When Meanings Meet: Investigating the Emergence and Quality of Shared Concept Spaces during Multilingual Language Model Training

Felicia Körner[1,2], **Max Müller-Eberstein**[3,4], **Anna Korhonen**[5], **Barbara Plank**[1,2]
[1]MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
[2]Munich Center for Machine Learning (MCML), Munich, Germany
[3]University of Tokyo, Japan
[4]IT University of Copenhagen, Denmark
[5]Language Technology Lab, University of Cambridge, United Kingdom
Correspondence: f.koerner@lmu.de

Abstract

Training Large Language Models (LLMs) with high multilingual coverage is becoming increasingly important—especially when monolingual resources are scarce. Recent studies have found that LLMs process multilingual inputs in shared concept spaces, thought to support generalization and cross-lingual transfer. However, these prior studies often do not use causal methods, lack deeper error analysis or focus on the final model only, leaving open how these spaces emerge *during training*. We investigate the development of language-agnostic concept spaces during pretraining of EuroLLM through the causal interpretability method of activation patching. We isolate cross-lingual concept representations, then inject them into a translation prompt to investigate how consistently translations can be altered, independently of the language. We find that *shared concept spaces emerge early* and continue to refine, but that *alignment with them is language-dependent*. Furthermore, in contrast to prior work, our fine-grained manual analysis reveals that some apparent gains in translation quality reflect shifts in behavior—like selecting senses for polysemous words or translating instead of copying cross-lingual homographs—rather than improved translation ability. Our findings offer new insight into the training dynamics of cross-lingual alignment and the conditions under which causal interpretability methods offer meaningful insights in multilingual contexts.

1 Introduction

Most Large Language Models (LLMs) are trained primarily on English, and even targeted multilingual training is typically imbalanced across languages (Zhao et al., 2024; Liu and Fu, 2024). Despite this, LLMs exhibit emergent cross-lingual alignment, enabling transfer of capabilities to other languages (Chirkova and Nikoulina, 2024).

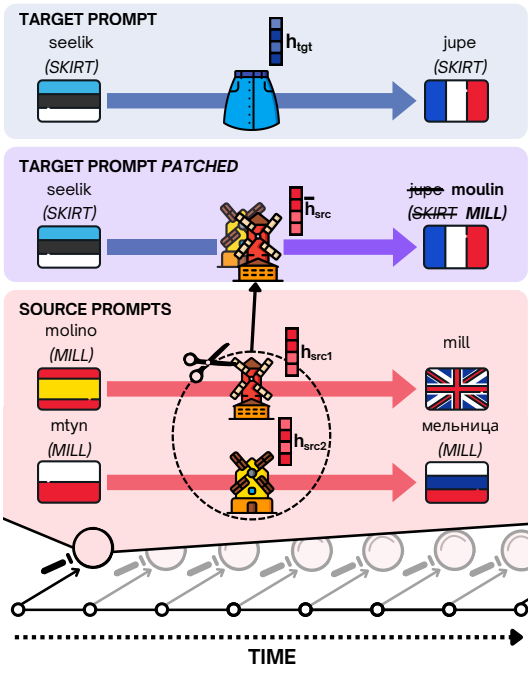

Figure 1: We extend cross-lingual concept patching (Dumas et al., 2025) through a systematic application over pretraining. The first row shows the vanilla translation of the concept SKIRT from Estonian → French. In the second, we replace the activation with averaged activations from Spanish → English and Polish → Russian for MILL. Though the representation comes from a *distinct* set of languages, it can induce MILL in French.

Since data is scarce for many languages (Joshi et al., 2020; Ranathunga and de Silva, 2022), this behavior is critical for narrowing the performance gap between English and other languages and to develop more inclusive language technologies (Peppin et al., 2025; Li et al., 2024). Understanding how and *when* cross-lingual alignment arises is crucial for determining the necessary factors for training multilingually performant LMs.

Recent interpretability work has found that models rely on shared concept[1] spaces to process mul-

[1]We use the term "concept" to describe a mental representation expressed by different words across languages.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3149–3169
March 24-29, 2026

tilingual input (Wendler et al., 2024). To causally demonstrate the existence of such spaces Dumas et al. (2025) propose *cross-lingual concept patching* (see Fig. 1 for an illustration). In cross-lingual concept patching, latent vectors are extracted from a prompt translating one concept (e.g. MILL in Fig. 1) across one set of language pairs (e.g. Spanish → English and Polish → Russian) and artificially applied during the translation of a *different* concept (e.g. SKIRT) across a *different* language pair (e.g. Estonian → French). If this patched inference produces the first concept (MILL) in the new language (French), this provides causal evidence for shared concept representations, as language-specific ones would not be able to change the concept while keeping the output language fixed.

In this work, we re-frame cross-lingual concept patching as a causal analysis of cross-lingual alignment. The key intuition is that *two factors* are involved in successfully inducing a concept change from concept representations averaged across languages: First, the languages the representations are drawn from must be aligned with the shared space. Second, the output language must be aligned with these representations, even when they stem from other languages.

However, it is not obvious how shared spaces develop, and how models learn to map them to a particular output language—a gap we aim to address in this study. One hypothesis is that models initially rely on language-specific representations and only gradually abstract these into a shared space over time. To test this, we apply cross-lingual concept patching across intermediate checkpoints during pretraining. Specifically, our contributions are:

- We introduce a systematic framework for cross-lingual concept patching: we carefully curate compatible source and target concept pairs (Section 2 and Section 4.2) and devise distinct language settings, including a control task (Section 4.4). We find that **alignment with shared spaces depends on training data proportion**.
- We conduct the first fine-grained investigation into the emergence of shared concept spaces during multilingual pretraining, studying intermediate checkpoints of EuroLLM, which provides a higher granularity of checkpoints than previously-studied multilingual LLMs (Section 5). We find that **shared concept spaces arise early in pretraining**.
- We apply cross-lingual concept patching to the largest and most diverse set of concepts to date (Section 4.2).
- We provide the first manual analysis of errors under cross-lingual concept patching (Section 5.3), uncovering nuance in its failure modes, and interpretation of its effect.

We share our implementation to facilitate further research.[2]

2 Related Work

Recent mechanistic interpretability work has proposed that multilingual LLMs process input in three stages: i) mapping from the input language to a shared concept space, ii) processing conceptual information, and iii) mapping the result to the output language (Zhao et al., 2024; Schut et al., 2025; Wendler et al., 2024; Tezuka and Inoue, 2025; Zhong et al., 2025).

In a similar vein, several studies have identified shared components of multilingual LLMs, largely through neuron analysis (Stanczak et al., 2022; Chen et al., 2024; Cao et al., 2024; Tang et al., 2024; Zhang et al., 2025; Wang et al., 2024b; Kojima et al., 2024), though often not causally. Recent work highlights the importance of causal interventions for robust mechanistic claims (Mueller et al., 2024), a perspective we adopt here. Methodologically, Dumas et al. (2025) is closest to our work; they introduce cross-lingual concept patching, causally demonstrating the existence of language-agnostic concept spaces. Feucht et al. (2025) use the same dataset to show that concept induction heads, used for copying concepts mono- and cross-lingually, are language-agnostic. Both works use the same method to demonstrate shared components at different granularity, but rely on a dataset of constructed concept pairs via LLM translation. In contrast, we re-frame the method to *analyze* how shared spaces develop. To do so, we introduce a systematic framework to compare results across different language settings and pretraining checkpoints, and propose a novel, carefully curated concept dataset extracted from human translations.

Despite growing evidence of shared spaces in multilingual models, little is known about how such spaces emerge throughout pretraining. Progress is

[2] https://github.com/mainlp/shared-concept-spaces

hampered by the limited availability of pretraining checkpoints for state-of-the-art multilingual LLMs. Until the recent release of Apertus (Hernández-Cano et al., 2025), BLOOM (Workshop et al., 2022) was the only model with publicly available checkpoints, yet provides only 4–8,[3] potentially obscuring finer-grained shifts. Studies of BLOOM use neuron analysis, probing, and, in some cases, concept-level analysis (Zeng et al., 2025; Wang et al., 2024a; Riemenschneider and Frank, 2025). In contrast, we study 26 pretraining checkpoints of EuroLLM, and provide additional results for Apertus 8B and OLMo-2 7B in Appendix A.5.

3 Methodology

We study model outputs under cross-lingual concept patching (Dumas et al., 2025), an activation patching (Vig et al., 2020) approach. Activation patching is a mechanistic interpretability method, whereby activations of a forward pass of a "target" prompt are overwritten, or "patched", with activations from a previous run of a different, "source", prompt. The effect of this intervention provides insight about the role of the patched component and the patch itself. We describe the variant of cross-lingual concept patching in the following subsections, and give more technical details in Appendix A.1.

3.1 Few-shot Concept Translation

As a testbed for cross-lingual concept patching, we use the task of few-shot, word-level translation. The task considers the translation of a concept C from an input language, ℓ^{in}, to an output language, ℓ^{out}. We denote a language-agnostic concept with capital letters, for example $C = \text{MILL}$. $C^{\ell^{\text{out}}}$ is the concept expressed as a word in the output language, for example C^{FR} ="moulin".

Each translation is formulated as a prompt in the format, "ℓ^{in}: $C^{\ell^{\text{in}}}$ - ℓ^{out}: $C^{\ell^{\text{out}}}$", where the output concept is omitted during inference. The prompt is further prepended with five few-shot examples. For example, a translation prompt from Spanish to English for $C = \text{MILL}$:

Español: "cultura" - English: "culture"
Español: "sentido" - English: "sense"
...
Español: "molino" - English: "

[3]Depending on model size.

3.2 Cross-lingual Concept Patching

To induce a change in translation, we apply a concept patching intervention during inference (see Fig. 1 for an illustration). First, we compute a "patch", an activation for C_{src}, the source concept that should be generated. This activation is extracted from a few-shot translation prompt in which C_{src} is the intended translation at the position of the last token of the word to be translated for some layer j and all subsequent ones. Extending our previous example, $C_{\text{src}} = \text{MILL}$, $\ell^{\text{in}}_{\text{src}}$ = Spanish, and $\ell^{\text{out}}_{\text{src}}$ = English, and we extract the activation at the last token of "molino". We repeat this for several source language pairs, and take the mean of the activations.

