Sparse Subnetwork Enhancement for Underrepresented Languages in Large Language Models

Anonymous ACL submission

Abstract

Large language models exhibit uneven performance across languages, with substantial gaps between high- and low-resource languages. We present a framework for enhancing monolingual capabilities of LLMs in underrepresented languages while preserving their general-purpose performance through targeted fine-tuning of language-specific subnetworks. Our approach identifies language-specific neurons using Language Activation Probability Entropy and fine-tunes only the weights associated with these neurons, a dedicated subnetwork, on target-language data. Experiments on LLAMA-3.1-8B and MISTRAL-NEMO-12B across 12 mid- and low-resource languages demonstrate that our method consistently outperforms full fine-tuning, FFN-only fine-tuning, LoRA adaptation, and random subset fine-tuning baselines while efficiently updating only up to 1% of model parameters. Beyond performance improvements, we observe enhanced favorable training dynamics, cross-lingual representational alignment, and systematic weight update changes. To facilitate future research, we release language-specific neuron identifications for over 100 languages as well as our adaptation pipeline, offering a cost-effective pathway for adapting state-of-the-art models to underrepresented languages.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across diverse tasks and languages, yet their performance remains unevenly distributed, with substantial gaps between high-resource and low-resource languages (Joshi et al., 2020; Huang et al., 2023). While scaling training data and model parameters has improved multilingual performance, this approach is still often impractical and ineffective for resource-constrained scenarios involving hundreds of languages (Robinson et al., 2023).

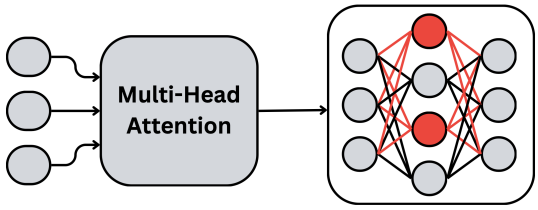


Figure 1: Sparse subnetwork adaptation method: attention blocks remain frozen (grey) while selected language-specific FFN neurons and their associated weights are fine-tuned (red). This approach enables parameter-efficient adaptation for underrepresented languages while preserving the model's general knowledge.

Recent advances in mechanistic interpretability have revealed that language models develop specialized internal structures for different languages, with distinct neurons and attention patterns activated for specific linguistic contexts (Tang et al., 2024; Zhao et al., 2024; Kojima et al., 2024). This observation suggests a promising alternative to full model retraining: **identifying and selectively enhancing language-specific components** while preserving general-purpose capabilities.

Current approaches to multilingual adaptation face a fundamental trade-off between target-language performance and retention of general capabilities. Full fine-tuning often leads to catastrophic forgetting (French, 1999), while parameter-efficient methods like adapters (Houlsby et al., 2019a), LoRA (Hu et al., 2021), and soft prompts (Vykopal et al., 2025) may not fully exploit the model's existing multilingual structure. Existing neuron-level approaches have shown promise but have primarily focused on task-specific adaptation rather than systematic language enhancement, and they have often fallen short in preserving existing knowledge (Zhu et al., 2024; Xu et al., 2024; Mondal et al., 2025).

In this work, we present a framework for enhancing monolingual capabilities of LLMs through targeted fine-tuning of language-specific subnetworks within feed-forward network (FFN) components. Our approach leverages Language Activation Probability Entropy (LAPE) (Tang et al., 2024) to identify neurons that exhibit strong language-specific activation patterns within LLAMA-3.1-8B and MISTRAL-NEMO-12B, then fine-tunes only the weights associated with these neurons using target-language data. This sparse fine-tuning updates up to 1% of model parameters while achieving substantial improvements in target-language performance and preserving general capabilities.

We make the following contributions:

- We introduce a systematic framework for efficient and effective language subnetwork identification and enhancement that preserves general-purpose performance while significantly improving target-language capabilities.
- We demonstrate the effectiveness of our approach across 12 mid- and low-resource languages, showing consistent improvements over full fine-tuning, FFN-only fine-tuning, LoRA fine-tuning, and random subset fine-tuning baselines.
- We show that target-language performance can be improved by increasing the amount of training data where possible, laying the foundation for optimal adaptation strategies across different languages.
- Through activation analysis, we show that targeted subnetwork fine-tuning improves cross-lingual alignment, while weight-change analysis reveals that its effects are concentrated in the later down-projection matrices of FFN layers, highlighting their central role in adapting to new languages.
- We release language-specific neuron identifications for over 100 languages, providing a foundation for future research in low-resource language adaptation.[1]

This work represents a step toward democratizing high-quality language model capabilities across diverse languages. **By providing both a practical framework and comprehensive resources, we aim to lower the barrier for adapting state-of-the-art models to underrepresented languages**.

[1]Our code and language-specific neuron dictionaries will be released upon acceptance under URL.

2 Related Work

Prior work on adapting multilingual language models to underrepresented languages can be grouped into three main directions: full fine-tuning, adapter-based approaches, and more recent neuron- or head-level methods.

2.1 Full Fine-Tuning Approaches

Full fine-tuning is the most direct strategy for language adaptation. Post-training on related languages has been used to mitigate overfitting (Neubig and Hu, 2018), while domain-adaptive fine-tuning improved contextualized models on specialized corpora (Han and Eisenstein, 2019). Language-specific fine-tuning on monolingual corpora (Gururangan et al., 2020; Chau et al., 2020), transliteration-based adaptation (Muller et al., 2021), and even small corpora such as Bible translations (Ebrahimi and Kann, 2021) have boosted performance on tagging, parsing, and NER. However, full fine-tuning is costly, data-hungry, and prone to catastrophic forgetting (French, 1999), motivating parameter-efficient alternatives.

2.2 Adapter-Based Approaches

Adapters (Houlsby et al., 2019b) introduce small trainable modules into frozen backbones, reducing computational cost (Strubell et al., 2019) and data requirements (Faisal and Anastasopoulos, 2022) while avoiding forgetting (French, 1999). MAD-X (Pfeiffer et al., 2020), UDapter (Üstün et al., 2020), and MAD-G (Ansell et al., 2021) extend adapters with task/language modularity and typological signals, while other work explores hierarchical designs (Faisal and Anastasopoulos, 2022), language balancing (Lee et al., 2022; Parović et al., 2022), and inference-aware training (Parović et al., 2023). Recent studies show adapter tuning can outperform continued pre-training for adaptation (Yong et al., 2023; Gurgurov et al., 2025b).

2.3 Neuron-Level Approaches

Recent work has explored fine-grained adaptation by targeting individual neurons or attention heads. Zhu et al. (2024) identify language-aware neurons in FFNs using Taylor scores (Xie et al., 2021) and fine-tune only the relevant subsets using parallel data for machine translation, outperforming both

adapters and full fine-tuning in high-resource settings. Similarly, Zhao et al. (2024) adapt ∼0.1% of neurons in FFNs and attention heads identified via neuron ablations, observing initial gains for a subset of languages but rapid saturation with additional data. Xu et al. (2024) propose NEFT, which compares pre- and post-fine-tuning representations to locate task-sensitive neurons, and show that fine-tuning just 5–10% of them consistently surpasses LoRA (Hu et al., 2021) on translation and summarization. Mondal et al. (2025) apply LAPE (Tang et al., 2024) for neuron identification and manipulation, but report limited improvements in cross-lingual transfer. Their setup, however, differs from ours in that they restrict analysis to only three relatively high-resource languages and adapt LoRA modules on top of FFNs rather than directly fine-tuning neurons inside the model, which makes their findings not directly comparable to ours. Complementary work shows that only a few attention heads encode translation mechanisms, and selectively adapting these heads matches full fine-tuning while preserving general abilities (Zhang et al., 2025).

2.4 Towards Targeted Adaptation

Neuron- and head-level methods demonstrate that precise adaptation can rival or exceed large-scale fine-tuning, yet prior work has emphasized task-specific gains rather than systematic language adaptation. In particular, it remains unclear to what extent language-specific information can be injected into carefully chosen subnetworks for underrepresented languages without harming general performance. Our work addresses this gap by leveraging LAPE, an information-theoretic method that has recently shown strong potential for precise language neuron identification (Tang et al., 2024) and manipulation (Gurgurov et al., 2025a), to identify language-specific subnetworks and fine-tune their associated weights on target-language data.

