	$\geq$ &$\surd$	$\geq$ &$\times$	$<$ &$\surd$	$<$ &$\times$
LLaMA2-chat-7B	27.3%	22.7%	3.0%	47.0%
Qwen-chat-7B	30.3%	19.7%	7.6%	42.4%
BLOOMZ-7B1	34.1%	15.9%	16.7%	33.3%

Table 7: Proportion of cases in which the concept recognition performance of language A either surpasses or underperforms language B, and whether the transfer from language A to language B is effective or not. "$\geq$" and "$<$" denote superiority and inferiority respectively, and "$\surd$" and "$\times$" represent successful and unsuccessful transfer.

		en	**zh**	**fr**	**es**	**pt**	**vi**	**ca**	**id**	**avg**
LLaMA2 -chat	7B	0	14	28	28	14	14	57	85	30
	13B	0	14	57	42	42	71	57	100	47
	70B	0	71	14	28	28	85	71	85	47
Qwen -chat	1B8	0	0	42	14	28	100	85	28	37
	7B	14	14	57	0	71	42	71	71	42
	14B	14	14	57	14	57	85	57	71	46
BLOOMZ	560M	14	14	100	0	57	85	14	100	48
	1B7	85	42	71	42	42	100	0	85	58
	7B1	100	14	100	71	57	100	42	85	71

Table 8: Proportions of different languages as targets of cross-lingual concept transfer. The displayed languages are those included both in LLaMA2-chat and BLOOMZ series' pre-training data.

G.2 Complete Results

Cross-lingual concept consistency of all models is presented in Figure 6.

G.3 Effect of Model Size

Despite larger models being able to capture more explicit concepts of human values (as shown in Figure 1 & **??**), the increase in model size does not steadily enhance cross-lingual concept consistency, as shown in Figure 2b.

H More Results of Cross-Lingual Concept Transferability

H.1 Transferability Beyond Language Performance

While the setting described in Section 3.4 may introduce bias of initial performance variations across languages, potentially leading to mono-directional transfer from high-performing languages to low-performing ones, our findings suggest that transferability is not solely determined by language performance, as detailed below.

Specifically, we calculated the proportion of cases where the concept recognition performance of language A either surpasses or underperforms language B, and whether the transfer from language A to language B is effective or not. The results are summarized in the Table 7, where "$\geq$" and "$<$" denote superiority and inferiority respectively, and "$\surd$" and "$\times$" represent successful and unsuccessful transfer. While effective transfers are mostly from languages with better performance (comparing the 1st and 3rd columns in the table, e.g., LLaMA2-chat-7B, 27.3% vs 3.0%), a comparison between the 1st and 2nd columns reveals that superior concept representations in language A do not necessarily ensure effective transfer to language B (e.g., LLaMA2-chat-7B, 27.3% vs 22.7%). Moreover, the results of BLOOMZ-7B1 further support this. For example, in comparison to the 1st column of BLOOMZ-7B1 ("$\geq$ &$\surd$" at 34.1%), reverse transfer from low-performing languages to high-performing languages also accounts for a considerable proportion (the 3rd column, "$<$ &$\surd$" at 16.7%). Notably, combining the results from Figure 1 and Figure 4 in the main content, it is evident that although BLOOMZ-7b1 encodes the most explict concepts in English, effective transfer from English to other languages is challenging.

In summary, although evaluating transferability based solely on changes in accuracy may pose limitations, the phenomenon that transfer is not solely determined by language performance indicates that this remains an open question. We plan to develop more robust and unbiased methodologies to further investigate cross-lingual transfer in our future research.

H.2 Complete Results

Cross-lingual concept transferability of all models is presented in Figure 7.

H.3 Effect of Multilinguality and Model Size

Table 8 provides a breakdown of the proportions of different languages as targets of cross-lingual concept transfer[7], providing a clearer illustration of the unidirectional transfer from dominant languages in LLaMA2- and Qwen-chat series. Conversely, the BLOOMZ series demonstrates a more balanced transfer pattern, showcasing a distinctly superior level of cross-lingual concept transferability.

Furthermore, Table 8 reveals that increasing the model size consistently improves in cross-lingual concept transferability, except for cases of LLaMA2-chat-13B and 70B, where similar levels of cross-lingual transfer are observed.

[7] If $\mathrm{Acc}^{l_1 \rightarrow l_2} \geq \mathrm{Acc}^{l_2}$, l_2 is considered as a target of the concept transfer between the two languages.