Next, we prompt the model to translate a semantically distinct target concept C_{tgt} across an entirely different target language pair. E.g., for $C_{\text{tgt}} = \text{SKIRT}$, $\ell^{\text{in}}_{\text{tgt}}$= Estonian, and $\ell^{\text{out}}_{\text{tgt}}$ = French:

Eesti: "korporatsioon" - Français: "fraternité"
Eesti: "lambatall" - Français: "agneau"
...
Eesti: "seelik" - Français: "

However, during inference of this target prompt, we artificially apply the averaged concept activations from the source prompts, at the equivalent position, and all subsequent layers. The goal of the intervention is to induce the source concept (MILL) in the target output language, French, even if the target prompt aims to translate the target concept (SKIRT). I.e., in our example: "Eesti: "seelik" - Français: "moulin".

Crucially, since the source languages are distinct from the target languages, and the source concept is distinct from the target concept, successfully inducing the source concept in the target language hinges on two representational criteria: First, concept representations must be *language-agnostic* in order for the mean over representations across languages to be meaningful. Second, the target output language must also be *aligned* with this language-agnostic space in order for the model to generate the source concept in the target output language.

4 Experimental Setup

4.1 Model

To investigate the emergence of shared concept spaces throughout training, and to link them to concrete training strategies, we analyze 26 pretraining checkpoints of the open-weight EuroLLM-1.7B

Category	Languages	Training Data Proportion
very-high	en	50% phase one, 32.5% phase two
high	es, fr	around 6%
med-high	zh	between 3–4%
med	ru, pl	between 2–3%
low	et, fi	around 1%
unseen but similar	yue	–
unseen	sw, cy	–

Table 1: Our languages of focus, selected from Multi-SimLex (Section 4.2) and categorized based on EuroLLM's training data proportions (Section 4.1).

(Martins et al., 2025, henceforth EuroLLM). EuroLLM is trained primarily on European languages; we categorize our languages of focus based on their proportion in EuroLLM's training data in Table 1.[4] The model is of particular interest to our study, as it offers a high granularity of intermediate checkpoints, and is trained in *two phases* with different multilingual training data compositions.

Specifically, its training data is composed of web data, code/math data, high-quality data (Wikipedia, arXiv, books, medical texts), and parallel data in two different configurations. In phase one (0–90% of training; 3.6T tokens), 77% of the data is *web*, and *parallel data* is primarily aligned with respect to English. In phase two (90%–100% of training; 0.4T tokens), *web* is reduced to 46.6%, while the *high-quality data* ratio is increased and upsampled from 9% to 34.4%. The amounts of *code/math* and *parallel data* remain similar, however, the parallel data is drawn from multi-aligned sources beyond English. As such, it is of particular interest to investigate how each training phase manifests in terms of cross-lingual concept alignment.

4.2 Data

To curate our dataset, we leverage Multi-SimLex (Vulić et al., 2020) as it provides a large, linguistically informed and diverse concept set in contrast to prior work, which focused on synthesized "picturable" concepts only (Feucht et al., 2025; Dumas et al., 2025). Multi-SimLex consists of 1,888 word pairs rated for lexical similarity, with human translations in 13 languages. The languages are typologically diverse, spanning both low- and high-resourced languages. We focus on the following 11 languages, covering eight languages included and three excluded from EuroLLM's training data (Section 4.1): English (en), Mandarin (zh), Welsh (cy), Estonian (et), Finnish (fi), French (fr), Polish (pl), Russian (ru), Spanish (es), Swahili (sw), Cantonese (yue). Multi-SimLex was originally created for word-pair similarity judgments, however, we do not retain the original word pairings. Instead, we treat the dataset as a multiway parallel lexicon, extracting a vocabulary of 2,147 concepts and their translations. Prior work has shown that word classes may behave differently—for example, function words are often language-specific (Schut et al., 2025). To avoid confounding effects of word class, and due to the extensive scope of our experiments across languages and checkpoints, we focus on a single class and follow prior work (Dumas et al., 2025; Feucht et al., 2025) in restricting analysis to nouns.

We greedily select 256 *compatible* source and target concept pairs from this pool, where we deem a pair compatible if there is no word overlap across all translations for the two concepts, for a total of 398 distinct concepts. For each source concept, we construct source prompts across all source language pairs, and for each target concept we construct target prompts across all target language pairs. So far, cross-lingual concept patching has been applied to 200 pairs constructed from a pool of about 100 concepts (Feucht et al., 2025; Dumas et al., 2025).

4.3 Evaluation

Activation patching evaluations typically measure the increase in next-token probability for the expected first token (Heimersheim and Nanda, 2024). However, as our results in the word-level translation experiments in Section 5.1 reveal, this metric is poorly suited for our task, where there are often multiple valid next-token sequences for a given concept. Unlike prior work, we do not expand the dataset with synonyms for evaluation, as we favor a precise fine-grained analysis, while covering a more diverse set of concepts and study of lower-resourced languages (see Appendix A.2). Finally, relying on the proxy of first-token probability may mask errors where the first-token appears correct, but the model outputs the wrong prediction. For example, consider "organ" and "organizer", which may share a first token, but are semantically distinct. To address these limitations, we evaluate the model on full token sequences and measure word-level translation accuracy.

In particular, given N test samples, y_i the *source*

[4] Cantonese is given a special category, though it is not included in EuroLLM's training data, the model translates it fairly well (Section 5.1). We attribute this to Cantonese's similarity to Mandarin.

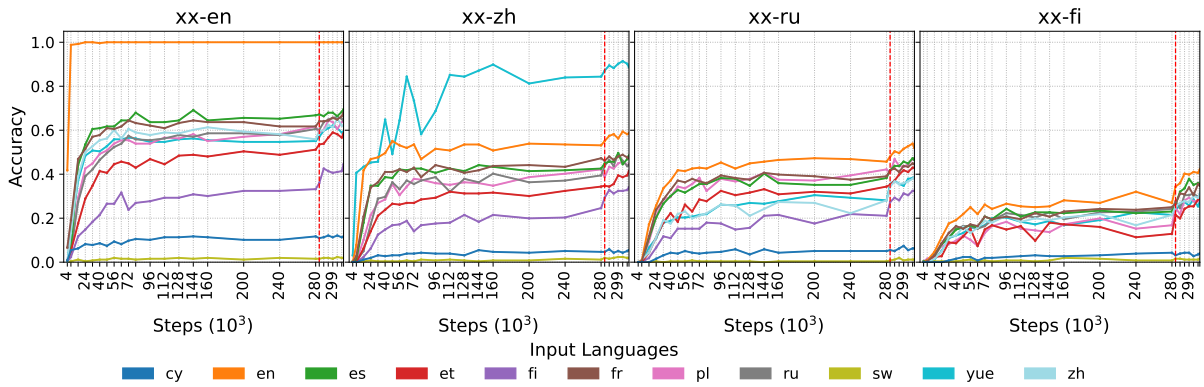

Figure 2: Mean word-level translation accuracy over checkpoints for source prompts used for patching, grouped by a selection of output languages. The red dotted line indicates the start of phase two of EuroLLM's training.

concept expressed in the *target* output language, and $\hat{y}_i$ the model prediction, we define mean word-level translation accuracy as:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\hat{y}_i = y_i\}.$$

This approach directly assesses whether patching meaningfully alters the translation, and furthermore, avoids cases where the first token appears correct but the final prediction is wrong. Finally, we conduct a multilingual manual evaluation (Section 5.3) to gain a deeper understanding of concept translation performance.

4.4 Language Settings

We use all languages in Table 1 except `cy` and `yue` as target input languages, treating `sw` as representative for unseen languages. We fix the target output languages to `en`, `ru`, and `zh`, covering three typologically distinct languages with diverse scripts.

For each target language pair $(\ell^{\text{in}}_{\text{tgt}}, \ell^{\text{out}}_{\text{tgt}})$, we define three sets of source language pairs used for patching: `seen`, `en_en` and `tgt`. `seen` is made up of languages in EuroLLM's training data, excluding `en` and the languages in the target pair. For example, for target pair `fr-en`, `seen` includes all ordered pairs of `es`, `zh`, `pl`, `ru`, `et`, `fi`, `yue`, for 42 total source language pairs.

`en_en` is a copying task, formulated in the same way as the translation task, except that both the input and output language are `en`. We include this as a control task and a strong baseline, reasoning that the `en_en` representations will be well-defined as English dominates EuroLLM's training data. `tgt` is an ablation, where only the source and target concepts differ, and the source language pair matches the target language pair. Both `en_en` and `tgt` consist of only a single language pair, hence, there are no concept-aligned representations to average over.

Given `msimlex` as the set of selected languages (Section 4.2), `seen`, `en_en` and `tgt` are defined:

$$\begin{aligned}
\texttt{ellm} &= \texttt{msimlex} \setminus \{\texttt{sw}, \texttt{cy}\} \\
\texttt{seen} &= \{(\ell_1, \ell_2) \mid \ell_1 \neq \ell_2, \\
&\qquad \ell_1, \ell_2 \in \texttt{ellm} \setminus \{\ell^{\text{in}}_{\text{tgt}}, \ell^{\text{out}}_{\text{tgt}}, \texttt{en}\}\} \\
\texttt{en_en} &= \{(\texttt{en}, \texttt{en})\} \\
\texttt{tgt} &= \{(\ell^{\text{in}}_{\text{tgt}}, \ell^{\text{out}}_{\text{tgt}})\}
\end{aligned}$$

As a baseline, we evaluate the unpatched translation of the source concept for the target language pair. I.e., if we aim to induce the source concept MILL for the target language pair `et-fr` through cross-lingual patching, we compare this to the vanilla translation of MILL for `et-fr`. We denote this as `src_unpatched`. Please see Appendix A.3 for example prompts to illustrate these settings.

5 Results

5.1 Word-Level Translation

We first evaluate EuroLLM's checkpoints on the unpatched word-level translation task to confirm our assumptions about the relationship between training data proportion (Table 1) and translation performance. Unsurprisingly, we find that translation accuracy (see Fig. 2 for a selection of languages, Fig. 9 in the App. for all) correlates with training data proportion. The model largely fails to translate words where the input or the output language is unseen, namely, `cy` and `sw`. The exception is `yue`, which is well-translated, in particular to or from `zh`, likely due to its similarity to `zh`.