3 Methodology

Our approach consists of two components: (*i*) identifying language-sensitive neurons, and (*ii*) selectively fine-tuning the corresponding subnetwork (see Figure 1). Together, these steps allow us to adapt LLMs to underrepresented languages while minimizing changes to unrelated parameters.

3.1 Neuron Identification

We leverage Language Activation Probability Entropy (LAPE) (Tang et al., 2024) to locate neurons that exhibit language-specific behavior. We focus on neurons within FFN components, since prior work has shown that they frequently encode rich factual knowledge (Geva et al., 2021; Dai et al., 2021; Wang et al., 2022) and multilingual information (Zhao et al., 2024).

For each neuron j in layer i, we compute the probability of activation under inputs from language k:

$$p^k_{i,j} = \mathbb{E}_{x\sim\mathcal{D}_k}\Big[\mathbf{1}\big(\phi(h_i(x)W^i_1)_j > 0\big)\Big],$$

where $h_i(x)$ is the hidden state entering layer i, W^i_1 is the input projection of the feed-forward network, ϕ is the activation function, and $\mathbf{1}$ is the indicator function.

The vector of probabilities across all languages $k = 1, \ldots, L$ is normalized:

$$\tilde{p}^k_{i,j} = \frac{p^k_{i,j}}{\sum_{\ell=1}^{L} p^\ell_{i,j}}.$$

The LAPE score is then given by Shannon entropy (Shannon and Weaver, 1998):

$$\text{LAPE}_{i,j} = -\sum_{k=1}^{L} \tilde{p}^k_{i,j} \log \tilde{p}^k_{i,j}.$$

We select neurons with the lowest $K\%$ of LAPE scores, indicating truly language-specific behavior due to their low-entropy activations, subject to two conditions:

1. $\max_k p^k_{i,j} \geq \tau_{\text{activity}}$: a neuron must reach at least τ_{activity} in some language; otherwise, uniformly weak neurons are discarded.
2. $\exists k : p^k_{i,j} \geq \tau_{\text{selectivity}}$: a neuron is kept only if its activation for at least one language exceeds a strong threshold $\tau_{\text{selectivity}}$, ensuring clear language preference.

The resulting subset forms a language-sensitive subnetwork $\mathcal{S}^k$.

3.2 Subnetwork Fine-Tuning

Given the identified subnetwork $\mathcal{S}^k$, we restrict parameter updates to the weights associated with these neurons. Let θ denote the full set of model parameters, and $\theta_{\mathcal{S}^k} \subset \theta$ the parameters belonging to the selected neurons.

Fine-tuning is then performed by minimizing the loss $\mathcal{L}$ on target-language data $\mathcal{D}_{\text{target}}$:

$$\min_{\theta_{\mathcal{S}^k}} \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{target}}} \mathcal{L}\big(f(x;\theta), y\big),$$

while keeping all parameters $\theta \setminus \theta_{\mathcal{S}^k}$ fixed.

This optimization aims at adapting only the language-relevant subnetwork while preserving the general-purpose behavior of the model.

4 Experimental Setup

Below, we describe the details of our training setup, the evaluation tasks chosen to assess model performance, and the baselines against which we compare.

4.1 Training Setup

We conduct experiments with LLAMA-3.1-8B-BASE and MISTRAL-NEMO-12B-BASE (Grattafiori et al., 2024), using GLOTCC (Kargaran et al., 2024) as the source of target-language training data. By default, each language subnetwork is trained on up to 100M tokens in the target language, with additional experiments varying the training size between 50M and 200M tokens for a subset of languages. We focus on 12 underrepresented languages (Joshi et al., 2020), reported in Table 1.

Subnetworks are identified for each language using 100MB of GLOTCC data, with thresholds $K = 5\%$ and activity thresholds of 0.95, following Gurgurov et al. (2025a). Each subnetwork is fine-tuned for one epoch with batch size 2, a learning rate of 1×10^{-4}, and weight decay of 0.01. We use AdamW (Loshchilov and Hutter, 2019) with gradient clipping (Mikolov et al., 2012) for stability. Validation is performed every 500 steps, and the best checkpoint is selected by validation loss on the corresponding FLORES-200 subset (Costa-Jussà et al., 2022).

4.2 Evaluation Tasks

We evaluate our method along two dimensions:

Target-language performance. We assess translation quality using the "valtest" splits of FLORES-200 (Costa-Jussà et al., 2022) and the "test" splits of OPUS-100 (Tiedemann, 2012), reporting results in SENTENCE BLEU (Post, 2018). For comprehension and classification, we evaluate on BELEBELE (Bandarkar et al., 2024) and SIB-200 (Adelani et al., 2023), using accuracy as the primary metric.

Lang.	# Neurons	# Trainable Param-s	%
Afrikaans	2581	31.7M	0.39
Welsh	6020	74.0M	0.92
Estonian	3204	39.4M	0.49
Icelandic	5033	61.8M	0.77
Georgian	1150	14.1M	0.18
Lithuanian	2978	36.6M	0.46
Latvian	3212	39.5M	0.49
Macedonian	1743	21.4M	0.27
Maltese	3015	37.0M	0.46
Nepali	4625	56.8M	0.71
Slovak	1940	23.8M	0.30
Slovenian	2046	25.1M	0.31

Table 1: Neuron, parameter counts for different language subnetworks, and percentages of model being trained within LLAMA-3.1-8B. The statistics for MISTRAL-NEMO-12B are available in Appendix A.

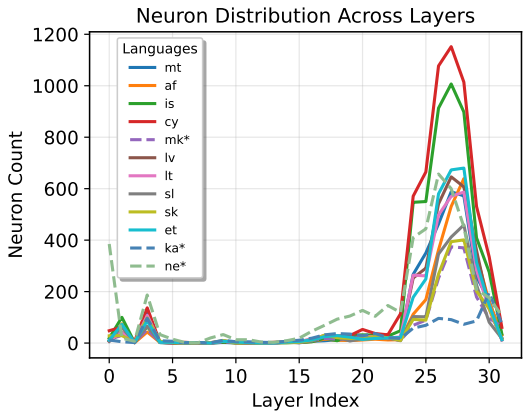


Figure 2: Neuron distributions over the layers of LLAMA-3.1-8B for all 12 languages. The distributions for MISTRAL-NEMO-12B are in Appendix A.

General capabilities. To measure the impact of adaptation on overall model performance, we report accuracy on several standard benchmarks before and after fine-tuning: MMLU (Hendrycks et al., 2021), HELLASWAG (Zellers et al., 2019), PIQA (Bisk et al., 2019), WINOGRANDE (Sakaguchi et al., 2019), and ARC (Easy and Challenge) (Clark et al., 2018).

For all tasks, we use deterministic generation with zero temperature to eliminate variance introduced by stochastic decoding. The prompts and hyperparameters specific to each evaluation are listed in Appendix C.2.

4.3 Baselines

We compare our sparse fine-tuning approach against four baselines:

- (i) full model fine-tuning,

Metric	Base Model	Full FT	FFN-Only FT	LoRA	Random Neurons	Target Neurons
Target Language Performance						
FLORES TG→EN	28.82	11.44	25.06	28.00	30.95	31.37
FLORES EN→TG	11.62	9.29	13.70	15.60	15.84	16.57
BELEBELE	66.52	21.22	38.35	56.61	59.82	66.84
SIB200	40.89	11.23	21.65	33.37	34.07	44.94
OPUS TG→EN	19.59	10.67	20.01	22.68	24.52	24.34
OPUS EN→TG	13.32	7.71	13.04	16.33	16.93	17.23
Avg.	29.56	11.88	21.81	28.41	29.89	33.02
General Capabilities						
ARC CHALLENGE	70.90	20.40	32.94	59.45	63.69	70.62
ARC EASY	85.79	23.86	46.77	77.12	80.83	84.72
HELLASWAG	40.38	13.49	22.56	35.02	34.90	40.19
PIQA	71.87	34.84	48.19	64.68	69.69	72.55
WINOGRANDE	51.40	38.15	48.34	50.76	51.24	52.41
MMLU	58.06	22.34	32.59	49.59	52.45	57.52
Avg.	63.07	25.51	38.57	56.10	58.80	63.00

Table 2: Average results for enhancing language subnetworks within LLAMA-3.1-8B, evaluated on both monolingual capabilities and general performance. All models were trained on 100M tokens in the target language. The downstream results for the individual languages are in Appendix C.

