ponents: extracting multilingual concept vectors from LLMs (§3.1) and evaluating their correlation with the corresponding concepts (concept recognition task in §3.2) to answer Q1; computing cross-lingual similarity of concept vectors (§3.3) and performing cross-lingual concept recognition (§3.4) to answer Q2 and Q3; and manipulating model behavior cross-lingually via concept vectors (§5) to answer Q4.

Our analysis covers 7 concepts related to human values: commonsense morality, deontology, utilitarianism, fairness, truthfulness, toxicity and harmfulness, given their significance for AI safety (Hendrycks et al., 2021; Bai et al., 2022; Askell et al., 2021; Touvron et al., 2023; Yu et al., 2024; Shen et al., 2023; Guo et al., 2023). To ensure the breadth and reliability of our findings, we have selected these 7 concepts for their diverse definitions and ethical attributes (Vida et al., 2023). Throughout this paper, we collectively refer to them as "value concepts" to reflect their diversity and keep consistent with existing AI alignment research (Bai et al., 2022; Askell et al., 2021; Hendrycks et al., 2021). For comprehensive definitions, ethical backgrounds and examples of these value concepts, please refer to Appendix A.

In addition to diverse human values, our experiments involve 16 languages[1] and 3 LLM families with different multilinguality. Specifically, we categorize the multilinguality of these 3 LLM families based on language distributions in their pre-training data into 3 groups: English-dominated LLMs (LLaMA2-chat series in our experiments), Chinese & English-dominated LLMs (i.e., Qwen-chat series), and LLMs with more balanced multilinguality (i.e., BLOOMZ series). Appendix D provides detailed language distributions of their pre-training data.

Through in-depth analysis spanning multiple tasks, value concepts, languages and LLMs, our key findings are as follows:

- LLMs encode concepts representing human values in multiple languages, and the expansion of model size and the richness of language resources both contribute to a more precise capture of these concepts (§4.2).
- The distribution of language resources significantly impacts the cross-lingual properties of these concepts. Specifically, an imbalance in language resources results in cross-lingual inconsistency (§4.3.1), distorted linguistic relationships (§4.3.2), and unidirectional cross-lingual transfer (§4.3.3) between high- and low-resource languages. The cross-lingual properties of value concepts are also intricately tied to the multilinguality of the models to be extracted (§4.3).
- The value alignment of LLMs can be effectively transferred across languages, with the dominant language as a source language (§5.2).

Drawing from these findings, we prudently consider the following suggestions for multilingual pre-training data of LLMs, which might contribute to enhancing multilingual AI safety and utility. First, despite the positive effect of dominant languages as sources for cross-lingual alignment transfer (§5.2), it is crucial to avoid an excessive prevalence of these languages to mitigate unfair cross-lingual patterns, such as inconsistent multilingual representations (§4.3.1), distorted linguistic relationships (§4.3.2), and monotonous transfer patterns (§4.3.3). These traits could potentially amplify the risk of multilingual vulnerability (§5.2) and undermine cultural diversity (Zhang et al., 2023; Cao et al., 2023). Furthermore, we encourage a more balanced distribution of non-dominant languages, particularly those with extremely limited resources, to foster more equitable cross-lingual patterns (§4.3.2 and §4.3.3).

2 Related Work

Representation Engineering Representation Engineering (RepE) introduced by Zou et al. (2023a) extracts abstract concepts as vectors from LLMs using positive and negative samples that describe specific concepts. The effectiveness of these vectors has been validated across dimensions such as correlation and manipulation. Specifically, correlation experiments have assessed the predictive power of the extracted vectors to classify out-of-distribution data as positive or negative, while manipulation experiments have evaluated the vectors' ability to

[1]We recognize that linguistic diversity can foster cultural variations, potentially resulting in diverse interpretations of the same value from different cultural backgrounds (Hershcovich et al., 2022; Hämmerl et al., 2023). For example, regarding deontology, some cultures prioritize individual responsibility while others emphasize social obligations (Cao et al., 2023; Hofstede, 1984). However, our work focuses on the multilingual representations of value concepts within LLMs and their universal cross-lingual patterns, leaving the exploration on cultural divergences in human values for our future research.