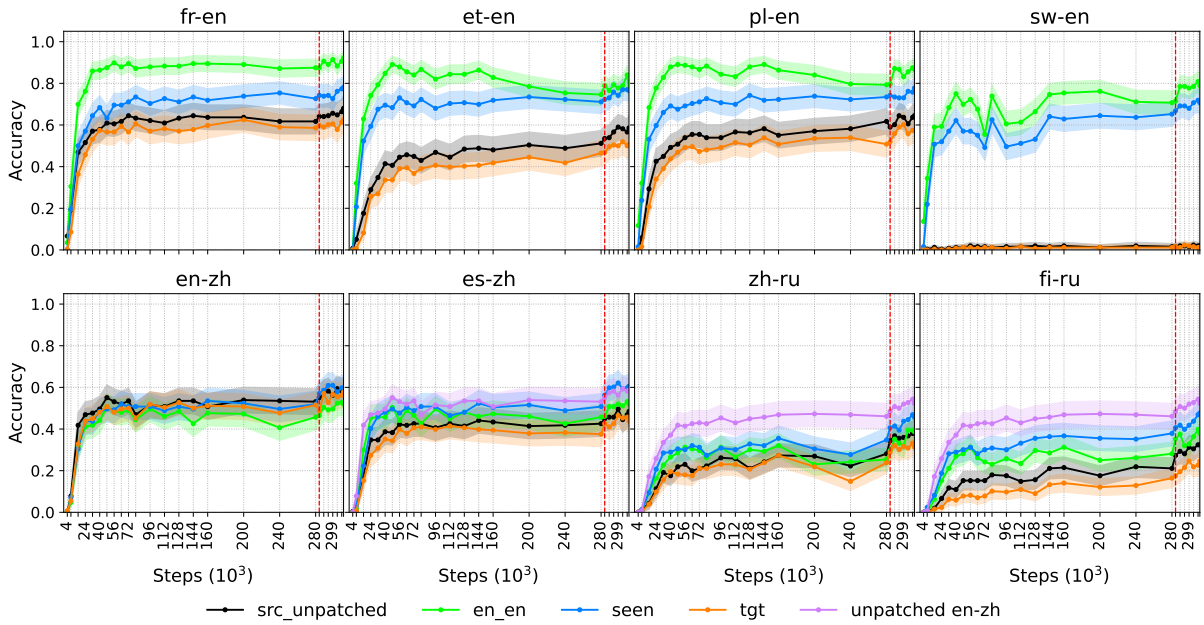

Figure 3: Mean word-level translation accuracy over checkpoints under different patching settings for a selection of target language pairs. We overlay `en-xx`, where `xx` is `zh` or `ru` on `es-zh`, and `fi-ru`, `zh-ru`, respectively. The red dotted line indicates the start of phase two of EuroLLM's training. We show 95% CI over 256 samples.

We interestingly observe that translation accuracy improves in the second phase of training, in particular for lower-resourced languages, such as `pl`, `ru`, `fi`, and `et`. We hypothesize that the multiway parallel data introduced in this phase helps align these languages, improving translation quality (Shen et al., 2025; Lin et al., 2025).

The `en_en` curve shows that the model quickly becomes proficient at copying input words—by the second checkpoint words are consistently copied. In fact, the model appears to prefer copying over translating in the early stages, even for the translation task (i.e., when processing the prompts for language pairs other than `en_en`). This behavior is in line with recent work suggesting that copying is learned before other tasks (Feucht et al., 2025).

Manual Analysis of Unpatched `fr-en` Using `fr-en` as a case study, we inspect the translations and find that many apparent errors can be attributed to dataset artifacts or model behavior. In particular, our analysis sheds light on the model's copying behavior. In some cases, these are loanwords, hence, the translation could be considered correct. E.g., when translating "balustrade", where the expected translation is "rail", the model consistently outputs "balustrade". However, cross-lingual homographs are also copied, resulting in an incorrect translation. For example, the model copies both "course" (expected: "racing"), and "coin" (expected: "corner").

Similarly, the results are confounded by polysemous words, which are also not captured by the translation dataset. For example, the model translates "femme" (expected: "wife") to "woman", or "pécheur" (expected: "sinner") to "fisherman". These factors motivate the manual analysis of errors under the `seen` patching setting (Section 5.3), to distinguish actual errors from reasonable outputs which are not captured by the automatic evaluation.

5.2 Cross-lingual Concept Patching

Patching from `seen` results in comparable or better accuracy than the unpatched translation across checkpoints for most target language pairs (Fig. 4), providing strong evidence that **language-agnostic spaces arise early in pretraining**.

Effect of Source Language Pair Groups `en_en` is initially the strongest patching setting across all target language pairs with `en` as the output language (Fig. 3, see App. Fig. 10, Fig. 11, Fig. 12 for full results for `en`, `zh`, `ru`, respectively). This aligns with our expectation that `en_en` provides well-aligned concept representations and the model is proficient at mapping these to `en` output. However, `seen` does not lag far behind. For target language pairs with `zh` or `ru` as the target output language, `en_en` is consistently on par with or *worse* than `seen`. This is surprising, as we expect both `zh` and `ru` to be well-aligned with `en`. In fact, un-

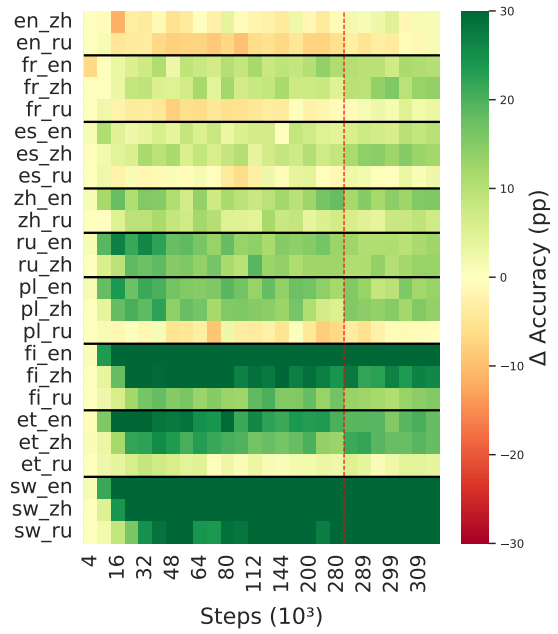

Figure 4: Net improvement (in percentage points, capped at ±35pp) of seen over unpatched translation accuracy. Each row shows a target language pair, grouped by input and ordered by output language; each column one checkpoint. The red dotted line indicates the start of phase two of EuroLLM's training.

patched en−ru and en−zh are an upper bound for the patched translation accuracy for target output ru and zh respectively. Overall, the similar behavior under the en_en and seen settings across checkpoints suggests that language-specific concept spaces do not strongly precede the emergence of shared spaces.

As an additional observation, accuracy under our ablation setting, tgt, is comparable to or lower than the unpatched translation, particularly for lower-resourced languages. Thus, when source and target language pairs are identical, patching results in a lower accuracy for the induced source concept, suggesting that information from the forward pass of the target prompt is not entirely overwritten.

Effect of Target Language Pair Under patching, the input language has much less of an effect on performance than in the unpatched setting. This aligns with recent work on multilingual processing, which suggests that models map from the input language to a shared space before mapping back to the output language (Wendler et al., 2024; Zhao et al., 2024). Since we patch at intermediate layers, the model's ability to map input from each language to the shared concept space becomes less important. Nevertheless, the input language does play a role. This is most visible for the target pair sw−en, where accuracy fluctuates significantly more than for other language pairs. Fig. 4 shows the impact of the target output language; the lower-resourced output language ru sees less gains from patching than zh, which, in turn, sees less gains than en. We hypothesize this reflects how well each output language is aligned with shared spaces, in accordance with theory on multilingual processing.

Concept-Specific Effects Since more frequent words have been found to be better aligned across languages (Peng and Søgaard, 2024), we analyze the effect of frequency on translation behavior under patching. We focus on the target output language en; assuming that frequency estimates derived from English fastText embeddings (Grave et al., 2018)[5] provide a better approximation of EuroLLM's token frequency distribution than analogous estimates for ru or zh, which are less represented in the pretraining corpus.

Under the seen setting, the model outputs the target word more often than a synonym for more frequent target words. This likely reflects the relative frequency of the synonym to the target word, and appears to hold for the unpatched setting, though less clearly. The most frequent words are mostly translated correctly early on, and only occasionally regress. For less frequent targets, the patched translation is more often incorrect, with some almost never translated correctly. The relationship between frequency and translation quality is far less pronounced for the unpatched case. These observations hold across target language pairs with output language of en (see Fig. 13 in the App. for categorical maps for es−en, fr−en, zh−en, ru−en).

Notably, for fr and es patching yields more incorrect translations for the least frequent concepts compared to unpatched. We attribute this to stronger unpatched translation for higher-resourced languages and poorly-aligned representations for rare concepts, making patching less effective.

5.3 Manual Error Analysis of seen Setting

To better understand the quality and evolution of the shared concept representations, and how the model fails to map them to the expected output, we conduct a novel error analysis. Specifically, outputs under seen are annotated by native speakers using the labels we propose in Table 2. Due to resource

[5] cc.en.300.bin

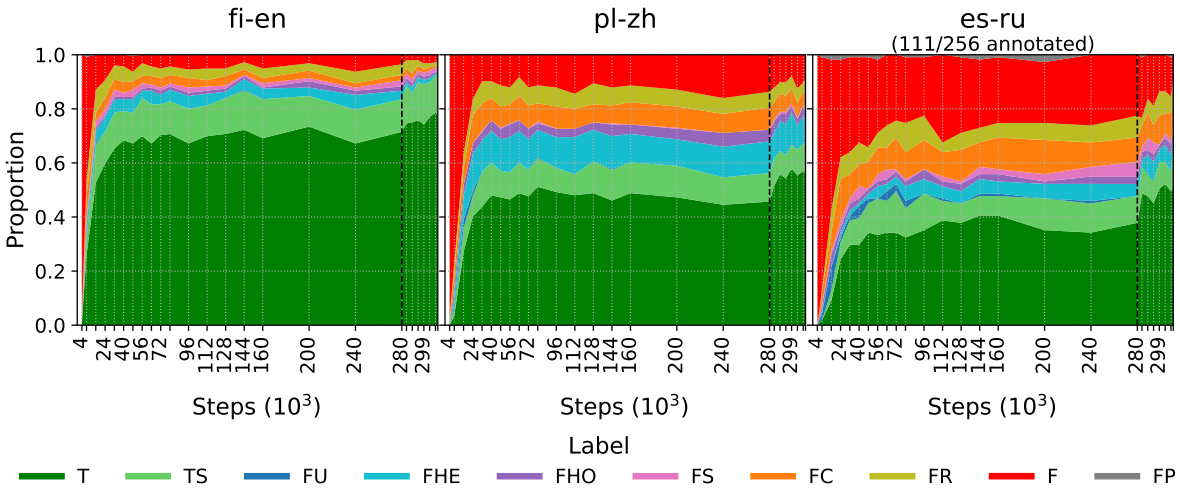

Figure 5: Area grid of label distribution for outputs under the seen patching setting for a selection of languages. For ru we show only the 111 concepts for which outputs that were labeled; en and zh were labeled in full. The black dotted line indicates the start of phase two of EuroLLM's training.

