The Role of Language Imbalance in Cross-lingual Generalisation: Insights from Cloned Language Experiments

Anonymous ACL submission

Abstract

Multilinguality is crucial for extending recent advancements in language modelling to diverse linguistic communities. To maintain high performance while representing multiple languages, multilingual models ideally align representations, allowing what is learned in one language to generalise to others. Prior research has emphasised the importance of parallel data and shared vocabulary elements as key factors for such alignment. In this study, we investigate an unintuitive novel driver of cross-lingual generalisation: language *imbalance*. In controlled experiments on perfectly equivalent cloned languages, we observe that the existence of a predominant language during training boosts the performance of less frequent languages and leads to stronger alignment of model representations across languages. Furthermore, we find that this trend is amplified with scale: with large enough models or long enough training, we observe that bilingual training data with a $^{90}/_{10}$ language split yields better performance on both languages than a balanced $^{50}/_{50}$ split. Building on these insights, we design training schemes that can improve performance in all cloned languages, even without altering the training data. As we extend our analysis to real languages, we find that infrequent languages still benefit from frequent ones, yet whether language imbalance causes cross-lingual generalisation there is not conclusive.

1 Introduction

In recent years, autoregressive language models (LMs) pretrained on massive text corpora have advanced the state of the art in NLP tasks across the board (Brown et al., 2020; Touvron et al., 2023a,b; Köpf et al., 2023). While most of the leading models are trained on English texts, multilingual capabilities are crucial to make these advances accessible to a broader user base with diverse linguistic backgrounds. Ideally, data in one language should improve these multilingual models' performance in others. Such multilingual models should thus display **cross-lingual generalisation**: by reusing circuits (Cammarata et al., 2020; Elhage et al., 2021) and aligning their internal representations across languages, they may generalise concepts learned in a language to another.[1]

How can such cross-lingual generalisation be achieved? This has been a focus of much prior work. One previously identified driver of cross-lingual generalisation is **parallel training data**; empirical evidence shows that training the model on either parallel sentence pairs (Lample and Conneau, 2019) or on corpora which are comparable across languages (Dufter and Schütze, 2020) improves generalisation. Another driver of cross-lingual generalisation is the availability of **anchor points**, i.e., vocabulary elements that are shared between languages; these can be naturally occurring subwords (e.g., *computer* in English and *computador* in Portuguese may share the subword *comp*; Pires et al., 2019; Wu and Dredze, 2019), shared special tokens (e.g., `mask` or `bos` symbols; Dufter and Schütze, 2020), or even artificially inserted "code-switching" augmentations (Conneau et al., 2020b; Reid and Artetxe, 2022; K et al., 2020; Feng et al., 2022). Beyond these two drivers, **limited model capacity** has been found to improve generalisation by Dufter and Schütze (2020), but to constrain multilingual capabilities by Chang et al. (2023).

In this work, we identify a surprising new factor that can boost cross-lingual generalisation abilities: **language imbalance**. We first conduct

[1] A circuit is typically defined as a subgraph of a neural network which performs a specific computation. E.g., a circuit could be responsible for computing "greater than" comparisons between numbers in English sentences (Hanna et al., 2024). If representations are aligned across languages (in terms of how they encode, e.g., numbers) and circuits are reused, a model should be able to apply what it learns in one language (e.g., "greater than" comparisons in English) to perform similar computations in another language (e.g., French).

experiments in a synthetic setting with perfectly equivalent cloned languages; this allows us to investigate LMs' generalisation abilities in isolation from the effects of languages' dissimilarities, giving us a rough upper bound on the generalisation we should expect to see between real language pairs. In this cloned language setting, we find that having a dominant main language improves generalisation, significantly boosting the performance of less frequent languages. Furthermore, we find that this effect becomes stronger when we either increase our model's size or when we train it for longer. Based on these insights, we design training curricula that improve performance in all cloned languages without any modifications to the training data. In the second part of our paper, we investigate to what extent our insights transfer to real language pairs. While we find that lower resource languages typically do benefit from higher resource ones, the impact of language imbalance on cross-lingual generalisation is much less clear in this more realistic setting. Overall, our results suggest an interesting attribute of model training dynamics: in some settings, having a main language can lead model components to be shared across languages.

2 Cross-lingual Generalisation

While natural languages differ widely in their typological properties, any pair of languages will share at least a few grammatical and syntactic patterns. Further, as their semantics reflect the underlying processes of our world, language pairs should also have similarities in the types of messages their users typically convey. Intuitively, this suggests that what is learned about a language L_{A} should be useful to model another language L_{B}, and vice versa. The extent of such generalisation depends not only on how similar the two languages are, but also on the employed learning algorithm. We analyse such generalisation here, with a focus on how language imbalance influences multilingual LMs.

Intuitively, if a model generalises well across languages, it should achieve better performance in each language (in terms of, e.g., perplexity) than a monolingual model trained on the same data. Concretely, a model trained on a multilingual dataset $\mathcal{D}_{\mathrm{multi}} = \mathcal{D}_{\mathrm{A}} \cup \mathcal{D}_{\mathrm{B}}$ containing languages L_{A} and L_{B} should perform better than monolingual models trained only on $\mathcal{D}_{\mathrm{A}}$ or $\mathcal{D}_{\mathrm{B}}$. This becomes clear when using definitions from information theory: $\mathcal{D}_{\mathrm{multi}}$ contains at least as much information about L_{A} as $\mathcal{D}_{\mathrm{A}}$. However, such a multilingual model could also perform worse. This could happen, for instance, if the data from different languages interfere with each other during optimisation through conflicting gradient update directions (Wang et al., 2020). It could also happen if the model has limited capacity: the multilingual model has to represent many languages, which intuitively requires more capacity than a single one, even if some parameters are shared across them (Conneau et al., 2020a; Pfeiffer et al., 2022).

In an attempt to make models better across many languages, many multilingual models these days are trained on somewhat balanced data (Scao et al., 2023; Faysse et al., 2024). In some of these cases, low-resource languages are upsampled to improve their performance under the model. As mentioned above, however, while balancing languages' appearance in a model's training set should intuitively improve performance, this is not necessarily true. In fact, (and perhaps surprisingly) some recent large language models trained in mostly English-focused settings perform reasonably well in a large sample of languages (Ahia et al., 2023; Blevins and Zettlemoyer, 2022; Briakou et al., 2023). These models' training data is typically highly imbalanced, with only a small fraction being composed of "non-English" languages. It is thus unclear whether multilingual models indeed benefit from training on datasets with balanced languages (Ye et al., 2023).

In smaller training scales, the benefits of multilingual training are better understood. In general, it has been found that low-resource languages tend to benefit from data in higher-resource languages, but high-resource languages benefit much less from each other (Conneau et al., 2020a; Chang et al., 2023). It is, however, unclear what causes cross-lingual generalisation in this case. Is the model in fact able to generalise better in the imbalanced setting? Or does the model generalise equally well in the balanced case, but its capacity bottlenecks performance in higher-resource languages, stopping us from observing performance gains?

We investigate the role of language imbalance in cross-lingual generalisation here. Notably, Wendler et al. (2024) recently showed that LMs seem to perform internal computations in an abstract "concept space" which is closest to their main language (English in this case); representations are then mapped back into the input language only in the models' final layers. Alabi et al. (2024) observe a similar trend when using language adapters.

3 Experimental Setup

In this section, we provide a brief overview over models, data, and metrics used; for more details, see App. A. Our code will be made available on GitHub. All of our experiments use GPT-2-style decoder-only transformers (Radford et al., 2019). We base our implementation on the Languini Kitchen codebase (Stanić et al., 2023), and unless otherwise noted, we use the `gpt-small` configuration with 85M non-embedding parameters, training on 1.2B tokens of English or French books. We use separate tokenisers for English and French. For some of our experiments, we treat their vocabularies as **disjoint** and do not merge them. If we merge subwords that occur in both vocabularies, we make this clear with the label **anchored**.

As our main evaluation metric, we report our models' perplexity (**PPL**) on the test set. Further, we define three metrics that allow for easy comparison of monolingual and multilingual models. Let t_A and t_B be the number of tokens a multilingual model is trained on in languages L_A and L_B, respectively. We define monolingual token equivalence (**MLTE**) as the number of tokens that would be required by a monolingual model, trained only in either language L_A or L_B, to achieve the same perplexity as the multilingual model does in that language. To determine MLTE, we fit a simple scaling law to predict perplexity from the number of training tokens (e.g., t_A) using the results from our trained monolingual models (see App. B for details). Analogously, we define monolingual PPL equivalence (**MLPE**) as the perplexity a monolingual model would reach when trained on the same number of L_A tokens (i.e., t_A) as a given multilingual model. Finally, we define token efficiency (**TEff**) as the fraction between MLTE and the number of tokens used for multilingual training, e.g., $\mathrm{TEff}_\mathrm{A} = \frac{\mathrm{MLTE}_\mathrm{A}}{t_\mathrm{A}}$. Intuitively, if TEff > 1, performance improves due to multilinguality, while if TEff < 1, multilinguality hurts performance.