- (ii) fine-tuning only the FFN components,
- (iii) fine-tuning a random subset of FFN neurons of the same size as the LAPE-selected subnetwork, and
- (iv) fine-tuning a LoRA (Hu et al., 2021) applied to GATE_PROJ and GATE_UP, configured to match the average parameter count used for most languages, with rank set to 62 and alpha set to 32, resulting in 57M trainable parameters (0.7%).

For full-model and FFN-only fine-tuning, we set the learning rate to 1×10^{-5} to account for the larger number of trainable parameters, while keeping all other hyperparameters identical to those of our method. For LoRA and the random-subset baseline, we fine-tune with exactly the same hyperparameters as our method.

5 Experimental Results

In this section, we present our experimental findings. We first report the results of language-specific subnetwork identification, analyzing the extent to which distinct subnetworks emerge across languages. We then evaluate the effectiveness of subnetwork fine-tuning, highlighting its impact on performance relative to baselines.

5.1 Subnetwork Identification

Table 1 summarizes the number of language-specific neurons, the corresponding trainable parameters, and the fraction of the model trained for each language-specific subnetwork within LLAMA-3.1-8B. The table shows substantial variation across languages: for instance, Welsh and Nepali subnetworks are relatively large, covering over 0.7% of the model parameters, whereas Georgian and Macedonian subnetworks are much smaller, accounting for less than 0.3%. This variation might reflect differences in language complexity and the amount of language-specific signal captured by GLOTCC. The subnetwork sizes for MISTRAL-NEMO-12B are in Appendix A.

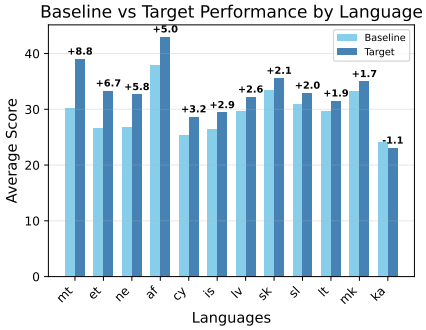
Figures 2 and 10 depict the distribution of the subnetwork neurons across layers for all 12 languages. **We observe that most language-specific neurons are concentrated in the upper layers, and related languages exhibit moderate overlap in their neuron sets** (such as Slovak and Slovenian, or Latvian and Lithuanian–see Appendix A for details). These patterns align with prior findings from Gurgurov et al. (2025a). Overall, the weights associated with these subnetworks account for roughly 0.2% to 1% of the total model parameters, highlighting the efficiency of this approach.

5.2 Subnetwork Fine-tuning

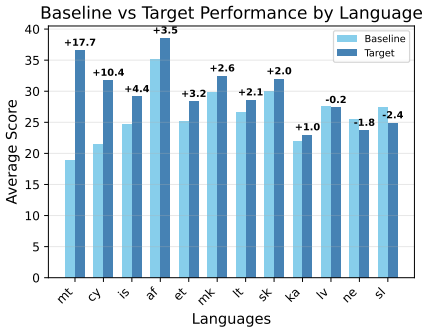
5.2.1 Average Task Performance

Table 2 reports average downstream task performance across the 12 target languages for different fine-tuning configurations, along with results on general-purpose benchmarks for LLAMA-3.1-8B. **Targeted sparse fine-tuning of language-specific subnetworks consistently yields the largest gains**

on target-language tasks, outperforming full-model, FFN-only, LoRA, and random-subset baselines, **while maintaining general capabilities as measured by MMLU and commonsense tasks**. For example, on FLORES, targeted subnetworks achieve average improvements of up to 5 points, while preserving or slightly improving performance on general benchmarks such as MMLU and WINOGRANDE. This contrasts with the findings of Mondal et al. (2025), who also employed LAPE for neuron identification but restricted their analysis to only three relatively high-resource languages and adapted LoRA modules on top of FFNs rather than directly fine-tuning neurons inside the model, leading to more limited improvements in cross-lingual transfer. By contrast, full-model and FFN-only fine-tuning not only fail to improve target-language performance but also substantially degrade general capabilities. The random neuron-subset and LoRA baselines show intermediate behavior, with modest target-language gains but noticeable reductions in overall performance, underscoring the advantage of precisely selecting language-specific subnetworks for sparse adaptation. Results for individual languages for both LLAMA-3.1-8B and MISTRAL-NEMO-12B are provided in Appendix C.



(a) LLAMA-3.1-8B



(b) MISTRAL-NEMO-12B

Figure 3: Average improvement scores across all tasks for the target fine-tuning for each of the 12 languages.

5.2.2 Language-Wise Variation

As shown in Figure 3, the magnitude of improvement varies considerably across languages. Languages such as Afrikaans, Maltese, and Welsh benefit the most from targeted fine-tuning, with average gains exceeding 5 points across all tasks, whereas languages such as Latvian and Macedonian see only marginal improvements. A likely explanation is that higher-resource languages are already well represented in pre-training, limiting the marginal utility of additional fine-tuning data. In some cases, the model may even have encountered overlapping material during pre-training, further reducing the benefit of adaptation (Lee et al., 2021). By contrast, lower-resource languages, on which the base model underperforms, gain substantially more from precise subnetwork adaptation. The only decreases we observe are for Georgian with LLAMA and Slovenian and Nepali with MISTRAL. Since both models show improvements on these languages overall (i.e., MISTRAL for Georgian and LLAMA for Slovenian and Nepali), and improvements are also evident on individual tasks (here we report averages), **the gains for these languages are clearly possible but may not materialize due to the relatively small number of subnetwork parameters allocated to them**. Finally, we note that some languages–particularly highly morphological ones or those with different scripts–do improve under our methods, but less than languages such as Afrikaans and Maltese. This may indicate that they require larger or smaller subnetworks (i.e., higher or lower K%). We leave the systematic exploration of optimal subnetwork sizes for different languages to future work.

5.2.3 Training Dynamics

Figure 4 illustrates the training dynamics for Afrikaans. **Sparse fine-tuning of target-language subnetworks generally converges faster and reaches lower validation loss than random-subset, FFN-only, full-model, or LoRA fine-tuning**. While full-model and FFN-only fine-tuning occasionally achieve lower loss on the target language (see Appendix B) due to the greater capacity of the updated networks, this is often accompanied by pronounced overfitting and a substantial drop in general capabilities.

5.2.4 Effect of Training Data Size

Figure 5 further examines the effect of varying training data size (50M, 100M, 150M, and 200M to-

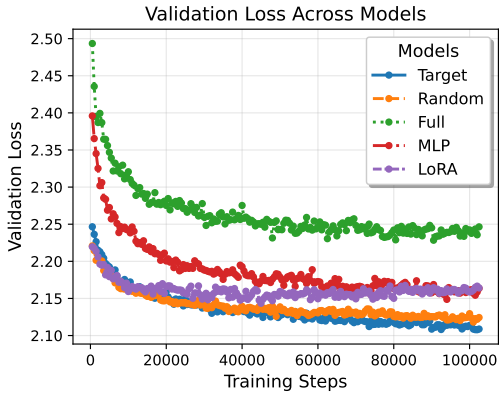


Figure 4: Validation loss curves for different fine-tuning configurations of LLAMA-3.1-8B on Afrikaans (all trained on 100M tokens). The plots for the other tested languages are in Appendix B.

