Language	ISO 639-1	Language Family	LLaMA2 Ratio(%)	BLOOMZ Ratio(%)
English	en	Indo-European	89.70	30.04
French	fr	Indo-European	0.16	12.90
Chinese	zh	Sino-Tibetan	0.13	16.17
Spanish	es	Indo-European	0.13	10.85
Portuguese	pt	Indo-European	0.09	4.91
Vietnamese	vi	Austro-Asiatic	0.08	2.71
Catalan	ca	Indo-European	0.04	1.10
Indonesian	id	Austronesian	0.03	1.24
Japanese	ja	Japonic	0.10	-
Korean	ko	Koreanic	0.06	-
Finnish	fi	Uralic	0.03	-
Hungarian	hu	Uralic	0.03	-
Tamil	ta	Dravidian	-	0.49
Telugu	te	Dravidian	-	0.19
Swahili	sw	Niger-Congo	-	0.01
Chichewa	ny	Niger-Congo	-	0.00007

Table 4: Language distributions of the 16 selected languages (including English), for LLaMA2-chat and BLOOMZ series. Languages ta, te, sw and ny are not included in the pre-training data of LLaMA2-chat series, and languages ja, ko, fi and hu are not included in the pre-training data of BLOOMZ series.

term or unrelated term. We inserted each of these three terms into the blank to form different complete sentences. In the intersentence class, each sentence containing a target term is followed by three associative sentences representing stereotypical, anti-stereotypical, and unrelated associations. We concatenated the preceding and subsequent three types of sentences to form different complete sentences. We only employed pairs of stereotypical and anti-stereotypical sentences to obtain positive and negative samples for this human value.

Truthfulness We used the TruthfulQA dataset (Lin et al., 2022), which consists of two tasks: generation and multiple-choice. Specifically, in the generation task, questions are accompanied by correct or incorrect responses. In the multiple-choice task, questions are accompanied by a set of candidate answers, some of which are correct and others incorrect. We concatenated the question and its corresponding correct response or answer as a positive example while the same question with its corresponding incorrect response or answer as a negative example.

Toxicity We utilized REALTOXICITYPROMPTS dataset (Gehman et al., 2020) consisting of naturally occurring prompts sampled from English web text and corresponding toxicity scores. We categorized prompts into non-toxic and toxic ones based on the scores, thereby forming positive and negative pairs.

Harmfulness We utilized the AdvBench dataset (Zou et al., 2023b) which contains harmful instructions eliciting LLMs to generate objectionable content. These harmful instructions are further combined with harmless instructions to form negative and positive pairs, as described in the work of Zou et al. (2023a).

After collecting and formatting these datasets, we divided each dataset of human values into the training and testing sets in an 8:2 ratio. The training set is used for obtaining concept vectors, as discussed in Section 3.1, while the testing set is employed for experiments, such as concept recognition in Section 3.2 and model control in Section 5. Table 3 presents the number of training and testing samples, as well as positive and negative examples of each human value.

C Impact of Translation Quality

Our primary experimental data rely on translations yielded by translation engines. However, the noise introduced by these translations has minimal impact on our research findings. Our exploration of universal cross-lingual characteristics in LLMs, such as cross-lingual consistency and transferability, suggests that overall patterns are likely pre-