4 Cloned Languages

In this section, we examine the model's capability to generalise across perfectly equivalent **cloned languages**. We create a cloned language by duplicating the language model's vocabulary; this allows us to encode each sequence in either the original language (using the original vocabulary) or in the cloned language (using the cloned vocabulary). This experimental paradigm was originally proposed by K et al. (2020) and Dufter and Schütze (2020).[2] Formally, let L_orig be an "original" language with a vocabulary of subword units Σ; we denote each subword $w \in \Sigma$. This language can be described by a probability distribution $p(\boldsymbol{w}_\text{orig})$, where $\boldsymbol{w}_\text{orig} \in \Sigma^*$. We clone language L_orig by creating multiple instantiations of it: L_1, L_2, $\dots L_N$. These languages have vocabularies Σ_n, each of which has symbols that are equivalent to the original ones.[3] Furthermore, these languages define probability distributions which are isometric to the original language. If we denote equivalence as $\boldsymbol{w}_n \mathrel{\mathring{=}} \boldsymbol{w}_\text{orig}$ for $\boldsymbol{w}_n \in \Sigma_n^*$ and $\boldsymbol{w}_\text{orig} \in \Sigma^*$, we have $\boldsymbol{w}_n \mathrel{\mathring{=}} \boldsymbol{w}_\text{orig} \implies p(\boldsymbol{w}_n) = p(\boldsymbol{w}_\text{orig})$.

Given dataset $\mathcal{D}_\text{orig} = \{\boldsymbol{w}_\text{orig}^{(k)}\}_{k=1}^K$ with $\boldsymbol{w}_\text{orig}^{(k)} \sim p(\boldsymbol{w}_\text{orig})$, we can now create a multilingual dataset $\mathcal{D}_\text{multi}$ by independently mapping each sequence to one of the cloned languages: For each $\boldsymbol{w}^{(k)}$, we first sample a language $L^{(k)} \sim p(L)$ from a categorical distribution over languages, then we map the sequence to $L^{(k)}$ by encoding it using the corresponding vocabulary. We can write $\mathcal{D}_\text{multi} = \bigcup_{n=1}^N \mathcal{D}_n$ where

$$\mathcal{D}_n = \left\{ \boldsymbol{w}_n^{(k)} \;\middle|\; \boldsymbol{w}_n^{(k)} \mathrel{\mathring{=}} \boldsymbol{w}_\text{orig}^{(k)} \text{ and } L^{(k)} = L_n \right\}$$

denotes the subset in language L_n.

Importantly, cloned languages are perfectly equivalent, having the same syntax, semantics, and distribution. They differ only in the symbols used to encode their vocabularies. Any generalisation we observe in this setting should thus serve as an upper bound on the potential to generalise across non-identical natural languages.[4] In other words, if our model cannot generalise across cloned languages, we would have strong reason to believe it shouldn't generalise across distinct languages. If we observe that a model can generalise across cloned languages, however, we may or may not observe the same to happen across non-cloned languages. We'll investigate the latter in Section 5.

[2] K et al. (2020) perform duplication on the character IDs, i.e., before tokenisation, while Dufter and Schütze (2020) adopt an approach equivalent to ours. Both of these works term L_2 a "fake" language. Since there is no distinction between L_1 and L_2, however, we call them cloned languages instead. Other related studies have investigated the effect of infinitely many cloned languages on LMs' performance (Huang et al., 2023; Chen et al., 2023), or employed duplicated vocabularies at the token level to study their impact on LMs' memorisation or performance (Kharitonov et al., 2021; Schäfer et al., 2024).

[3] Unless otherwise noted, these vocabularies are defined as disjoint sets in our experiments, meaning that no anchor points exist across languages.

[4] As for most of our experiments we consider cloned languages' alphabets to be disjoint, in practice our results only upper bound the cross-lingual generalisation of models with no anchor points (i.e., with disjoint vocabularies).

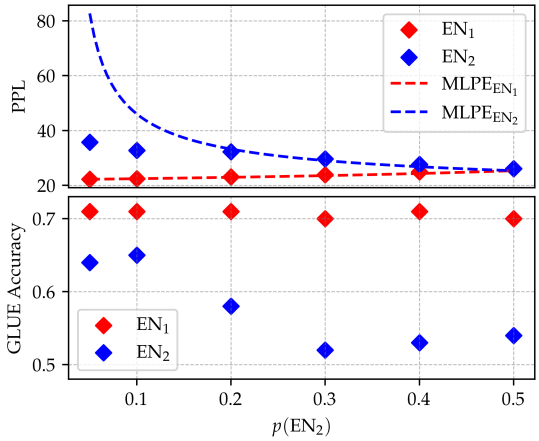

Figure 1: LM performance by imbalance ratio. (top) LM perplexity. (bottom) LM accuracy on GLUE; models were fine-tuned in EN_1 and evaluated on either EN_1 and EN_2.

4.1 Generalisation

Due to the equivalence of cloned languages, one may expect language models to easily generalise across them. In that case, training a multilingual model on datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ would lead to similar performance to training a monolingual model on the original dataset $\mathcal{D}_{\text{orig}}$ (note that $|\mathcal{D}_{\text{orig}}| = |\mathcal{D}_1| + |\mathcal{D}_2|$). We perform this experiment here, training either monolingual models on English (EN), or multilingual models on cloned English (EN_1 and EN_2), setting $p(\mathrm{EN}_1) = p(\mathrm{EN}_2) = 0.5$. Perhaps surprisingly, when training in this balanced multilingual setting, language modelling performance is significantly worse than in the monolingual setting (see Table 1, rows 2 & 4). In fact, one would obtain better performance training two monolingual models for half as many steps than training on this combined data. Training data in one language seems to hurt performance in the other language instead of boosting it. This indicates that the model is not able to generalise well across languages in this setting.

Takeaway 1. *The model does not generalise well across cloned languages given a 50/50 data split.*

4.2 Language Imbalance

How does the balance of the languages' data affect generalisation performance? Will the multilingual model still underperform its monolingual equivalents when trained on an uneven language distribution? When varying the ratio of EN_1 to EN_2 data shown during training (while keeping the total number of training steps constant), we observe that the rarer "lower resource" language, here always EN_2, benefits from the presence of a dominant "main language". Fig. 1 (left) shows that, under higher imbalance, the model's performance on EN_2 becomes much better than that of a monolingual model trained on the same amount of EN_2 data. For example, when training in the 90/10 regime, we obtain a $\mathrm{TEff}_{\mathrm{EN}_2}$ of over 2 (see Table 1, row 5). Do these improvements translate to better cross-lingual generalisation on downstream tasks as well? We test this by fine-tuning models on the GLUE benchmark (Wang et al., 2019) in EN_1 only, and evaluating them on EN_1 and EN_2. We observe that models trained under higher language imbalance indeed have significantly better EN_2 zero-shot performance (see Fig. 1 right). Together, these results suggest that cross-lingual generalisation is occurring.

Is this generalisation attained due to the model's internal computations being shared across languages? To answer this question, we analyse how language imbalance affects the cross-lingual alignment of our models' representations. Looking at the cosine similarity of equivalent subwords $w_1 \overset{\circ}{=} w_2$ in EN_1 and EN_2, we find that similarity steadily increases with higher imbalance: in the 50/50 setting, embeddings are not aligned (exhibiting an average cosine similarity of 0.02), while, e.g., in the 90/10 setting, equivalent subwords are much more aligned, showing a similarity of 0.28 (details in App. C). Comparing the cosine similarity of hidden states when the LM is given equivalent sequences $\boldsymbol{w}_1 \overset{\circ}{=} \boldsymbol{w}_2$, we also observe stronger alignment for a model trained in the imbalanced 90/10 regime, compared to the 50/50 counterpart (see App. F). Interestingly, the cosine similarity between gradients is also higher in the imbalanced setting: when processing equivalent sequences, the gradients with respect to $\boldsymbol{w}_1$ or $\boldsymbol{w}_2$ have an average cosine similarity of 0.53 for the model trained in the 90/10 setting, compared to 0.07 in the 50/50 setting (see full plots of similarities per model component in App. G). This suggests that the gradient updates with respect to one language may benefit the optimisation process of that language's cloned counterpart more when training under higher imbalance.

Takeaway 2. *Language imbalance improves generalisation and leads to representations which are more aligned across cloned languages.*

4.3 Many Languages

How does this trend transfer to settings with more than two languages? In such cases, sharing circuits across languages might be even more crucial due to the model's limited capacity. Instead of cloning

Run Type	Row	Training Data				PPL		TEff	
		# Tokens	$p(\text{EN}_1)$	$p(\text{EN}_2)$	$p(\text{EN}_3), \ldots, p(\text{EN}_{10})$	EN_1	EN_2	EN_1	EN_2
Monolingual	1	1.2B	100%	0%	0%	21.9	-	1	-
	2	$0.5 \times$ 1.2B	100%	0%	0%	25.3	-	1	-
	3	$0.1 \times$ 1.2B	100%	0%	0%	48.4	-	1	-
2 languages	4	1.2B	50%	50%	0%	26.1	26.1	0.89	0.89
	5	1.2B	90%	10%	0%	22.5	32.8	1.00	2.08
10 languages	6	1.2B	10%	10%	10%, ..., 10%	35.5	35.7	1.69	1.67
	7	1.2B	50%	$\frac{1}{18}$	$\frac{1}{18}, \ldots, \frac{1}{18}$	24.6	33.4	1.15	3.56
Schedule	8	1.2B	100% ↳ 0%	0% ↱ 100%	0%	>1B	31.4	-	0.47
	9	1.2B	90% ↳ 10%	10% ↱ 90%	0%	26.5	24.4	0.83	1.18
2x data	10	$2 \times$ 1.2B	50%	50%	0%	23.3	23.3	0.73	0.73
	11	$2 \times$ 1.2B	90% ↳ 10%	10% ↱ 90%	0%	22.8	20.4	0.83	1.60
3x data	12	$3 \times$ 1.2B	50%	50%	0%	22.2	22.2	0.64	0.64
	13	$3 \times$ 1.2B	90% ↳ 10%	10% ↱ 90%	0%	21.5	19.3	0.77	1.63

Table 1: Performance of language models trained on different compositions of EN_1 and EN_2. a% ↳ b% indicates an immediate decrease from a% down to b% halfway during training. Analogously, a% ↱ b% indicates an immediate increase.

the language only once, we now clone it nine times, obtaining in total 10 languages. In Table 1 (rows 6 & 7), we report the performance when sampling the languages in a balanced way and when having a much stronger main language.