Label	Description	Example
T	exact match (automatic)	poverty → poverty
TS	synonym	outlander → foreigner
FS	differ in one aspect of grammar to a T or TS	outlander → foreigners actor → actress
FHE	hypernym relative to target	liquor → drink
FHO	hyponym relative to target	insect → honeybee
FR	co-hyponym to target	skirt → dress
FU	untranslated	culture → cultura
FC	conceptually similar to target	archer → arrow
F	wrong	aunt → sheep

Table 2: Label definitions for manual error analysis. Labels are mutually exclusive, and ordered by priority, i.e. FS overrides FC; F only applies if no other labels do (see 6 in the App.) T and TS are considered correct.

constraints, we annotate a subset of outputs for ru, corresponding to 111 concepts (see Appendix A.4 for more information on annotation). Results are provided in Fig. 5.

We consider the categories F* (excluding F) "partially" correct. For all language settings, outputs are more often partially correct than fully wrong, and the proportion of F stays mostly stable, or decreases. This is evidence that the averaged concept representations already encode conceptual and syntactic information prior to inducing the desired concept, and that they are refined throughout training.

We do not observe language-specific patterns on the target input side. However, the distribution of categories is dependent on the output language, e.g. FS makes up a larger proportion of the errors for ru, which likely reflects its complex morphology.

In general, FHO is less common than FHE, aligning with our intuition that poorly-aligned concept spaces are more likely to be mapped to general terms. We also note that FR is a relatively large class, which may reflect the language model training objective, i.e., the model learns that a word plays a particular syntactic role.

Interestingly, for all target language pairs, in phase two the proportion of T increases more than TS (which remains stable), suggesting increased cross-lingual alignment through parallel data, resulting in more specific representations that induce the target word rather than a synonym.

Linking back to errors seen in Section 5.1, we find that patching shifts sense selection for polysemous words. For example, for fr-en 'avocat' may be translated to 'attorney' (expected) or 'avocado'. Under seen patching, the model output is shifted from 'avocado' to 'attorney'. For es-en, 'mañana' is translated to 'morning' in the unpatched case and 'tomorrow' under patching. Since we average over representations for many language pairs, polysemy for particular pairs is lost, leading to a selection of the dominant shared sense. Similarly, patching discourages the model from copying cross-lingual homographs. For example, for fr-en, the model copies 'tour' in the unpatched case, whereas under patching, the model outputs the expected 'tower'.

6 Conclusion

We present the first fine-grained investigation into the emergence of shared concept spaces during multilingual pretraining, showing that shared concept spaces *emerge early and remain relatively stable throughout training*. Our novel dataset and error analysis reveals that intermediate concept representations encode meaningful semantic information even before they can be mapped to specific concepts, suggesting that these spaces are progressively refined during training. Comparing across language settings, we observe that alignment with concept spaces depends on language similarity and training data composition.

These findings have important implications for multilingual language model training, in particular in low-resource settings. Since language-agnostic concept spaces emerge early and are relatively stable, multilingual training can focus on aligning languages to these spaces. Notably, our results indicate that relatively small amounts of high-quality, multi-aligned data improve alignment compared to en-only pivot data (phase one vs. phase two in EuroLLM). This suggests a promising path for closing the performance gap between high- and low-resource languages: improving alignment with concept spaces through targeted, multi-aligned data.

Acknowledgments

FK and BP are supported by the ERC Consolidator Grant DIALECT 101043235. MME is supported by the Carlsberg Foundation, grant CF-25-0624. We are grateful to Dr. Mateusz Klimaszewski for giving us access to the EuroLLM checkpoints, and Dr. Kamil Deja for giving feedback on an earlier draft. We thank Pingjun Hong and Darja Jepifanova for providing their native speaker expertise, and the members of MaiNLP for many useful discussions.

Limitations

In this work, we conduct a novel fine-grained investigation into how language-agnostic spaces emerge throughout multilingual LLM training. We mainly focus on only a single, relatively small (1.7B) model, yet provide additional supporting evidence on OLMo-2 7B and Apertus 8B (Appendix A.5). This choice is due to the limited availability of multilingual pretraining checkpoints at the time of writing. As noted in Section 2, BLOOM (Workshop et al., 2022) was, until the recent release of Apertus (Hernández-Cano et al., 2025), the only other state-of-the-art multilingual LM for which checkpoints are available, and offers only a very coarse granularity (4−8 checkpoints vs. our 26). We repeat experiments for target language pairs xx−en for pretraining checkpoints of Apertus 8B and OLMo-2 7B (Walsh et al., 2025), a model trained on English data, with inadvertent multilingual ability, in Appendix A.5. These experiments suggest that our findings may hold for other multilingual models. However, further research, and in particular transparent and open multilingual pretraining is needed to understand the generalizability of our findings and the impact of model size and pretraining strategies on the development of shared concept spaces.

Furthermore, though we study a larger and more diverse set of concepts than previous work, we focus only on nouns. Further work is needed to understand how representations for other word classes behave and whether our findings apply.

Since we use Multi-SimLex as the pool for our concepts and their translations, we inherit its biases. In particular, the expected translation for a particular word may be misleading. For example, as discussed in Section 5.1, the translation for "wife" in French in the dataset is "femme", but this word also means "woman". However, we introduce a manual error analysis (Section 5.3), capturing such biases. Because we also study lower-resourced languages, this analysis strengthens the evaluation compared to previous studies, which either relied on automatic machine translation (Wendler et al., 2024) or linguistic resources (Dumas et al., 2025) to evaluate output. Such tools are often less reliable for lower-resourced languages (Peppin et al., 2025). Our approach therefore improves the evaluation methodology itself, allowing for a more equitable treatment of lower-resourced languages.

Finally, we focus only on the task of word-level translation. This is a fairly simple task, and language-agnostic spaces may form and behave differently for more complex tasks like sentence-level translation, open-ended generation, or reasoning. The focus on word-level translation is motivated by two factors. First, activation patching relies on a simple, controlled task (see discussion of confounding "degrees of freedom" in Heimersheim and Nanda, 2024). Second, we study pretraining checkpoints, specifically without any instruction tuning, which should improve the model's ability to follow more complex task instructions (Wei et al., 2022). Therefore, focusing on this simple task reduces the

confounding factor of whether the model can perform the task. This is confirmed by our preliminary experiments in Section 5.1, which show that even early checkpoints can follow the task, in particular for high-resourced language pairs. We leave investigation of how language-agnostic spaces used for more complex multilingual tasks emerge and behave throughout multilingual pretraining to future work.

Ethical Considerations

We acknowledge the use of ChatGPT for paraphrasing and lexical suggestions, with output checked carefully to avoid changes in meaning. In addition, ChatGPT and Github Copilot provided coding assistance.

References

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pengfei Cao, Yuheng Chen, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. 2024. One mind, many tongues: A deep dive into language-agnostic knowledge neurons in large language models. *CoRR*, abs/2411.17401.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Journey to the center of the knowledge neurons: discoveries of language-independent knowledge neurons and degenerate knowledge neurons. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Nadezhda Chirkova and Vassilina Nikoulina. 2024. Zero-shot cross-lingual transfer in instruction tuning of large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 695–708, Tokyo, Japan. Association for Computational Linguistics.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2025. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31822–31841, Vienna, Austria. Association for Computational Linguistics.

Sheridan Feucht, Eric Todd, Byron C Wallace, and David Bau. 2025. The dual-route model of induction. In *Second Conference on Language Modeling*.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *CoRR*, abs/2404.15255.

Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert i Llaquet, Barna Pásztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Durech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolcec, Yixuan Xu, Michael Aerni, Badr AlKhamissi, and 82 others. 2025. Apertus: Democratizing open and compliant llms for global language environments. *CoRR*, abs/2509.14233.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

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. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6919–6971, Mexico City, Mexico. Association for Computational Linguistics.

Nahyun Lee, Yeongseo Woo, Hyunwoo Ko, and Guijin Son. 2025. Controlling language confusion in multilingual LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1026–1035, Vienna, Austria. Association for Computational Linguistics.

Jiahuan Li, Shujian Huang, Aarron Ching, Xinyu Dai, and Jiajun Chen. 2024. PreAlign: Boosting cross-lingual transfer by early establishment of multilingual alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10246–10257, Miami, Florida, USA. Association for Computational Linguistics.

Peiqin Lin, Andre Martins, and Hinrich Schuetze. 2025. A recipe of parallel corpora exploitation for multilingual large language models. In *Findings of the*

Association for Computational Linguistics: NAACL 2025, pages 4038–4050, Albuquerque, New Mexico. Association for Computational Linguistics.

Junhua Liu and Bin Fu. 2024. Responsible multilingual large language models: A survey of development, applications, and societal impact. *arXiv:2410.17532v1*.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm: Multilingual language models for europe. *Procedia Computer Science*.

Aaron Mueller, Jannik Brinkmann, Millicent L. Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. 2024. The quest for the right mediator: A history, survey, and theoretical grounding of causal interpretability. *CoRR*, abs/2408.01416.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Qiwei Peng and Anders Søgaard. 2024. Concept space alignment in multilingual LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5511–5526, Miami, Florida, USA. Association for Computational Linguistics.

Aidan Peppin, Julia Kreutzer, Alice Schoenauer Sebag, Kelly Marchisio, Beyza Ermis, John Dang, Samuel Cahyawijaya, Shivalika Singh, Seraphina Goldfarb-Tarrant, Viraat Aryabumi, Aakanksha, Wei-Yin Ko, Ahmet Üstün, Matthias Gallé, Marzieh Fadaee, and Sara Hooker. 2025. The multilingual divide and its impact on global ai safety. *arXiv:2505.21344v1*.

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 823–848, Online only. Association for Computational Linguistics.

Frederick Riemenschneider and Anette Frank. 2025. Cross-lingual generalization and compression: From language-specific to shared neurons. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13470–13491, Vienna, Austria. Association for Computational Linguistics.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual LLMs think in english? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Yingli Shen, Wen Lai, Shuo Wang, Ge Gao, Kangyang Luo, Alexander Fraser, and Maosong Sun. 2025. From unaligned to aligned: Scaling multilingual LLMs with multi-way parallel corpora. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7357–7379, Suzhou, China. Association for Computational Linguistics.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Karolina Stanczak, Edoardo Ponti, Lucas Torroba Hennigen, Ryan Cotterell, and Isabelle Augenstein. 2022. Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1589–1598, Seattle, United States. Association for Computational Linguistics.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Hinata Tezuka and Naoya Inoue. 2025. The transfer neurons hypothesis: An underlying mechanism for language latent space transitions in multilingual LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31742–31792, Suzhou, China. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart,

and Anna Korhonen. 2020. Multi-SimLex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897.

