ilarly, a larger model size and larger seed dictionary generally improve the concept alignment. On Group 2 and 3, mT0 and Aya101 show better before-align performance compared to other models. In some cases, results are extremely good. Llama2-13B with prompt-based embeddings exhibits a P@1 score of 59.27% before alignment for French, for example. This means that the model has induced perfect alignment of 3/5 concepts in the absence of any explicit supervision. It is interesting to see the gap between the red and black dashed lines. The size of this gap indicates how much of the (alignable part of the) concept space was *not* aligned, with given seed dictionary. For vanilla word embeddings, the gaps are relatively small, but for prompt-based embeddings the gaps tend to be much larger, again indicating that prompting somewhat breaks the implicitly learned concept alignment.

Abstract vs. Physical We analyze performance differences across abstract and physical concepts. To make a fair comparison, we randomly down-sample[6] physical concepts and compare retrieval performance across the two classes. In this section, we report P@1 with models that have comparable model sizes (7B/13B) in each family; results for the other models can be found in the Appendix. As shown in Table 2, all models generally have better alignment performance on abstract concepts compared to physical concepts.

		fr	ro	eu	fi	ja	th
Llama2-13B	Abstract	**63.48**	**46.06**	**17.42**	**21.00**	**26.01**	**2.15**
	Physical	50.12	33.41	14.08	18.62	23.39	1.91
BLOOMZ-7B1	Abstract	**64.92**	**33.41**	**27.92**	**10.74**	**10.26**	**1.19**
	Physical	52.51	27.45	18.62	6.44	4.53	0.00
mT0-xxl (13B)	Abstract	**59.90**	**49.88**	**7.88**	**38.90**	**38.42**	**34.37**
	Physical	46.78	41.29	5.73	37.23	36.28	28.64
Aya101 (13B)	Abstract	**58.47**	**52.27**	**27.68**	**40.81**	**36.28**	**32.70**
	Physical	44.63	36.75	18.38	30.79	26.73	29.12

Table 2: The results (P@1) for abstract and physical concepts. We report after-align results for prompt-based embedding and comparable sizes (13B/7B) of each model family.

What explains this very consistent finding? One hypothesis would be that physical nouns are more ambiguous, since they often source metaphor and metonymy. However, our words for abstract concepts have more senses in WordNet (2.94) than our words for physical concepts (1.96); see Table 3. Instead, we found another, simpler explanation. Frequency statistics (obtained from the English Wikipedia dump of 2023-04-13) relevant that the abstract concept words are considerably more frequent than the physical concept words, which makes sense, as abstract concepts apply very generally across contexts and domains.

	Abstract	Physical
avg # of senses	2.94	1.96
median # of senses	2	1
avg # of counts	103,934	28,762
median # of count	12,787	5,122

Table 3: Number of senses and frequency of words.

3 Discussion and Related Work

Related Work The idea that distributional representations facilitate cross-lingual alignment goes back to explicit semantic analysis (Gabrilovich and Markovitch, 2007), but the idea of training multilingual, neural language models also has a long history. Such models have traditionally used explicit alignment objectives, e.g., either from word alignments, bilingual dictionary seeds (Lample et al., 2018; Li et al., 2024), or by training on mixed corpora constructed using such resources (Gouws and Søgaard, 2015; Workshop et al., 2022; Chai et al., 2024). Cross-lingual generalization has been studied in different NLP tasks, including question answering (Artetxe et al., 2020), commonsense reasoning (Ponti et al., 2020; Lin et al., 2022), code generation (Peng et al., 2024), and knowledge transfer and consistency (Xu et al., 2023; Qi et al., 2023). Cross-lingual word alignment also has a long history by examining bilingual lexicon induction (Xing et al., 2015; Søgaard et al., 2018; Li et al., 2023). For concept understanding specifically, previous works have examined concept understanding in LLMs by definition matching (Xu et al., 2024), hypernym/hyponym detection (Liao et al., 2023; Shani et al., 2023), and relation discovery (Gu et al., 2023). However, they are limited to the English language only.

Linear Alignment We saw that concepts are represented in similar ways across languages in multilingual LLMs, as shown in the upper bound. This indicates structural similarities and facilitates cross-lingual transfer. Prompt-based embeddings exhibit significantly lower linearity compared to word embeddings, and the gaps between before and after

[6] See Appendix for numbers without down-sampling.