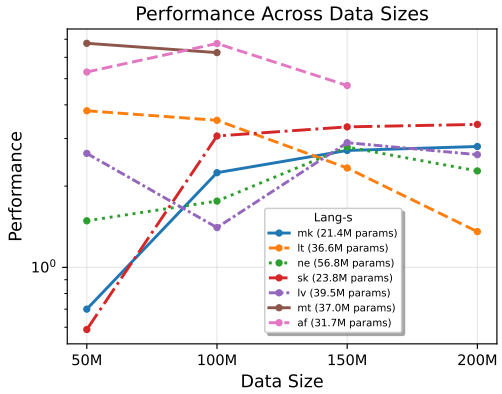


Figure 5: Overall improvement scores against various data sizes for a subset of 5 languages when sparsely fine-tuning LLAMA-3.1-8B.

kens) for a subset of seven languages (Macedonian, Latvian, Nepali, Slovak, Afrikaans, Maltese, and Lithuanian) using LLAMA-3.1-8B. **We observe relatively consistent improvement trends with increasing data across most tested languages, irrespective of subnetwork size, suggesting that additional data could further enhance performance.** This differs from the findings of Zhao et al. (2024), who fine-tune an ablation-identified subset of neurons using between 100 and 800 documents–likely comprising fewer tokens–and report saturation after 400 documents. However, for Lithuanian we similarly observe a consistent decrease with larger training token sizes, which may indicate that the additional data is not helpful (Lee et al., 2021) or that the capacity of the identified subnetwork is insufficient to effectively utilize it. Due to computational constraints and the limited availability of data for some languages (e.g., Maltese and Afrikaans, with at most 100–150M tokens), we leave the exploration of larger-scale training to future work.

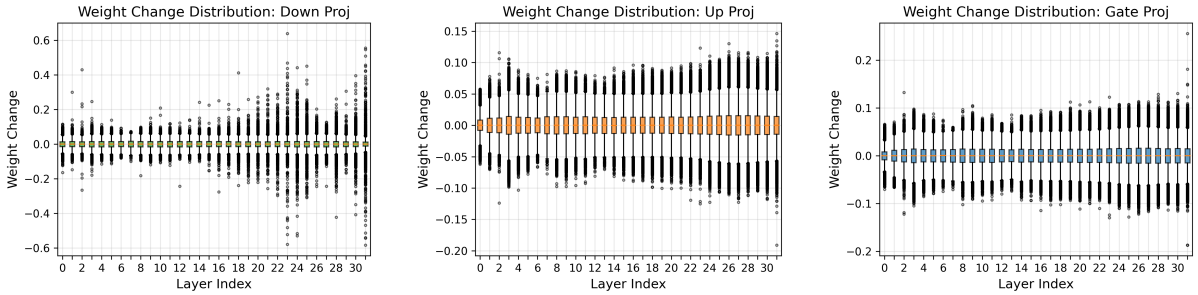
6 Effects of Sparse Fine-Tuning

In this section, we present additional analyses focusing on two aspects: the changes in model weights induced by our sparse fine-tuning approach and the evolution of representation similarities before and after fine-tuning.

6.1 Weight Change Analysis

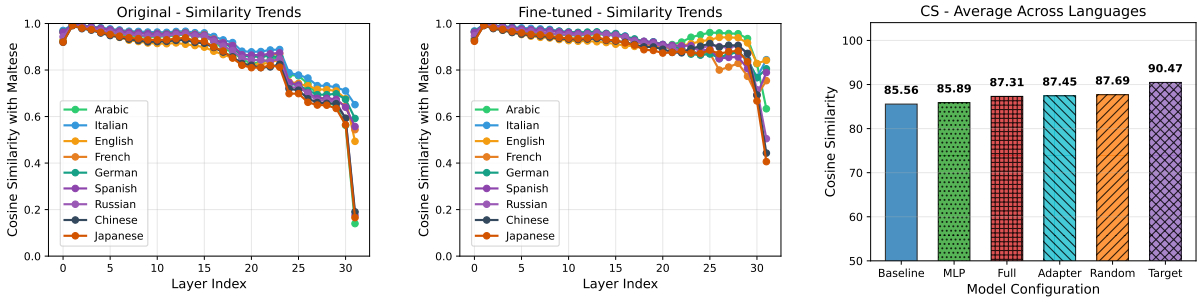
Our analysis of the weight changes during fine-tuning reveals a notable asymmetry between the up/gate projections and the down projection. As shown in Figure 6, the distribution of weight changes for the `gate_proj` and `up_proj` remains relatively consistent across all layers. The small size of the box plots indicates that most weight changes are exceptionally small and tightly clustered around zero. This suggests that the fine-tuning process is highly precise, making only minor adjustments to most weights. In contrast, the weight changes in the `down_proj` show an increase in magnitude across all layers, with especially large shifts in the later layers. This suggests that individual weights undergo more substantial alterations there, aligning with the greater prominence of language-specific neurons in those parts of the model. **This finding suggests a functional distinction in how different weight components of the FFN layers contribute to language fine-tuning.**

This asymmetry can be explained by the distinct roles of these projections within the SwiGLU-based architecture (Shazeer, 2020). The `up_proj` and `gate_proj` work together to create a gated feature space, refining the model's ability to select and activate an appropriate subset of existing features for the new language. Updating these weights primarily involves slightly re-weighting established feature-extraction capabilities rather than fundamentally altering the feature space itself. Conversely, the `down_proj` is responsible for an important integrative step, taking the activated features and synthesizing a new representation. Our observation that its weight changes are substantially larger in later layers indicates this is where the most significant language-specific re-wiring occurs, as these deeper layers learn to reorganize high-level concepts to adapt to the new language.



(a) Down Projection (b) Up Projection (c) Gate Projection

Figure 6: Per-layer changes in weight values for three considered weight types for Nepali in LLAMA-3.1-8B. The plots for the other 11 languages follow the same patterns and will be made available in our GitHub repository.



(a) Original Llama-3.1-8B (b) Fine-tuned (100M tokens) (c) Average cosine similarity

Figure 7: Cosine similarity for post-FFN hidden state activations of the original and sparsely fine-tuned LLAMA-3.1-8B between Maltese (left, middle) and average across target languages (right). Full results in Appendix D.

6.2 Representation Analysis

Targeted neuron fine-tuning not only improves performance in the target language but also enhances cross-lingual alignment. To investigate this, we examine whether parallel sentences become closer in embedding space after adaptation. Figure 7 presents cosine similarities between post-layer hidden state activations of FLORES-200 sentences in Maltese and several high-resource languages (Arabic, Italian, English, French, German, Spanish, Russian, Chinese, Japanese) across layers. We observe a consistent increase in similarity for all languages after adaptation, especially in the later layers, as the early layers already exhibit relatively high similarity.

The average results across all languages in Figure 7c show that targeted sparse fine-tuning achieves the strongest cross-lingual alignment, with an average cosine similarity of 90.47 between parallel sentences. This represents a substantial improvement over all considered fine-tuning baselines. These findings indicate that selectively updating neurons identified as important for the target language produces more aligned representations across languages, suggesting that sparse adaptation can improve cross-lingual generalization without compromising overall model structure. Consistent with these results, Sundar et al. (2025) also report improved cross-lingual alignment when manipulating language-relevant activations.

7 Conclusion

We presented a framework for enhancing the capabilities of LLMs in underrepresented languages through targeted fine-tuning of language-specific subnetworks. By identifying neurons with strong language selectivity using LAPE and adapting only their associated weights, our approach achieves substantial gains in low- and mid-resource languages while preserving general-purpose performance and updating less than 1% of parameters. Beyond improved accuracy, we showed that targeted adaptation enhances cross-lingual alignment and yields favorable training dynamics. Our release of language-specific neuron identifications for over 100 languages provides a scalable and cost-effective resource for extending state-of-the-art models to underrepresented languages, helping narrow the gap in global language coverage.

Limitations

While our framework improves performance across diverse languages, it relies on accurate identification of language-specific neurons, which can be sensitive to dataset quality and language coverage. Our experiments primarily focus on mid- and low-resource languages, leaving open the question of how the method scales to extremely low-data settings or to languages that are highly typologically distant. Furthermore, we do not explore varying the capacity of identified language-specific subnetworks through the parameter K%. Understanding how many parameters are needed to effectively capture language-specific knowledge, and how much data is required to leverage them, remains an important direction for future work. Such analyses, conducted at the per-language or per-language-family level, have significant potential.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2023. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. *Preprint*, arXiv:2309.07445.