Evan Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, and 23 others. 2025. 2 OLMo 2 furious (COLM's version). In *Second Conference on Language Modeling*.

Hetong Wang, Pasquale Minervini, and Edoardo Ponti. 2024a. Probing the emergence of cross-lingual alignment during LLM training. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12159–12173, Bangkok, Thailand. Association for Computational Linguistics.

Weixuan Wang, Barry Haddow, Wei Peng, and Alexandra Birch. 2024b. Sharing matters: Analysing neurons across languages and tasks in llms. *CoRR*, abs/2406.09265.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, and 375 others. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv:2211.05100v4*.

Hongchuan Zeng, Senyu Han, Lu Chen, and Kai Yu. 2025. Converging to a lingua franca: Evolution of linguistic regions and semantics alignment in multilingual large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10602–10617, Abu Dhabi, UAE. Association for Computational Linguistics.

Shimao Zhang, Zhejian Lai, Xiang Liu, Shuaijie She, Xiao Liu, Yeyun Gong, Shujian Huang, and Jiajun Chen. 2025. How does alignment enhance llms' multilingual capabilities? a language neurons perspective. *ArXiv*, abs/2505.21505.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? In *Advances in Neural Information Processing Systems*, volume 37, pages 15296–15319. Curran Associates, Inc.

Chengzhi Zhong, Qianying Liu, Fei Cheng, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2025. What language do non-English-centric large language models think in? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26333–26346, Vienna, Austria. Association for Computational Linguistics.

A Appendix

A.1 Technical Details

We build on the implementation of concept patching published by Dumas et al., 2025[6]. Specifically, we modify the prompt construction to produce concept-aligned prompts for both source and target prompt sets across all tested language pairs. We also adapt it to prevent overlap between few-shot examples on the source and target side. Furthermore, we update it to use nnsight 0.5.5[7] in order to enable multi-token generation under patching.

We generate five next tokens, and parse output up to the first generated quotation mark. If an initial substring matched the target or a synonym, we count this as correct, to avoid penalizing overly long continuations (for example: "absence of desire", when the expected word is "absence").

Please refer to the implementation at `https://github.com/mainlp/shared-concept-spaces` for more information.

Choice of Layer Dumas et al., 2025 find that the choice of layer to patch from doesn't significantly affect whether patching can induce the source concept, so long as it is not one of the later layers. We conduct initial experiments for target language pairs `xx-en` to confirm this, and find that we can patch up to about layer 14, after which the concept can no longer be reliably induced. Table 3 shows the Pearson correlation coefficient between the tested layers (6, 8, 9, 11, 12, 14, 16) and layer 10 computed over the full delta accuracy curves for `seen` over `unpatched` across checkpoints. A high correlation (>0.9) indicates that the trajectory of the patching effect over checkpoints is consistent between layers.

[6] `https://github.com/Butanium/llm-lang-agnostic`

[7] `https://nnsight.net`

tgt pair	L6	L8	L9	L11	L12	L14	L16
es-en	0.93	0.96	0.97	0.97	0.95	0.92	0.08
fr-en	0.97	0.99	0.99	0.99	0.98	0.95	-0.06
zh-en	0.93	0.95	0.98	0.99	0.97	0.93	0.00
pl-en	0.95	0.98	0.98	0.99	0.97	0.94	0.63
ru-en	0.99	0.99	1.00	1.00	0.99	0.97	0.74
et-en	0.99	0.99	0.99	1.00	0.99	0.97	0.67
fi-en	0.99	1.00	1.00	1.00	1.00	0.99	0.68
sw-en	0.99	1.00	1.00	1.00	1.00	1.00	0.91

Table 3: Pearson correlation of the difference in translation accuracy curves (seen vs. unpatched) over checkpoints between each tested layer and layer 10, for target language pairs xx-en. A high correlation (>0.9) indicates a consistent trend between checkpoints.

A.2 Synonym Expansion

We attempt to follow prior work in expanding our translation dataset with synonyms from BabelNet (Navigli and Ponzetto, 2010), which would capture more correct translations of a particular concept. However, we find the resulting expansions to be overly generous, even after filtering by quality tags (as provided by BabelNet) and for nouns. This is likely partially due to the nature of our concept set, which includes abstract and infrequent terms (e.g., "afterworld", "acetylcholine"). For example, "acid" is expanded to many different street names for the drug "LSD". Such expansions are difficult to filter out automatically, especially considering our treatment of low-resourced languages, such as Welsh and Swahili. For such languages, linguistic resources and tools are typically of lesser quality (Peppin et al., 2025). Furthermore, too generous expansions across 11 tested languages results in very few compatible source and target concept pairs where a compatible pair has no overlapping words across all languages. Therefore, we instead adapt our evaluation as described in Section 5.3, manually inspecting model outputs to gain a more accurate understanding of the effect induced by cross-lingual concept patching.

A.3 Example Prompts

Here, we provide some example prompts for the different settings. For target language pair `fr-en`, a full target prompt for the target concept "triumph" is shown below.

Français: "août" - English: "august"
Français: "animal" - English: "animal"
Français: "critique" - English: "criticism"
Français: "base" - English: "base"
Français: "cage" - English: "cage"
Français: "triomphe" - English: "

In the following sections, we show examples of source prompts for the source concept "curtain" in different settings, given this target prompt.

unpatched Setting As a baseline, we run inference on the target prompt *without* any intervention. This prompt is exactly the same as the one used in `tgt` setting, the difference is that here, the prompt is run without intervention.

Français: "pollution" - English: "pollution"
Français: "dépression" - English: "depression"
Français: "structure" - English: "structure"
Français: "action" - English: "action"
Français: "os" - English: "bone"
Français: "rideau" - English: "

`tgt` Setting Source Prompts In this ablation setting, the source and target language pairs are the same, the prompts differ only in the few-shot examples and the concept to be translated.

Français: "pollution" - English: "pollution"
Français: "dépression" - English: "depression"
Français: "structure" - English: "structure"
Français: "action" - English: "action"
Français: "os" - English: "bone"
Français: "rideau" - English: "

`en_en` Setting Source Prompts In this control setting, the model must *copy* the English word, rather than translate it.

English: "pollution" - English: "pollution"
English: "depression" - English: "depression"
English: "structure" - English: "structure"
English: "action" - English: "action"
English: "bone" - English: "bone"
English: "curtain" - English: "

`seen` Setting Source Prompts The language pairs in `seen` are in EuroLLM's training data, `excluding en` and the languages in the target pair. For this example, `seen` includes all ordered pairs of `es`, `zh`, `pl`, `ru`, `et`, `fi`, `yue`, for 42 total source language pairs. Below, we show a few sample prompts.

Suomi: "saastuminen" - Polski: "zanieczyszczenie"
Suomi: "masennus" - Polski: "depresja"
Suomi: "rakenne" - Polski: "struktura"
Suomi: "toiminta" - Polski: "akcja"
Suomi: "luu" - Polski: "kość"
Suomi: "verho" - Polski: "

Polski: "zanieczyszczenie" - Español: "contaminación"
Polski: "depresja" - Español: "depresión"
Polski: "struktura" - Español: "estructura"
Polski: "akcja" - Español: "acción"
Polski: "kość" - Español: "hueso"
Polski: "zasłona" - Español: "

Русский: "загрязнение" - Suomi: "saastuminen"
Русский: "депрессия" - Suomi: "masennus"
Русский: "структура" - Suomi: "rakenne"
Русский: "действие" - Suomi: "toiminta"
Русский: "кость" - Suomi: "luu"
Русский: "занавеска" - Suomi: "

A.4 Manual Annotation

The in-house annotators (one per output language) are native speakers of the respective languages. All three annotators label a small subset of the English outputs (120 examples total). We exclude `T` from this set as it is automatically tagged, and stratify over the remaining classes, except for `F`. `F` is the majority class and typically most straightforward to tag, as many outputs, in particular in the very early stages of training (first two steps), are gibberish, or immediately recognizable as unrelated. Therefore, we down-sample `F` to 20/120. We compute a Fleiss' Kappa of 0.63 between the three annotators, indicating good annotator agreement. Fig. 6 shows the decision tree used by annotators. `FP` is used to indicate a parsing failure, where the model failed to output a prediction in the expected format (ending in a quotation mark).

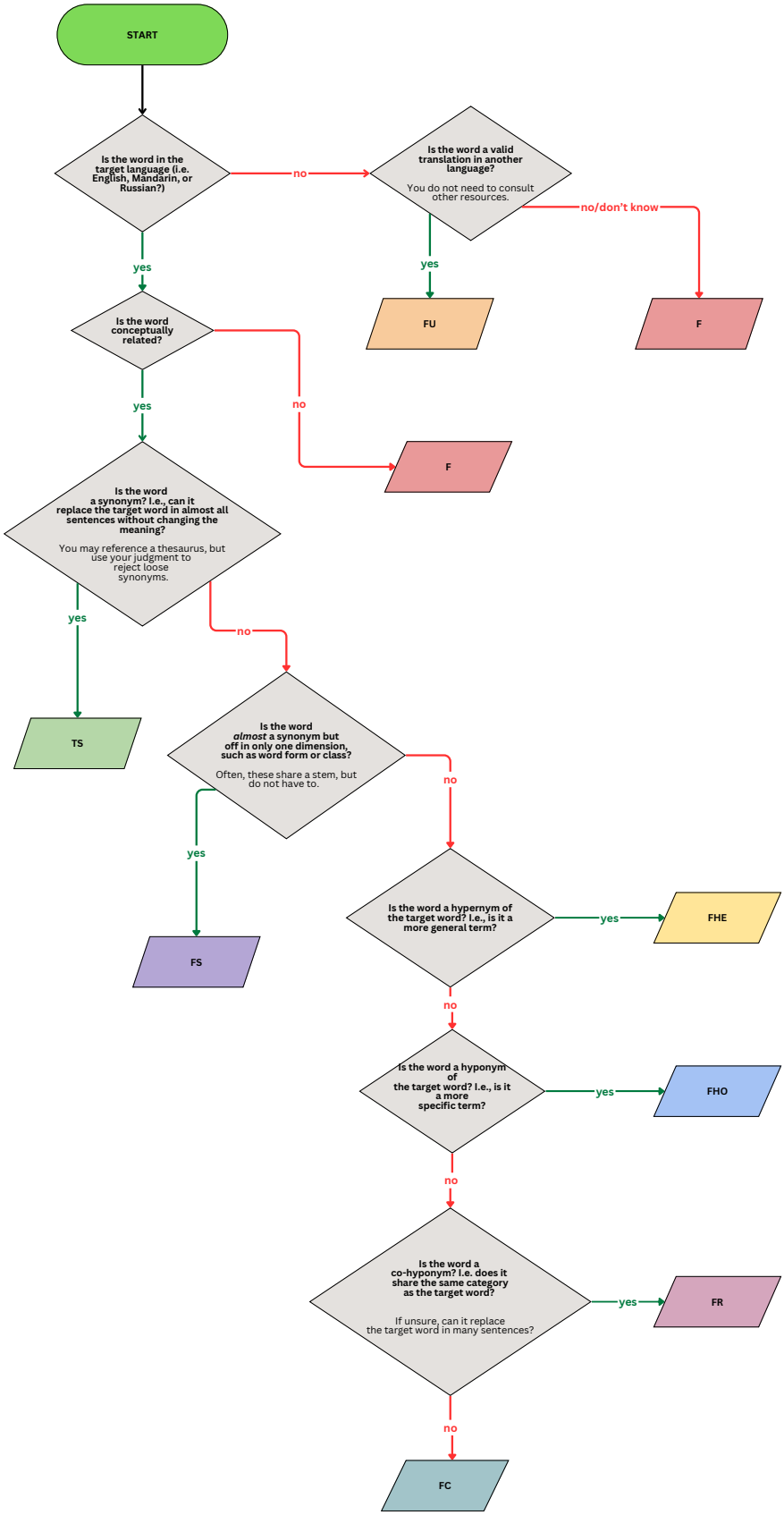

Figure 6: The decision tree used to annotate errors as described in Section 5.3. `T` is not included, as this was tagged automatically.

