity and convergence of moral judgments. To assess how well the models' predictions match empirical moral norms obtained from the World Values Survey and PEW worldwide surveys, the researchers use Pearson correlation tests. By assessing the models' performance across multiple cultural clusters and probing them for fine-grained moral norms, this quantitative analysis is further developed and biases towards Western moral norms are identified. Furthermore, correlation tests are used to evaluate the fine-tuning procedure to investigate the trade-offs between enhancing cultural moral knowledge and the precision of English moral norms. The conclusion, however, also critically examines the drawbacks and compromises that were made, including the possible loss of accuracy about English moral standards and the creation of additional biases. Furthermore, it shows the authors' dedication to ongoing development and ethical consideration in AI research that they have identified the trade-offs and limitations involved in fine-tuning language models for cultural sensitivity and have called for future research to address these challenges.