Interestingly, when sampling uniformly, we obtain TEff ≈ 1.7; performance is thus better than with a monolingual model trained on an equivalent amount of monolingual data (compare rows 6 & 3). This differs from our observations for the bilingual setting, where uniform language sampling performed worse than the equivalent monolingual models. Presumably, modelling this many languages effectively with limited model capacity may lead the model to share its circuits, improving cross-lingual generalisation (Dufter and Schütze, 2020). The limit of infinite languages (in which a model never observes the same language more than once) was analysed by Huang et al. (2023); interestingly, LMs still seem to learn the language, to some extent, even in that setting.

In the imbalanced setting where we sample a stronger "main language" 50% of the time, we observe even stronger performance on all languages. Despite the model seeing only roughly 67M tokens in each of the rarer languages (1/18 of all steps), it achieves **better** performance in these languages than in the uniform setting with 120M tokens (1/10 of all steps) per language. In fact, on the rarer languages, the model achieves TEff ≈ 3.6, matching the performance of a monolingual model trained on 240M tokens.

Takeaway 3. *When training on many cloned languages, sampling a main language disproportionately improves generalisation.*

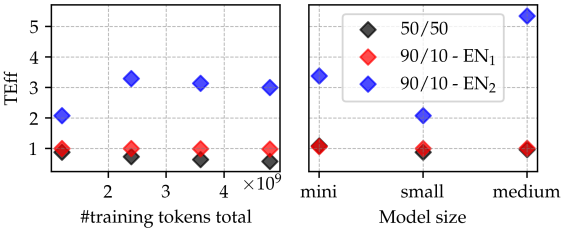

Figure 2: TEff as we train LMs with (left) more data, or (right) larger architectures. mini, small and medium denote GPT sizes in Languini (Stanić et al., 2023), with 11M, 85M, and 303M non-embedding parameters.

4.4 Effect of scaling

Model and data size are crucial factors for the performance of LMs. Here, we investigate how the previously identified trends are affected by scaling the model architecture or training data. Fig. 2 (left) shows that the effect of imbalance on cross-lingual generalisation appears to increase when we train on twice as much data (2.4B tokens instead of 1.2B), reaching TEff > 3; this corresponds to a "chinchilla-optimal" setup for our GPT small model (Hoffmann et al., 2022). At the same time, the TEff of the 50/50 setting seems to be decreasing under prolonged training. This might be caused by the heightened importance of model capacity under longer training, which may have a stronger impact on performance when representations are less aligned across languages. Overall, the disparity in effectiveness between the imbalanced and balanced settings grows with longer training. Remarkably, when training for 4.8B tokens, the 90/10 setting yields better performance in both languages, compared to the 50/50 setting.

When decreasing the model size, we also

observe higher performance benefits in the imbalanced setting (see Fig. 2 right), potentially due to the capacity argument described above. Interestingly, however, the effect of imbalance appears to be significantly stronger for larger models as well. When training a larger model with around 300M parameters (GPT medium in Languini; Stanić et al., 2023), in the $^{9}/_{10}$ setting, we achieve better performance on both languages than under the $^{50}/_{50}$ split. This might be because larger models generally exhibit better generalisation ability than smaller ones (Brown et al., 2020).

Takeaway 4. *Longer training and larger models lead to stronger performance benefits due to language imbalance.*

4.5 Language Sampling Schedule

Knowing that language imbalance boosts generalisation, how can we use this insight to train better models? Is there a way to leverage our insights in order to improve performance on two languages, even with the same training data? In Table 1 (rows 8, 9, 11, and 13), we report results when training with a language sampling schedule that ensures a language imbalance throughout all of training, but which still leads to an overall $^{50}/_{50}$ split between EN_1 and EN_2 data seen by the model. We sample EN_1 with a higher probability during the first half of training. Then, we sample EN_2 more often to achieve a marginal split of $^{50}/_{50}$.

When showing exclusively EN_1 at first, and then showing only EN_2 ($^{100\searrow 0}/_{0\nearrow 100}$; row 8), we observe bad overall performance. By the end of training, perplexity on EN_1 is very high, presumably due to catastrophic forgetting (McCloskey and Cohen, 1989; French, 1999). Further, EN_2 does not seem to benefit from the EN_1 data, achieving very low performance, which might be due to the lower learning rate in the second half of training. [5]

On the other hand, if we avoid catastrophic forgetting, making sure that the model encounters at least some samples of both languages at every point in training, via a $^{90\searrow 10}/_{10\nearrow 90}$ split (first sampling languages with ratio $^{9}/_{10}$, and then switching to $^{1}/_{90}$ after half of training), we can mitigate these issues. On our standard training set (1.2B tokens, row 9), we observe almost equivalent performance to uniform language sampling on EN_1, but significantly improved performance on EN_2. Under longer training, these benefits become more pronounced: this language schedule improves performance on both languages compared to the simple $^{50}/_{50}$ setting (compare row 10 vs 11 and row 12 vs 13).

Takeaway 5. *Compared to uniform language sampling, an imbalanced ratio throughout training can lead to better results on all languages, even if the overall language split remains balanced.*

5 Real Languages

To verify whether the insights from our cloned-language experiments hold in a more natural setting, we now run experiments with multilingual models on English (EN) and French (FR).

5.1 Generalisation

In the cloned setting, we observed no significant generalisation when training on a balanced language mix (i.e., TEff < 1, representations were unaligned, and zero-shot GLUE accuracy on EN_2 was bad). Similarly, when sampling EN and FR data uniformly, we also obtain TEff < 1. A multilingual model's performance is thus worse than a monolingual model trained only in the same EN or FR data (see Table 2, row 7). Notably, prior work has identified anchors (shared vocabulary items across languages) help generalisation (Dufter and Schütze, 2020; Pires et al., 2019; Wu and Dredze, 2019). We thus experiment with similarly merging vocabulary items shared between EN and FR, and confirm this helps performance (compare Table 2, row 7 vs 11). We run more experiments analysing the impact of anchor points in both cloned and real languages, see App. D. Note that, with an anchored vocabulary, generalisation across EN and FR is not necessarily upper bounded by our results on disjoint cloned languages. In fact, in the $^{50}/_{50}$ setting, we observe a marginally higher TEff for EN–FR models with an anchored vocabulary than for EN_1–EN_2 models where we used disjoint vocabularies (compare Table 1 row 4 and Table 2 row 11).

5.2 Language Imbalance

Analogous to the cloned setting, we observe that an imbalanced $^{\mathrm{EN}}/_{\mathrm{FR}}$ ratio leads to improved performance (TEff > 1), on the rarer language (see Table 2, rows 7-9 & 11-13). This is the case for both, a $^{9}/_{10}$ and a $^{1}/_{90}$ $^{\mathrm{EN}}/_{\mathrm{FR}}$ ratio. Fig. 3 shows PPL and TEff in EN and FR as a function of the language imbalance. We observe that large

[5] Chen et al. (2023) find that an equivalent setting can still be beneficial when using many more languages: they periodically reinitialise the learned embeddings (which is equivalent to switching to a new cloned language) and obtain models that are better adaptable to new languages.

Run Type	Row	Training Data # Tokens	Training Data p(EN)	Training Data p(FR)	PPL EN	PPL FR	TEff EN	TEff FR
Monolingual	1	1.2B	100%	0%	21.9	-	1	-
	2	0.5 × 1.2B	100%	0%	25.3	-	1	-
	3	0.1 × 1.2B	100%	0%	48.4	-	1	-
	4	1.2B	0%	100%	-	16.0	-	1
	5	0.5 × 1.2B	0%	100%	-	18.4	-	1
	6	0.1 × 1.2B	0%	100%	-	34.1	-	1
Multilingual disjoint vocabs	7	1.2B	50%	50%	26.4	19.4	0.85	0.82
	8	1.2B	90%	10%	22.5	31.9	1.00	1.05
	9	1.2B	10%	90%	43.5	16.4	1.10	0.97
	10	1.2B	90% ↳ 10%	10% ↱ 90%	29.1	20.5	0.60	0.66
Multilingual anchored vocabs	11	1.2B	50%	50%	26.0	19.0	0.91	0.88
	12	1.2B	90%	10%	22.5	29.0	1.00	1.27
	13	1.2B	10%	90%	39.5	16.5	1.33	0.96
	14	1.2B	90% ↳ 10%	10% ↱ 90%	28.9	19.3	0.61	0.83
	15	1.2B	90%↳10% ↱50% →50%	10%↱90% ↳50% →50%	26.4	18.5	0.85	1.00
	16	1.2B	95%↳35% →35% →35%	5%↱65% →65% →65%	26.3	18.7	0.86	0.95
2x data	17	2 × 1.2B	50%	50%	23.0	16.9	0.79	0.76
	18	2 × 1.2B	90% ↳ 10%	10% ↱ 90%	26.1	17.1	0.44	0.70
3x data	19	3 × 1.2B	50%	50%	21.8	16.0	0.70	0.67
	20	3 × 1.2B	90% ↳ 10%	10% ↱ 90%	25.1	16.2	0.35	0.63

Table 2: Performance of language models trained on different compositions of EN and FR. a%→b%→c% →d% indicates a four stage language schedule, switching immediately between, e.g., c% and d% after 75% of training.

imbalances generally seem to yield TEff > 1; the worst TEff is reached with a balanced $^{\text{EN}}/_{\text{FR}}$ ratio. These trends are in line with our findings in the cloned setting. However, especially with disjoint vocabularies, the observed performance benefits due to generalisation are less significant. Presumably, this is due to EN and FR not being equivalent and thus generally allowing less generalisation.

