Steering into New Embedding Spaces: Analyzing Cross-Lingual Alignment Induced by Model Interventions in Multilingual Language Models

Anonymous ACL submission

Abstract

Aligned representations across languages is a desired property in multilingual large language models (mLLMs), as alignment can improve performance in cross-lingual tasks. Typically alignment requires fine-tuning a model, which is computationally expensive, and sizable language data, which often may not be available. A data-efficient alternative to fine-tuning is model interventions — a method for manipulating model activations to steer generation into the desired direction. We analyze the effect of a popular intervention (finding experts) on the alignment of cross-lingual representations in mLLMs. We identify the neurons to manipulate for a given language and introspect the embedding space of mLLMs pre- and post-manipulation. We show that modifying the mLLM's activations changes its embedding space such that cross-lingual alignment is enhanced. Further, we show that the changes to the embedding space translate into improved downstream performance on retrieval tasks, with up to 2x improvements in top-1 accuracy on cross-lingual retrieval.

1 Introduction

Large language models (LLMs) exhibit impressive performance on a variety of tasks from text summarization to zero-shot common-sense reasoning (Raffel et al., 2020; Liu and Lapata, 2019; Bosselut et al., 2019; Richardson and Heck, 2023) and are increasingly deployed in a variety of fields ranging from health to entertainment. Despite these capabilities, to ensure that deployed LLMs align with human values, are non-toxic, and do not hallucinate, they often must be adapted post pre-training.

Model interventions have emerged as data- and compute-efficient tools for model adaptation, whereby targeted updates are applied to model activations after pre-training (Rodriguez et al., 2024; Li et al., 2023; Rimsky et al., 2024). One such method is *finding experts* (Suau et al., 2022, 2024) which manipulates the activations of *expert* neurons responsible for encoding a broadly defined concept (e.g., a word or style of text) to steer model generations into a desired direction. This approach has been successfully used in a variety of domains, ranging from achieving gender parity (Suau et al., 2022), reducing toxicity (Suau et al., 2024),studying geopolitical biases (Faisal and Anastasopoulos, 2023) and multilingual capabilities (Kojima et al., 2024) in mLLMs.

While model interventions successfully control model generations, we do not fully understand their implications for model performance. Two observations are relevant: First, model intervention methods increase perplexity on a fixed dataset post-intervention (Suau et al., 2024) meaning that the intervention introduces changes in how the model represents language. Second, work on mLLMs (Kojima et al., 2024) has shown that intervening on experts for a given language increases the probability of generating that language in the model output (expected outcome) and improves prompt-based machine translation (surprising outcome). These performance gains suggest that the intervention may increase the alignment between representations of different languages.

In this work, we focus on representational changes in mLLMs with an emphasis on cross-lingual alignment for two reasons. First, gains in mLLM performance are largely attributed to better alignment of multilingual representations (Wu et al., 2024; Lample et al., 2018). This has generated a lot of interest in improving multilingual alignment (Chaudhary et al., 2020; Efimov et al., 2023; Lample and Conneau, 2019; Liu et al., 2025). Second, datasets with the same text in multiple languages are available for a variety of tasks, which enables us to study the impact of the intervention in a controlled way across multiple languages.

Specifically, we examine changes in the embedding space of mLLMs introduced by the finding ex-

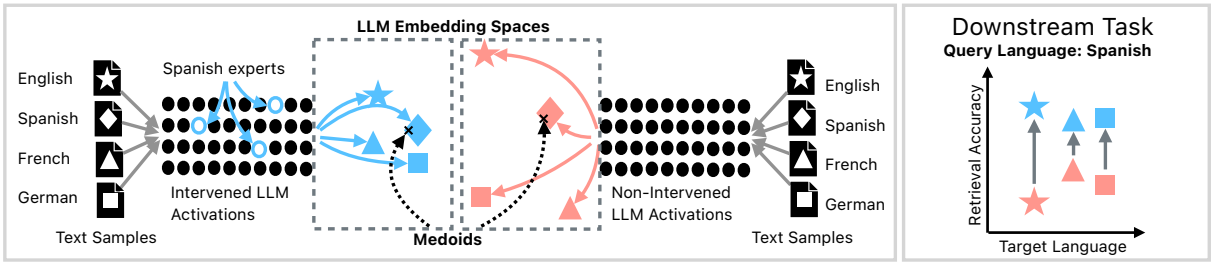


Figure 1: When text from multiple languages is embedded using an LLM with an intervention on expert Spanish neurons, the resulting text embeddings cluster more closely around the medoid of the embedding space (left). This modified model is better at matching Spanish queries to translated text compared to an unmodified model (right).

perts intervention and link these changes to downstream task performance (see Fig. 1). We hypothesize that this intervention increases cross-lingual alignment in mLLMs and present results supporting this hypothesis. We find that:

1. The intervention projects all languages into a new space with new properties in the mLLM .

2. Some of the new properties are less desirable, as evidenced by an increase in perplexity post-intervention.

3. Other properties of the new space are desirable. Specifically, the distances between the language embeddings are reduced in the new space, i.e., the cross-lingual representations are more aligned (Section 4). This translates into a performance gain on cross-lingual retrieval with up to 2x improvement in top-1 accuracy (Section 5), while preserving within-language similarity (Section 6).

2 Related Work

Model interventions. Model interventions are a family of approaches that manipulate model activations to control generations (Li et al., 2024b; Turner et al., 2024; Rodriguez et al., 2024). Suau et al. (2022) propose a method to identify neurons in pre-trained transformer models that are most predictive of a particular concept (*expert neurons*) and show that setting the activations of these experts to their mean value can induce the presence of the target concept in model generations. Suau et al. (2024) find the expert neurons for toxic language and steer the LLM to generate less toxic text by dampening these neurons, while Turner et al. (2024) achieve detoxification by using a contrastive prompt. Rimsky et al. (2024) propose a method to control generations by leveraging the differences in residual stream activations between pairs of positive and negative examples. In mLLMs, Kojima et al. (2024) use this approach to produce more target language tokens in open-ended generation. However, prior work does not analyze the changes these interventions introduce in the representational space of mLLMs nor does it explore the impact of the interventions on cross-lingual alignment.

Aligning multilingual representations in mLLMs. Research on LLM representation alignment falls into two broad categories: 1) Improving model performance on downstream tasks via post-training methods such as prompt-based techniques (Huang et al., 2023; Tanwar et al., 2023), fine-tuning, or continuous pre-training (Zhang et al., 2023; Li et al., 2024a). 2) Understanding where and how representation alignment is achieved in mLLMs. For example, Wendler et al. (2024) show that English-dominated mLLMs like Llama-2 use English as a pivot language and Zhao et al. 2024 systematically evaluate factor contributing to successful cross-lingual transfer in such models.

3 Methods

We seek to understand the impact of network interventions on the representational space of mLLMs with a focus on cross-lingual alignment. We consider three open-source mLLMs: Aya-8B (instruction fine-tuned) (Aryabumi et al., 2024), PolyLM-13B (chat version) (Wei et al., 2023), and Bloom-7B (base) (Scao et al., 2022). Since our aim is to draw conclusions about cross-lingual alignment, we want to make sure that we know what languages were seen in pre-training and include mLLMs for which a detailed description of pre-training datasets is available, excluding LLMs such as Mistral (Jiang et al., 2023), Llama (Touvron et al., 2023), and

Gemma (Team et al., 2024). We begin by identifying and intervening on the language experts in the mLLMs and then study cross-lingual alignment in the embedding space and downstream task performance pre- and post-intervention.

3.1 Probing dataset construction

Following Kojima et al. (2024), we use the Flores200 dataset (NLLB Team, 2022) to find the expert neurons for a particular target language (i.e., the language specifically targeted by the intervention). Flores200 is a machine translation dataset containing short paragraphs sampled from Wikimedia [1] and subsequently translated into 204 languages by skilled human translators. We limit our investigations to the intervention on five target languages — English, German, French, Spanish, and Japanese. These languages are well represented in pre-training data of the models we are considering, ensuring the existence of expert neurons.

3.2 Identifying expert neurons

Expert neurons for a given language are identified following Suau et al. (2024). This *finding experts* approach consists of two steps – first, expert neurons are identified for a particular concept of interest (in our case, a particular target language) and then an intervention is performed to change the activations of these neurons. Expert neurons are those that can classify sentences as being from the positive set (containing the target language) vs. the negative set (that does not contain the target language), as measured by the area under the ROC curve. For each of the five languages under consideration, we use the Flores200 dev split for the target language as the positive set, and the dev splits for the other four languages plus Chinese as the negative set. We include Chinese to increase variety in the character systems in the negative set but we do not consider it for the positive set.

3.3 Intervening on expert neurons

For the intervention, we select the k neurons with the highest expertise (i.e., highest AUROC). We select the value for k that balances generating text in the target language with a low perplexity on the language-specific Wikipedia text (See Appendix B for further details). The activation for these neurons is set to their respective mean value calculated over the positive sentences (Suau et al., 2022).

[1] https://commons.wikimedia.org/wiki/Main_Page

For almost all target languages, the probability of generating that language increases post-intervention (Fig. 2, top), suggesting that the intervention is successful. The only exception is English in the Aya-8B model, where the intervention reduces the likelihood of generating English. We believe that the intervention steers the model away from the default configuration, and English is the default language for that model.

Interestingly, despite Bloom-7B's training set containing neither German nor Japanese, the intervention results in generating both languages with high probability. Our hypothesis is that the Bloom-7B pre-training data contains some amount of German and Japanese data that is large enough to enable expert discovery and controlled generation.

While we are successfully able to increase the accuracy of target language generation through the interventions, consistent with prior work (Suau et al., 2024), we observe an increase in perplexity post-intervention as the number of activated neurons increases, see Fig. 2 (bottom).

For our analyses, we set k to 100 experts for Bloom-7B and 2000 for PolyLM-13B and Aya-8B. For brevity, we present the results for the intervention on Spanish (randomly chosen) in the main text. The results for the other languages are in the respective appendices.

4 The intervention shifts the embedding space increasing cross-lingual alignment

We begin our investigation by quantifying the differences induced by the intervention into the embedding space. For this analysis, we intervene on each of the five target languages discussed in Section 3.3 and examine the effect of the intervention on the representations of 22 languages (the union of all languages present in the pre-training across the three language models). We exclude Arabic and Chinese from consideration due to the lack of conformity in the scripts used[2]. Note that not all of these languages are part of the pre-training for every model under consideration but we present them for consistency (clearly indicating in all figures if

[2] Arabic data are represented in both the Arabic and Latin scripts, while Chinese data are written using both Simplified and Traditional scripts. This decision is motivated by prior work showing that a discrepancy in the encoding can influence performance (Blaschke et al., 2025) and several models under consideration do not provide information on which encoding was used in the pre-training. Appendix A contains the full list of languages and the language codes.

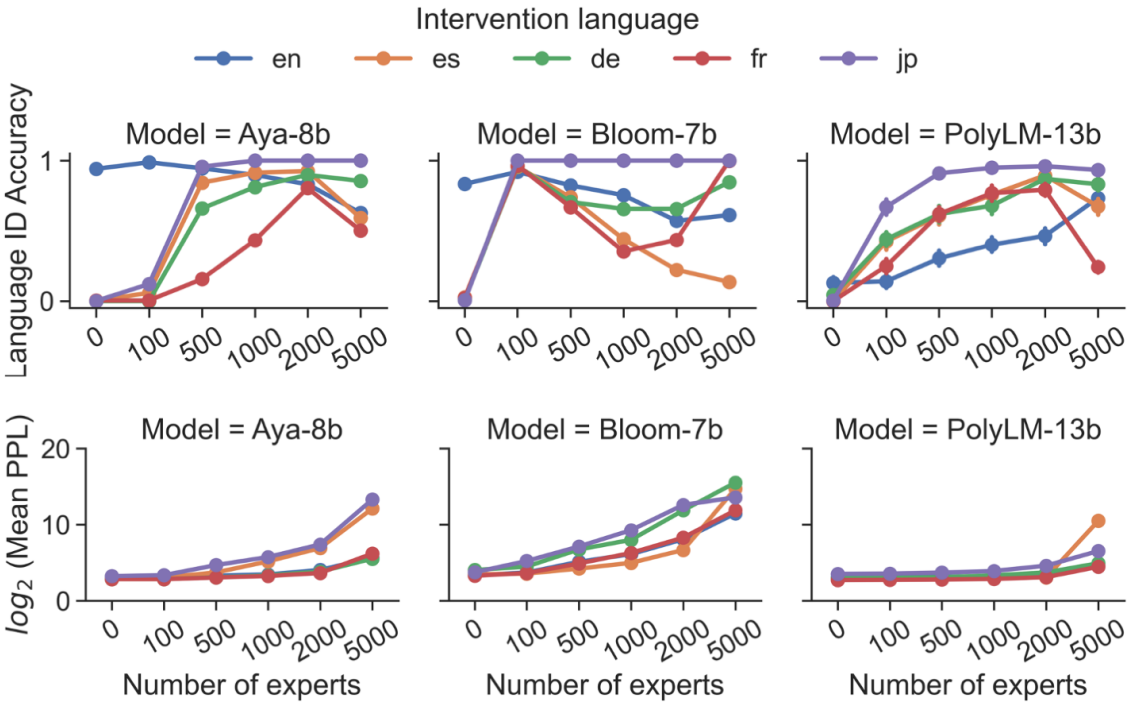


Figure 2: Language ID accuracy and Log perplexity for the intervention on five target languages. The x-axis shows the number of activated experts (0 indicates the original model).

the languages were seen by the model during the pre-training).

For each of the 22 languages, we embed the Flores200 test set (1012 sentences per language) with the original and intervened models' last layer. To characterize the changes in the embedding space, we calculate two types of distances: (1) the pairwise cosine distance between the embeddings of the 22 languages for the intervened and unintervened spaces and (2) the cosine distance between the mean of the embedding for each of the 22 languages and the medoid of each space (pre- and post-intervention) (Table 1). We project the multidimensional embeddings into a two-dimensional space using UMAP (McInnes et al., 2020) to visualize how the embedding space changes.

Our findings are as follows. The intervention pulls the embeddings of all languages into a new space (see Fig. 3 for Spanish and Appendix C for other languages), rather than moving them closer to the embeddings of the target language in the unintervened space. The increase in perplexity post-intervention discussed in Section 3.3 also supports this finding.

The post-intervention embeddings for the different languages are closer to each other compared to the pre-intervention embeddings, as indicated by the reduced pairwise cosine distances between the languages. Specifically, the distances are reduced because the post-intervention embeddings are pulled closer to the medoid of the embedding space. As a result of the shift, all languages are closer to the target post-intervention.

We notice that all distances under consideration are reduced less post-intervention for PolyLM-13B compared to the other models. We hypothesize that this relates to the specific data distribution and training procedure used for PolyLM-13B. Unfortunately, since we do not have access to the data the three models under consideration were trained on, we cannot test this hypothesis in this work. We return to this point in Section 9.

