embeddings. This suggests that some of the implicitly learned concept alignments are broken by prompt-based methods. Prompt-based embeddings, which are now commonly used in different retrieval scenarios, seem to be less effective in extracting cross-lingually alignable embeddings, compared to vanilla embeddings. Results are generally good, but the old problem of generalization across typological distance (Singh et al., 2019) rears its ugly face again, with Basque, Finnish, Japanese and Thai exhibiting generally lower overall performance for both experimental set-ups. Furthermore, abstract concepts exhibit better alignment than physical concepts. We suspect that it is because abstract concepts are more frequent and occur in more diverse contexts.

2 Experiments

Concepts We collect English noun synsets from WordNet (Miller, 1995). For each synset, its first (most frequent) lemma name is used as the surface form of the corresponding concept. We use WordNet's hierarchical structure to filter out top-level concepts (top-5 levels) to avoid too general concepts. WordNets in other languages, such as French WordNet (Sagot and Fišer, 2008), Basque WordNet (Gonzalez-Agirre et al., 2012), or Romanian WordNet (Dumitrescu et al., 2018), have similar structure and were all aligned in the Open Multilingual WordNet project (OMW) (Bond et al., 2016). To produce a repository of parallel semantic concepts, we collect synsets with shared ID across different WordNets, after removing duplicate concepts. In total, we obtain 4,397 parallel concepts across 7 different languages (English, French, Romanian, Basque, Finnish (Lindén and Niemi, 2014), Japanese (Bond et al., 2009) and Thai (Thoongsup et al., 2009)); had we included more languages, the number of parallel concepts would have been prohibitively small. The 4,397 concepts were divided into abstract (e.g., happiness) and physical (e.g., vehicle) concepts. See Table 1 for data characteristics, and Figure 1 for examples of parallel concepts.

	Abstract	Physical	Total
Train	1500	1500	3000
Test	419	978	1397
Total	1919	2478	4397

Table 1: The statistics of the parallel concept dataset. We use 1000, 2000, or 3000 concepts for training.

To create a **seed dictionary** (training data) for supervised alignment, we randomly sample 3,000 parallel concepts,[1] including 1,500 abstract concepts and 1,500 physical concepts. The 3,000 concepts are used to induce the linear mapping.

LLMs We experiment with four different LLM families with varying sizes: Llama2 (7B, 13B, 70B) (Touvron et al., 2023), mT0 (1.2B, 3.7B, 13B), BLOOMZ (1B7, 3B, 7B1) (Muennighoff et al., 2022), and Aya101 (13B) (Üstün et al., 2024). We use *two different* concept space extraction methods (vanilla and prompt-based). The vanilla method simply uses the last token representation as the concept embedding for decoder-only models (Llama2 and BLOOMZ); and the average embedding of the last hidden layer of the encoder as the concept embedding for encoder-decoder models (mT0 and Aya101)[2]. The prompt-based extraction method exploits the fact that all these models were instruction-tuned. The template we use for prompt-based extraction is adapted from Li and Li (2023) and shown as follows:

Summarize concept [text] in one [lang] word:

where [text] and [lang] will be replaced by the corresponding concept (in the source language) and the language name (in adjectival form), e.g.,"summarize concept "動物" in one Japanese word" for the concept *animal*[3]. The prompt-based concept embedding is that of the last hidden state.

Alignment and Retrieval We rely on Procrustes Analysis (Schönemann, 1966), a form of statistical shape analysis, to discover good linear transformations (e.g., translation, rotation, and scaling) between concept spaces in different languages. Suppose X and Y are two matrices of size $n \times d$ (n is the seed dictionary size, d is the embedding size) such that the ith row in X is an embedding of concept c_i in one language, and the ith row in Y is c_i's embedding in the other language. The linear transformation is derived through singular value decomposition (SVD) of YX^T:

$$W^* = \underset{W \in O_{d(\mathbb{R})}}{\arg\min} ||WX - Y||_F = UV^T \quad (1)$$

where $U\Sigma V^T = \text{SVD}(YX^T)$. With W^*, we transform source language concept embeddings X into

[1] See Appendix for results with 1,000 or 2,000 concepts.
[2] This is decided by preliminary experiment results.
[3] The [text] of a concept can be made up of multiple words.