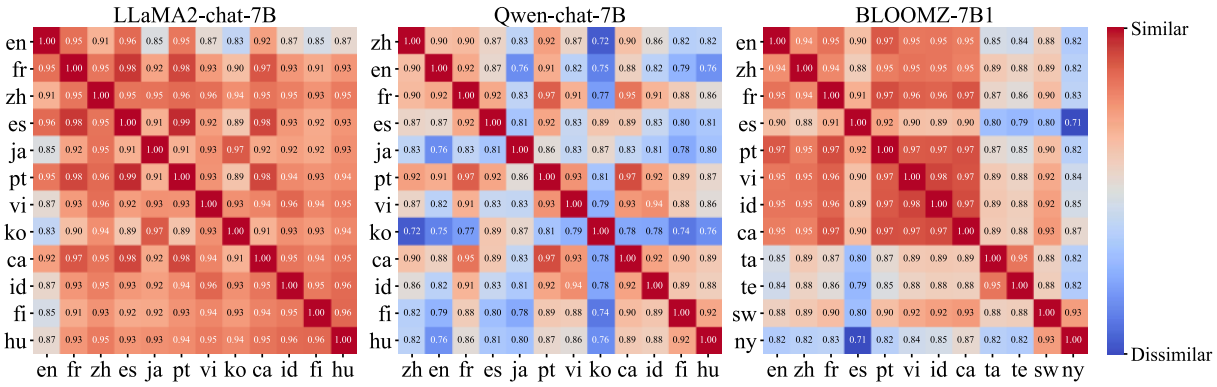

Figure 3: Cross-lingual similarity of concept vectors across all language pairs, averaged over all value concepts. The languages included in each model's pre-training data are presented and sorted based on their proportions in the corresponding model's pre-training data. For Qwen-chat series, we conjecture its language inclusion based on multilingual concept recognition accuracy (§4.2) and display its primary languages, zh and en, at the forefront.

Appendix E.1 compares the PCA-based method with the mean-based method outlined in §3.1. It reveals that both methods produce concept vectors of comparable precision, with the mean-based technique holding a slight edge. The consistent performance across various extraction techniques confirm the effectiveness of concept vectors in capturing conceptual information. Appendix E.2 demonstrates that even a small number of training samples can effectively extract representations of value concepts in LLMs. For detailed results on each value concept and additional discussions, please refer to Appendix E.3 and E.4.

4.3 Q2 & Q3: How Consistent and Transferable are Value Concepts across Languages, and What is the Impact of LLMs' Multilinguality?

Through computing cross-lingual similarity of concept vectors (§3.3) and recognizing cross-lingual concepts (§3.4), we investigated the cross-lingual consistency and transferability of these value concepts (Q2). Moreover, analyzing these concepts on LLMs trained with different multilingual data distributions provides insights into the multilinguality of LLMs (Q3).

4.3.1 Trait 1: Inconsistency of Concept Representations between High- and Low-Resource Languages

Figure 3 illustrates the cross-lingual similarity of concept vectors captured by the three 7B-sized models. We find that different multilinguality leads to different patterns of cross-lingual concept consistency. In the case of LLaMA2-chat-7B, the absolute dominance of English results in the model learning relatively independent concept representations for English, showing concept representation inconsistency between English and other languages, while higher cross-lingual concept consistency is observed among other languages. BLOOMZ-7B1's cross-lingual concept consistency exhibits a very different pattern: the four languages with the lowest proportions (ta, te, sw, ny, accounting for 0.50%, 0.19%, 0.015%, and 0.00007% of pre-training data, respectively) show the lowest concept consistency (similarity) with other languages, while languages with relatively higher proportions (en with the highest percentage of 30.04%, and ca with the lowest percentage of 1.10%) demonstrate higher concept consistency with each other.[2] For Qwen-chat-7B, we do not observe significant cross-lingual consistency between the main languages (zh, en) and other languages. In summary, cross-lingual concept inconsistency is more likely to occur between high- and low-resource languages.

Additionally, Figure 2b illustrates the trends in cosine similarity across different model layers. We observe that the peak of cross-lingual consistency appears in the intermediate layers, with lower similarity near the input and output layers. This observation is consistent with previous research (Chi et al., 2021; Bhattacharya and Bojar, 2023), suggesting that middle layers of multilingual models encode a higher degree of language-independent information, while language-specific information is more prominent near the input and output layers.

The findings from Steck et al. (2024) suggest that

[2]We observe inconsistency between Spanish and other languages in BLOOMZ-7B1. We would like to explore this in our future work.