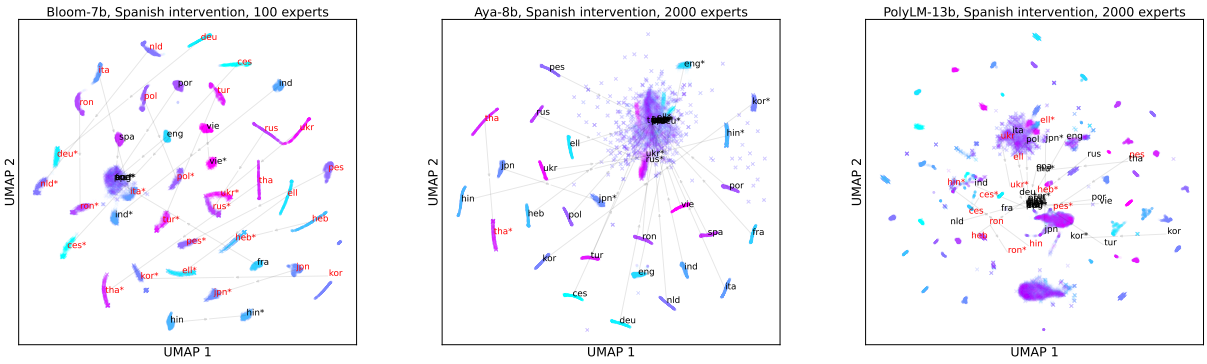
Taken together, these findings suggest that the intervention projects language embeddings into a new space where they are more aligned. In the following sections, we explore if this change translates into downstream task performance.

5 Cross-lingual similarity is enhanced in the new space

We now ask if the increased alignment post-intervention translates to downstream task performance. We use **cross-lingual** retrieval as our downstream task: Given a sentence (query) in one lan-

Model	Language	Distance (all languages)		Distance to Medoid		Distance to Target	
		Pre	Post	Pre	Post	Pre	Post
Aya-8B	Target	–	–	$0.62_{\pm 0.03}$	$0.14_{\pm 0.03}$	–	–
	Non-Target	$0.72_{\pm 0.00}$	$0.19_{\pm 0.04}$	$0.58_{\pm 0.01}$	$0.12_{\pm 0.01}$	$0.77_{\pm 0.01}$	$0.2_{\pm 0.01}$
Bloom-7B	Target	–	–	$0.60_{\pm 0.21}$	$0.04_{\pm 0.01}$	–	–
	Non-Target	$0.72_{\pm 0.00}$	$0.17_{\pm 0.06}$	$0.5_{\pm 0.03}$	$0.11_{\pm 0.01}$	$0.78_{\pm 0.03}$	$0.13_{\pm 0.01}$
PolyLM-13B	Target	–	–	$0.72_{\pm 0.04}$	$0.43_{\pm 0.09}$	–	–
	Non-Target	$0.85_{\pm 0.00}$	$0.56_{\pm 0.09}$	$0.72_{\pm 0.01}$	$0.43_{\pm 0.02}$	$0.86_{\pm 0.01}$	$0.54_{\pm 0.02}$

Table 1: Cosine distances between 22 languages under consideration, mean distance to the target of the intervention, and and the distance to the medoid of the embedding space are reduced post-intervention. Distance (all languages) refers to pairwise cosine distance between the embeddings of 22 languages; distance to target refers to the distance between the intervention target and the remaining 21 languages. Pre refers to pre-intervention and post to post-intervention. Distances are means and standard errors of the mean over the five intervention targets.



(a) Bloom-7b (b) Aya-8b (c) PolyLM-13b

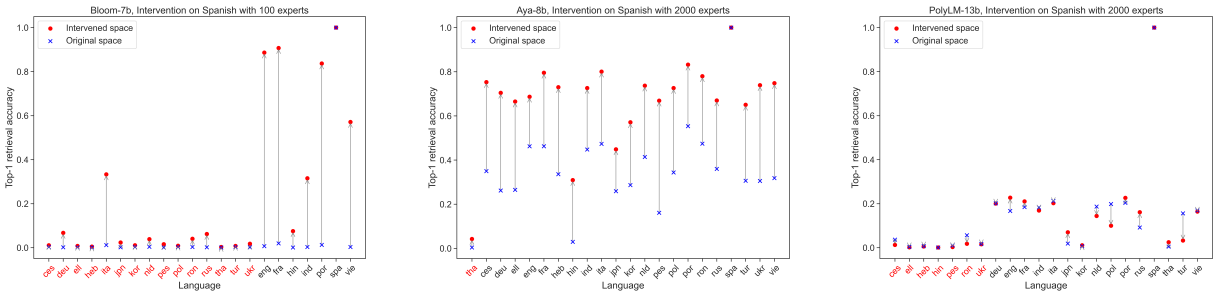
Figure 3: UMAP embeddings for 22 languages in the Bloom-7B model (left), Aya-8B (middle), and PolyLM-13B (right). The embeddings post-intervention are marked with '*' for each language. The dots represent individual sentences in the pre-intervention space; the crosses represent individual sentences in the post-intervention space. The languages that are not in the training set for a given model are marked in red.

guage (query language), and a set of sentences (candidates) in a different language (candidate language), which of the candidates is a translation (match)? Our main experiments are carried out on the Flores200 test split (NLLB Team, 2022) as it allows us to test cross-lingual retrieval across multiple combinations of query and candidate languages. As the dev split of the Flores200 dataset was used to identify language experts, we also present results on the validation split of Tatoeba (Tiedemann, 2012) and the test split of BUCC-18 (Hu et al., 2020) for an independent validation of our findings (see Appendix H).

For each sentence, we compute pre- and post-intervention embeddings by averaging over the last hidden state of the mLLM, producing vectors with dimensions matching the model's hidden size. To identify the closest matching sentence, we compute inner products between the query (e.g., in Spanish), and all candidates (e.g., in French). We select the candidate with the highest inner product as the match, and then measure top-1 accuracy.

Top-1 retrieval accuracy improves post-intervention for retrieval with the target language. We first examine if the increased proximity to the target language in the post-intervention embedding space translates into top-1 retrieval accuracy improvement when the target is used as the retrieval query for the 22 candidate languages under consideration.

We find that top-1 retrieval accuracy improves post-intervention when using the target as the query language (see Fig. 4 for the Spanish intervention and Appendix E for the remaining four languages). This finding is consistent across most target languages and models. Candidate languages present in the pre-training data generally demonstrate larger gains post-intervention. The pattern of improvement differs based on the model. Specifically, for Aya-8B a successful intervention results in con-



(a) Bloom-7b (b) Aya-8b (c) PolyLM-13b

Figure 4: Top-1 retrieval accuracy for the intervention on Spanish for 22 languages in the Bloom-7B model (left), Aya-8B (middle), and PolyLM-13B (right). The languages that are not in the training set for a given model are marked in red.

Model	Query language	$r(\text{acc}_{\text{post}}, d_{\text{post}})$	$r(\text{acc}_{\text{pre}}, d_{\text{pre}})$	$r(d_{\text{post}}, d_{\text{pre}})$	$r(\Delta\text{acc}, d_{\text{pre}} - d_{\text{post}})$
Aya-8B	es	-0.51 [-0.88 -0.18]	-0.89 [-0.98 -0.55]	0.48 [0.32 0.78]	0.86 [0.49 0.96]
	fr	-0.64 [-0.91 -0.45]	-0.86 [-0.97 -0.57]	0.51 [0.27 0.89]	0.89 [0.84 0.97]
	en	-0.94 [-0.97 -0.85]	-0.80 [-0.96 -0.34]	0.65 [0.44 0.92]	0.10 [-0.69 0.60]
	de	-0.89 [-0.96 -0.75]	-0.87 [-0.98 -0.44]	0.33 [-0.74 0.76]	0.52 [0.12 0.95]
	jp	-0.02 [-0.62 0.34]	-0.96 [-0.99 0.34]	0.27 [-0.18 0.62]	0.89 [0.30 0.99]
Bloom-7B	es	-0.97 [-0.99 -0.95]	-0.83 [-0.99 -0.54]	0.79 [0.71 0.98]	0.1 [-0.98 0.88]
	fr	-0.98 [-0.99 -0.94]	-0.89 [-0.99 -0.38]	0.75 [0.62 0.99]	0.23 [-0.98 0.83]
	en	-0.89 [-0.99 -0.60]	-0.89 [-0.99 -0.44]	0.97 [0.96 0.99]	0.23 [-0.90 0.86]
	de	-0.90 [-0.99 -0.74]	-0.50 [-0.96 0.34]	0.95 [0.86 0.99]	-0.72 [-0.97 0.22]
	jp	-0.90 [-0.99 -0.80]	NA[3]	-0.48 [-0.90 0.97]	0.64 [-0.70 0.93]
PolyLM-13B	es	-0.44 [-0.91 -0.38]	-0.84 [-0.96 -0.65]	0.70 [0.44 0.91]	0.10 [-0.31 0.57]
	fr	-0.44 [-0.82 -0.35]	-0.90 [-0.99 -0.62]	0.66 [0.20 0.93]	0.30 [-0.18 0.82]
	en	-0.86 [-0.98 -0.53]	-0.84 [-0.98 -0.52]	0.99 [0.96 0.99]	0.28 [-0.33 0.62]
	de	-0.01 [-0.51 0.81]	-0.95 [-0.99 -0.57]	0.15 [-0.56 0.57]	-0.04 [-0.71 0.34]
	jp	-0.52 [-0.91 0.92]	-0.96 [-0.99 0.00]	0.73 [0.56 0.96]	0.25 [-0.25 0.57]

Table 2: Pearson correlations (r) between top-1 retrieval accuracy (acc) and mean pairwise cosine distance in the embedding space d. Subscripts indicate the space from which embeddings are sampled: pre = original model; post = intervened model. Numbers in brackets represent bootstrapped 95% confidence intervals. Correlations that are not statistically significant (p-values >0.05) are shown in gray.

sistent improvements in top-1 accuracy for the majority of candidate languages (median=32%; max=74%). For Bloom-7B, top-1 accuracy gains are large (up to 89%) for a small number of candidate languages, with moderate improvements for other languages (median=14%). For PolyLM-13B, the improvements are small (median=0.5%; max=12%).

To better understand how the increased alignment in the embedding space influences cross-lingual retrieval, we look at the mean pairwise cosine distances between the query and candidate languages and explore how this correlates with retrieval accuracy (Fig. 4). Table 2 shows average correlations between post-intervention top-1 retrieval accuracy (acc_{post}) and mean query-candidate language distance both pre- and post-intervention (d_{pre}, d_{post}), average correlations between d_{pre} and d_{post}, and average correlations between improvement in accuracy ($\Delta\text{acc} = \text{acc}_{\text{post}} - \text{acc}_{\text{pre}}$) and change in distance between pre- and post-intervention embeddings ($d_{\text{pre}} - d_{\text{post}}$). When calculating averages, we only include candidate languages seen in pre-training for each model; we note that the general pattern stays the same but the correlations are somewhat weaker if all 22 languages are considered for all models.

We find that in this setting, when the query and intervention-target language are the same, the distance between query/target and match language is predictive of top-1 cross-lingual retrieval accuracy in both pre- and post-intervention spaces. As discussed in Section 4, all language embeddings move closer to the target's embeddings post-intervention, which explains the gains in cross-lingual retrieval accuracy. The distances in the unintervened and intervened space are positively correlated—language embeddings that are closer

to the target pre-intervention are also closer to the target post-intervention. However, the magnitude of the performance gain in the intervened space does not correlate with the reduction in distance between the match and target languages across the two spaces, suggesting that the increased alignment post-intervention cannot be simply explained by a reduction in distances.

Top-1 retrieval accuracy improves post-intervention for retrieval with the non-target languages. In Section 4, we found that the distances between almost all languages decrease post-intervention—not just the distances to the intervention target. We next examine if these reduced cosine distances between languages *other than* the intervention target translate into improved top-1 retrieval accuracy when using these languages as the query language. For example, we study if intervening on Spanish experts improves Dutch-English retrieval (in this case, neither the query nor candidate language is the intervention-target language). We find that, perhaps surprisingly, improvements observed when the query language is the intervention target (see Fig. 4 and Table 2) carry over to query languages other than the intervention-target language (see Fig. 5 for the Spanish intervention and Appendix F for the remaining four languages). For example, the intervention on Spanish expert neurons for Bloom-7B results in retrieval improvement when English, French, and Portuguese are the query language. The same intervention improves retrieval when querying Hebrew with Persian or when querying Czech with Greek in the Aya-8B model and when querying Russian with Portuguese in PolyLM-13B. These are examples of larger improvements, but many other languages follow the same pattern with smaller gains. Generally, the patterns in improvement are consistent with those seen when Spanish is the query language. Languages that are in the pre-training set have larger accuracy gains. Bloom-7B has large improvements for a small number of languages and no drops in performance. Aya-8B has relatively large improvements for a majority of languages but also has a drop in performance for some. As noted previously, PolyLM-13B performance is uneven—the improvement varies by language with languages in the pre-training set generally having larger improvements.

6 Within-language similarity is preserved in the new space

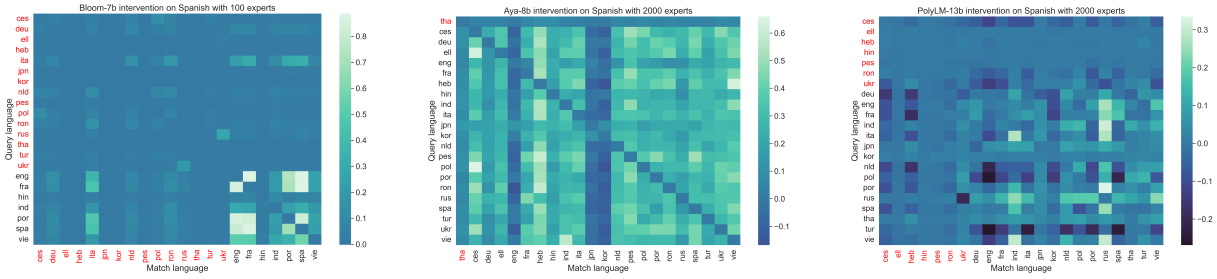
As observed in Section 4, all languages move toward the medoid of the embedding space post-intervention, which raises the question of whether language-specific similarities preserved in the new space. To answer this question, we evaluate performance on a **paraphrase** retrieval task which tests whether a sentence in the intervened space can be matched with its paraphrase in the intervened space. We utilize the PAWS-X dataset (Hu et al., 2020), which provides paired sentences across seven languages, including all five of our intervention targets. From the test split, we retain only the paraphrase pairs, excluding non-paraphrases and sentences from other languages. This transforms our evaluation into a within-language sentence retrieval task, where the goal is to match each sentence with its paraphrase from a pool of candidates for that language.