A.5 Results for Other Models

We repeat experiments for target language pairs xx-en for Apertus 8B (Hernández-Cano et al., 2025), and OLMo-2 7B (Walsh et al., 2025) (henceforth simply Apertus and OLMo-2, respectively). Apertus is a recently released multilingual model, including 1,811 languages in its pretraining data. OLMo-2 was trained on Dolma, an English corpus (Soldaini et al., 2024), hence, we do not list it in our discussion of available pretraining checkpoints in Section 2. Despite this, OLMo-2 is known to have some multilingual capabilities (cf. Lee et al., 2025), likely due to the "accidental" inclusion of multilingual data in its pretraining corpus (Blevins and Zettlemoyer, 2022). For consistency with the main results, we patch from layer 10 for these models, even though they are deeper, i.e. have more layers. Since we patch all subsequent layers as well, patching earlier for the deeper models is a conservative choice; we see in preliminary experiments in Appendix A.1 that trends are fairly robust to layer choice.

For both models, we see evidence of a shared concept space, as seen in the successful translation under the seen setting (see Fig. 7 for OLMo-2, and Fig. 8 for Apertus). This is more surprising for OLMo-2, which is trained on English data. However, we do see this English-centric training reflected in the en_en performance, which is generally stronger than that under seen setting, especially in the beginning of training. This is in contrast to EuroLLM and Apertus, where the difference in en_en and seen is much less pronounced.

Though we commend the rare release of multilingual pretraining checkpoints by the developers of Apertus, the first available checkpoint is already at 210B tokens consumed (compared to roughly 48B at the first checkpoint of EuroLLM, and 1B at the first checkpoint of OLMo-2). This low granularity of early checkpoints appears to mask the development of shared concept spaces —at the first checkpoint cross-lingual concept patching is successful, and does not improve much during training. However, the unpatched translation also does not improve much over checkpoints (with the exception of en, fi, sw, which are presumably lower-resourced), indicating that the model has already learned to translate our concepts by the first checkpoint. For OLMo-2, there is an increase in translation accuracy under seen over training checkpoints, suggesting that cross-lingual alignment increases over training. We hypothesize that this reflects the small amounts of multilingual data in OLMo-2's pretraining data, such that alignment may occur much later, as multilingual data is consumed by chance.

Overall, these results suggest that our main finding applies to other models: shared spaces develop early, in particular for models trained with multilingual data. Furthermore, multilingual pretraining data appears to drive development of this space, such that patching from seen is more successful for Apertus and EuroLLM than it is for OLMo-2, where the latter is trained on much fewer multilingual data.

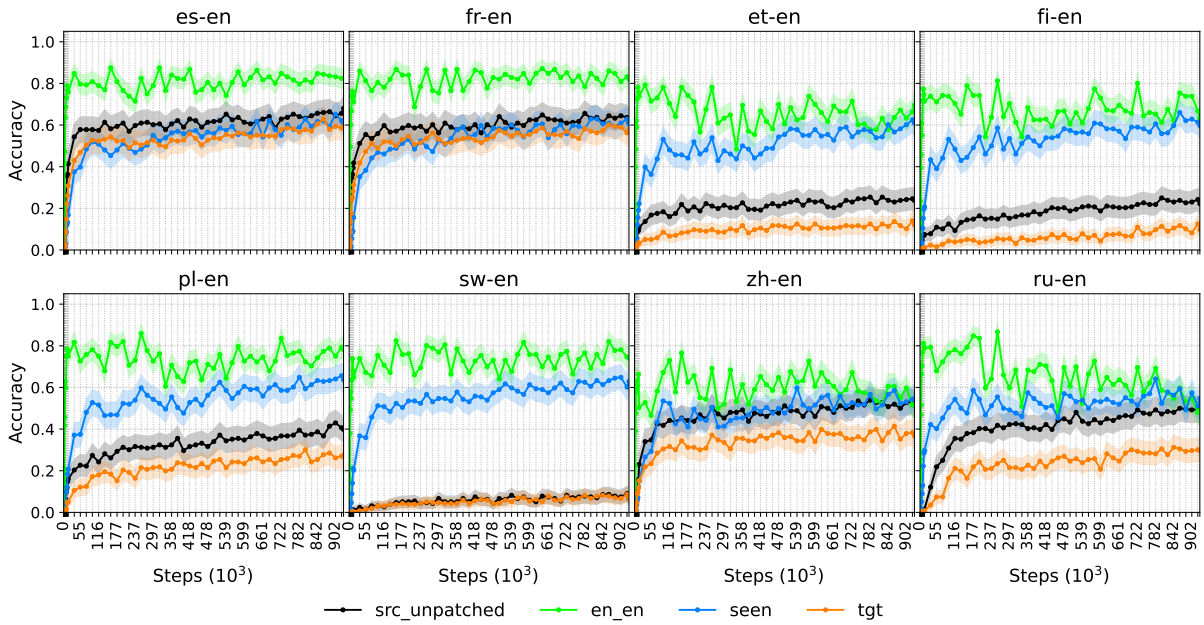

Figure 7: Mean word-level translation accuracy over checkpoints of OLMo-2 7B under patching for target output language en. We show 95% CI over 256 samples.

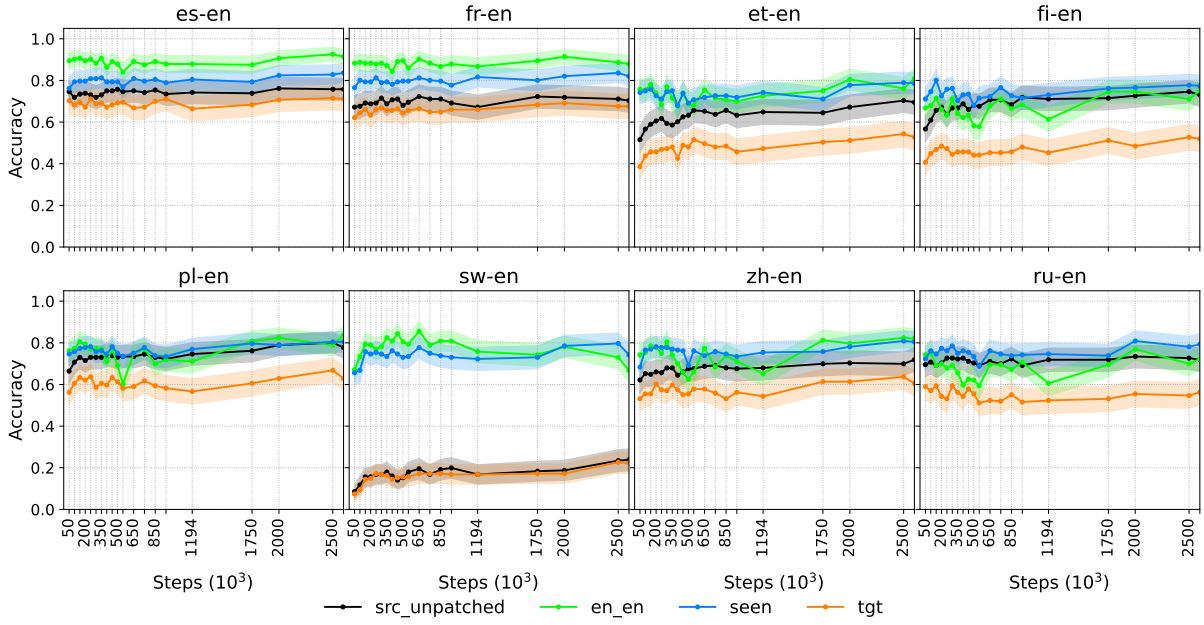

Figure 8: Mean word-level translation accuracy over checkpoints of Apertus 8B under patching for target output language en. We show 95% CI over 256 samples.

