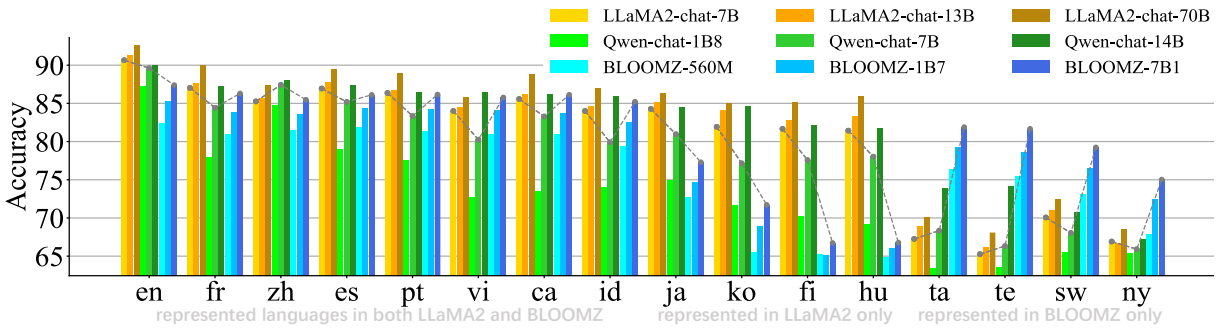

Figure 1: Multilingual concept recognition accuracy (%) of LLaMA2-chat, Qwen-chat and BLOOMZ series, averaged across all value concepts. The performance of the three 7B-sized models are connected with dashed lines for performance comparison. "Represented languages" refer to the languages present in the pre-training corpus.

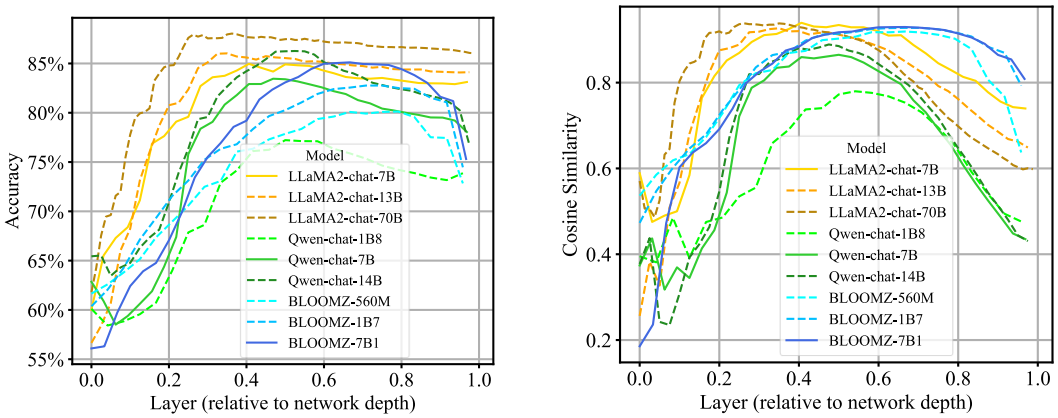

Figure 2: (a) Multilingual concept recognition accuracy across different model layers. (b) Cross-lingual similarity of concept vectors across different model layers. Results are averaged across languages included both in LLaMA2-chat and BLOOMZ series' pre-training data, as well as across all human values.

16 languages, regardless of the model series, while other tasks explore only the languages covered in the pre-training data.

4.2 Q1: Do LLMs Encode Concepts Representing Human Values in Multiple Languages?

Figure 1 illustrates the multilingual concept recognition accuracy of the three LLM families, averaged across all value concepts. We first observe that all three models achieve notable accuracy across all represented languages and even the smallest models surpass $\tau = 65\%$ accuracy in them. It's important to note that the accuracy of 65% is a conservative statistic and represents a lower bound, derived from the smallest model (BLOOMZ-560M) on the poorest-performing language (ny, accounting for only 0.00007% in pre-training data). However, results from larger models are significantly higher. For example, BLOOMZ-7B1 achieves accuracy exceeding 81% on the majority of seen languages (10 out of 12). In addition to BLOOMZ-7B1, other model families with equivalent model sizes also demonstrate similarly high performance. Overall, these results confirm that LLMs effectively encode value concepts in a multilingual context.

We also observe a certain level of recognition accuracy in some unrepresented languages. We conjecture that the ability of models in capturing these languages may stem from cross-lingual transfer from other languages. Additionally, as mentioned in Section 4.1, Qwen's technical report only mentions the inclusion of en and zh in its pre-training data. We conjecture the inclusion of 10 other languages (fr,es,pt,vi,ca,id,ja,ko,fi,hu) based on its significant performance in these languages.

Although previous results represent the best performance across all layers, Figure 2a presents the concept recognition accuracy across different model layers. We observe that middle layers encode more abstract information related to human values, aligning with the findings of Li et al. (2023).