The paraphrase retrieval task reveals two key findings about embedding spaces before and after the intervention. First, the top-1 paraphrase retrieval accuracy remains largely unchanged after the intervention (see Table 3), indicating that the new embedding space preserves within-language similarity. Second, when attempting retrieval between intervened and unintervened embeddings of the same language — i.e., using the embeddings from the unintervened model as the query and the embeddings from the intervened model as candidate matches — accuracy drops significantly. This decline supports the observation in Section 4 suggesting that the intervention projects embeddings into a distinctly different space from their original unintervened representations. This finding also aligns with the increase in perplexity observed post-intervention – the intervened space of a given language is *not* the same space as the original space of this language.

7 Intervention on random neurons does not provide an improvement on downstream tasks

In our analyses so far, we have attributed the changes in the embedding space to the intervention

[3]The cosine distances between Japanese and other languages are identical in Bloom-7b in the unintervened space and thus the correlation coefficient is not defined. This is likely due to the fact that Japanese is not in Bloom 7b's pre-training.



(a) Bloom-7b (b) Aya-8b (c) PolyLM-13b

Figure 5: (Top-1 accuracy$_{\text{post-intervention}}$ − Top-1 accuracy$_{\text{pre-intervention}}$) for 22 languages after intervening on Spanish expert neurons in the Bloom-7B model (left), Aya-8B (middle), and PolyLM-13B (right). The languages that are not in the training set for a given model are marked in red.

Model	Top-1 Accuracy		
	(Pre)	(Post)	(Mixed)
Bloom-7B	0.80	0.80	0.33
Aya-8B	0.85	0.86	0.64
PolyLM-13B	0.52	0.56	0.41

Table 3: Top-1 accuracy results for the paraphrase retrieval task following the intervention on Spanish. The results for other languages can be found in Appendix D. Pre = both the query and candidate embeddings are from the original model; Post = both the query and the candidate embeddings are from the intervened model; Mixed = query is from the original model and candidates are from the intervened model.

on expert neurons. Before we conclude, we consider an alternative possibility — that the expert neurons do not play a role in increasing alignment post-intervention, but instead alignment is achieved by fixing the activations of a number of neurons in a network. To address this, we assign the activation levels of the language expert neurons used in prior sections to the same number of neurons chosen randomly in the network and repeat our analyses on these models.

We find that intervening on random neurons produces markedly different results compared to activating language experts (see Appendix G). The embedding space after the intervention on random neurons does not have the same properties as described in Section 4, which translates into the performance on downstream tasks for all models. Specifically, for the Aya-8B and PolyLM-13B top-1 cross-lingual retrieval accuracy drops for all languages post-intervention on random neurons compared to pre-intervention. Interestingly, for Bloom-7B, there is mostly no change for all target languages except French, which surprisingly improves post-intervention on random neurons. However, the gains are significantly smaller compared to those after intervening on French experts. Similar to the intervention on language experts, within-language paraphrase retrieval shows only small changes post-intervention. When they occur, these changes tend to be negative (i.e., the performance drops) after intervening on random neurons and positive after intervening on the actual language experts.

8 Conclusions

We present a novel analysis of the impact of the finding-experts intervention on cross-lingual alignment in mLLMs. We find that intervening on language experts projects model embeddings into a new space where languages are more aligned than in the original space but still preserve within-language similarity. These findings provide an explanation for the increase in perplexity observed post-intervention in prior work (Suau et al., 2022). The increase in cross-lingual alignment results in up to 2x improvement in top-1 retrieval accuracy. Additionally, we show that the correlation between cross-lingual alignment and cross-lingual retrieval is high and statistically significant. Importantly, these improvements are not restricted to the intervention-target language. For instance, an intervention on Japanese results in a large top-1 retrieval accuracy improvement for English-Portuguese for Bloom-7B. We find that the three models we study show markedly different patterns both in the changes to the embedding space and downstream tasks. We leave it to future to work to determine the causes of these differences, though we hypothesize that they are due to the pre-training differences.

9 Limitations

There are several limitations that need to be considered when interpreting our results.

The major limitation is that we are working with pre-trained models and we have only limited information on training data and procedure. Specifically, for Bloom-7B and PolyLM-13B, we have the information on the proportion of each language in the pre-training set. For Aya-8B, only information on which languages were seen in the pre-training (but no proportions) is available. We observe different performance gains post-intervention for the three models under consideration and while we hypothesize that these differences are due to training data and/or procedure, we do not have enough information to test this hypothesis. Future work should explore the effect of intervention on alignment in a more controlled setting where the models are trained from scratch on a publicly available dataset manipulating language proportions in the training data to better understand what is driving the difference.

We have studied only one approach out of a family of approaches to controllable generations (Rimsky et al., 2024; Suau et al., 2024; Rodriguez et al., 2024). Each approach in the family comes with its differences – in the way the neurons targeted by the intervention are discovered, how the changes are introduced to the activations, how many neurons are intervened on, etc. We do not fully understand how these design decisions impact the representation space. For example, it is possible that some of these approaches are more beneficial for alignment while others introduce changes that are more beneficial for other tasks (or not at all). The comparison of approaches is beyond the scope of current work and we leave it for future investigations.

References

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Verena Blaschke, Masha Fedzechkina, and Maartje ter Hoeve. 2025. Analyzing the effect of linguistic similarity on cross-lingual transfer: Tasks and experimental setups matter. *arXiv preprint arXiv:2501.14491*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Aditi Chaudhary, Karthik Raman, Krishna Srinivasan, and Jiecao Chen. 2020. Dict-mlm: Improved multilingual pre-training using bilingual dictionaries. *ArXiv*, abs/2010.12566.

Pavel Efimov, Leonid Boytsov, Elena Arslanova, and Pavel Braslavski. 2023. The impact of cross-lingual adjustment of contextual word representations on zero-shot transfer. In *Advances in Information Retrieval*, pages 51–67, Cham. Springer Nature Switzerland.

Fahim Faisal and Antonios Anastasopoulos. 2023. Geographic and geopolitical biases of language models. In *Proc. of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, D de las Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. 2023. Mistral 7b (2023). *arXiv preprint arXiv:2310.06825*.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. *NAACL*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *NeurIPS*, arXiv:1901.07291.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2024a. Improving in-context learning of multilingual generative language models with cross-lingual alignment. *NAACL*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024b. Inference-time intervention: Eliciting truthful answers from a language model. *NeurIPS*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yihong Liu, Mingyang Wang, Amir Hossein Kargaran, Ayyoob ImaniGooghari, Orgest Xhelili, Haotian Ye, Chunlan Ma, François Yvon, and Hinrich Schütze. 2025. How transliterations improve crosslingual alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2417–2433, Abu Dhabi, UAE. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction. *Preprint*, arXiv:1802.03426.

et al NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation. *EMNLP*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Christopher Richardson and Larry Heck. 2023. Commonsense reasoning for conversational ai: A survey of the state of the art. *ArXiv*, abs/2302.07926.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Pau Rodriguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, Marco Cuturi, and Xavier Suau. 2024. Controlling language and diffusion models by transporting activations. *arXiv preprint arXiv:2410.23054*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Xavier Suau, Pieter Delobelle, Katherine Metcalf, Armand Joulin, Nicholas Apostoloff, Luca Zappella, and Pau Rodriguez. 2024. Whispering experts: Neural interventions for toxicity mitigation in language models. In *Forty-first International Conference on Machine Learning*.

Xavier Suau, Luca Zappella, and Nicholas Apostoloff. 2022. Self-conditioning pre-trained language models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 4455–4473. PMLR.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual LLMs are better cross-lingual in-context learners with alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6292–6307, Toronto, Canada. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Joerg Tiedemann. 2012. Parallel data, tools and interfaces in opus. *LREC*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. Polylm: An open source polyglot large language model. *Preprint*, arXiv:2307.06018.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *Preprint*, arXiv:2402.10588.

Di Wu, Yibin Lei, Andrew Yates, and Christof Monz. 2024. Representational isomorphism and alignment of multilingual large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14074–14085, Miami, Florida, USA. Association for Computational Linguistics.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *Preprint*, arXiv:2306.10968.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *arXiv preprint arXiv:2401.01055*.

A List of languages studied

The following languages are considered in this work:

#	Language	ISO 639-1	ISO 639-3
1	Thai	th	tha
2	Czech	cs	ces
3	German	de	deu
4	Greek	el	ell
5	English	en	eng
6	French	fr	fra
7	Hebrew	he	heb
8	Hindi	hi	hin
9	Indonesian	id	ind
10	Italian	it	ita
11	Japanese	ja	jpn
12	Korean	ko	kor
13	Dutch	nl	nld
14	Persian	fa	pes
15	Polish	pl	pol
16	Portuguese	pt	por
17	Romanian	ro	ron
18	Russian	ru	rus
19	Spanish	es	spa
20	Turkish	tr	tur
21	Ukrainian	uk	ukr
22	Vietnamese	vi	vie

Table 4: ISO 639-1 and ISO 639-3 Language Codes

B Selecting expert neurons

The value of k is determined as follows. For each of the five languages, we sweep over expert set sizes ranging from 100 to 5000. For each setting of language and number of experts, we run free-form generation to generate 256 sentences over eight random seeds (for a total of 2048 sentences) using the beginning of sentence (BOS) token as the prompt. We perform generation with temperature=1 and top_p=0.9 [4]. We then use lang-id (Lui and Baldwin, 2012) to measure the probability of the text generated in the target language.

To calculate the perplexity of Wikipedia text in the target language for the original and intervened models, we use the Wikimedia dump from 2023-11-01 [5]. Paragraphs of text shorter than 100 characters are removed and the remaining paragraphs are concatenated together. Finally, a corpus of 10 million tokens is selected from the concatenated paragraphs. The context length is set to the model's maximum input size (in tokens) and a stride (i.e., sliding the context window) of 512 tokens is used to speed up the perplexity measurement.

[4] The other hyper-parameters are set to default for $\texttt{transformers}==4.41$

[5] https://huggingface.co/datasets/wikimedia/wikipedia

C UMAP embeddings for four intervention-target languages

Note: The embeddings post-intervention are marked with '*' for each language. The languages that are not in the training set are marked in red.