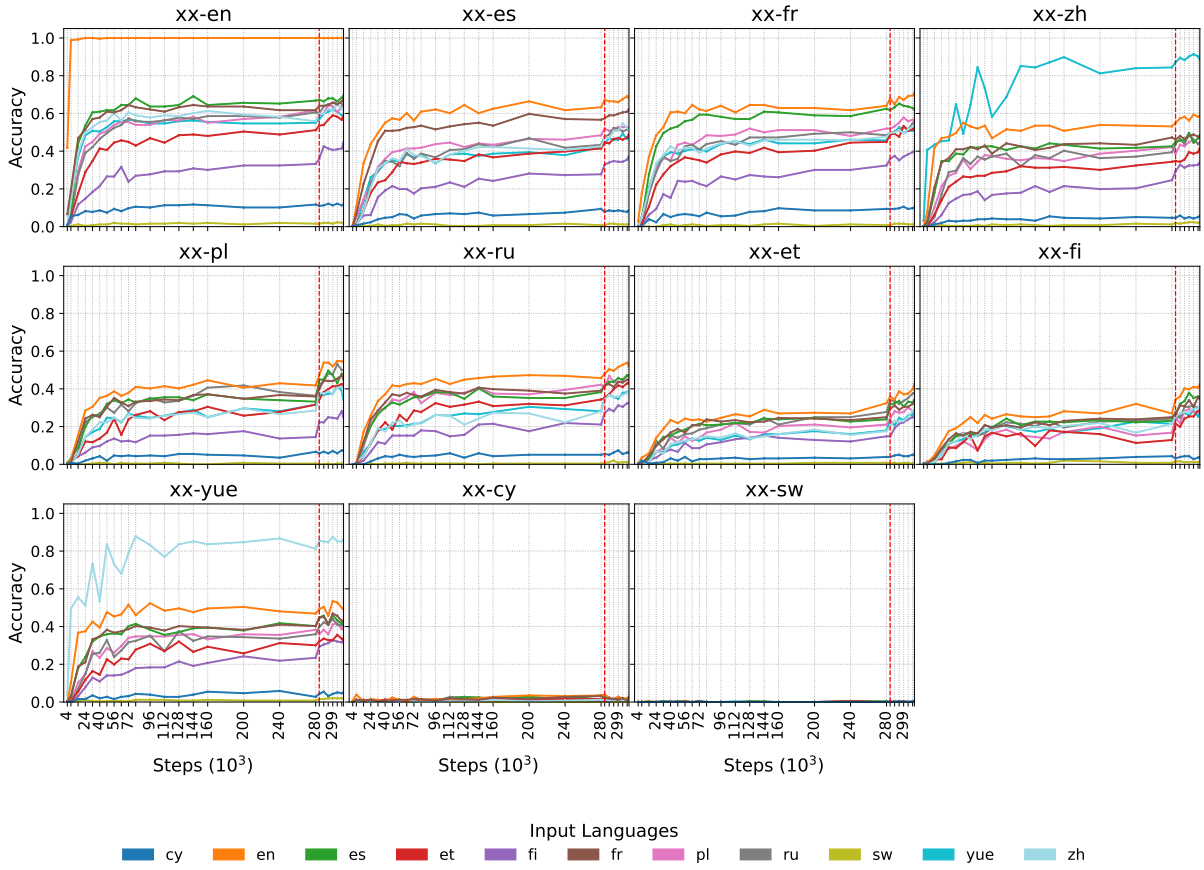

Figure 9: Mean word-level translation accuracy over checkpoints for source prompts used for patching, grouped by output language for all language pairs. The red dotted line indicates the start of phase two of EuroLLM's training.

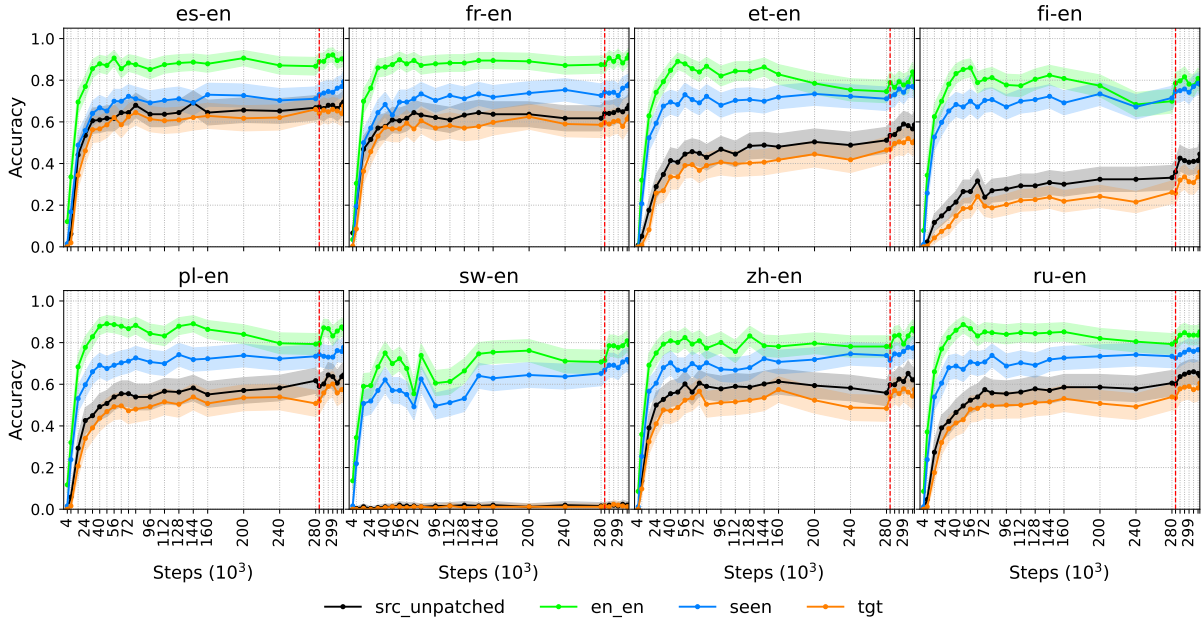

Figure 10: Mean word-level translation accuracy over checkpoints under patching for target output language en. The red dotted line indicates the start of phase two of EuroLLM's training. We show 95% CI over 256 samples.

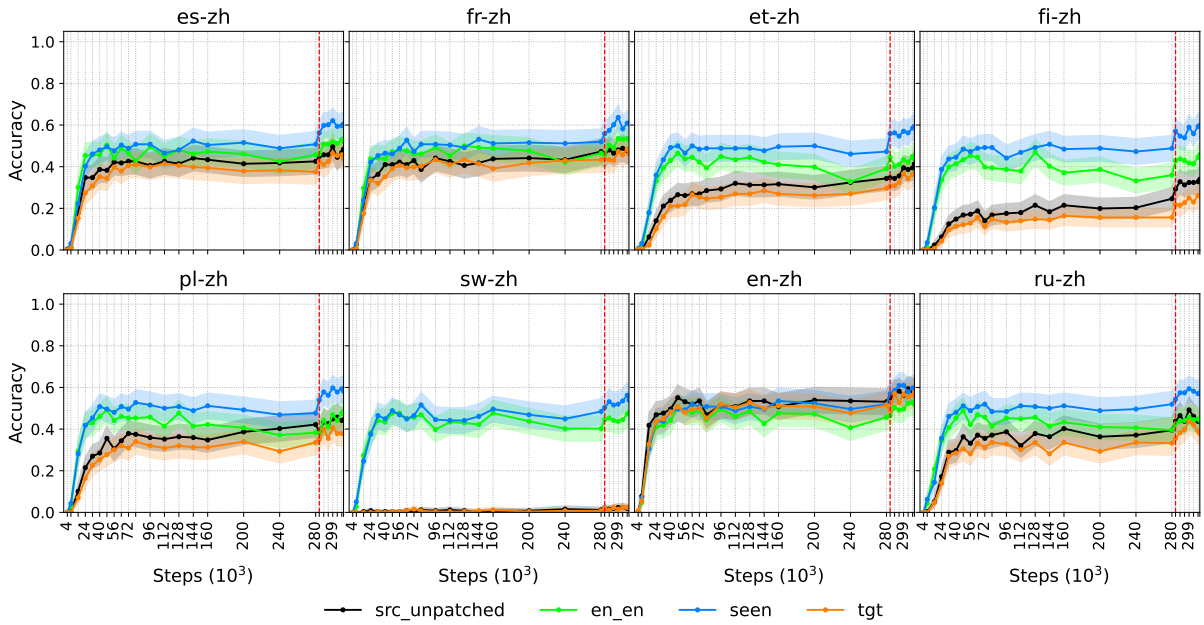

Figure 11: Mean word-level translation accuracy over checkpoints under patching for target output language zh. The red dotted line indicates the start of phase two of EuroLLM's training. We show 95% CI over 256 samples.

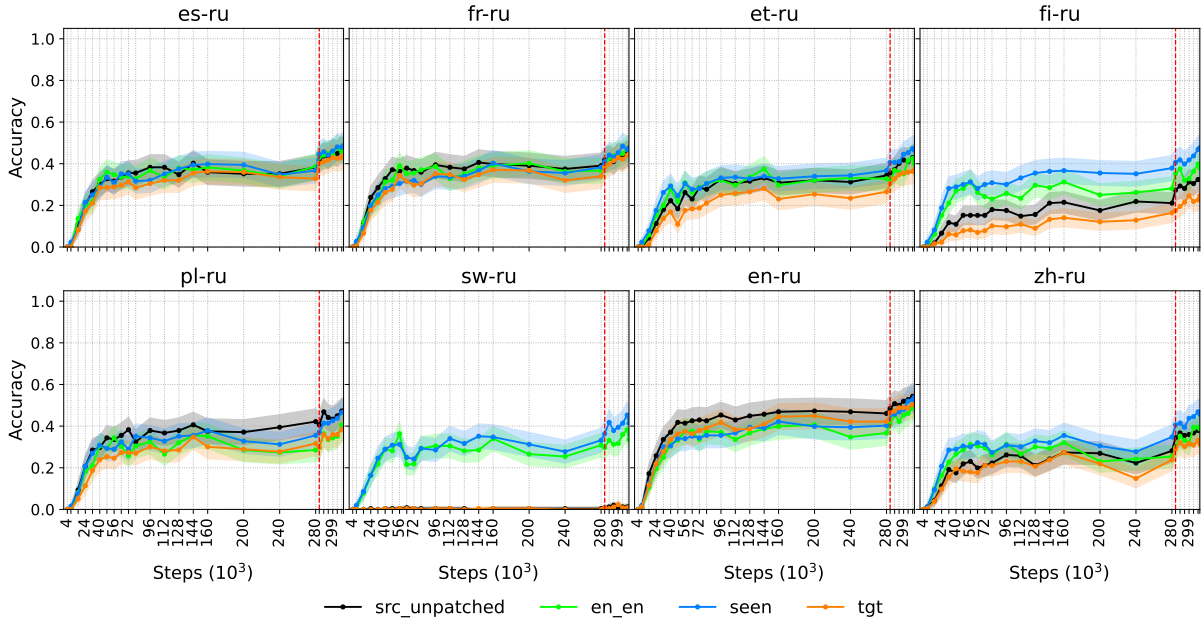

Figure 12: Mean word-level translation accuracy over checkpoints under patching for target output language ru. The red dotted line indicates the start of phase two of EuroLLM's training. We show 95% CI over 256 samples.