Does imbalance again improve generalisation due to a better alignment of the model's representations in the two languages? As in the cloned language setting, we investigate the cosine similarity between the models' hidden states when processing "equivalent" sequences in the two languages. For real languages, however, we do not have access to perfectly equivalent sequences. Instead, we mimick this scenario using parallel translated sequences in the two languages, which should contain roughly similar properties. Differently from the cloned language setting, we do not observe higher hidden state similarities for models trained on imbalanced data (see App. F). Further, we find that gradient similarities barely differ across balanced and imbalanced settings when using disjoint vocabularies. For the anchored vocabulary they are even marginally higher in the balanced setting (see App. G). We thus do not find evidence that the improved TEff in the imbalanced setting is caused by a stronger alignment of model updates across languages in this setting. A possible reason for this discrepancy could be that, at the scales of our experiments, LMs tend to rely on language specific surface-level features (which are shared by cloned languages, but not by distinct real languages) and show less understanding of complex semantics which might be more generalisable. Future research might thus consider investigating these trends at larger scales.

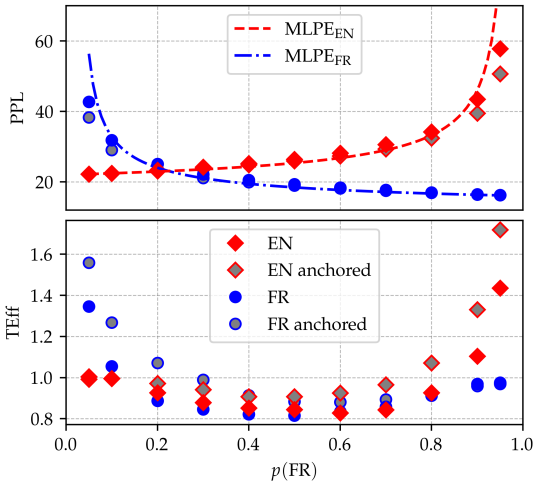

Figure 3: LM performance on EN and FR by imbalance ratio.

Takeaway 6. *Imbalanced multilinguality boosts the performance of real low-resource languages. However, this effect is weaker here than for cloned languages. Further, for real languages, we do not find evidence of language imbalance leading to representations which are more cross-lingually aligned.*

5.3 Effect of Scaling

In the cloned setting, we observed that prolonging training significantly decreased TEff in the $^{50}/_{50}$ setting. We hypothesised that this might be caused by a stronger influence of the limited model capacity with longer training, and poor sharing of representations between languages. As EN and FR are distinct languages that require at least some language specific representations, we might expect this trend to be even more pronounced for these languages. However, compared to the cloned setting, prolonging training leads to a smaller decline in TEff in the $^{50}/_{50}$ setting here. Presumably, the anchored vocabulary allows for better generalisation compared to the cloned setting, despite the languages being distinct.

Further, unlike in the cloned setting, longer training significantly decreases the TEff of the lower-resource language in the imbalanced setting here (see Fig. 4). In fact, the $^{90}/_{10}$ TEff even falls below 1, approaching the TEff of the $^{50}/_{50}$ setting. This suggests that language imbalance might not improve generalisation across distinct real languages. Still, when scaling up the model, we observe an increase of almost 2x in the TEff of the lower-resource language (see Fig. 4). This is in line with our cloned languages observations, although the effect is weaker.

Takeaway 7. *Performance benefits for real low-resource languages tend to decrease or vanish with longer training. Larger models, however, appear to yield higher performance benefits in both the balanced and imbalanced setting.*

5.4 Language Sampling Schedule

For equivalent cloned languages, we found that an imbalanced language sampling schedule can lead to improvements upon simple uniform sampling. If this held for real languages as well, it could have important practical implications for future multilingual LM training. However, whereas a $^{90}\searrow\,{}^{10}/_{10}\,\nearrow{}^{90}$ schedule yielded strong performance on cloned languages, matching or outperforming the $^{50}/_{50}$ setting, this is not the case for EN and FR (see Table 2, row 10 vs 7 and row 14 vs 11). Furthermore, in line with the observations above, longer training does not make this schedule more effective, but instead increases its gap to the performance of the $^{50}/_{50}$ setting (see rows 17-20).

The discrepancy between these results and the ones in cloned languages might be explained by the reduced effect of imbalance on the generalisation and representation alignment in real languages. The schedules may be enough to force LMs to share circuits across cloned languages, but not across real ones. To investigate if this negative result was a particular property of our chosen schedule, we explore other more complex scheduling options. [6] In general, none of the tested schedules appears to outperform the $^{50}/_{50}$ setting (see rows 15, 16) on both languages. However, more complex 4-stage schedules, can obtain better performance on one language while incurring a slight performance drop in the other. Intriguingly, this allows trading off the performance of different languages without altering the training data.

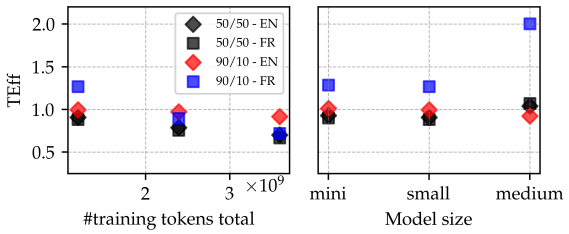

Figure 4: TEff of models on EN and FR with anchored vocab as we train them with (left) more data, or (right) larger architectures.

Takeaway 8. *For real languages, we do not find improvements on all languages due to the tested language schedules. However, they allow for trading off performance in different languages.*

6 Conclusion

We ran experiments to measure cross-lingual generalisation in both a controlled setting with cloned English languages, as well as with English and French. In both settings, we find that, without vocabulary overlap, our models do not show strong cross-lingual generalisation when trained on a balanced language set. However, when training on an imbalanced mix of languages, we observe increased performance compared to monolingual settings. For cloned languages, we find that this can be explained by a higher alignment of the model's representations across languages, which indicates circuit reuse and improved cross-lingual generalisation. Yet, at the scales of our experiments, such a correlation is less evident in real languages. While our findings allow us to design an imbalanced language schedule that yields improved performance in the cloned setting, further research is required to extend these improvements to real-world settings.

[6] Future research might design these more carefully, also analysing the interplay of language- and learning rate schedule

Limitations

There are several limitations of our work, many of which present opportunities for future research.

Data and model size. While we conduct experiments with varying data (up to 4.8B tokens) and model size (up to 336M parameters), it is uncertain whether the identified trends also apply at the scale of modern large language models. Additionally, for more capable models, cross-lingual generalisation might be relevant in different aspects, with, e.g., semantics playing a larger role. As the semantic content communicated in different languages might be easily transferable, this might impact generalisation dynamics.

Languages. We only run experiments on English and French, two Indo-European languages. Further work could consider more languages and investigate the impact of language similarity in results more broadly.

Model architecture. We run most of our experiments on a Transformer decoder (we also measure embedding alignment for simpler Word2Vec models). Future research could analyse the effects of architecture in more depth to better understand the drivers of representation alignment. Conneau et al. (2020b), e.g., find that shared parameters in the top layers lead to better cross-lingual transfer. In our Word2Vec experiments, we do not observe improvements in representation alignment due to language imbalance (see Fig. 6), presumably due to no parameters being shared between the two languages. Would this change when adding a shared layer to the Word2Vec model?

Downstream performance. In our evaluation we mainly rely on perplexity as a metric, with a single experiment on GLUE accuracy. It might be insightful to analyze effects on downstream task performance more broadly.

Quantifying generalisation. In this work, we mainly measure cross-lingual generalisation by comparing the performance of multilingual models with that of monolingual models trained on the same amount of data in the given language. If a multilingual model on languages L_A and L_B requires fewer L_A tokens to reach a given perplexity on L_A than a monolingual model, we speak of cross-lingual generalisation, knowing that performance on L_A must have been boosted by data in language L_B. Future work could formalise this measure and aim to model/quantify the relationship between the number of training tokens seen in a language L_B and performance in another language L_A, depending on model size, language imbalance, language similarity, anchor points, and other factors. An accurate model of these relationships could be of substantial practical value.

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? Tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.

Jesujoba O. Alabi, Marius Mosbach, Matan Eyal, Dietrich Klakow, and Mor Geva. 2024. The hidden space of transformer language adapters. *arXiv preprint arXiv:2402.13137*.

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.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Nick Cammarata, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov, Ludwig Schubert, Chelsea Voss, Ben Egan, and Swee Kiat Lim. 2020. Thread: Circuits. *Distill*.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K Bergen. 2023. When is multilinguality a curse? language modeling for 250 high-and low-resource languages. *arXiv preprint arXiv:2311.09205*.

Yihong Chen, Kelly Marchisio, Roberta Raileanu, David Adelani, Pontus Lars Erik Saito Stenetorp, Sebastian Riedel, and Mikel Artetxe. 2023. Improving language plasticity via pretraining with active forgetting. *Advances in Neural Information Processing Systems*, 36:31543–31557.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034. Association for Computational Linguistics.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437. Association for Computational Linguistics.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Manuel Faysse, Patrick Fernandes, Nuno Guerreiro, António Loison, Duarte Alves, Caio Corro, Nicolas Boizard, João Alves, Ricardo Rei, Pedro Martins, et al. 2024. CroissantLLM: A truly bilingual french-english language model. *arXiv preprint arXiv:2402.00786*.

