	en	**fr**	**zh**	**es**	**pt**	**vi**	**ca**	**id**	**ja**	**ko**	**fi**	**hu**	**Avg**
mean	**97.5**	90.2	**91.0**	**91.7**	**92.0**	84.9	**90.2**	**86.4**	**87.4**	**82.7**	**83.4**	**81.4**	**88.2**
pca	96.7	92.7	90.7	91.7	89.2	85.9	90.2	83.2	86.9	80.7	82.2	81.2	87.6

Table 5: Comparison of multilingual concept recognition accuracy between PCA-based and mean-based concept extraction methods.

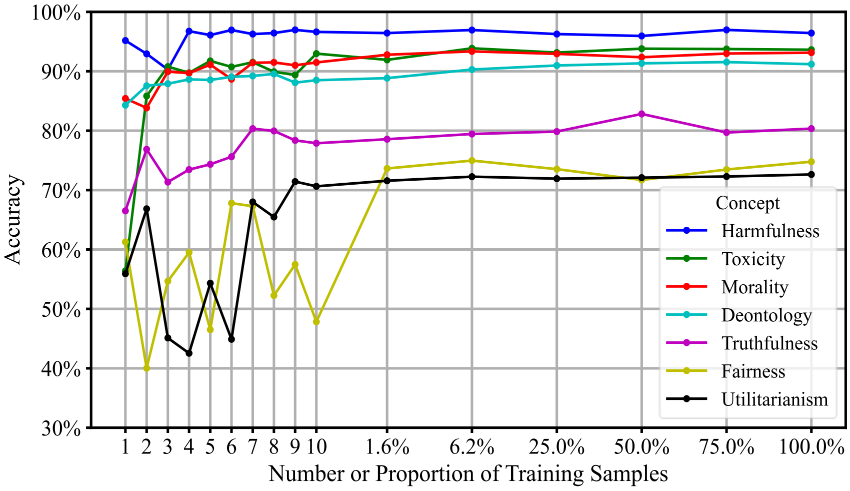

Figure 5: English concept recognition accuracy with varying numbers of training samples for collecting concept vectors. The result are based on LLaMA2-chat-13B. We calculate the average accuracy across all layers to ensure the results of different settings are comparable.

	same	**different**
LLaMA2-chat-7B (en-en)	1.00	0.56
Qwen-chat-7B (en-en)	1.00	0.49
BLOOMZ-7B1 (en-en)	1.00	0.49
LLaMA2-chat-7B (en-fr)	0.95	0.54
Qwen-chat-7B (en-fr)	0.92	0.44
BLOOMZ-7B1 (en-fr)	0.95	0.53

Table 6: Cosine similarity between concept vectors representing either the same or different values across languages.

G More Results of Cross-Lingual Concept Consistency

G.1 Cosine Similarity between Concept Vectors can Reflect Their Correlation

Steck et al. (2024) discussed the limitations and potential issues with using cosine similarity as a measure of semantic similarity, particularly in the context of embeddings learned from linear models. They highlight that cosine similarity can sometimes produce arbitrary and non-unique results, implying that a high average cosine similarity might raise concerns when dealing with unrelated representations.

In our paper, cosine similarity is calculated on concept vectors across different languages to measure their consistency. It is worth recalling that these concept vectors are computed by averaging a set of difference vectors. This averaging process inherently filters out irrelevant information to some extent, thereby mitigating the unpredictable impact on cosine similarity results.

Furthermore, we attempt to evaluate the effectiveness of cosine similarity outcomes in our specific context. Specifically, we compute the cosine similarity between concept vectors of different values in English (e.g., $cosine(\boldsymbol{v}_{c1}^{en}, \boldsymbol{v}_{c2}^{en})$) and cross-lingually between English (en) and French (fr) for both the same (e.g., $cosine(\boldsymbol{v}_{c1}^{en}, \boldsymbol{v}_{c1}^{fr})$) and different (e.g., $cosine(\boldsymbol{v}_{c1}^{en}, \boldsymbol{v}_{c2}^{fr})$) human values. The averaged results presented in Table 6 indicate that, compared to the same human values, the concept representations of unrelated human values exhibit significantly lower cosine similarity. This observation holds true both within a single language and across languages. These findings suggest that, at least in our context, high cosine similarity tends to indicate high relevance, while low cosine similarity often signifies irrelevance to a considerable extent.