C.1 Bloom-7B

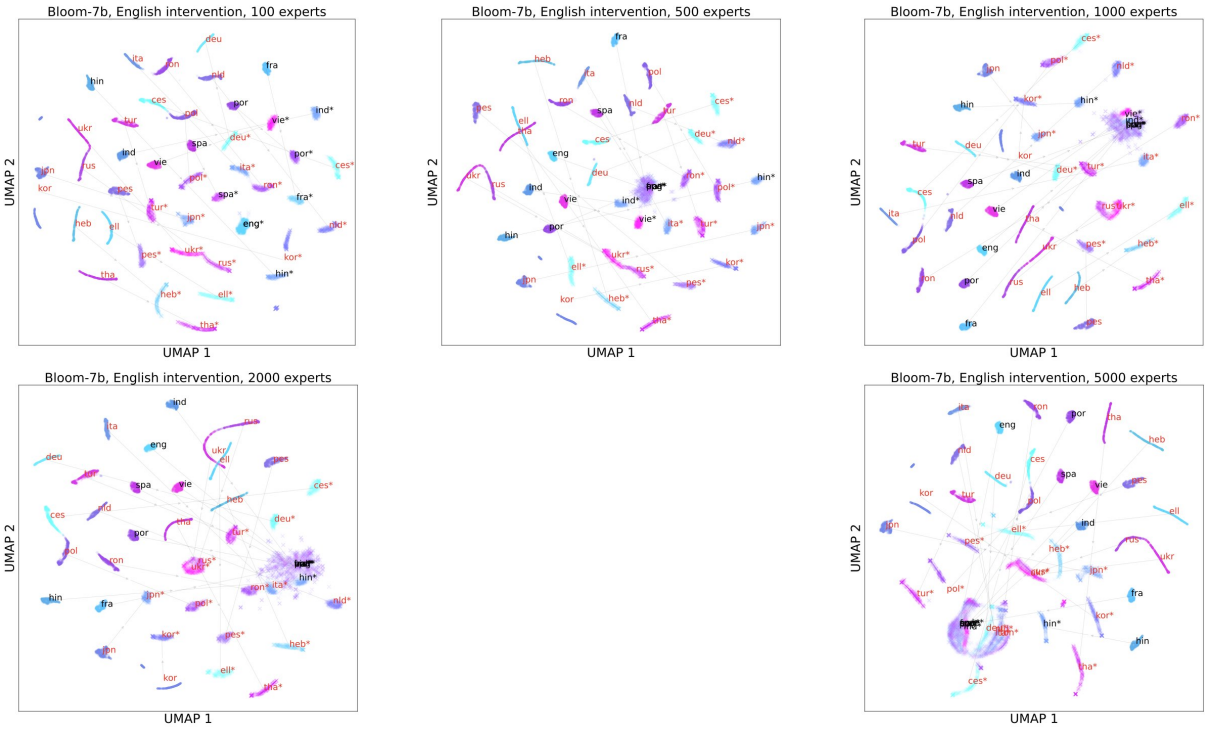


Figure 6: UMAP embeddings for intervention on English expert neurons

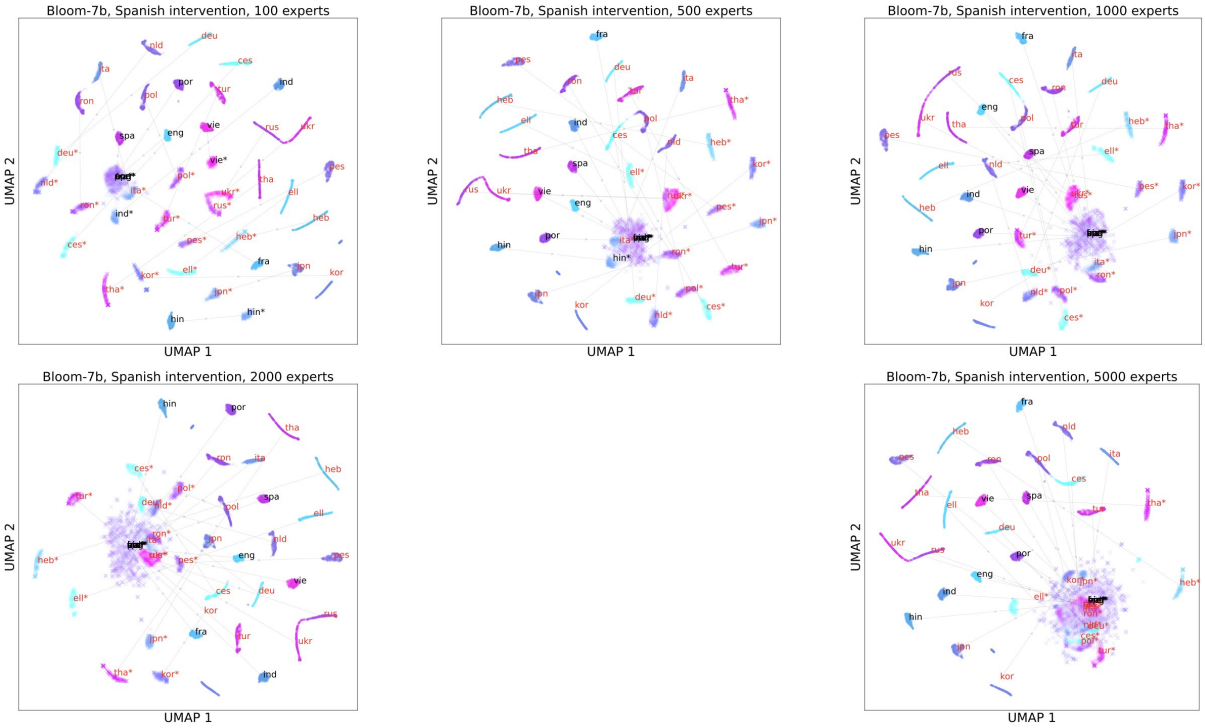


Figure 7: UMAP embeddings for intervention on Spanish expert neurons

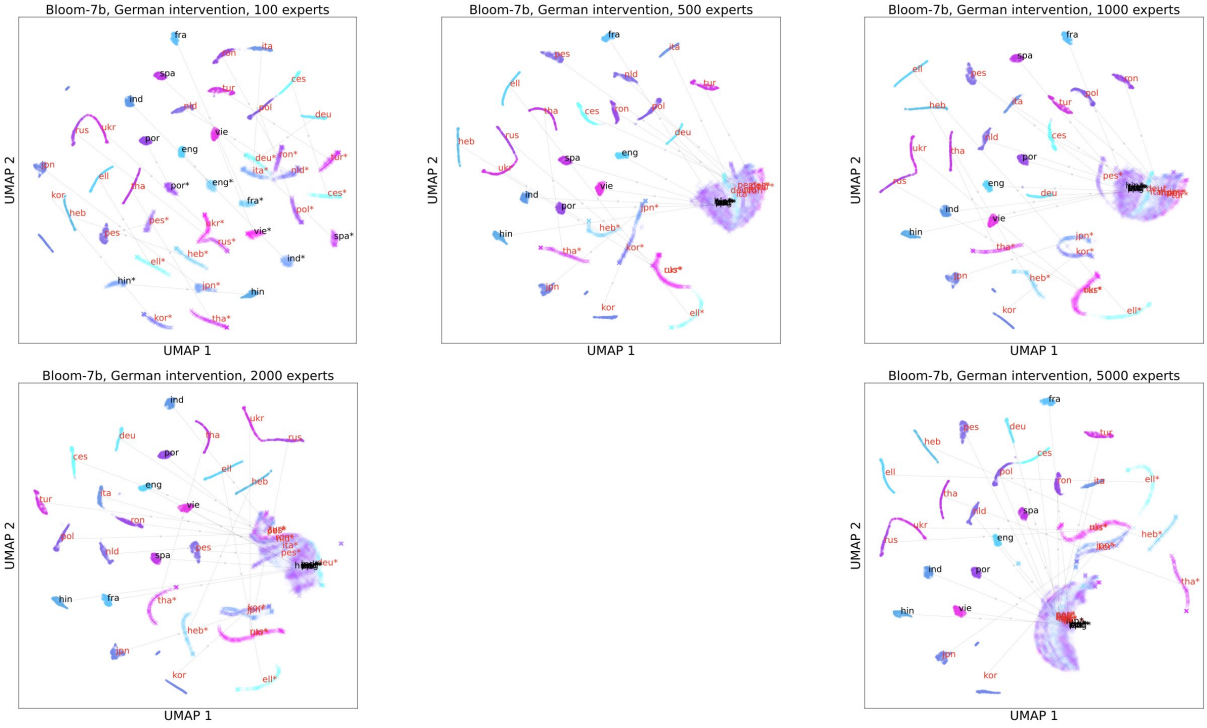


Figure 8: UMAP embeddings for intervention on German expert neurons

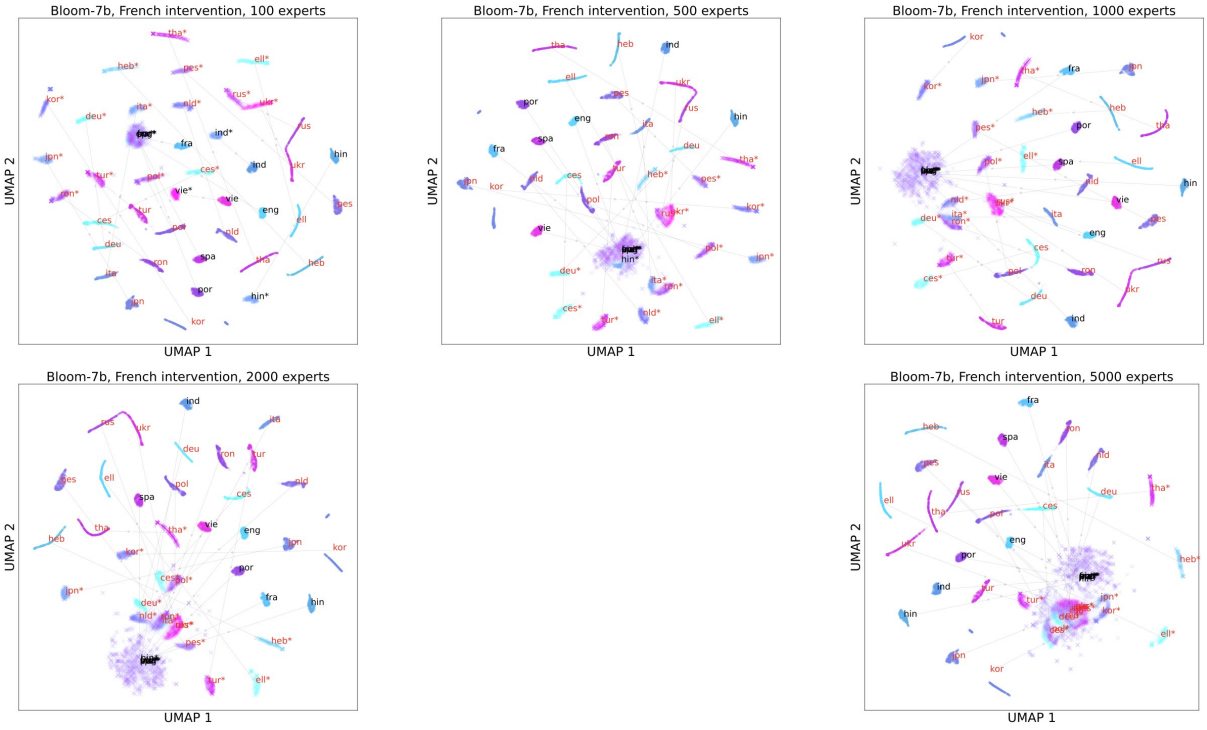


Figure 9: UMAP embeddings for intervention on French expert neurons

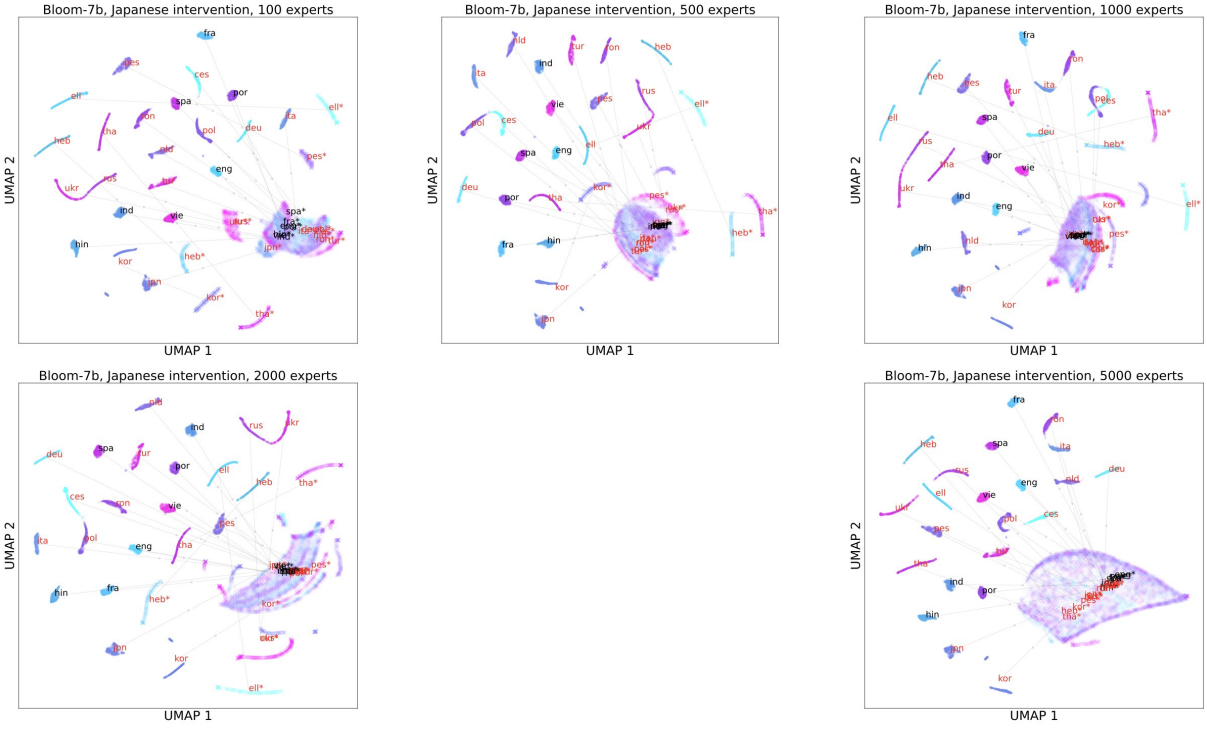


Figure 10: UMAP embeddings for intervention on Japanese expert neurons

C.2 Aya-8B

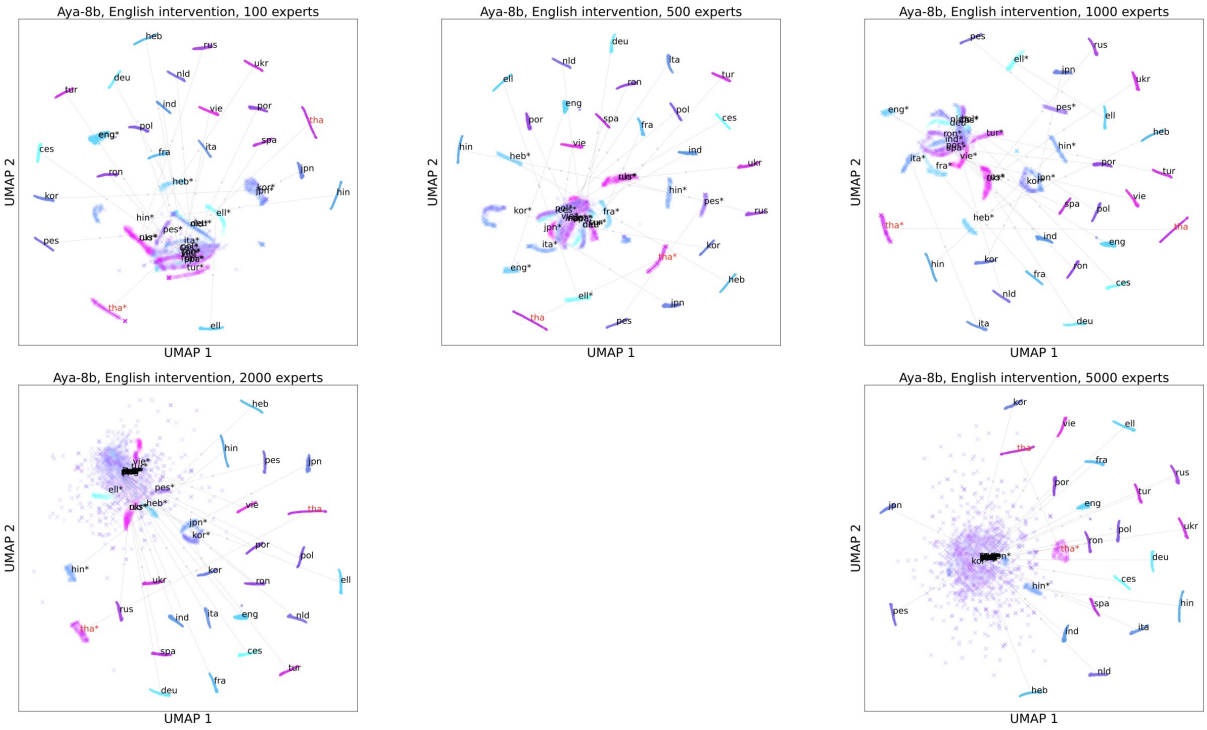


Figure 11: UMAP embeddings for intervention on English expert neurons

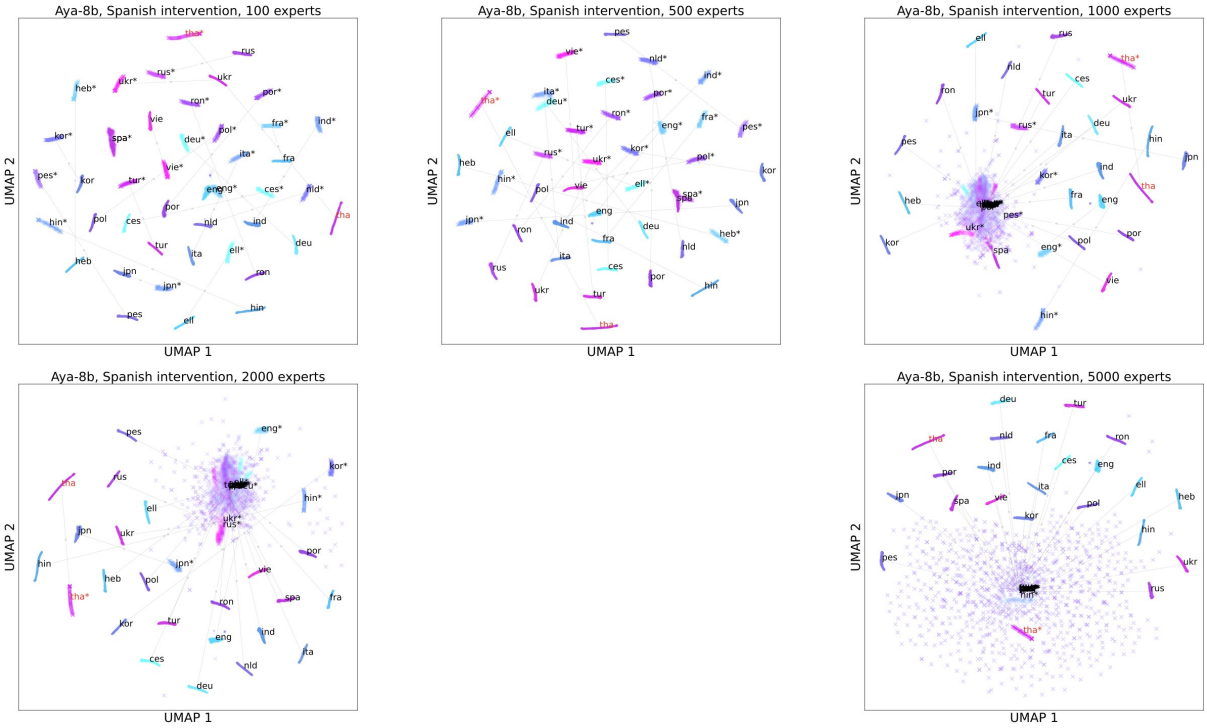


Figure 12: UMAP embeddings for intervention on Spanish expert neurons

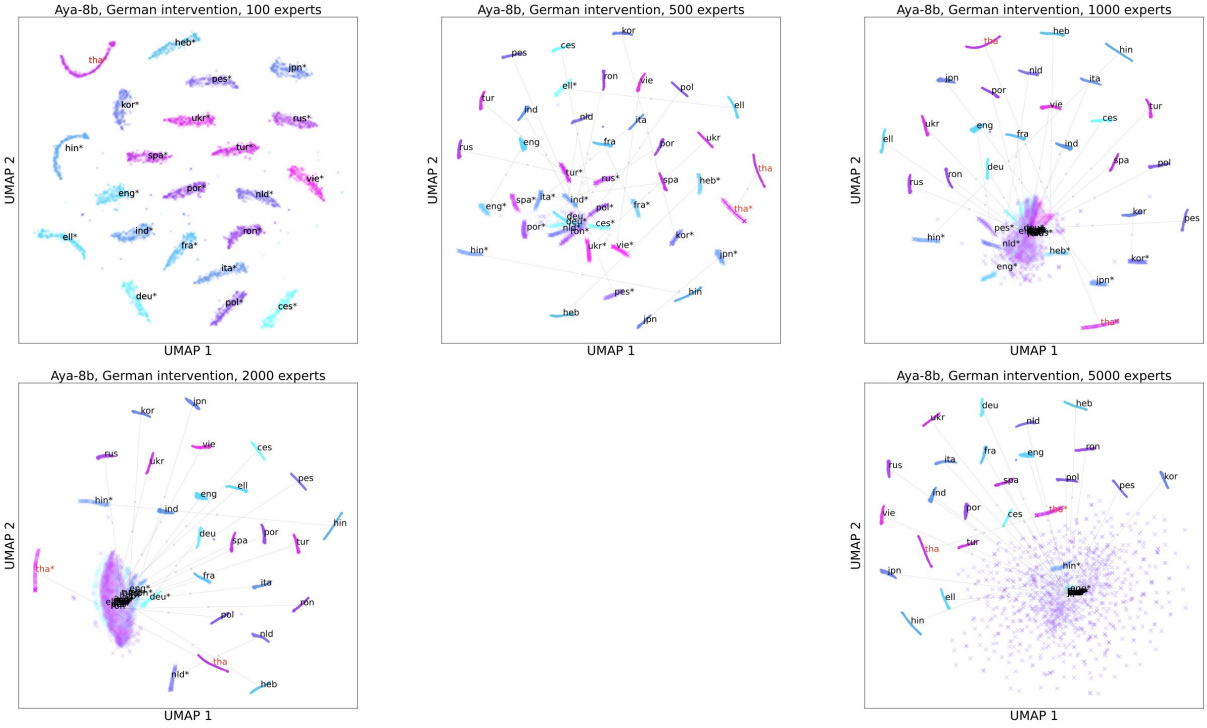


Figure 13: UMAP embeddings for intervention on German expert neurons

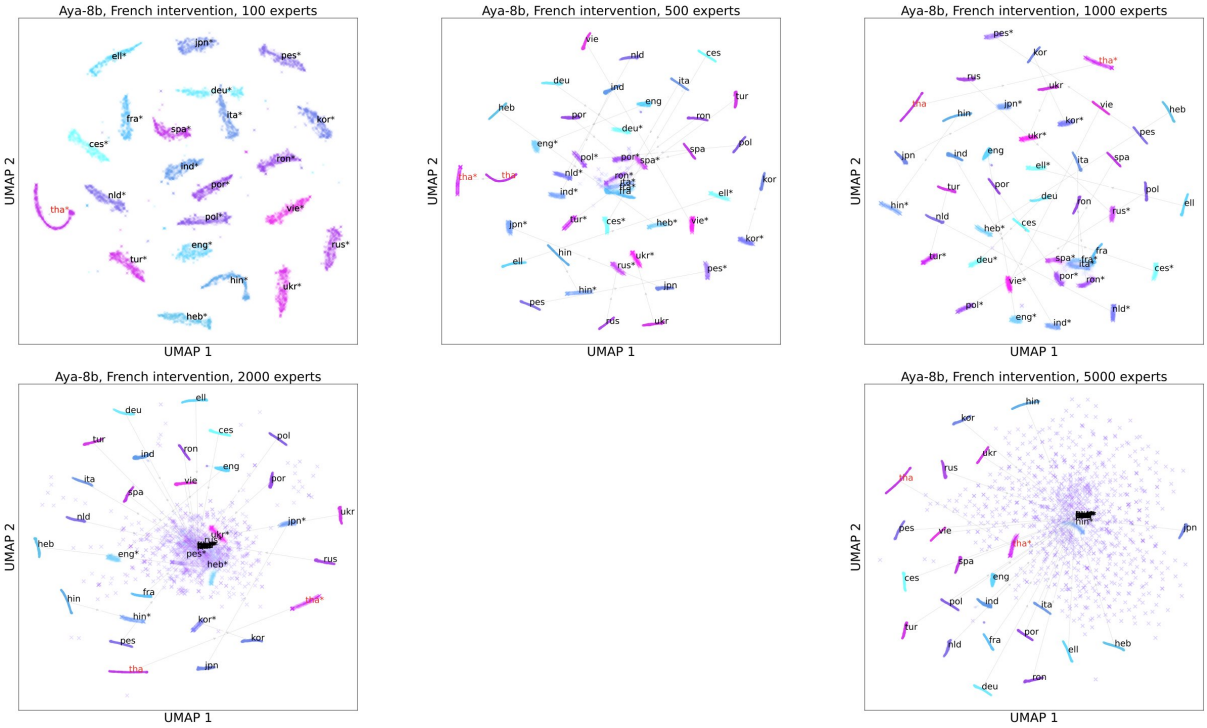


Figure 14: UMAP embeddings for intervention on French expert neurons