Yukun Feng, Feng Li, and Philipp Koehn. 2022. Toward the limitation of code-switching in cross-lingual transfer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5966–5971. Association for Computational Linguistics.

Robert M. French. 1999. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Aric Hagberg, Pieter Swart, and Daniel Chult. 2008. Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science Conference*.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2024. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems*, 36.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*.

Qian Huang, Eric Zelikman, Sarah Li Chen, Yuhuai Wu, Gregory Valiant, and Percy Liang. 2023. Lexinvariant language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: An empirical study. In *International Conference on Learning Representations*.

Richard M. Karp. 1978. An algorithm to solve the mxn assignment problem in expected time O(mn log n). Technical Report UCB/ERL M78/67, EECS Department, University of California, Berkeley.

Eugene Kharitonov, Marco Baroni, and Dieuwke Hupkes. 2021. How BPE affects memorization in transformers. *arXiv preprint arXiv:2110.02782*.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, San Diego, CA, USA.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Minh Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Alexandrovich Glushkov, Arnav Varma Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Julian Mattick. 2023. OpenAssistant conversations - democratizing large language model alignment. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

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

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, Workshop Track Proceedings*, Scottsdale, Arizona, USA.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

PleIAs. 2024. French-PD-Books dataset. `https://huggingface.co/datasets/PleIAs/French-PD-Books`. Accessed in 01/2024, Hugging Face Datasets library.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Machel Reid and Mikel Artetxe. 2022. PARADISE: Exploiting parallel data for multilingual sequence-to-sequence pretraining. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 800–810, Seattle, United States. Association for Computational Linguistics.

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, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-

Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. BLOOM: A 176B-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Anton Schäfer, Thomas Hofmann, Imanol Schlag, and Tiago Pimentel. 2024. On the effect of (near) duplicate subwords in language modelling. *arXiv preprint arXiv:2404.06508*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Aleksandar Stanić, Dylan Ashley, Oleg Serikov, Louis Kirsch, Francesco Faccio, Jürgen Schmidhuber, Thomas Hofmann, and Imanol Schlag. 2023. The languini kitchen: Enabling language modelling research at different scales of compute. *arXiv preprint arXiv:2309.11197*.

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. 2023a. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Zirui Wang, Zachary C Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do Llamas work in English?

On the latent language of multilingual transformers. *arXiv preprint arXiv:2402.10588*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 833–844.

Jiacheng Ye, Xijia Tao, and Lingpeng Kong. 2023. Language versatilists vs. specialists: An empirical revisiting on multilingual transfer ability. *arXiv preprint arXiv:2306.06688*.

A Experimental Setup

Model. We use a GPT-2-style decoder-only transformer architecture in our experiments (Radford et al., 2019). Unless otherwise noted, we instantiate our model with 12 layers and a hidden size of 768, which results in 85M non-embedding parameters; this corresponds to Languini's `gpt-small` configuration. We follow previous work and train our models with sequence length 512, batch size 128, the Adam optimiser (Kingma and Ba, 2015), and a cosine learning rate schedule from 6e-4 to 6e-6 with 500 warmup steps.

Data. For the English settings, we use Languini's default datasets to train and evaluate our models. These are English books from a filtered version of the books3 subset from the Pile (Gao et al., 2020). The train set consists of a total of 23.9B tokens, while the test set contains i.i.d. books, with a total of roughly 11M tokens. This data is tokenised into a vocabulary of size 16k, obtained using a BPE tokeniser trained with SentencePiece (Gage, 1994; Sennrich et al., 2016; Kudo and Richardson, 2018). For our experiments in French, we use the French-PD-Books dataset (PleIAs, 2024), to which we apply the preprocessing pipeline of the Languini Kitchen, but for French. We train a separate BPE tokeniser on this French dataset, using a 16k-sized vocabulary. Depending on the experiment, the French and English vocabularies are either kept separate (disjoint) or merged (anchored). Unless otherwise noted, we train our models for 18,265 steps—i.e., the first 1.2B tokens in our dataset; this corresponds to a GPT small model trained for 6h on an RTX 3090 GPU, the Languini GPT small 6h setting (Stanić et al., 2023). For experiments where we compare hidden representations or gradients on parallel French–English or cloned English sequences, we use data from the Europarl parallel corpus (Koehn, 2005).

Evaluation. When evaluating PPL (from which we also compute MLPE, MLTE and TEff) on the held-out test set, we want to ensure sufficient context for all predictions. To this end, we use a sliding window with steps of 128: we fill in a 512 tokens context, ignore the model's outputs on the initial 384, and evaluate it only using the last 128 tokens.

B Fitted Scaling Laws

To predict the performance of monolingual models depending on the amount of tokens they are trained on, we fit a power law curve to predict the relationship between number of training tokens and perplexity for models of all three sizes and for both languages (see Fig. 5).

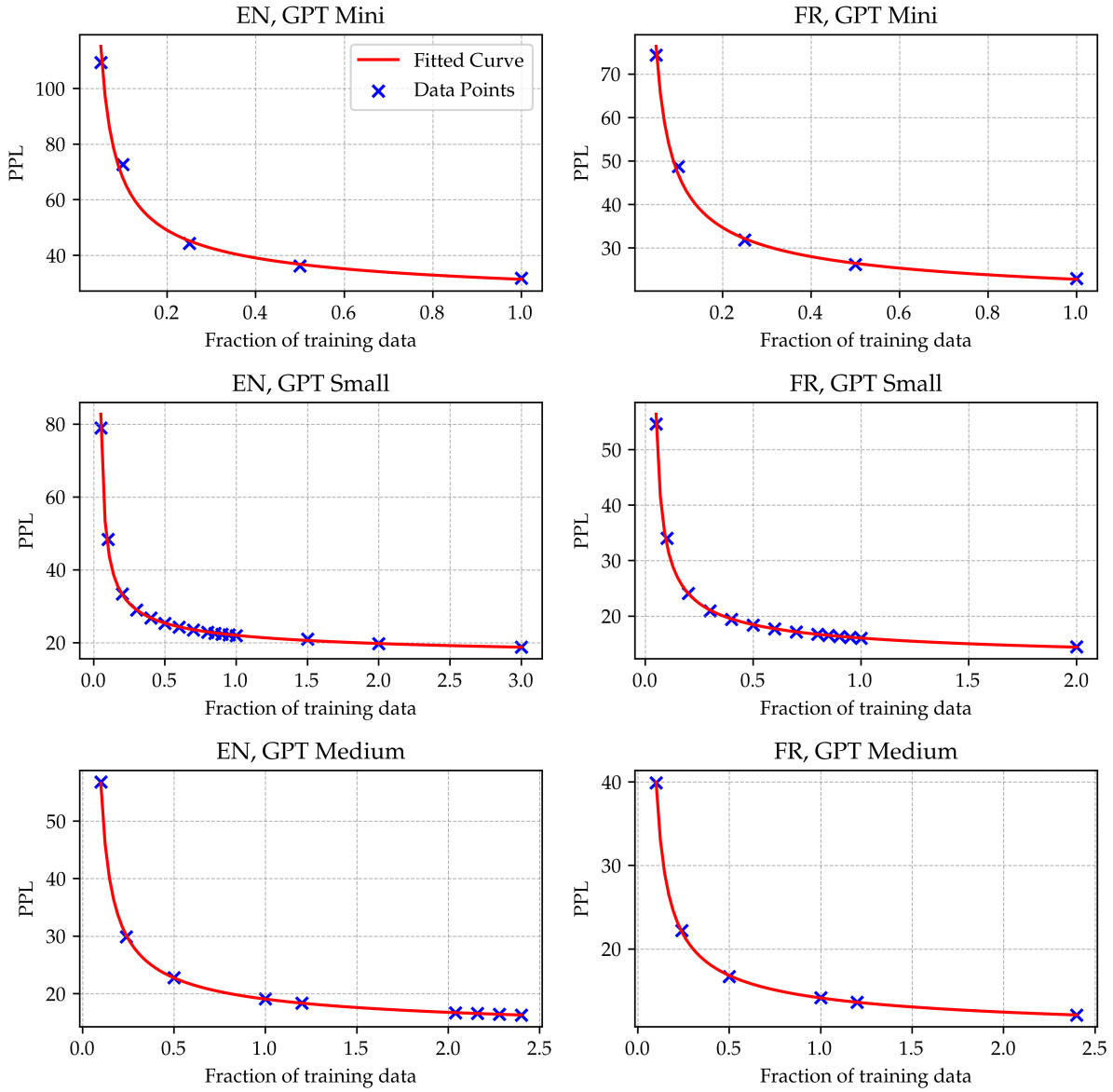

Figure 5: Fitted power laws curves predicting perplexity depending on the fraction of training tokens (compared to our standard 1.2B tokens) for different languages and model sizes.

C Alignment of EN_1 and EN_2 Representations

While, under balanced language sampling, embeddings of corresponding subwords are not much more aligned than embeddings of random pairs, we observe an increase in cosine similarity with increasing language imbalance: from 0.02 for 50/50 to 0.28 for 95/5 (see Fig. 6). Fig. 7 shows that this alignment is higher for frequent subwords. This seems natural: at initialisation, subword embeddings are random and not aligned. Then, they become more and more aligned over the course of training.

Interestingly, the embeddings of a simple word2vec (Mikolov et al., 2013) model do not show stronger alignment under higher imbalance. This might be due to a lack of shared parameters between the languages (Conneau et al., 2020b).

