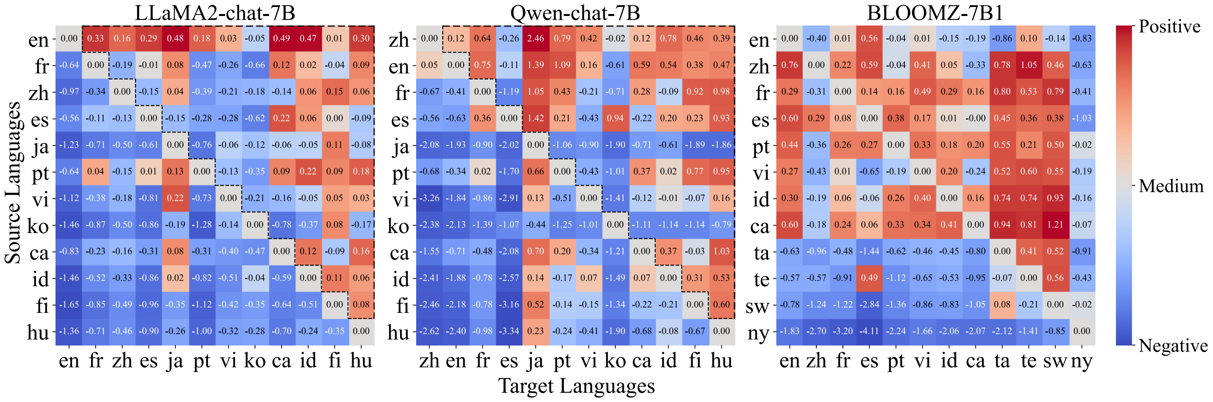

Figure 4: Cross-lingual concept transferability across all language pairs, averaged over all value concepts. Languages are sorted based on their percentages in the pre-training data.

ing with previous findings on multilingual factual knowledge (Qi et al., 2023).

4.3.3 Trait 3: Unidirectional Concept Transfer from High- to Low-Resource Languages

For a given source language l_1 and target language l_2, we compute $\text{Acc}_c^{l_1 \rightarrow l_2} - \text{Acc}_c^{l_2}$ (the difference in accuracy scores) to measure the transferability of concept c from l_1 to l_2 (§3.4). We average differences in accuracy scores over all value concepts to measure the overall transferability. If the average difference is greater than 0, it indicates positive transferability from l_1 to l_2.

We present the cross-lingual concept transferability of the three 7B-sized models in Figure 4. It provides insights into the influence of LLMs' multilinguality. Firstly, based on the results of LLaMA- and Qwen-chat-7B, we observe a pattern of monotonic concept transfer from the dominant languages to other languages. This pattern also exhibits an upper triangular cross-lingual transferability (the dashed triangular in Figure 4), indicating that cross-lingual concept transfer from high- to low-resource languages is more prevalent. In contrast, BLOOMZ-7B1 exhibits a relatively balanced bidirectional cross-lingual concept transferability, while for languages with extremely low resources, the tendency of unidirectional transfer persists.

While evaluating transferability based solely on changes in accuracy may introduce biases due to initial performance variations across languages, potentially amplifying the observed unidirectional transfer, Appendix H.1 indicates that transferability is not solely determined by language performance. For comprehensive results on each value concept and further discussions, please refer to Appendix H.2 and H.3.

5 Q4: Is Value Alignment of LLMs Controllable across Languages?

LLaMA2-chat models, trained with alignment techniques such as RLHF, exhibit value alignment capabilities like rejecting harmful instructions. In this section, we employed the Representation Engineering (RepE) methodology (Zou et al., 2023a) to bypass such defense and further explored the potential for cross-lingual control of value alignment.

5.1 Cross-Lingual Value Alignment Control

To control a LLM to exhibit behavior aligned with a value concept c, a straightforward RepE-style method is multiplying the previously extracted concept vector $\boldsymbol{v_c}$ by a control strength s and adding it to the hidden states of multiple layers L within the target model. This procedure is iteratively applied to each token, formulated as $\boldsymbol{h_i^{'}} = \boldsymbol{h_i} + s \cdot \boldsymbol{v_c}$, where $\boldsymbol{h_i}$ and $\boldsymbol{h_i^{'}}$ denote the original and perturbed hidden state of i-th token, respectively.[5] In a cross-lingual scenario, we leverage the concept vector $\boldsymbol{v_c^l}$ of the source language l to control the model's behavior across various target languages. To determine appropriate control strength s and control layers L for cross-lingual control, we first conduct hyperparameter search to choose the combination that demonstrates the most effective control on language l. Subsequently, we employ this combination for cross-lingual control across all target languages and evaluate the control effect on each of them.

In our experiments, a successful control is steering the LLM to follow a harmful instruction rather

[5] Reflecting on §3.1, each layer has its specific concept vector, and the perturbation is executed across multiple layers L. We omit the detail here for simplicity.