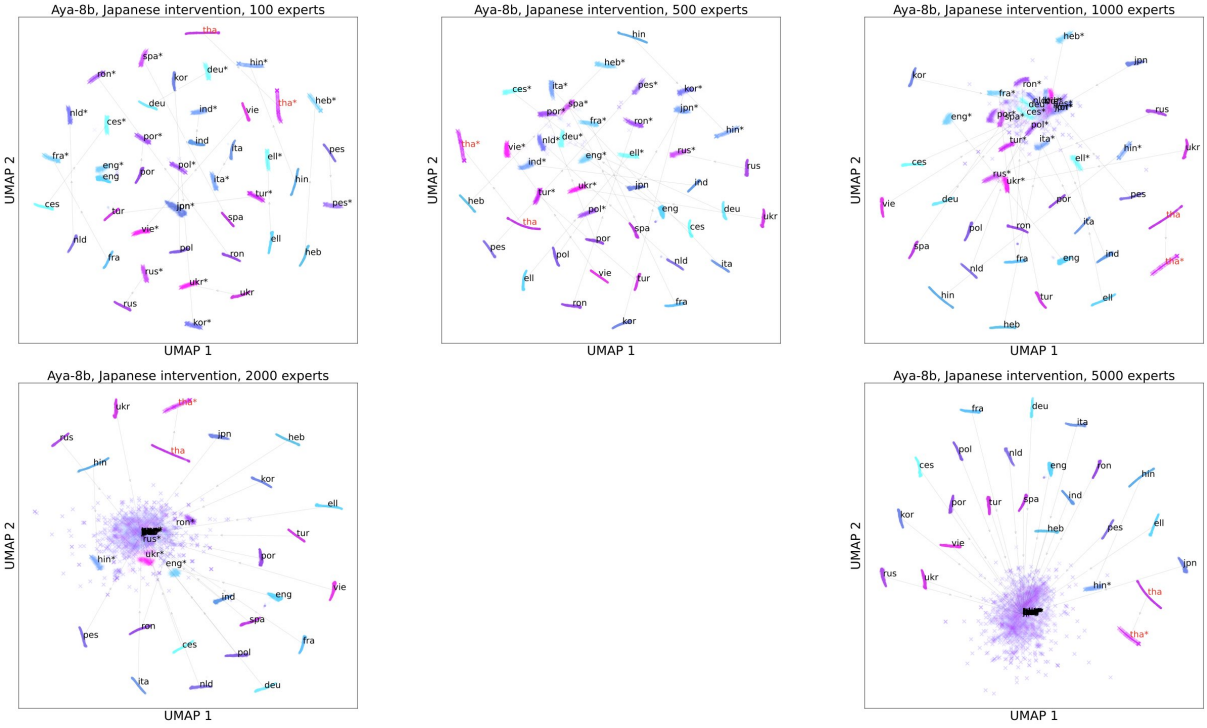


Figure 15: UMAP embeddings for intervention on Japanese expert neurons

C.3 PolyLM-13B

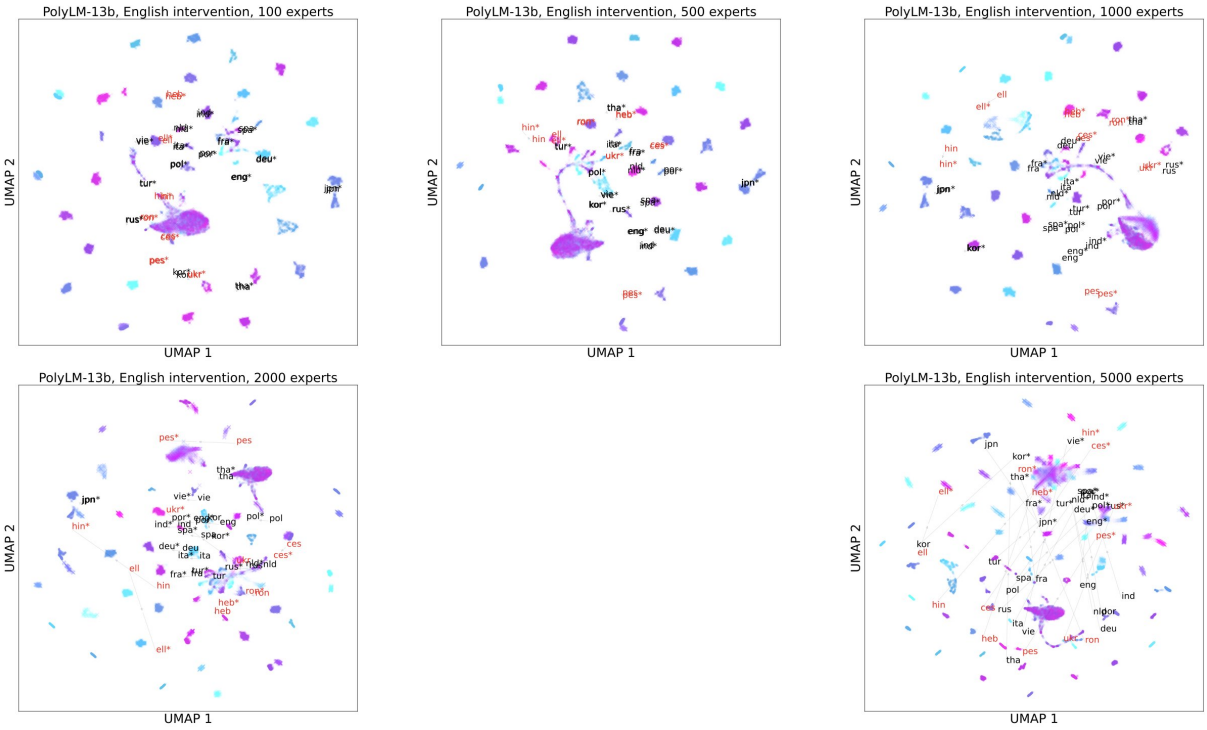


Figure 16: UMAP embeddings for intervention on English expert neurons

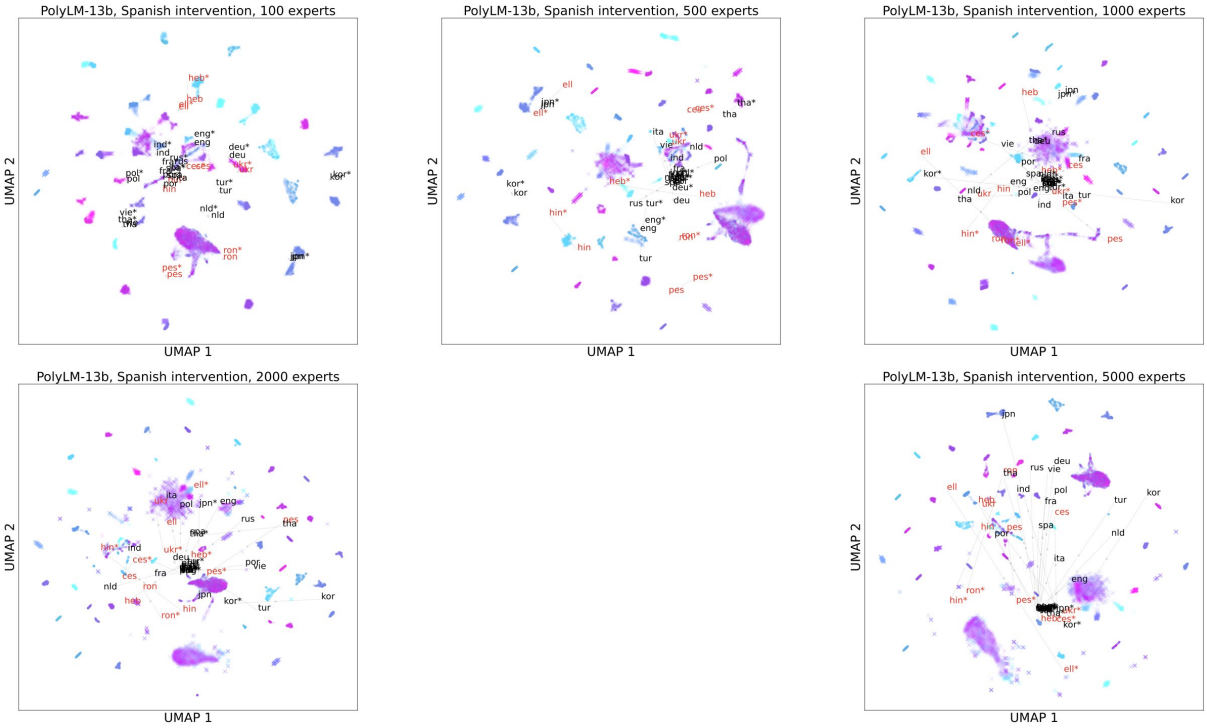


Figure 17: UMAP embeddings for intervention on Spanish expert neurons

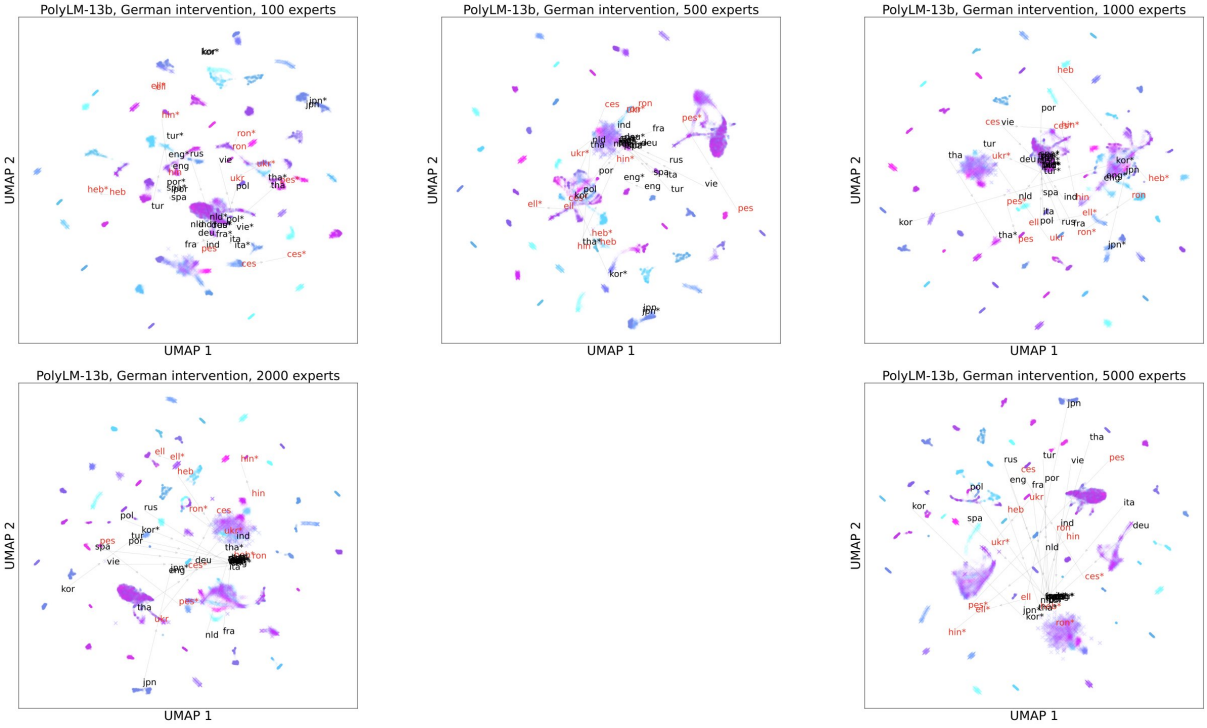


Figure 18: UMAP embeddings for intervention on German expert neurons

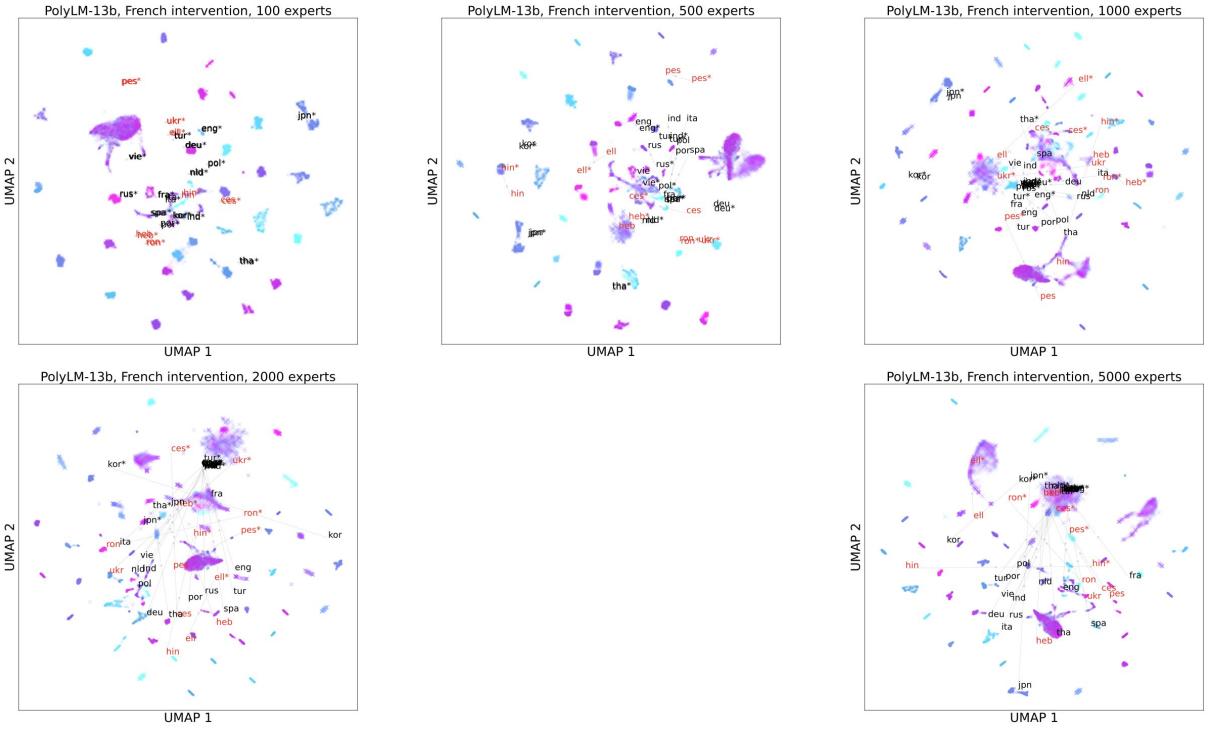


Figure 19: UMAP embeddings for intervention on French expert neurons

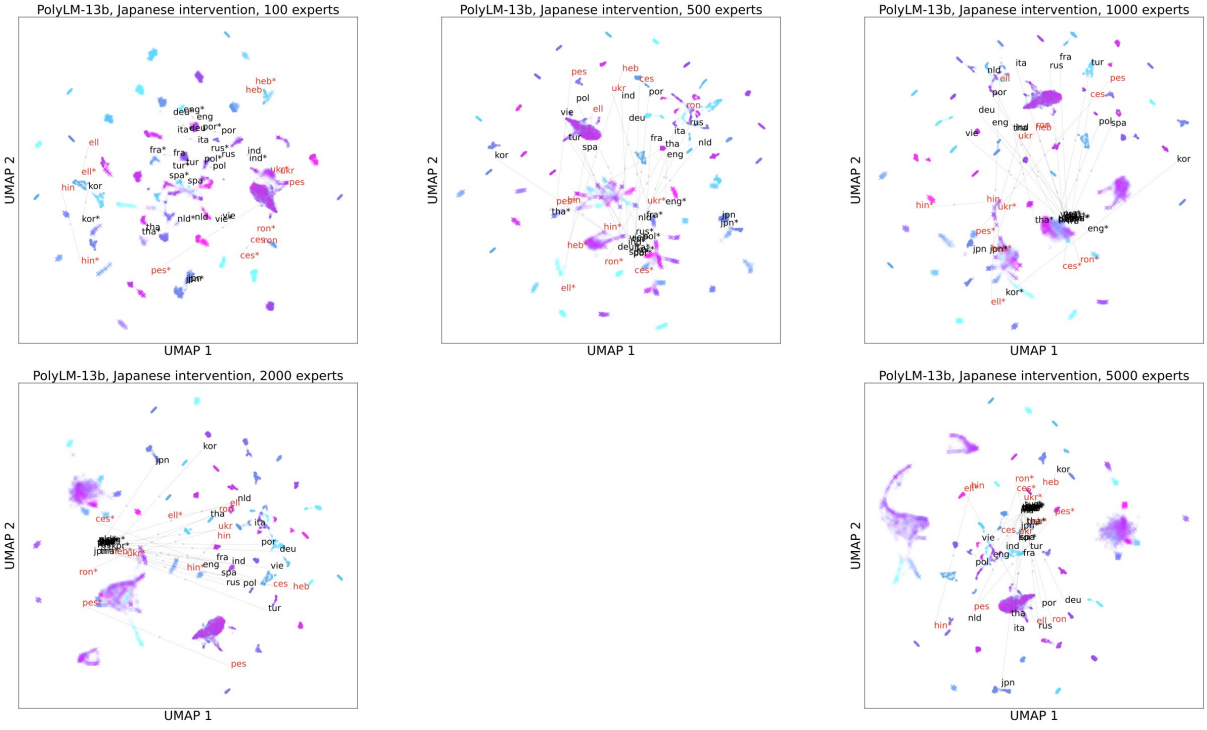


Figure 20: UMAP embeddings for intervention on Japanese expert neurons

D Paraphrase retrieval accuracy for four intervention-target languages

Model	Language	Top-1 Accuracy (Pre)	(Post)	(Mixed)
Bloom-7B	en	0.80	0.80	0.71
	fr	0.80	0.80	0.26
	de	0.72	0.75	0.22
	ja	0.47	0.59	0.07
Aya-8B	en	0.87	0.87	0.56
	fr	0.83	0.83	0.75
	de	0.82	0.82	0.62
	ja	0.70	0.76	0.55
PolyLM-13B	en	0.55	0.53	0.48
	fr	0.52	0.50	0.44
	de	0.50	0.55	0.39
	ja	0.57	0.57	0.32