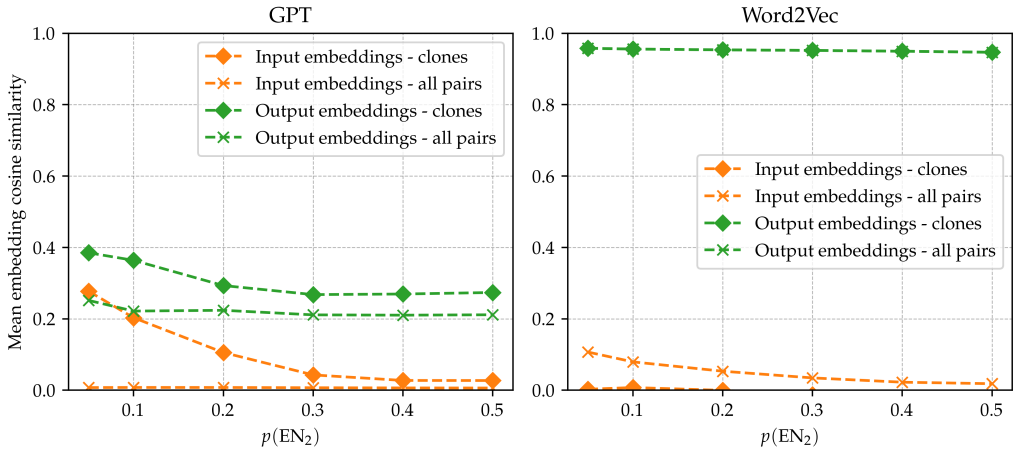

Figure 6: Embedding cosine similarity of corresponding duplicate subwords from EN_1 and EN_2 and random pairs to control for anisotropy. Left: our GPT model. Right: Word2vec embeddings trained on the same data (computed with Gensim).

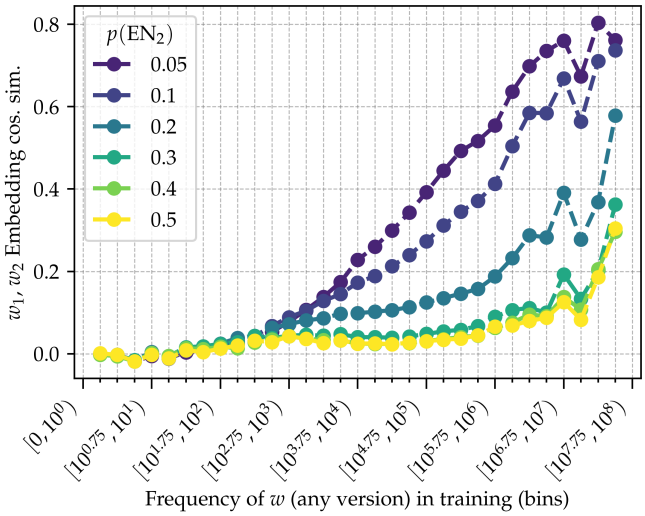

Figure 7: Embedding cosine similarity of corresponding cloned subwords $w_1 \stackrel{\circ}{=} w_2$ from EN_1 and EN_2, by frequency.

D Anchor Points

D.1 Anchors on Cloned Languages

As described earlier, previous works found that anchor points—i.e., lexical items which overlap between languages—can lead to better generalisation and alignment of representations (Dufter and Schütze, 2020; Pires et al., 2019; Wu and Dredze, 2019). In our cloned setting, we can investigate this in a controlled manner by varying the number of vocabulary elements we duplicate. While in the experiments described above we created EN_2 by duplicating the entire vocabulary, we now duplicate only a fraction. The remaining vocabulary is shared between EN_1 and EN_2. In this experiment, we observe that a small number of anchor points already significantly boosts model performance (see Fig. 8), which indicates improved generalisation.

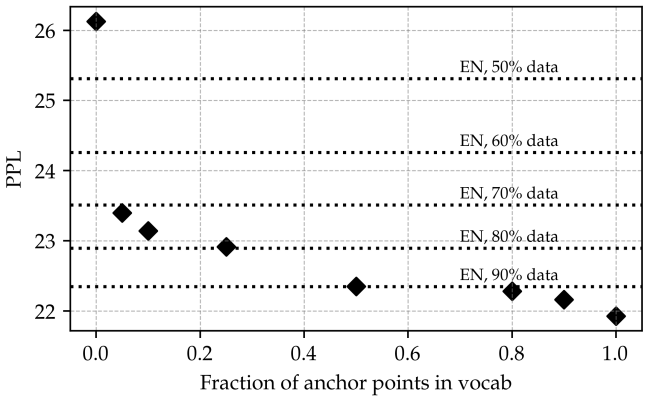

Figure 8: Perplexity by percentage of anchor points, i.e., overlap between EN_1 and EN_2 vocabularies. Models trained on balanced $\mathrm{EN}_1/\mathrm{EN}_2$ split.

D.2 Anchors on Real Languages

English and French vocabularies naturally overlap, having common subwords. These shared elements potentially act as anchors, facilitating better cross-lingual generalisation. However, the effectiveness of such anchor points may be moderated by semantic differences; for instance, a shared subword might carry a different meaning or connotations in English and French, affecting its utility as an anchor. Despite these nuances, anchor points appear to boost generalisation between real languages: when we merge the EN and FR vocabularies, we obtain better performance on both languages (compare Table 2, row 7 vs 11) as well as higher alignment of gradients (see App. G). This aligns with our findings from the cloned language setting (see App. D.1). Given these benefits, it is natural to use an anchored (i.e., merged) vocabulary when possible.[7]

[7] In practice, this is usually achieved by training a tokeniser on multilingual data, instead of merging monolingually trained vocabularies.

E Larger Models and More Data

Fig. 9 and Fig. 10 contain results for the full array of model- and dataset size combinations we ran for cloned languages and for English and French, respectively.

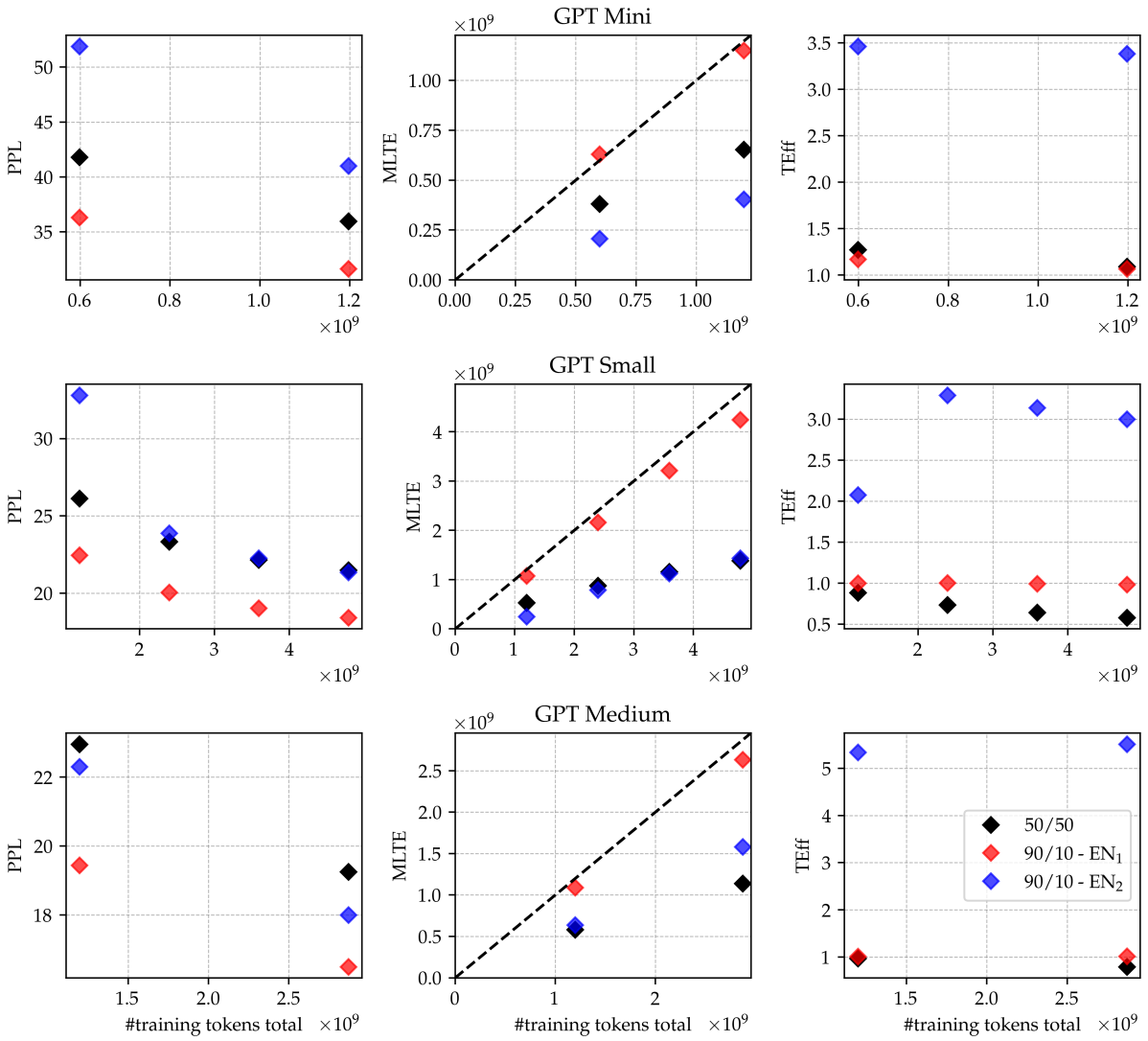

Figure 9: Performance with balanced and imbalanced EN_1 and EN_2 data for different configurations of model- and dataset size

