served when similar noise affects all languages simultaneously. For example, despite the "translationese effect" which could potentially enhance the similarity between non-English texts and English, significant cross-lingual inconsistencies remain between English and other languages in the LLaMA2-chat-7B series, as illustrated in Figure 3.

D Language Distribution

Table 4 displays language distributions of the 16 selected languages (including English) in both the LLaMA2-chat and BLOOMZ series' pre-training data. For the Qwen-chat series, English and Chinese constitute a significant portion of its pre-training data, although detailed language distribution is not publicly accessible.

Based on the language distributions in their pre-training data, we categorize the multilinguality of these 3 LLM families into 3 groups: English-dominated LLMs (LLaMA2-chat series in our experiments), Chinese & English-dominated LLMs (i.e., Qwen-chat series), and LLMs with balanced multilinguality (i.e., BLOOMZ series).

E More Results of Multilingual Concept Recognition

E.1 Extracting Concept Vectors based on PCA

To further enhance the robustness of our results, we also employed the PCA-based method and compared it with the mean-based approach outlined in Section 3.1 (refer to Hämmerl et al. (2023) or Zou et al. (2023a) for details on the PCA-based method). Table 5 presents the multilingual concept recognition accuracy (Section 3.2) for the concept of deontology on LLaMA2-chat-7B. The results suggest that the mean-based method extracts more distinct concept vectors across languages compared to the PCA-based method, consistent with the conclusions of Zou et al. (2023a).

E.2 Varying the Size of $\mathcal{T}_c^{\text{train}}$

We employed varying amounts of training samples to extract concept vectors, and the recognition performance for each human value is illustrated in Figure 5. Surprisingly, optimal accuracy can be achieved for all human values even with few training samples, consistent with the findings by Li et al. (2023), suggesting that the concept vectors for human values are readily extractable in LLMs. Furthermore, we observe notable differences in the recognition accuracy of different human values, indicating different degrees of difficulty in capturing them. Specifically, harmfulness, toxicity, commonsense morality, and deontology are relatively explicitly encoded human values. In contrast, LLMs encounter a greater challenge in recognizing concepts like truthfulness, fairness and utilitarianism.

E.3 Complete Results

Complete results of multilingual concept recognition are provided in Table 9.

E.4 Multilingual Performance Reflects Multilinguality

As shown in Figure 1, the performance distributions of different models across all languages reflect their multilinguality. Specifically, while all three model families perform best in English, the LLaMA2-chat series exhibits significant performance disparities between English and non-English languages. The Qwen-chat series, while excelling at English, also outperforms other languages in Chinese. In contrast, the BLOOMZ series demonstrates the smallest performance gap between English and non-English, reflecting a more balanced multilinguality.

F Computing Pearson Correlation Coefficients Considering Differences in Language Resources

This method begins by categorizing languages into high- and low-resource based on their proportions in the LLM pre-training data. Specifically, for the LLaMA2-chat series, English is designated as a high-resource language, while the remaining languages are considered as low-resource languages. In the case of BLOOMZ series, the low-resource languages include ta, te, sw, and ny, while the rest are considered as high-resource languages. For the Qwen-chat series, en and zh are treated as high-resource languages. We then partition the scores of cross-lingual concept consistency and linguistic similarity among all language pairs into two groups: those between high-resource languages and all languages, and those among low-resource languages themselves. Subsequently, we compute the Pearson correlation coefficients separately for these two sets and report the average result. In this way, imbalance of language distributions between high- and low-resource languages is mitigated when computing the Pearson correlation between cross-lingual concept consistency and linguistic similarity.