Table 5: Top-1 accuracy results for the paraphrase retrieval task for four intervention languages. Pre= both the query and the candidate embeddings are from the original unintervened model; Post= both the query and the candidate embeddings are from the intervened model; Mixed = query embedding is from the original model and the candidates are from the intervened model.

E Top-1 cross-lingual retrieval accuracy for four intervention-target languages (query language is the same as the intervention target)

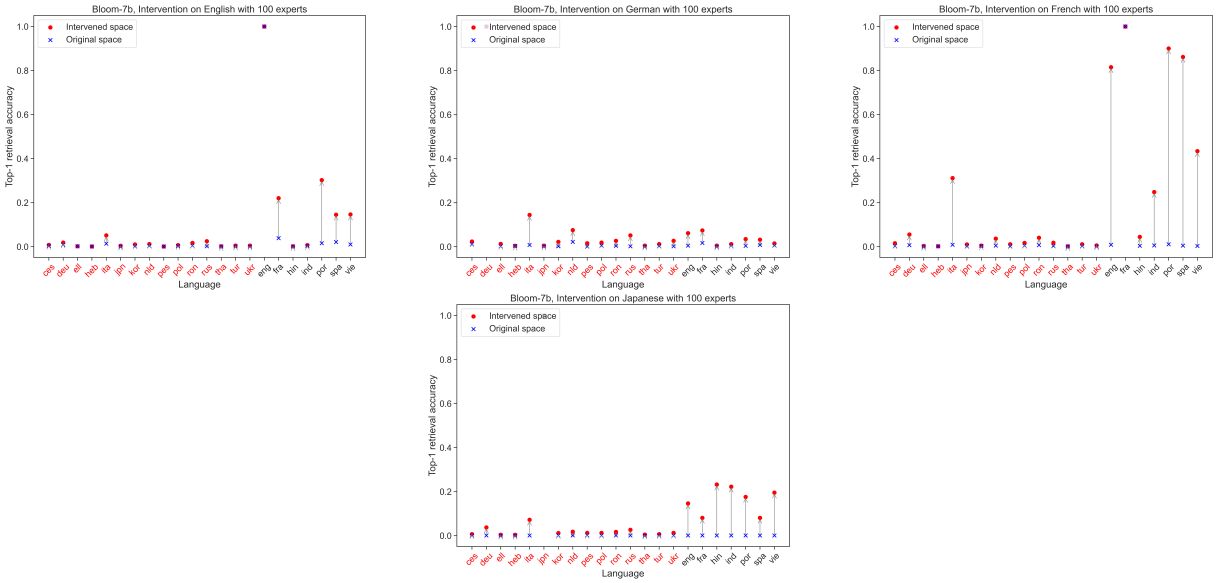


Figure 21: Top-1 retrieval accuracy for 22 languages in the Bloom-7B model. The language of the intervention is provided in the caption to each subfigure. The languages that are not in the training set for a given model are marked in red.

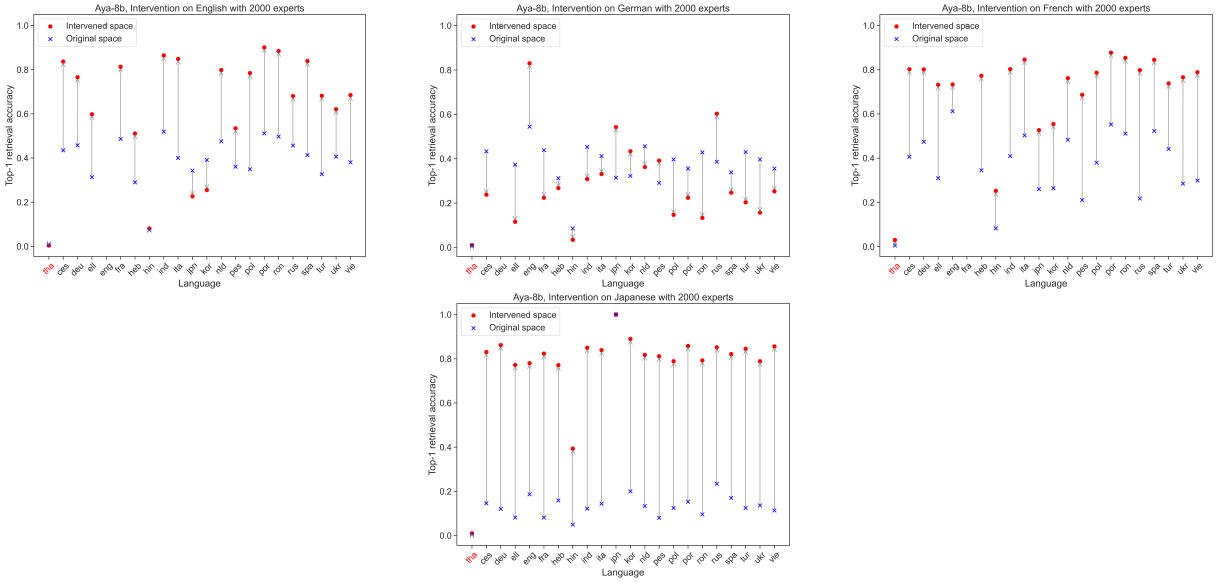


Figure 22: Top-1 retrieval accuracy for 22 languages in the Aya-8B model. The language of the intervention is provided in the caption to each subfigure. The languages that are not in the training set for a given model are marked in red.