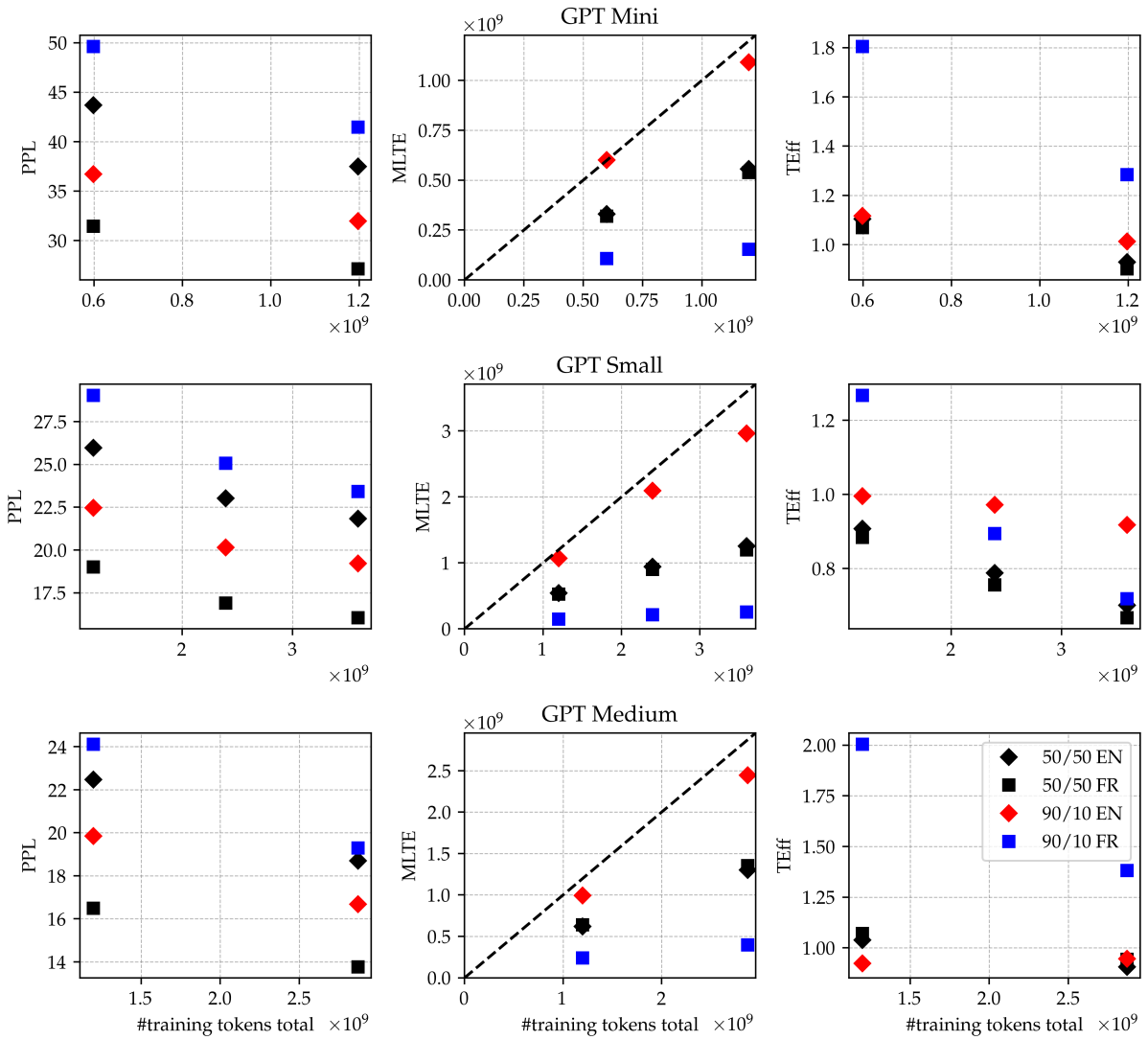

Figure 10: Performance with balanced and imbalanced EN and FR data for different configurations of model- and dataset size. Using anchored vocabulary.

F Hidden State Similarity

Here, we compare the hidden states of our model when processing parallel sequences, both in cloned languages (see Table 3) and in English and French (see Table 4). I.e., for a given trained model and parallel sequences $\boldsymbol{w}_a$ and $\boldsymbol{w}_b$, we first feed $\boldsymbol{w}_a$ through the model, then $\boldsymbol{w}_b$, and finally compute the cosine similarities for the hidden states of pairs of corresponding tokens from $\boldsymbol{w}_a$ and $\boldsymbol{w}_b$ (see App. H for details on how these pairs are determined). We use 500 parallel sequences obtained from the Europarl parallel corpus (Koehn, 2005). For cloned languages, we observe that hidden states of the model trained under higher language imbalance generally have higher cosine similarity than the those of the model trained in a balanced setting. For English and French such a trend is less clear. Interestingly, however, an anchored vocabulary seems to lead to slightly higher similarities of the hidden states.

Training Data		Layer											
$p(\mathrm{EN}_1)$	$p(\mathrm{EN}_2)$	1	2	3	4	5	6	7	8	9	10	11	12
50%	50%	0.55	0.79	0.83	0.88	0.85	0.83	0.78	0.66	0.56	0.46	0.25	-0.21
90%	10%	0.86	0.93	0.96	0.96	0.96	0.96	0.96	0.95	0.94	0.90	0.67	0.11
Δ		0.31	0.14	0.13	0.09	0.11	0.14	0.18	0.28	0.38	0.44	0.42	0.32

Table 3: Hidden states' cosine similarity when LM is fed equivalent inputs in cloned languages. Similarity is computed per token (i.e., comparing pairs of equivalent tokens).

	Training Data		Layer											
	$p(\mathrm{EN})$	$p(\mathrm{FR})$	1	2	3	4	5	6	7	8	9	10	11	12
Disjoint	50%	50%	0.68	0.80	0.84	0.88	0.86	0.84	0.80	0.75	0.62	0.53	0.34	-0.15
	90%	10%	0.71	0.83	0.88	0.87	0.86	0.84	0.81	0.74	0.69	0.57	0.40	-0.17
	Δ		0.03	0.03	0.04	-0.01	0.00	0.00	0.01	0.00	0.07	0.04	0.06	-0.03
Anchored	50%	50%	0.73	0.84	0.88	0.91	0.89	0.88	0.85	0.78	0.71	0.61	0.36	0.10
	90%	10%	0.78	0.87	0.89	0.91	0.89	0.87	0.84	0.77	0.72	0.63	0.28	0.06
	Δ		0.05	0.03	0.01	0.00	0.00	-0.01	0.00	-0.01	0.00	0.02	-0.08	-0.04

Table 4: Hidden states' cosine similarity for parallel inputs in EN and FR for anchored and disjoint vocabularies. We first match which tokens correspond to each other in the two languages, and then compare their representations (see App. H).

G Gradient Similarity

Here, we compare the cosine similarity of trained models' gradients with respect to parallel sequences in two different (possibly cloned) languages. For cloned languages, the alignment between gradients is significantly higher for the model trained in the imbalanced 90/10 setting (see Fig. 11). For EN and FR data, this does not seem to be the case, whether the vocabulary is anchored (see Fig. 12) or disjoint (see Fig. 13). However, under the anchored vocabulary, the gradient similarities appear to be generally higher, suggesting better cross-lingual representation alignment.

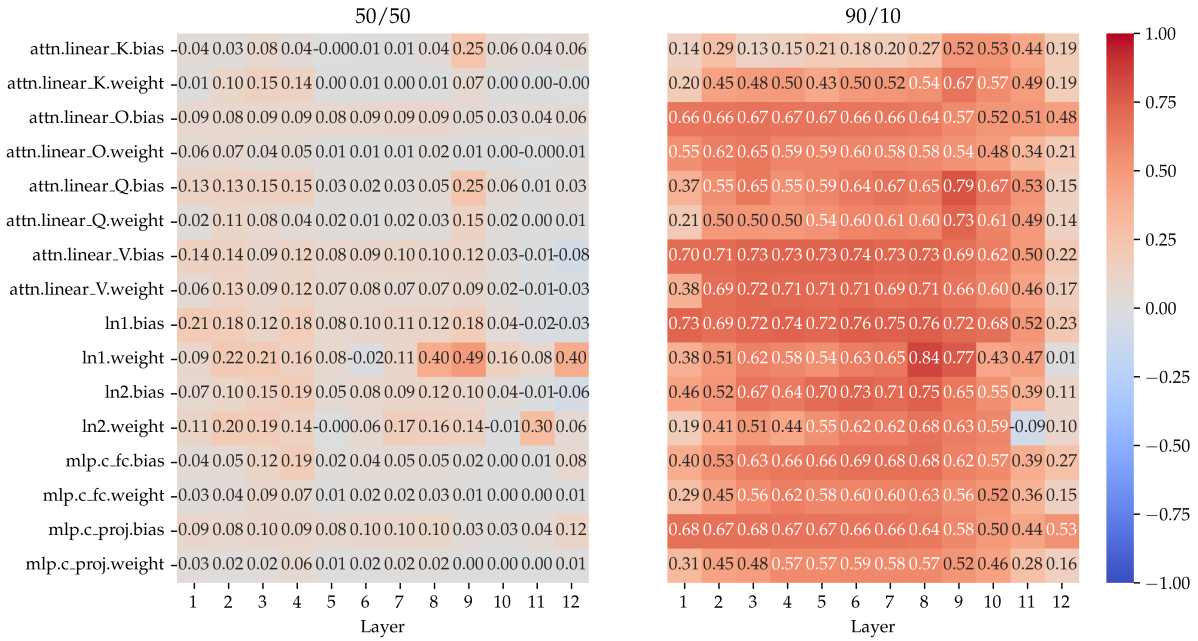

Figure 11: Similarity of gradients with respect to parallel sequences in EN_1 and EN_2 for models trained in balanced and imbalanced settings. Macro average for 50/50: 0.07. Macro average for 90/10: 0.53.

