them into the model. Similarly, we obtain text representations $\mathcal{R}_c^{\text{test}} = [(\hat{\boldsymbol{r}}_{0+}, \hat{\boldsymbol{r}}_{0-}), (\hat{\boldsymbol{r}}_{1+}, \hat{\boldsymbol{r}}_{1-}), ...]$ by taking the last token's representation of each corresponding text. Furthermore, we calculate the dot product between the previously acquired vector $\boldsymbol{v}_c$ and these text vectors, resulting in classification scores $\mathcal{S}_c^{\text{test}} = [(s_{0+}, s_{0-}), (s_{1+}, s_{1-}), ...]$, where $s_{i\pm} = \boldsymbol{v}_c \cdot \hat{\boldsymbol{r}}_{i\pm}$. The inequality $s_{i+} - s_{i-} = \boldsymbol{v}_c \cdot (\hat{\boldsymbol{r}}_{i+} - \hat{\boldsymbol{r}}_{i-}) > 0$ holding indicates that the direction of $\boldsymbol{v}_c$ aligns with that of the test vector $\hat{\boldsymbol{r}}_{i+} - \hat{\boldsymbol{r}}_{i-}$, signifying a successful concept recognition. We calculate the accuracy of the concept distinction for each concept on the test data as Acc_c:

$$\text{Acc}_c = \frac{\sum_{i=0}^{\hat{N}-1} \mathbb{I}(s_{i+} > s_{i-})}{\hat{N}} \quad \hat{N} = |\mathcal{T}_c^{\text{test}}| \quad (2)$$

A high accuracy ($\text{Acc}_c > \tau$) indicates the presence of a specific value concept in the model.

This process is performed for each language l, resulting in Acc_c^l. The results provide insights into whether the model effectively encodes the value concept c in the context of language l.

Note that each layer has a recognition accuracy, using the concept vector of that layer. Unless specified otherwise, we report the best accuracy.

3.3 Calculating Cross-Lingual Similarity of Concept Vectors

Through calculating cross-lingual similarity of concept vectors, we explore the extent to which LLMs encode consistent representations for the same value concept in different languages, namely, the cross-lingual consistency of multilingual value concepts. Specifically, given two languages l_1 and l_2, we calculate the cosine similarity of their concept vectors $\boldsymbol{v}_c^{l_1}$ and $\boldsymbol{v}_c^{l_2}$. Appendix G.1 highlights the effectiveness of employing cosine similarity to assess the correlation between concept vectors.

3.4 Recognizing Cross-Lingual Concepts

To investigate the cross-lingual transferability of a specific value concept across languages, we propose a method for cross-lingual concept recognition. Given two languages, l_1 and l_2, we calculate how accurately $\boldsymbol{v}_c^{l_1}$ and $\boldsymbol{v}_c^{l_2}$ can be used to recognize the concept c in language l_2, resulting in $\text{Acc}_c^{l_1 \to l_2}$ and $\text{Acc}_c^{l_2}$. The inequality $\text{Acc}_c^{l_1 \to l_2} \geq \text{Acc}_c^{l_2}$ being true signifies the successful transfer of concept c from l_1 to l_2. Conversely, we calculate $\text{Acc}_c^{l_2 \to l_1}$ and $\text{Acc}_c^{l_1}$ to explore the transferability of concept c from l_2 to l_1. While evaluating transferability based solely on accuracy changes might imply a unidirectional transfer from high- to low-performing languages, Appendix H.1 indicates that transferability is not solely determined by language performance.

4 Experiments

We conducted extensive experiments with the proposed framework on 7 human values, 16 languages and 3 LLM families to answer questions Q1, Q2 and Q3. We leave the question Q4 to §5.

4.1 Experimental Setup

Human Value Datasets We explored the following values: commonsense morality, deontology, utilitarianism, fairness, truthfulness, toxicity and harmfulness. We utilized 3 subsets of ETHICS dataset (Hendrycks et al., 2021) for commonsense morality, deontology, and utilitarianism. Regarding fairness, truthfulness, toxicity, and harmfulness, we chose the StereoSet (Nadeem et al., 2021), TruthfulQA (Lin et al., 2022), REALTOXICITYPROMPTS (Gehman et al., 2020), AdvBench (Zou et al., 2023b) dataset, respectively.

Appendix B details the sources, data splits, and positive and negative examples for each value.

Examined Languages and LLMs We translated the aforementioned human value datasets from English into 15 non-English languages using Google Translate. These languages belong to various language families, including Indo-European (Catalan, French, Indonesian, Portuguese, Spanish), Niger-Congo (Chichewa, Swahili), Dravidian (Tamil, Telugu), Uralic (Finnish, Hungarian), Sino-Tibetan (Chinese), Japonic (Japanese), Koreanic (Korean) and Austro-Asiatic (Vietnamese). The impact of translation quality on our results is discussed in Appendix C.

Our experiments involved three multilingual LLM families, including the LLaMA2-chat series (7B, 13B, 70B) (Touvron et al., 2023), Qwen-chat series (1B8, 7B, 14B) (Bai et al., 2023) and BLOOMZ series (560M, 1B7, 7B1) (Scao et al., 2022). Appendix D provides detailed language distributions of their pre-training data. Notably, not all selected languages are included in the pre-training data of these model families. Specifically, both LLaMA2 and BLOOMZ cover 12 of these languages, though their selections do not fully overlap. In contrast, Qwen's technical report only mentions the inclusion of English and Chinese. For the multilingual concept recognition task, we consider all