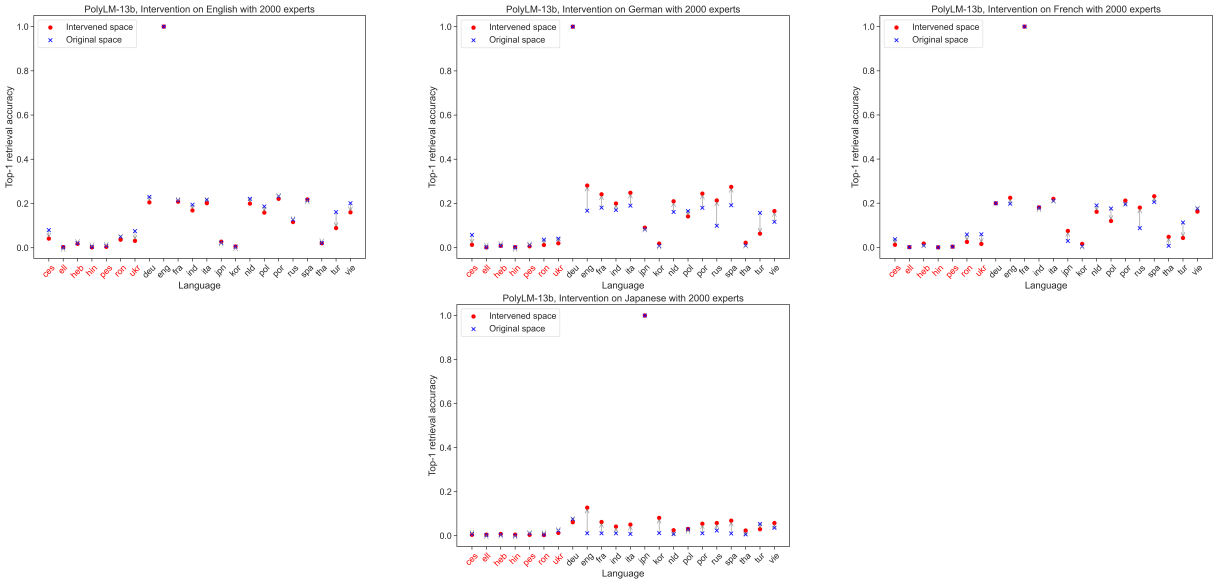


Figure 23: Top-1 retrieval accuracy for 22 languages in the PolyLM-chat-13b model. The language of the intervention is provided in the caption to each subfigure. The languages that are not in the training set for a given model are marked in red.

F Top-1 cross-lingual retrieval accuracy for four intervention-target languages (query language is different from the intervention target)

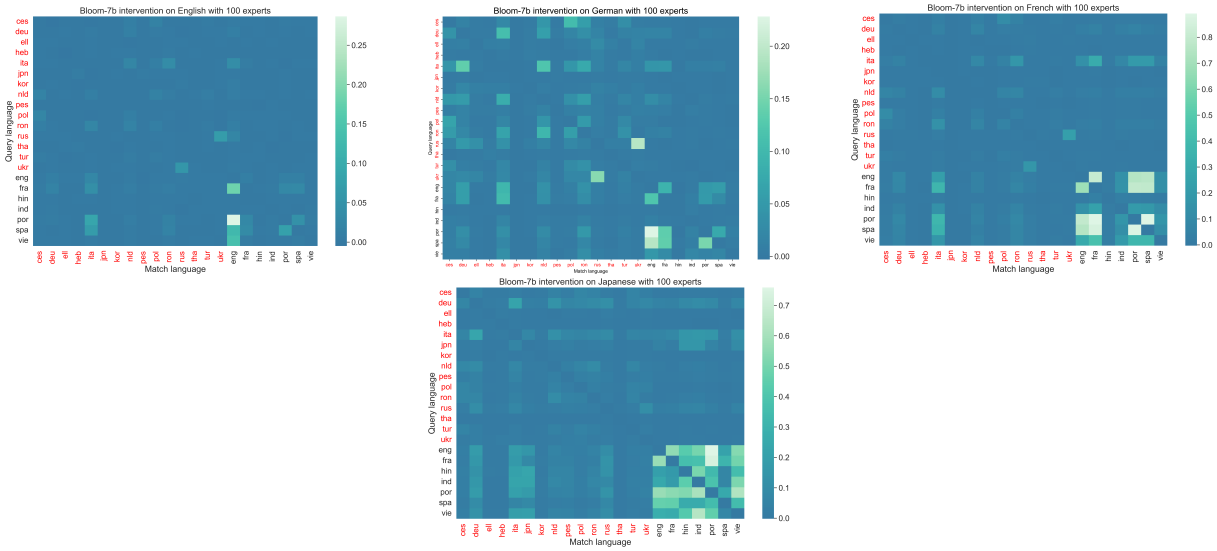


Figure 24: $(\text{Top-1 accuracy}_{\text{post-intervention}} - \text{Top-1 accuracy}_{\text{pre-intervention}})$ for Bloom-7B. The language of the intervention is provided in the caption to each subfigure. The languages that are not in the training set are marked in red.

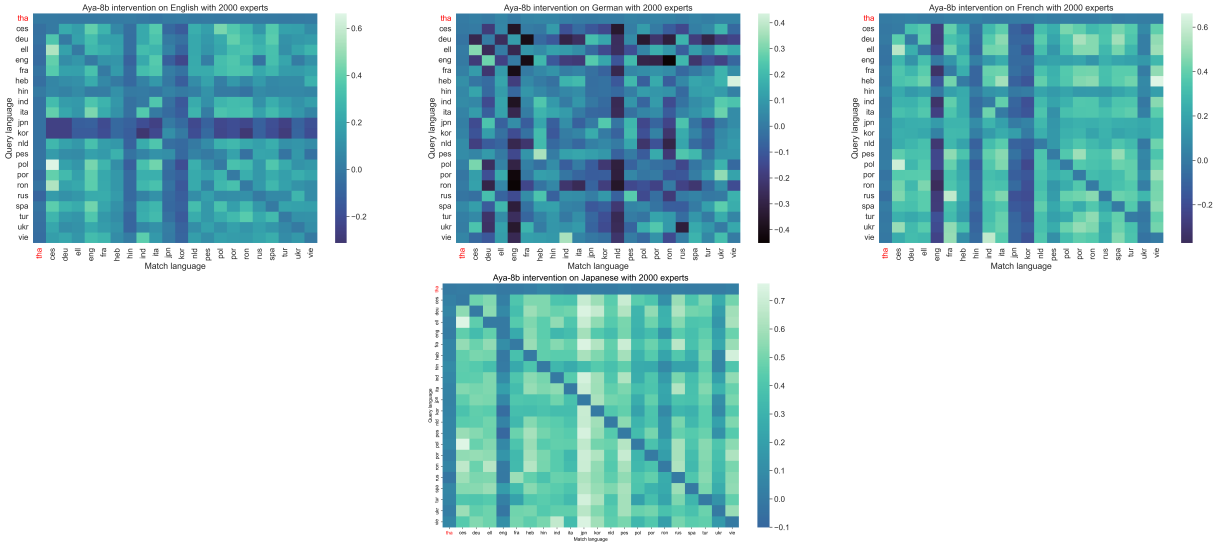


Figure 25: $(\text{Top-1 accuracy}_{\text{post-intervention}} - \text{Top-1 accuracy}_{\text{pre-intervention}})$ for Aya-8B. The language of the intervention is provided in the caption to each subfigure. The languages that are not in the training set are marked in red.

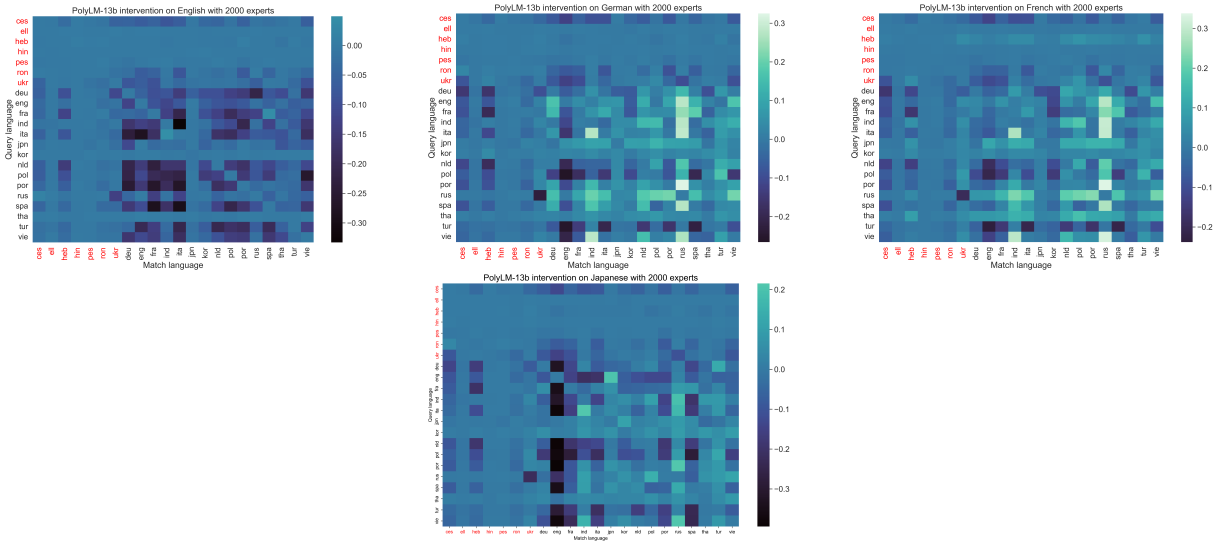


Figure 26: $\left(\text{Top-1 accuracy}_{\text{post-intervention}} - \text{Top-1 accuracy}_{\text{pre-intervention}}\right)$ for PolyLM-13B. The language of the intervention is provided in the caption to each subfigure. The languages that are not in the training set are marked in red.

G Results for the interventions on random neurons

G.1 Top-1 paraphrase retrieval accuracy after the intervention on random neurons

Model	Language	Accuracy		
		Pre	Post	mixed
Bloom-7B	en	0.80	0.80	0.78
	es	0.80	0.79	0.62
	fr	0.80	0.71	0.00
	de	0.72	0.72	0.62
	ja	0.47	0.45	0.35
PolyLM-13B	en	0.55	0.55	0.51
	es	0.53	0.53	0.48
	fr	0.53	0.54	0.50
	de	0.50	0.54	0.45
	ja	0.60	0.58	0.23
Aya-8B	en	0.87	0.81	0.00
	es	0.85	0.73	0.01
	fr	0.83	0.70	0.01
	de	0.82	0.70	0.00
	ja	0.70	0.44	0.00

Table 6: Top-1 accuracy results for the paraphrase retrieval task for five intervention languages for the intervention on random neurons. Pre= both the query and the candidate embeddings are from the original unintervened model; Post= both the query and the candidate embeddings are from the intervened model; Mixed = query embedding is from the original model and the candidates are from the intervened model.