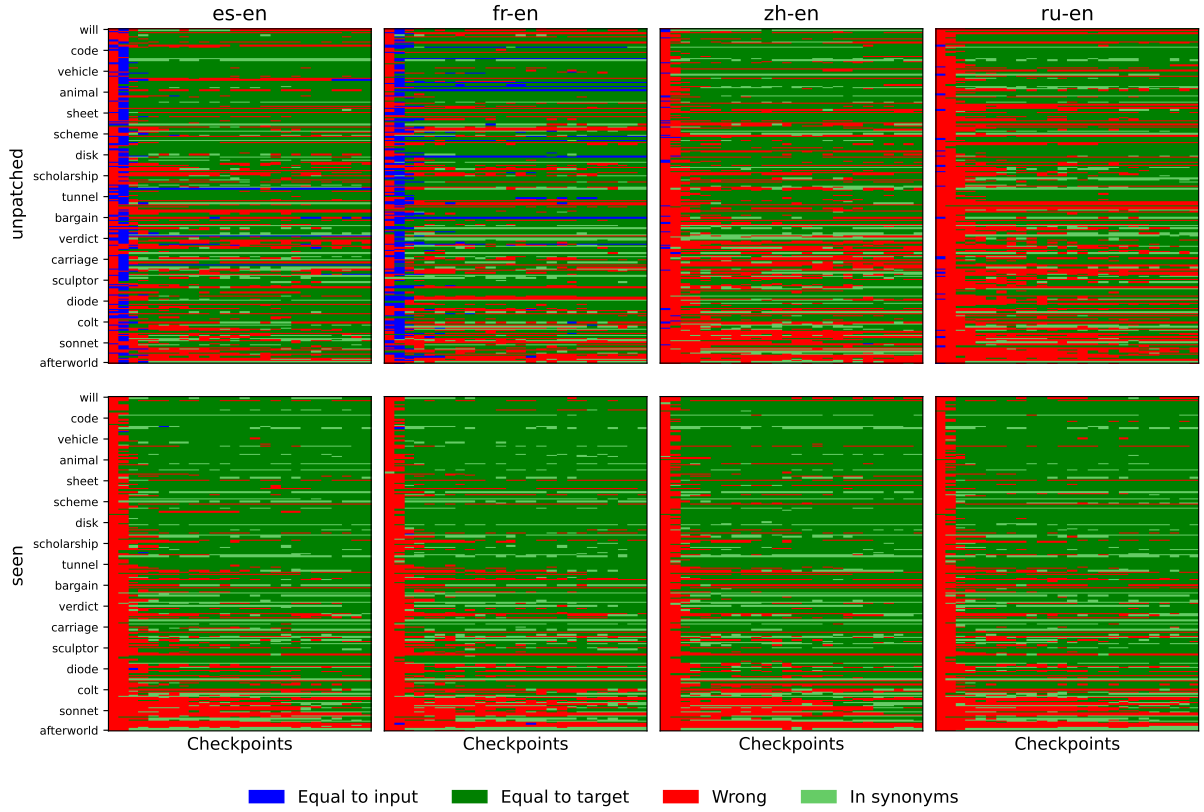

Figure 13: Translation categorical maps under the unpatched and seen setting for es−en, fr−en, zh−en, ru−en sorted by target word frequency. Each row represents a target concept, with a selection of concepts shown as ticks. Each column represents one checkpoint.

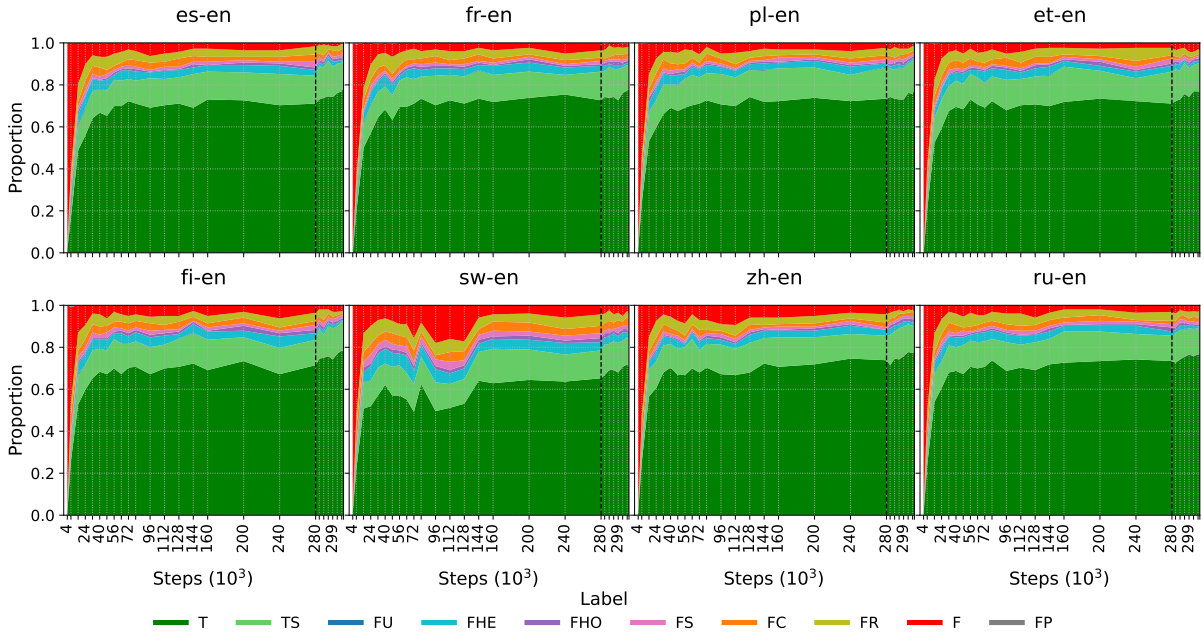

Figure 14: Area grid of label distribution for outputs under the seen patching setting for target output language en. The black dotted line indicates the start of phase two of EuroLLM's training.

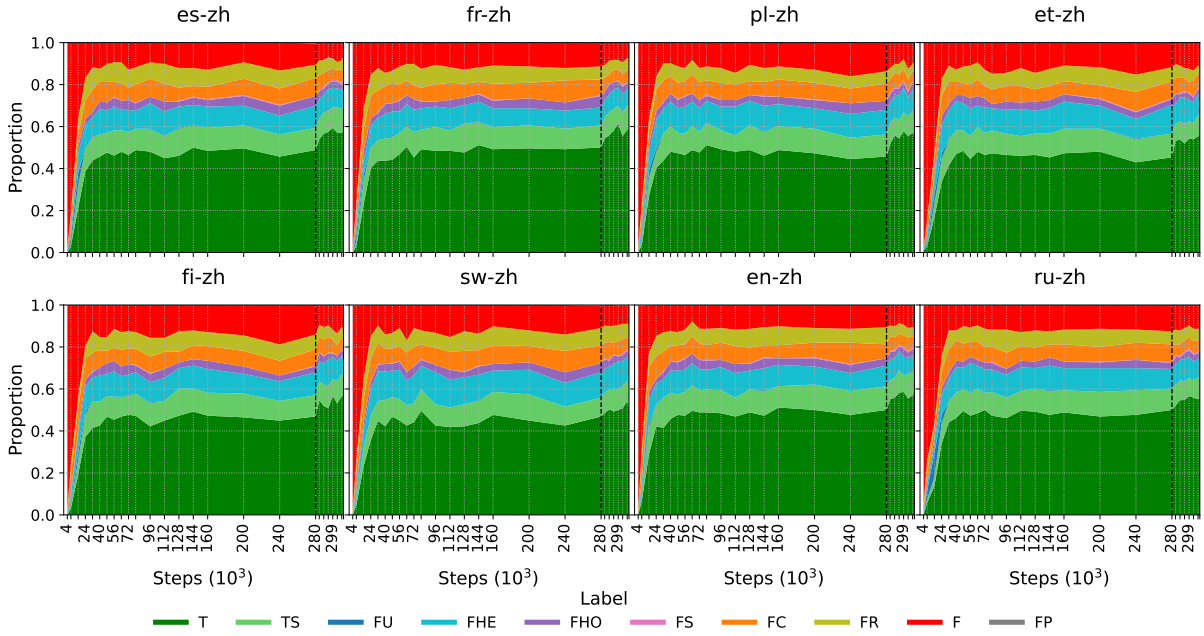

Figure 15: Area grid of label distribution for outputs under the seen patching setting for target output language zh. The black dotted line indicates the start of phase two of EuroLLM's training.

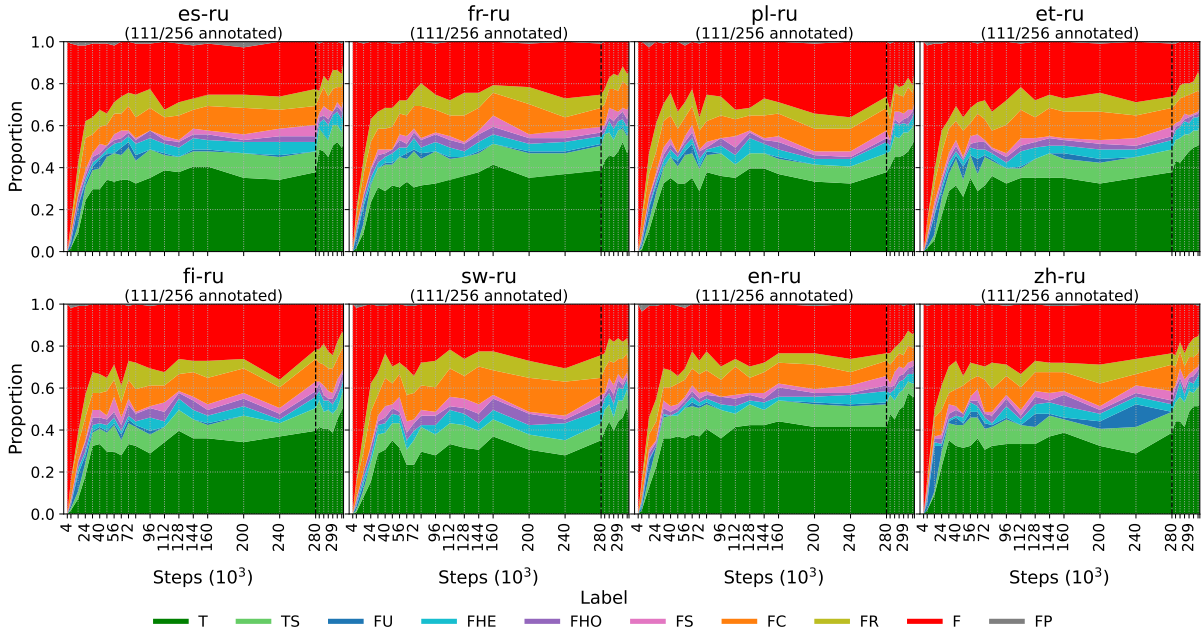

Figure 16: Area grid of label distribution for outputs under the seen patching setting for target output language ru. We show only the subset of outputs that were labeled (corresponding to 111/256 concepts). The black dotted line indicates the start of phase two of EuroLLM's training.