Concept Space Alignment in Multilingual LLMs

Qiwei Peng and **Anders Søgaard**
University of Copenhagen
Denmark
{qipe, soegaard}@di.ku.dk

Abstract

Multilingual large language models (LLMs) seem to generalize somewhat across languages. We hypothesize this is a result of implicit vector space alignment. Evaluating such alignment, we see that larger models exhibit *very* high-quality linear alignments between corresponding concepts in different languages. Our experiments show that multilingual LLMs suffer from two familiar weaknesses: generalization works best for languages with similar typology, and for abstract concepts. For some models, e.g., the Llama-2 family of models, prompt-based embeddings align better than word embeddings, but the projections are less linear – an observation that holds across almost all model families, indicating that some of the implicitly learned alignments are broken somewhat by prompt-based methods.

1 Introduction

Cross-lingual word embeddings are typically induced by supervised or unsupervised alignment of the word vector spaces of monolingual language models. Compression in multilingual models, i.e., parameter efficiency, can also drive *implicit* alignment (Devlin et al., 2019; Pires et al., 2019; Conneau et al., 2020), but until recently, the mappings could still be much improved by supervised or unsupervised alignment (Hu et al., 2021; Pan et al., 2021). Multilingual large language models (LLMs) are increasingly used for different tasks and demonstrate impressive ability in understanding different languages, but it is unclear whether this is a result of improved, implicit alignment, or of something else, e.g., linguistic overlap or semi-parallel subsets of training data.

LLMs have shown promising capability to comprehend *English* concepts (Liao et al., 2023; Xu et al., 2024). Our paper sets out to evaluate concept alignment in multilingual LLMs. We aim to investigate two things: First, is there a linear mapping between corresponding concepts in different languages? Second, how does a learned linear mapping generalize to new concepts? We explore both questions by revisiting a set of techniques used in early work on bilingual dictionary induction (Kementchedjhieva et al., 2018; Ruder et al., 2018; Søgaard et al., 2018; Kementchedjhieva et al., 2019). We evaluate multilingual LLMs *as if* they were bilingual dictionary induction algorithms by doing nearest neighbor search – with cross-domain local scaling (Lample et al., 2018) – and evaluating their retrieval precision (precision@k). We first derive concept embedding in their standard way (last token or average). Since many of these models were *instruction fine-tuned*, we also compare *prompt-based embeddings* to standard techniques based on (low-level) word embeddings. We then compare their precision to retrieval rates after *explicit* concept space alignment. We perform analyses with and without leakage, across multiple languages, and across both abstract and physical concepts.

en	failure, purchase, lizard, blink
fr	défaillance, emplette, lézard, ciller
ro	eșec, cumpărare, șopârlă, clipire
eu	porrot, erosketa, musker, betikara
fi	epäonnistuminen, osto, lisko, räpytys
ja	しくじり, 買いあげ, リザード, まばたき
th	ความล้มเหลว, การซื้อ, ปอมข่าง, การกะพริบตา

Figure 1: Examples of four parallel WordNet concepts, aligned across 7 languages.

Contributions Our findings across experiments with 10 LLMs and six languages suggest that linear alignment can be induced in multilingual LLMs (if sufficiently big) to map concepts across different languages. Compared to vanilla embeddings, prompt-based concept embeddings exhibit significantly lower linearity, and the gaps between before and after alignment are larger for prompt-based