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. *Preprint*, arXiv:1911.11641.

Ethan C. Chau, Lucy H. Lin, and Noah A. Smith. 2020. Parsing with multilingual bert, a small corpus, and a small treebank. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, page 1324–1334. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2021. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*.

Abteen Ebrahimi and Katharina Kann. 2021. How to adapt your pretrained multilingual model to 1600 languages. *Preprint*, arXiv:2106.02124.

Fahim Faisal and Antonios Anastasopoulos. 2022. Phylogeny-inspired adaptation of multilingual models to new languages. *Preprint*, arXiv:2205.09634.

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

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. *Preprint*, arXiv:2012.14913.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daniil Gurgurov, Katharina Trinley, Yusser Al Ghussin, Tanja Baeumel, Josef van Genabith, and Simon Ostermann. 2025a. Language arithmetics: Towards systematic language neuron identification and manipulation. *Preprint*, arXiv:2507.22608.

Daniil Gurgurov, Ivan Vykopal, Josef van Genabith, and Simon Ostermann. 2025b. Small models, big impact: Efficient corpus and graph-based adaptation of small multilingual language models for low-resource languages. *Preprint*, arXiv:2502.10140.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. *Preprint*, arXiv:1904.02817.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language

understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019a. Parameter-efficient transfer learning for nlp. *Preprint*, arXiv:1902.00751.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019b. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *Preprint*, arXiv:2305.07004.

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.

Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. 2024. GlotCC: An open broad-coverage commoncrawl corpus and pipeline for minority languages. *Advances in Neural Information Processing Systems*.

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. *arXiv preprint arXiv:2404.02431*.

Jaeseong Lee, Seung-won Hwang, and Taesup Kim. 2022. FAD-X: Fusing adapters for cross-lingual transfer to low-resource languages. 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 2: Short Papers)*, pages 57–64, Online only. Association for Computational Linguistics.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Tomáš Mikolov and 1 others. 2012. Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April*, 80(26).

Soumen Kumar Mondal, Sayambhu Sen, Abhishek Singhania, and Preethi Jyothi. 2025. Language-specific neurons do not facilitate cross-lingual transfer. In *The Sixth Workshop on Insights from Negative Results in NLP*, pages 46–62, Albuquerque, New Mexico. Association for Computational Linguistics.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. *Preprint*, arXiv:1808.04189.

Marinela Parović, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2022. BAD-X: Bilingual adapters improve zero-shot cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1791–1799, Seattle, United States. Association for Computational Linguistics.

Marinela Parović, Alan Ansell, Ivan Vulić, and Anna Korhonen. 2023. Cross-lingual transfer with target language-ready task adapters. *Preprint*, arXiv:2306.02767.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Nathaniel R. Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. Chatgpt mt: Competitive for high- (but not low-) resource languages. *Preprint*, arXiv:2309.07423.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *Preprint*, arXiv:1907.10641.

Claude E Shannon and Warren Weaver. 1998. *The mathematical theory of communication*. University of Illinois press.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Anirudh Sundar, Sinead Williamson, Katherine Metcalf, Barry-John Theobald, Skyler Seto, and Masha Fedzechkina. 2025. Steering into new embedding spaces: Analyzing cross-lingual alignment induced by model interventions in multilingual language models. *Preprint*, arXiv:2502.15639.

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. *arXiv preprint arXiv:2402.16438*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ivan Vykopal, Simon Ostermann, and Marian Simko. 2025. Soft language prompts for language transfer. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10294–10313, Albuquerque, New Mexico. Association for Computational Linguistics.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022. Finding skill neurons in pre-trained transformer-based language models. *arXiv preprint arXiv:2211.07349*.

Wanying Xie, Yang Feng, Shuhao Gu, and Dong Yu. 2021. Importance-based neuron allocation for multilingual neural machine translation. *arXiv preprint arXiv:2107.06569*.

Haoyun Xu, Runzhe Zhan, Derek F Wong, and Lidia S Chao. 2024. Let's focus on neuron: Neuron-level supervised fine-tuning for large language model. *arXiv preprint arXiv:2403.11621*.

Zheng-Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Indra Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. Bloom+1: Adding language support to bloom for zero-shot prompting. *Preprint*, arXiv:2212.09535.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *Preprint*, arXiv:1905.07830.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Xiucheng Li, Yang Xiang, and Min Zhang. 2025. Exploring translation mechanism of large language models. *arXiv preprint arXiv:2502.11806*.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *Preprint*, arXiv:2402.18815.

Shaolin Zhu, Leiyu Pan, Bo Li, and Deyi Xiong. 2024. Landermt: Detecting and routing language-aware neurons for selectively finetuning llms to machine translation. *arXiv preprint arXiv:2409.19523*.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. Udapter: Language adaptation for truly universal dependency parsing. *Preprint*, arXiv:2004.14327.

Appendix

A Language Neuron Overlap and Individual Distributions

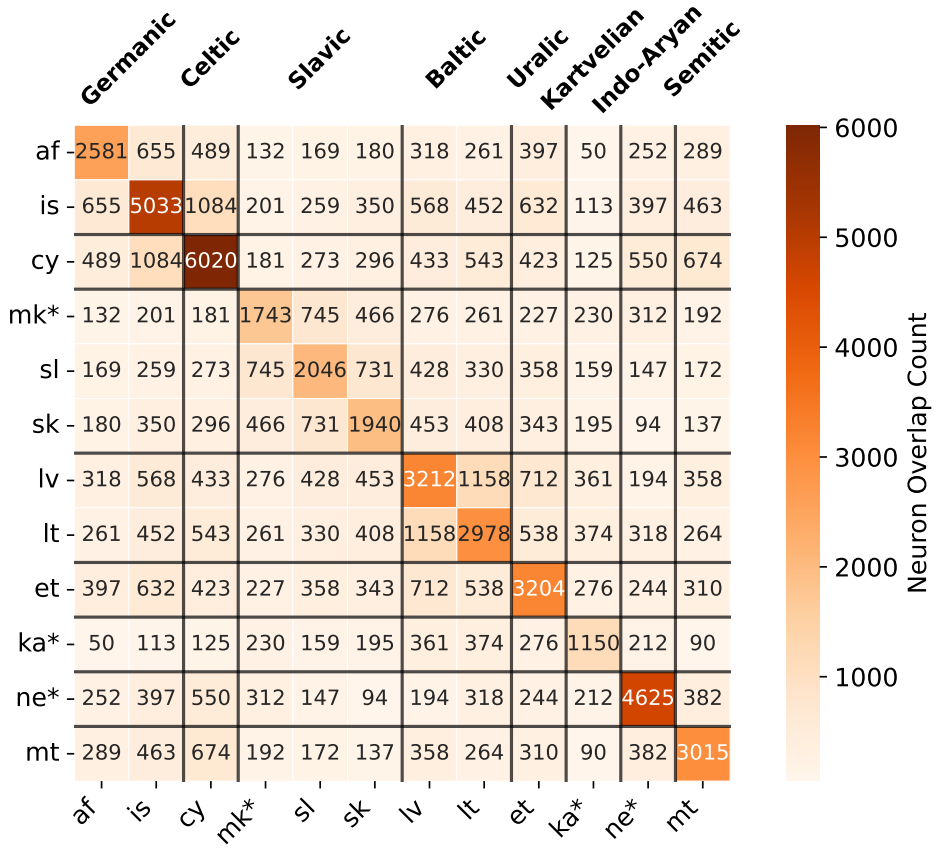


Figure 8: Language neuron overlap for all 12 languages for LLAMA-3.1-8B.