G.2 UMAP embeddings for the interventions on random neurons

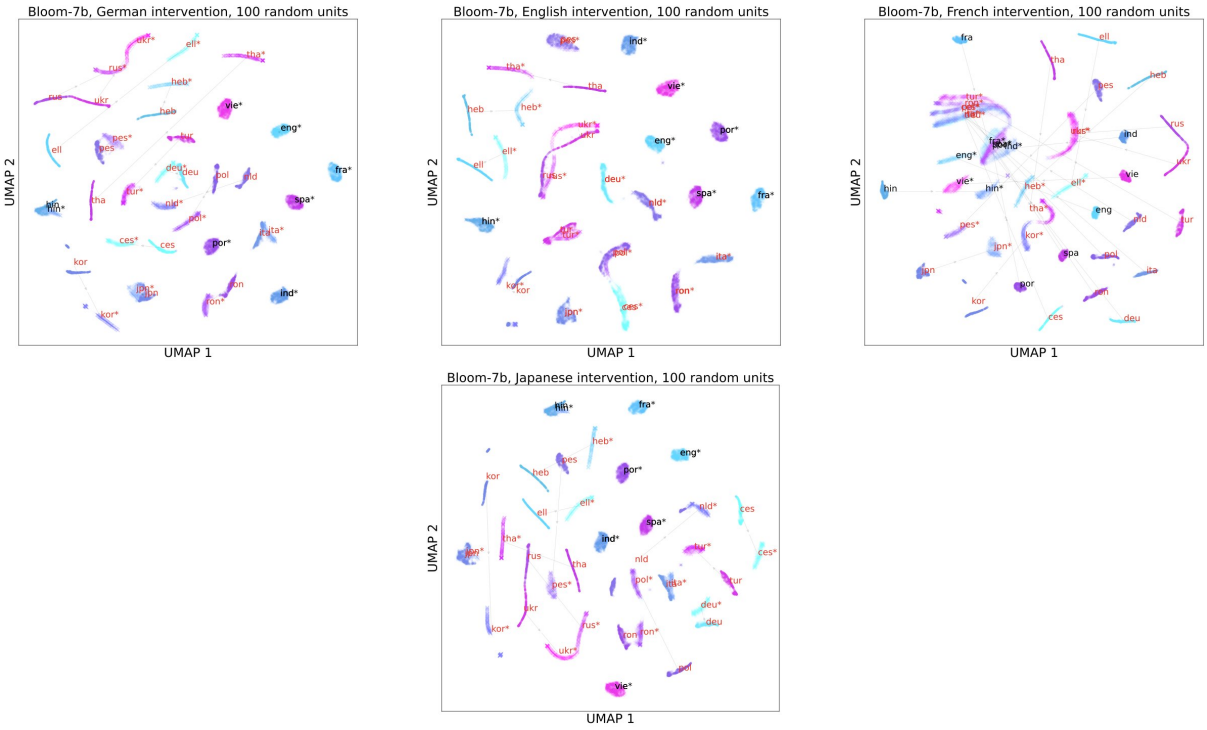


Figure 27: UMAP embeddings for interventions on random neurons for Bloom-7B

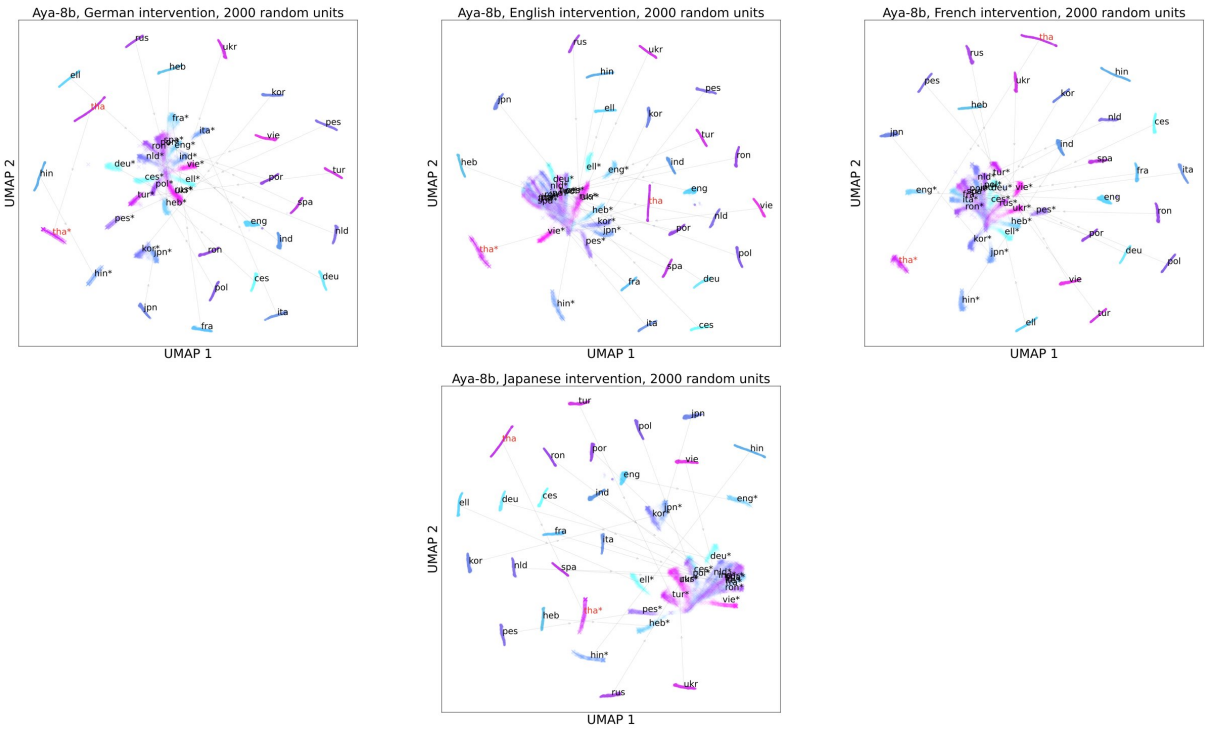


Figure 28: UMAP embeddings for interventions on random neurons for Aya-8B

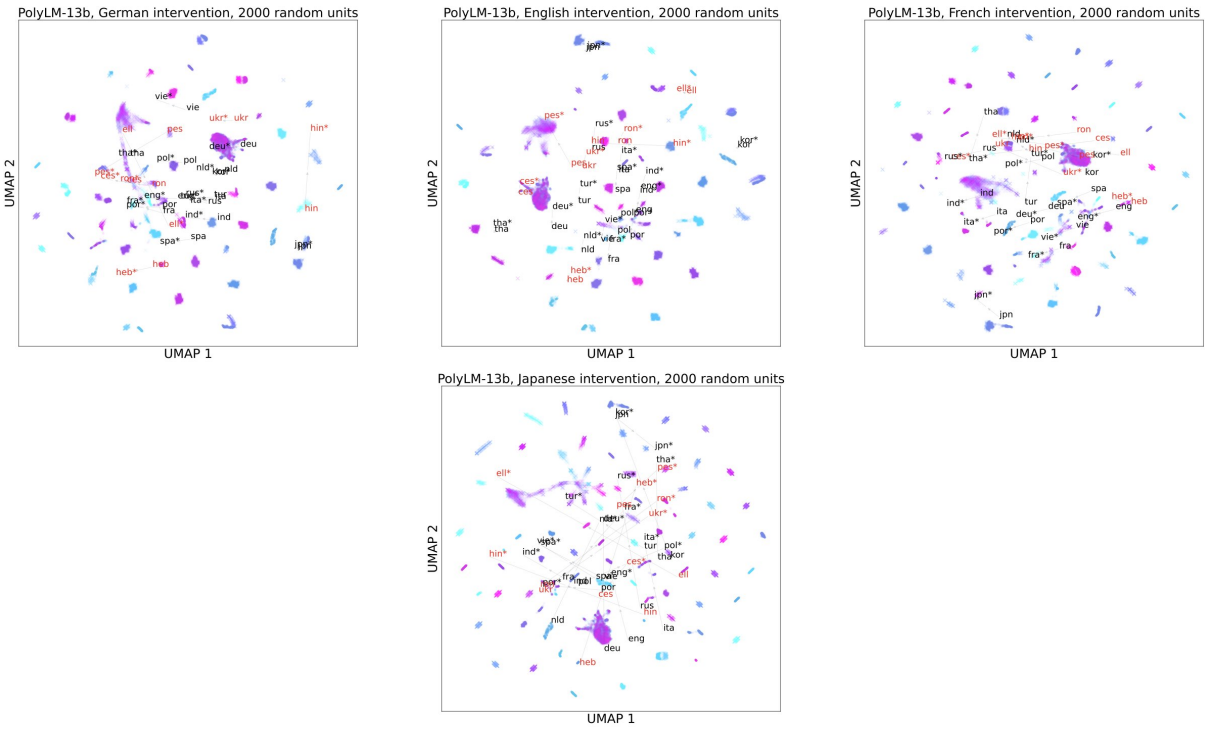


Figure 29: UMAP embeddings for interventions on random neurons for PolyLM-13B

G.3 Top-1 Retrieval accuracy for interventions on random neurons

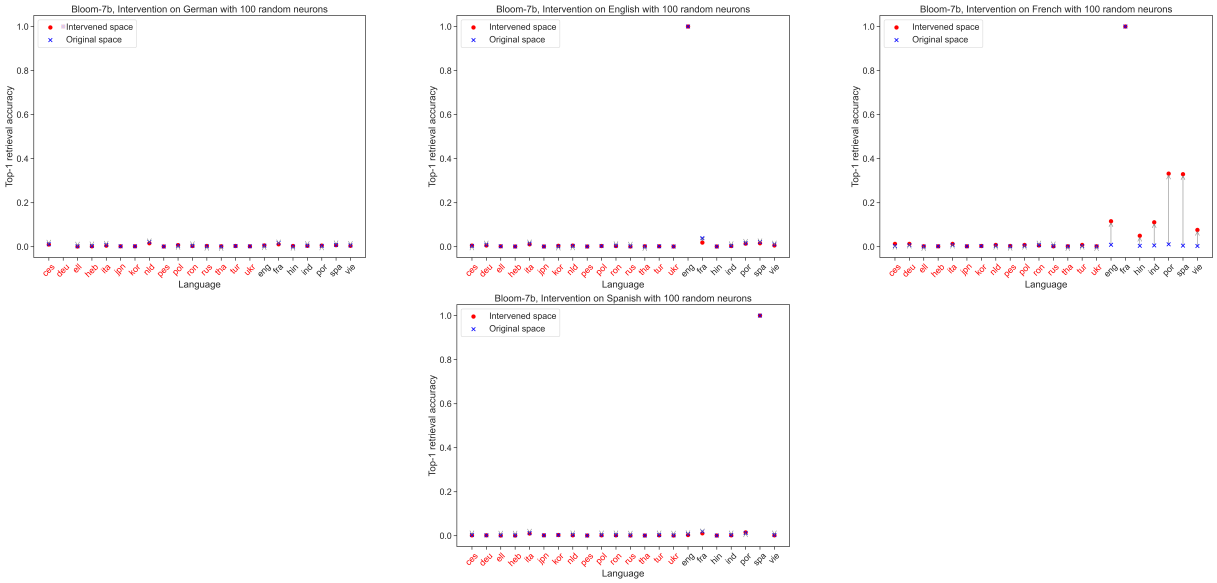


Figure 30: Top-1 retrieval accuracy for 22 languages in the Bloom-7B model with the intervention on 100 random neurons. The language of the intervention is provided in the caption to each subfigure. The languages that are not in the training set are marked in red.

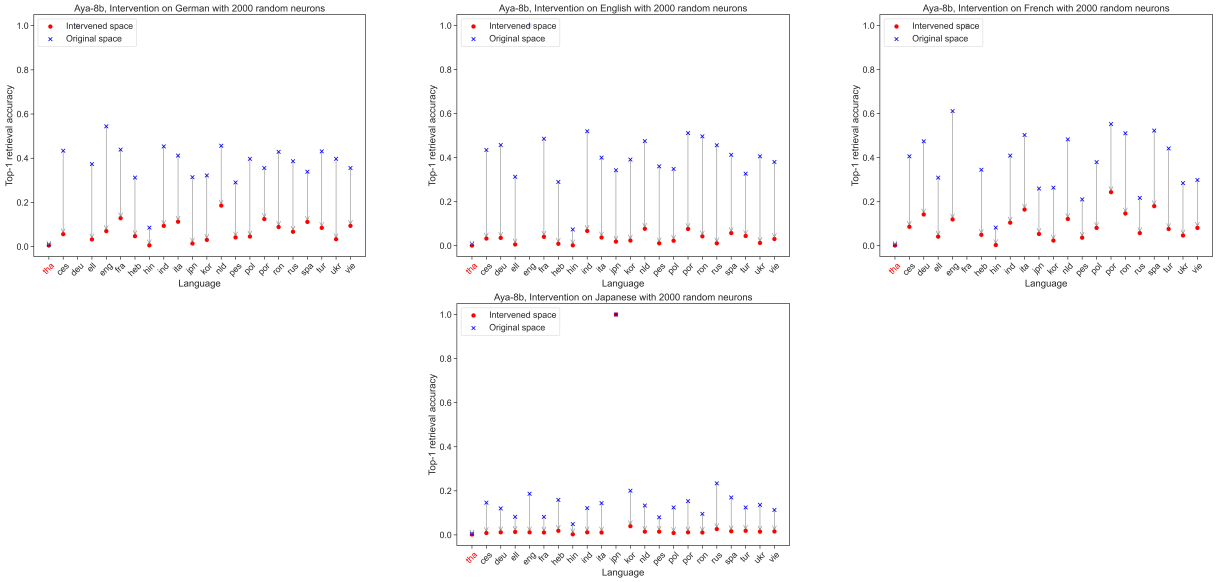


Figure 31: Top-1 retrieval accuracy for 22 languages in the Aya-8B model with the intervention on 2000 random neurons. The language of the intervention is provided in the caption to each subfigure. The languages that are not in the training set are marked in red.

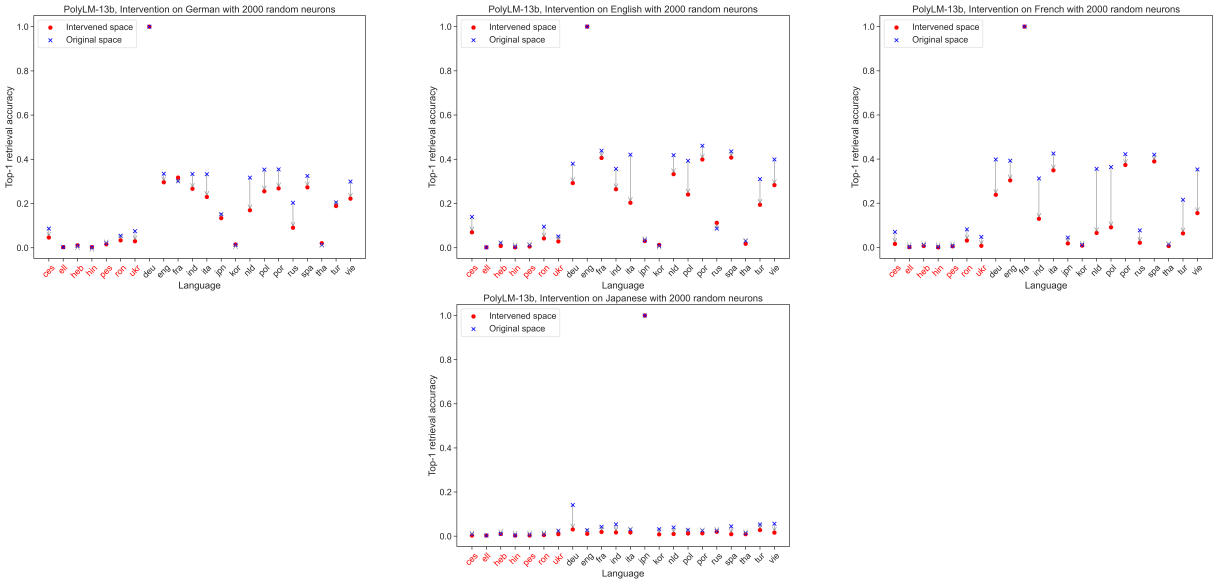


Figure 32: Top-1 retrieval accuracy for 22 languages in the PolyLM-13b-chat model with the intervention on 2000 random neurons. The language of the intervention is provided in the caption to each subfigure. The languages that are not in the training are marked in red.

H Cross-lingual retrieval results on BUCC-18 and Tatoeba

Model	Language	Top-1 Accuracy Pre	Post
Tatoeba			
Aya-8B	es	0.114	0.415
	fr	0.087	0.251
	de	0.119	0.444
	jp	0.034	0.307
Bloom-7B	es	0.008	0.551
	fr	0.011	0.434
	de	0.006	0.032
	jp	0.002	0.043
PolyLM-13B	es	0.082	0.178
	fr	0.067	0.130
	de	0.029	0.171
	jp	0.000	0.040
BUCC-18			
Aya-8B	fr	0.012	0.073
	de	0.017	0.332
Bloom-7B	fr	0.000	0.287
	de	0.000	0.02
PolyLM-13B	fr	0.006	0.286
	de	0.008	0.281

Table 7: Top-1 retrieval for the intervetion on five target languages for Tatoeba and BUCC-18. Pre= original model; Post= intervened model.

I Computational budget

All experiments were run on 8 A100(80GB) GPUs. The total approximate running time for 90 GPU/hours Aya-8B, 120 GPU/hours for PolyLM-13B, and 110 GPU/hours for Bloom-7B.

J License and Attribution

All datasets used in this work are supported by public licenses. PAWS-X, Tatoeba, BUCC are part of the XTREME benchmark licensed under Apache; Flores200 is licensed under Creative Commons. The pre-trained models used in this work are also supported by public licenses Bloom-7B (RAIL 1.0), Aya-8B (Creative Commons), and PolyLM-chat-13B (Apache).