alignment are larger for prompt-based embeddings. Both things suggest that some of the implicitly learned concept alignment is broken by the prompt-based method. On the other hand, prompt-based embeddings demonstrate larger improvements with explicit post-hoc alignment while supervised alignment struggles to improve on vanilla word embeddings.

Difference in Languages The degree of isomorphism to English is similar across languages, as indicated by the upper bounds on performance. All concept spaces are (almost) equally alignable. However, the induced maps generalize much better across typologically related (Indo-European) languages: French and Romanian. Generalization is considerably poorer for the other two groups.

Types of Concepts Though previous works show that physical concepts do better than abstract ones in bilingual dictionary induction (Kementchedjhieva et al., 2019), as well as in related tasks such as hypernym detection (Liao et al., 2023), we show that abstract concepts tend to align better across different languages, as shown in Table 2. This, however, was explained by a spurious correlation with frequency. It would be interesting to control for frequency in future error analysis.

4 Conclusion

We evaluated concept alignment on multilingual LLMs by revisiting the traditional bilingual dictionary induction task, but with semantic concepts rather than words. Our experiments show that multilingual LLMs exhibit high-quality, linear concept alignment across different languages. However, the ability of supervised maps to generalize varied across different models, languages, and ways of obtaining embeddings.

Limitations

Because of the small overlap between multilingual WordNets, we only include six (6) test languages. While this is too small a set of languages to draw universally applicable conclusions. Fortunately, the set includes both Indo-European and non-Indo-European languages, as well as both Latin script and non-Latin script languages. We also limited ourselves to studying nouns; for how linear alignment generalizes to other parts of speech, see Kementchedjhieva et al. (2018) and Hartmann and Søgaard (2018).

Ethical Considerations

We do not anticipate any risks in the work. In this study, our use of existing artifacts is consistent with their intended purposes. Semantic concepts are collected from previously published and publicly available resources (WordNets). Aya101[7], BLOOMZ, and mT0 models have Apache-2.0 License[8]. Llama2 models are under the LLAMA 2 Community License[9].

Acknowledgement

We would like to thank all anonymous reviewers for their insightful comments and feedback. This work was supported by DisAI - Improving scientific excellence and creativity in combating disinformation with artificial intelligence and language technologies, a project funded by European Union under the Horizon Europe, GA No. 101079164.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. 2009. Enhancing the japanese wordnet. In *Proceedings of the 7th Workshop on Asian Language Resources (ALR7)*, pages 1–8.

Francis Bond, Piek Vossen, John Philip McCrae, and Christiane Fellbaum. 2016. Cili: the collaborative interlingual index. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 50–57.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, et al. 2024. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. *arXiv preprint arXiv:2401.07037*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

[7] https://huggingface.co/CohereForAI/aya-101#model-summary

[8] https://github.com/bigscience-workshop/xmtf/blob/master/README.md

[9] https://ai.meta.com/llama/license/