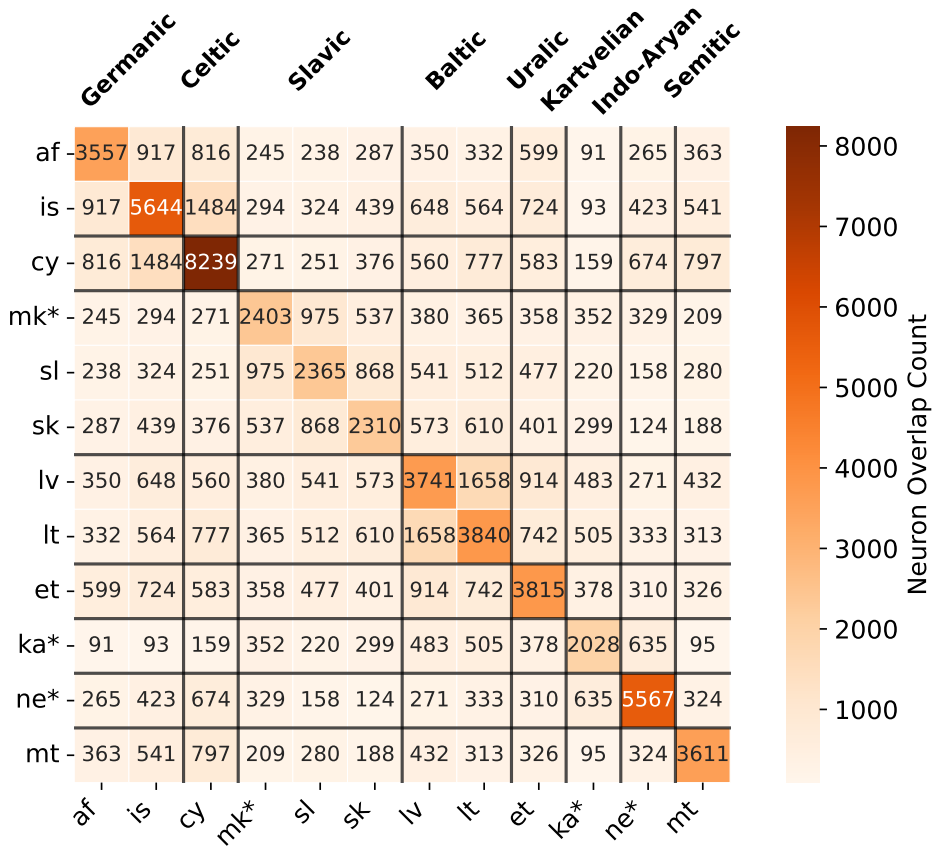


Figure 9: Language neuron overlap for all 12 languages for MISTRAL-NEMO-12B.

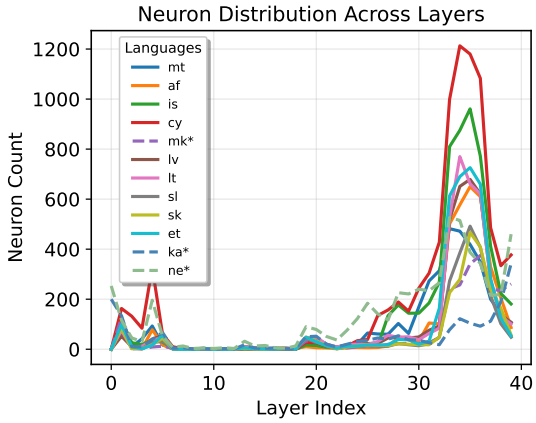


Figure 10: Neuron distributions over the layers of MISTRAL-NEMO-12B for all 12 languages.

Lang.	# Neurons	# Trainable Param-s	%
Afrikaans	3557	54.6M	0.45
Welsh	8239	126.6M	1.03
Estonian	3815	58.6M	0.48
Icelandic	5644	86.7M	0.71
Georgian	2028	31.1M	0.25
Lithuanian	3840	59.0M	0.48
Latvian	3741	57.5M	0.47
Macedonian	2403	36.9M	0.30
Maltese	3611	55.5M	0.45
Nepali	5567	85.5M	0.70
Slovak	2310	35.5M	0.29
Slovenian	2365	36.3M	0.30

Figure 11: Neuron and parameter counts for different language subnetworks, and percentages of the model being trained within MISTRAL-NEMO-12B.

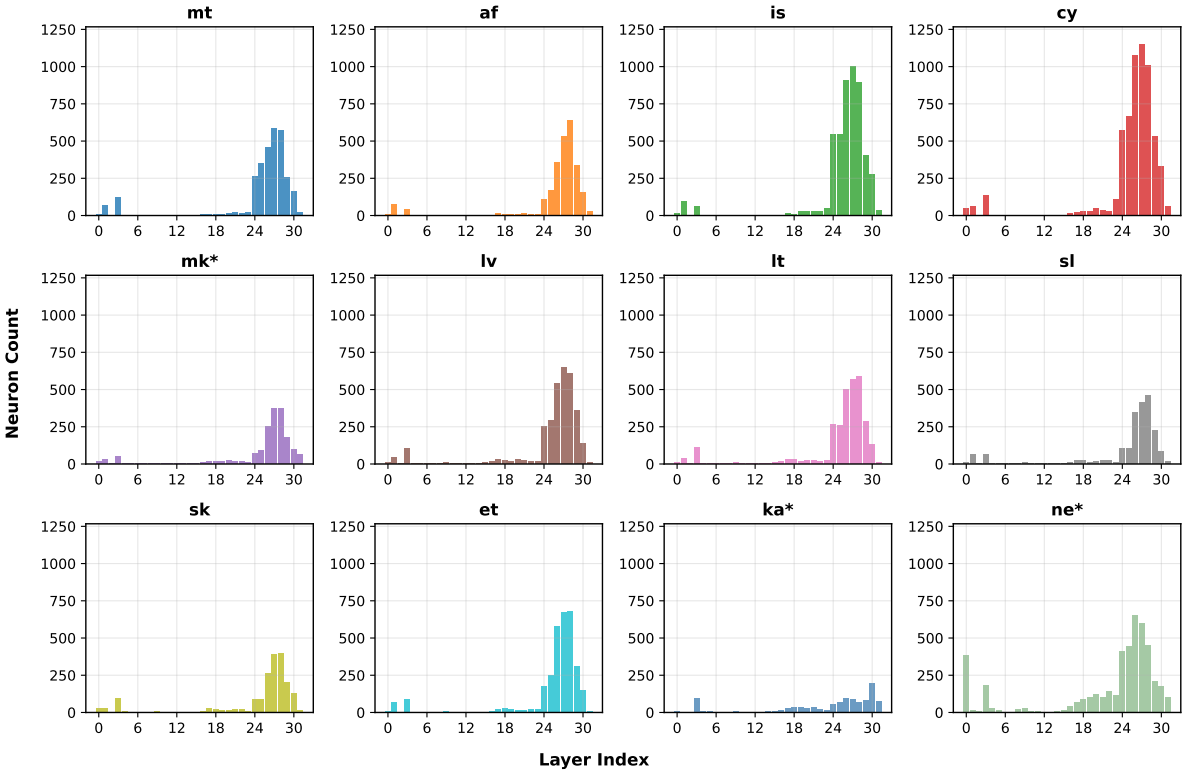


Figure 12: Language neuron distributions for all 12 languages for LLAMA-3.1-8B-12B.