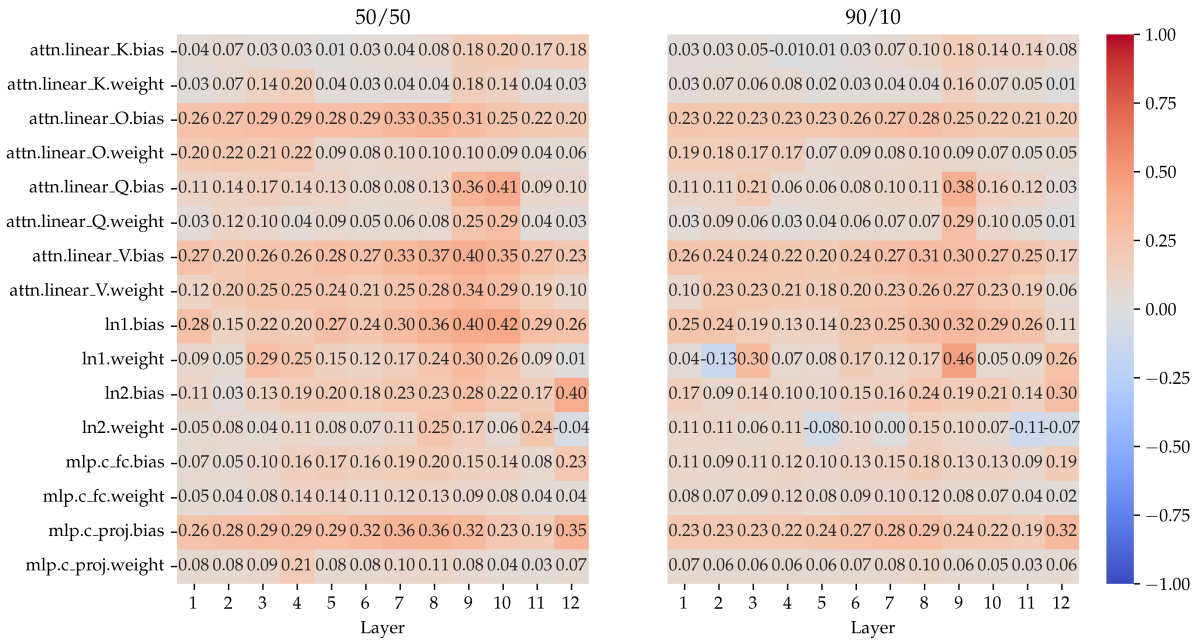

Figure 12: Similarity of gradients with respect to parallel sequences in EN and FR for models with anchored (i.e., merged) vocabulary, trained in balanced and imbalanced settings. Macro average for 50/50: 0.17. Macro average for 90/10: 0.14.

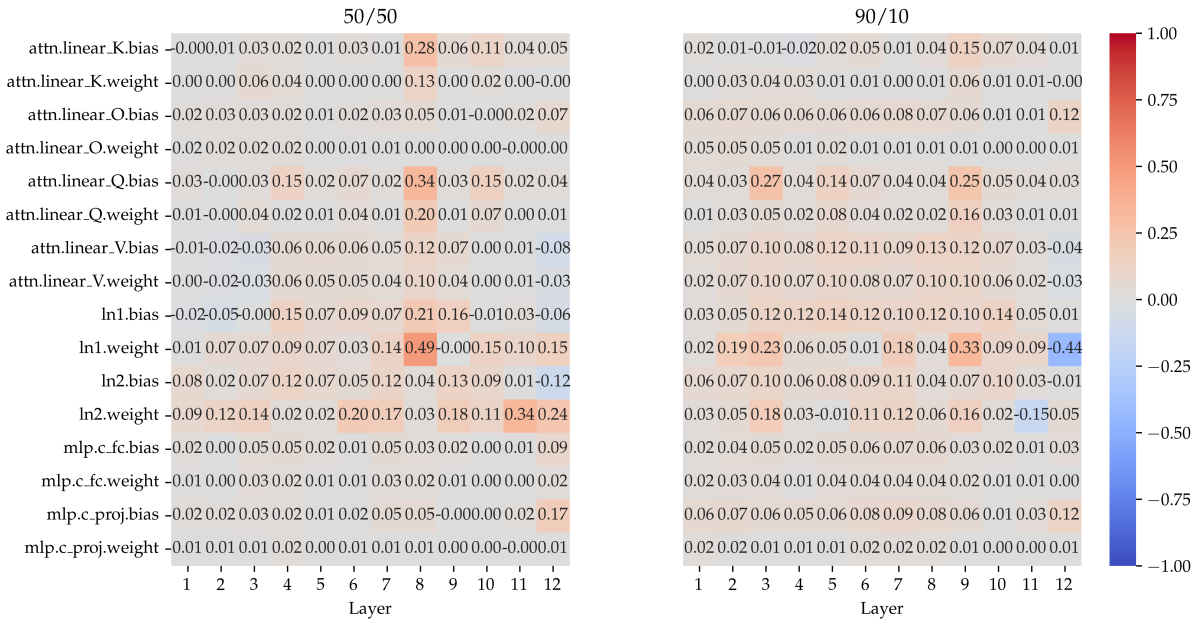

Figure 13: Similarity of gradients with respect to parallel sequences in EN and FR for models with disjoint vocabularies, trained in balanced and imbalanced settings. Macro average for 50/50: 0.04. Macro average for 90/10: 0.05.

H Matching Corresponding Tokens

In our experiments in §5.2, we employ parallel sequences in different languages and compare both their hidden states' and their gradients' similarity.

When comparing gradients (see App. G), we adopt a setup that is analogous to the training process as we aim to understand how one language might affect optimisation of the other: we compute gradients with respect to a full sequence in each language, and then compare these sequence-level aggregated gradients. Analogously, during training, gradient updates are also aggregated for entire sequences. (In fact, during training, these updates are also aggregated for an entire batch, but we use a batch size of 1 for this evaluation.)

However, when comparing hidden states, we compare the individual representations of corresponding tokens in the two sequences. We first compute the cosine similarity of each equivalent token pair, and only then average over the sequence dimension; this provides us with a more informative signal. For parallel sequences $\boldsymbol{w}_{\text{EN}_1} \overset{\circ}{=} \boldsymbol{w}_{\text{EN}_2}$ in cloned languages, it is clear which token corresponds to which: At each given position t, we know that $w_{\text{EN}_1,t} \overset{\circ}{=} w_{\text{EN}_2,t}$ so we can simply compare the hidden states position by position (see Table 3).

Yet, this might not be the case for real languages EN and FR, e.g., due to differing word order or tokenisation. To ensure that we still compare the hidden states of tokens that approximately correspond to each other in the respective languages, we match them based on their cosine similarity scores. Concretely, we create a bipartite graph where the nodes consist of the tokens of the two sequences. For every pair of tokens $w_{\text{EN},t}$ and $w_{\text{FR},t'}$ we add an edge which is weighed by the mean cosine similarity of their hidden states across all layers. We then compute a maximum weight full matching in this graph.[8] Such a matching maximises the average similarity across all token pairs. Indeed, the resulting token pairs appear to approximately correspond to each other (see Fig. 14). We can then compare the hidden states of these pairs (see Table 4).

Notably, the cosine similarities of hidden states of corresponding EN and FR tokens $w_{\text{EN},t}$ and $w_{\text{FR},t'}$ computed in this way generally appear to be slightly higher than for corresponding tokens $w_{\text{EN}_1,t} \overset{\circ}{=} w_{\text{EN}_2,t}$ of cloned languages (compare Table 4 (disjoint) and Table 3). This might seem unexpected, given that $w_{\text{EN}_1,t}$ and $w_{\text{EN}_2,t}$ are perfectly equivalent but $w_{\text{EN},t}$ and $w_{\text{FR},t'}$ are generally not. Could this be an artifact of the employed matching strategy which always maximises the average similarity, potentially matching tokens that have very high similarity but are completely unrelated? If this is the case, we should also obtain higher similarity scores in the cloned setting when using the described matching strategy instead of comparing position by position. After running this experiment, we find that using the matching strategy the similarities under the $^{50}/_{50}$ cloned language split are indeed marginally higher, although only in the last layers. Under the $^{90}/_{10}$ split, however, we observe no notable changes. It thus seems that the proposed matching strategy does not artificially inflate similarity scores too strongly.

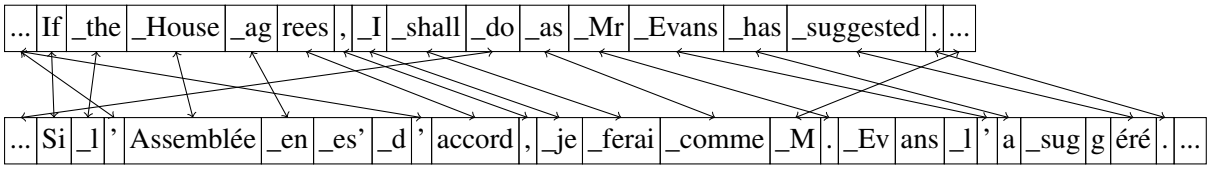

Figure 14: Computed matching for an example sentence using a model trained under $^{50}/_{50}$ split with anchored vocabulary. Pointers to "..." denote a match with a token earlier or later in the sequence.

[8] We compute the matching using the NetworkX (Hagberg et al., 2008) implementation of the algorithm proposed by Karp (1978).