		en	fr	zh	es	pt	vi	ca	id	ja	ko	fi	hu	Avg
LLaMA2 -chat-7B	No-Control	0.97	1.94	6.80	1.94	6.80	4.85	8.74	5.83	3.88	10.68	14.56	4.85	6.44
	LS-Control	97.09	99.03	95.15	99.03	97.09	97.09	90.29	98.06	97.09	100.0	99.03	99.03	97.35
	En-Control	97.09	94.17	94.17	97.09	91.26	96.12	**91.26**	88.35	**99.03**	95.15	95.15	91.26	93.91
LLaMA2 -chat-13B	No-Control	0.97	0.97	5.83	1.94	5.83	5.83	27.18	8.74	2.91	10.68	15.53	6.80	8.38
	LS-Control	88.35	99.03	97.09	98.06	99.03	98.06	98.06	100.0	98.06	97.09	98.06	100.0	98.41
	En-Control	88.35	**99.03**	95.15	**98.06**	97.09	**98.06**	93.20	94.17	**99.03**	**97.09**	90.29	87.38	95.32
LLaMA2 -chat-70B	No-Control	0.00	1.94	4.85	0.97	6.80	2.91	27.18	11.65	2.91	20.39	18.45	10.68	9.89
	LS-Control	74.76	87.38	68.93	55.34	90.29	79.61	98.06	92.23	63.11	84.47	95.15	96.12	82.79
	En-Control	74.76	**95.15**	**70.87**	**92.23**	79.61	**95.15**	63.11	73.79	**92.23**	74.76	72.82	63.11	79.35

Table 2: Following rates on LLaMA2-chat series under different control methods. "No-Control": no control is applied; "LS-Control": language-specific control with each language controlling itself; "En-Control": cross-lingual control with English as the source language. "Avg" denotes the average results excluding English.

than rejecting it. We compute the Following rate, representing the proportion of harmful instructions the model follows, to assess the effectiveness of model control. Specifically, we utilize the multilingual negative testing data (harmful instructions) for the concept of harmfulness (§4.1), calculating the Following rate in each language. Please refer to Appendix I for details of hyperparameter search and model control evaluation.

5.2 Results

Cross-lingual value alignment control results are presented in Table 2. First, without applying any control (No-Control), LLaMA2-chat series refrains from responding to almost all harmful instructions in English. However, simply translating these prompts into other languages partially circumvents the models' defense, exposing LLMs' multilingual vulnerability (Deng et al., 2023; Shen et al., 2024; Yong et al., 2023). Surprising, we observe larger models are more prone to responding to non-English harmful instructions, potentially due to their enhanced instruction-following capabilities.

Second, we discover that cross-lingual control from English to other languages (En-Control) can achieve control effectiveness comparable to that of LS-Control. While LS-Control achieves performance through language-specific optimization of hyperparameters, En-Control simply adopts hyperparameters found in English, highlighting the ease of achieving cross-lingual control with English as a source language in English-dominated LLMs.

6 Discussions and Suggestions

Drawing our empirical observations and findings, we prudently consider the following suggestions for the configuration of multilingual pre-training data for LLMs, which might contribute to enhancing multilingual AI safety and utility. First, despite the positive effect of dominant languages as sources for cross-lingual alignment transfer (§5.2), it is essential to avoid an excessive prevalence (exemplified by LLaMA2's pre-training data, which comprises about 90% English data). Our analysis suggests that such excessive dominance can lead to unfair cross-lingual patterns, manifested as inconsistent multilingual representations (§4.3.1), distorted linguistic relationships (§4.3.2), and monotonous transfer patterns (§4.3.3). These tendencies could potentially further amplify the risk of multilingual vulnerability (§5.2) and undermine cultural diversity (Zhang et al., 2023; Cao et al., 2023). Furthermore, we encourage a more balanced distribution of non-dominant languages, particularly those with extremely limited resources, to foster more equitable cross-lingual patterns (§4.3.1 and §4.3.3).[6]

7 Conclusion

We have presented a systematic exploration of multilingual concepts embedded in LLMs, focusing specifically on human value-related concepts (i.e., value concepts). Through our extensive analysis spanning 7 human values, 16 languages, and 3 LLM families, we have obtained many interesting findings. Specifically, we empirically verify the presence of multilingual value concepts in LLMs and identify the cross-lingual characteristics of these concepts arising from language resource disparities. Furthermore, our experiments on cross-lingual control illuminate the multilingual vulnerability of LLMs, as well as the feasibility of cross-lingual manipulation over value alignment of LLMs. With these findings, we prudently present several suggestions for collecting multilingual pre-training data for advanced multilingual AI.

[6]These suggestions are based on our findings, which might be biased by factors like variations in language performance (§3.4) and other unobserved ones.