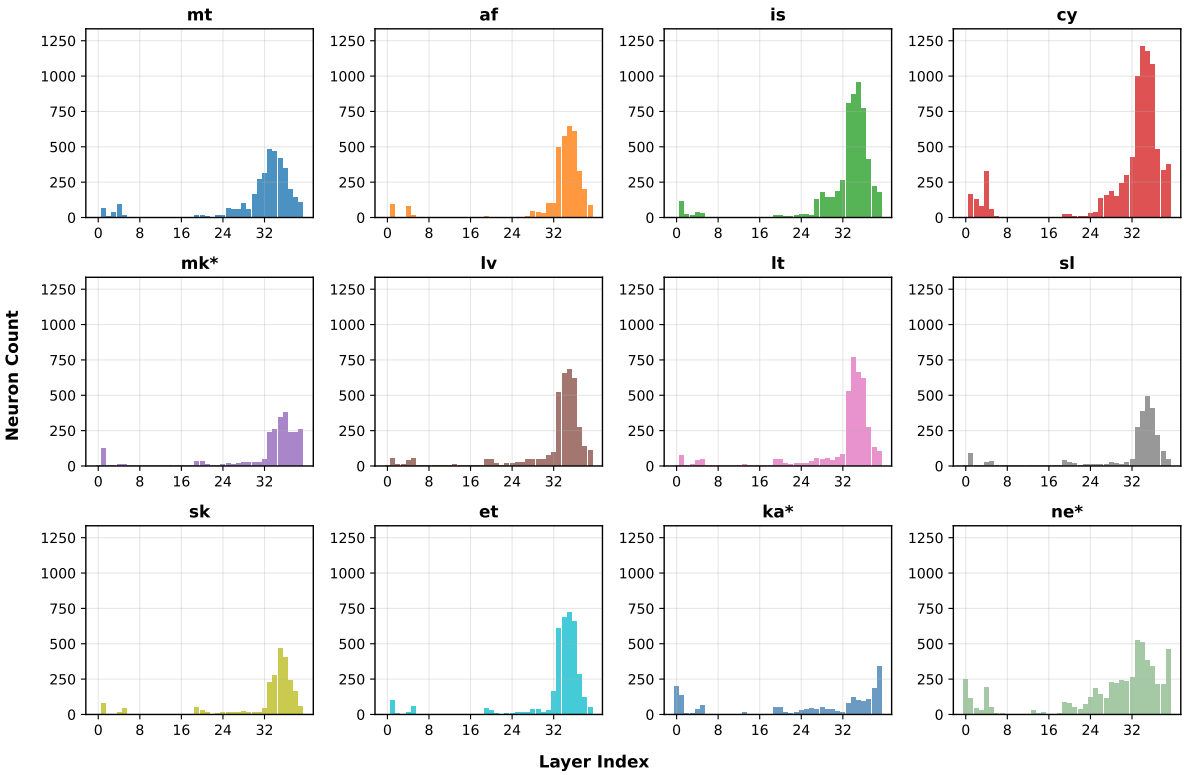
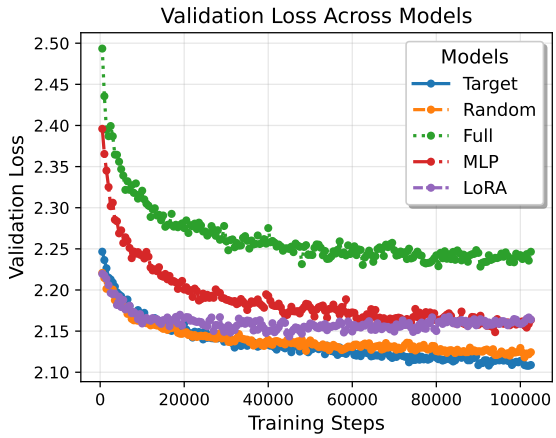
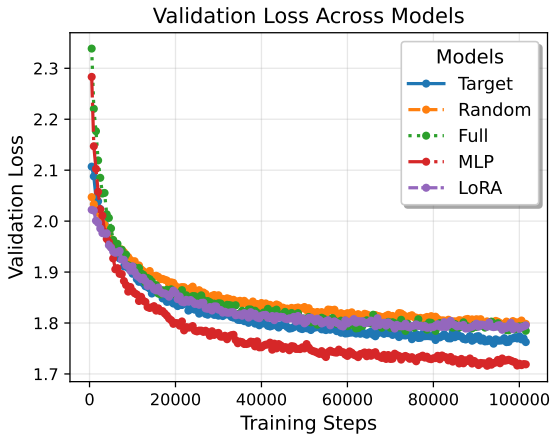


Figure 13: Language neuron distributions for all 12 languages for MISTRAL-NEMO-12B.

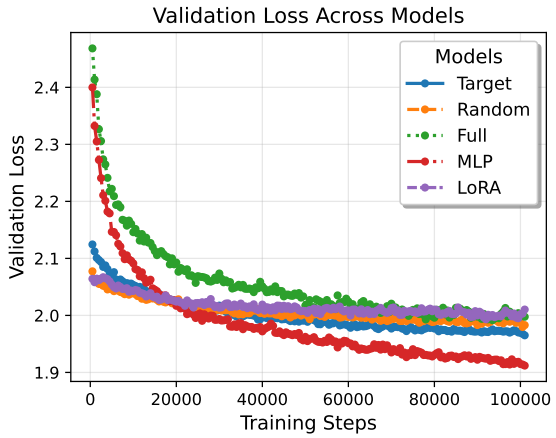
B Validation Losses



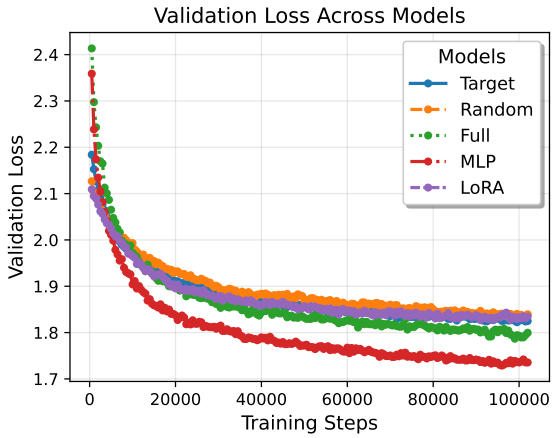
(a) Afrikaans validation loss curves for LLAMA-3.1-8B.



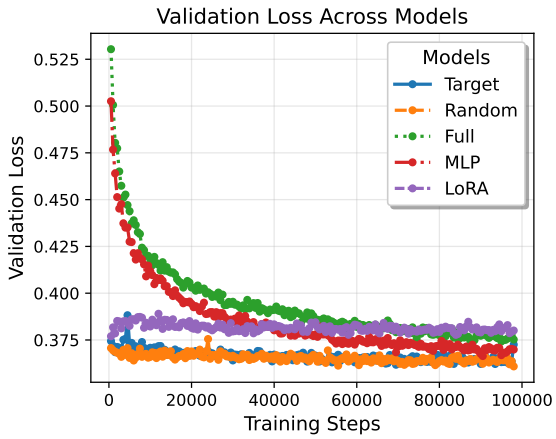
(b) Welsh validation loss curves for LLAMA-3.1-8B.



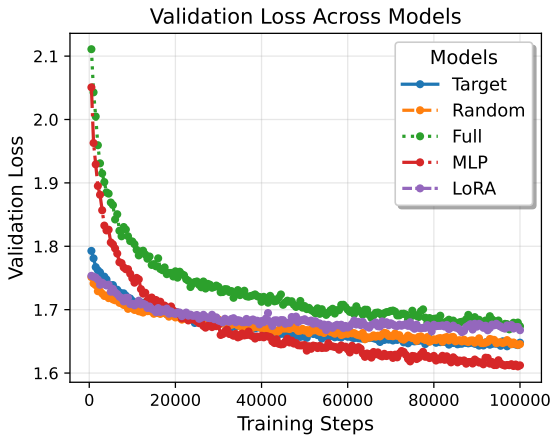
(c) Estonian validation loss curves for LLAMA-3.1-8B.



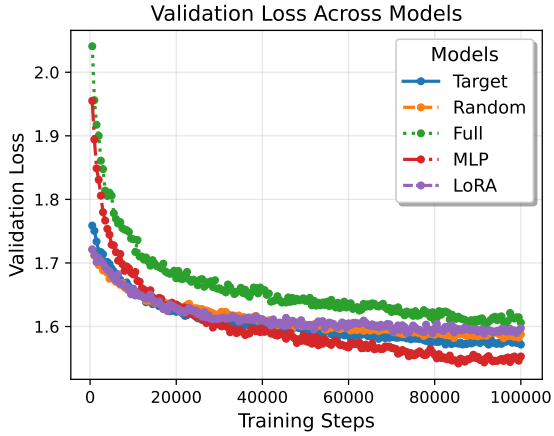
(d) Icelandic validation loss curves for LLAMA-3.1-8B.



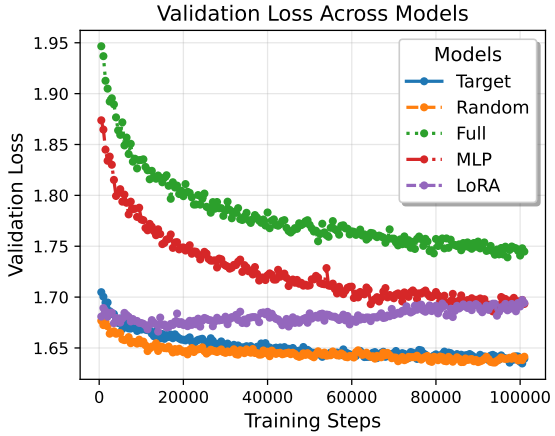
(e) Georgian validation loss curves for LLAMA-3.1-8B.



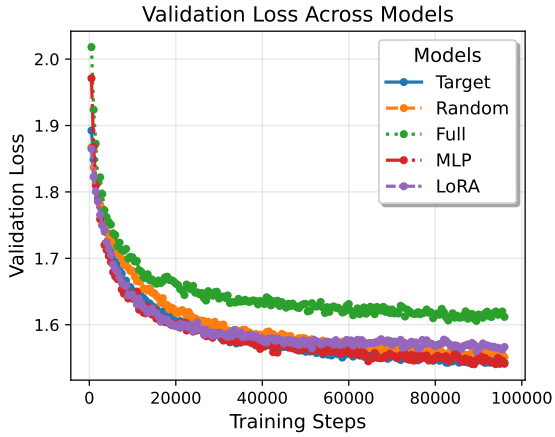
(f) Lithuanian validation loss curves for LLAMA-3.1-8B.



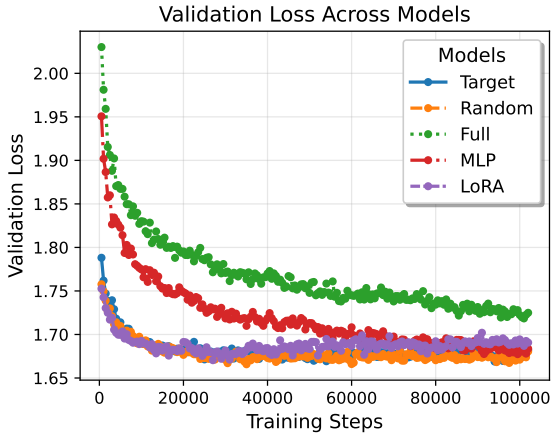
(a) Latvian validation loss curves for LLAMA-3.1-8B.



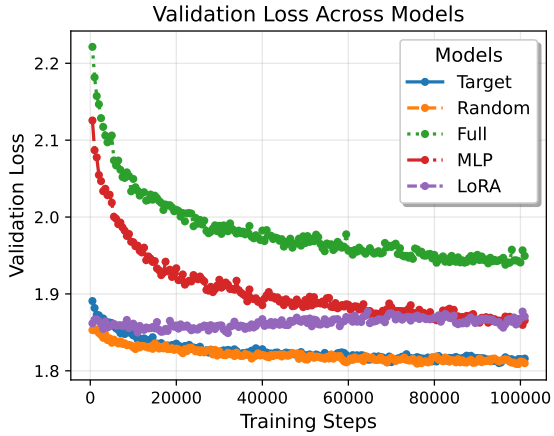
(b) Macedonian validation loss curves for LLAMA-3.1-8B.



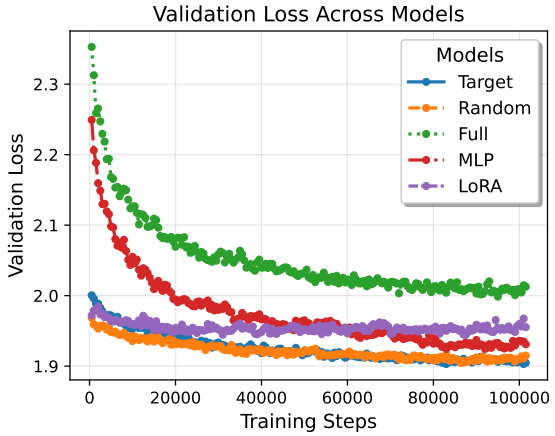
(c) Maltese validation loss curves for LLAMA-3.1-8B.



(d) Nepali validation loss curves for LLAMA-3.1-8B.



(e) Slovak validation loss curves for LLAMA-3.1-8B.



(f) Slovenian validation loss curves for LLAMA-3.1-8B.

C Downstream Results

Metric	Base	Full	MLP	LoRA	Random	Target
Monolingual Capabilities						
FLORES afr Latn→en	45.88	17.35	45.58	48.16	49.34	49.43
FLORES en→afr Latn	28.19	20.91	31.96	30.97	32.52	32.11
BELEBELE afr Latn	72.11	31.22	47.67	65.00	68.89	73.11
SIB200 afr Latn	35.29	27.45	30.39	31.37	45.59	44.12
OPUS afr Latn→en	25.49	6.62	25.96	34.13	34.18	34.66
OPUS en→afr Latn	21.06	7.25	14.83	20.76	21.63	24.50
Avg.	38.01	18.47	32.73	38.40	42.03	42.99
General Performance						
arc challenge	70.90	25.75	35.45	64.55	62.21	71.24
arc easy	85.79	26.14	53.33	79.12	81.75	85.26
hellaswag	40.38	14.68	26.13	35.22	36.04	38.72
piqa	71.87	39.17	50.00	73.07	72.03	71.71
winogrande	51.38	49.17	50.43	50.91	47.04	51.46
mmlu	58.07	27.63	37.30	51.60	54.74	56.43
Avg.	63.07	30.42	42.11	59.08	58.97	62.47

Table 3: Evaluation results comparing different fine-tuning configurations across various tasks for Afrikaans for LLAMA-3.1-8B. Best performing configurations are highlighted in dark green, second-best in light green.

Metric	Base	Full	MLP	LoRA	Random	Target
Monolingual Capabilities						
FLORES cym Latn→en	39.81	32.37	40.48	39.82	42.40	43.90
FLORES en→cym Latn	14.78	23.90	29.11	23.46	21.77	25.57
SIB200 cym Latn	37.25	1.47	25.49	49.51	20.10	35.78
OPUS cym Latn→en	18.16	23.03	26.13	26.82	29.74	23.65
OPUS en→cym Latn	17.16	8.96	13.80	17.05	19.04	14.36
Avg.	25.43	17.95	27.00	31.33	26.61	28.65
General Performance						
arc challenge	70.90	16.05	22.74	56.52	54.52	68.23
arc easy	85.79	16.84	30.70	74.56	76.14	84.04
hellaswag	40.38	12.31	25.27	29.93	30.88	40.79
piqa	71.87	3.97	49.46	58.98	66.97	68.39
winogrande	51.62	20.28	49.57	49.17	51.46	50.51
mmlu	58.07	22.27	25.67	45.79	48.14	56.69
Avg.	63.10	15.29	33.90	52.49	54.68	61.44

Table 4: Evaluation results comparing different fine-tuning configurations across various tasks for Welsh for LLAMA-3.1-8B. Best performing configurations are highlighted in dark green, second-best in light green.

Metric	Base	Full	MLP	LoRA	Random	Target
Monolingual Capabilities						
FLORES ekk Latn→en	25.18	9.88	20.61	25.09	26.72	27.58
FLORES en→ekk Latn	2.71	2.90	4.28	11.15	12.49	10.82
BELEBELE ekk Latn	66.67	15.00	42.00	56.22	58.11	65.00
SIB200 ekk Latn	40.20	12.25	30.88	36.76	37.75	55.88
OPUS ekk Latn→en	16.27	8.37	19.50	21.66	21.78	23.32
OPUS en→ekk Latn	8.40	3.19	12.00	16.01	17.08	17.19
Avg.	26.57	8.60	21.55	27.81	28.99	33.30
General Performance						
arc challenge	70.90	22.07	36.45	58.86	63.55	72.58
arc easy	85.79	21.40	55.09	77.89	81.58	84.74
hellaswag	40.38	13.92	22.14	36.11	37.06	41.82
piqa	71.87	40.10	54.35	59.63	72.96	75.24
winogrande	51.38	41.83	50.20	51.14	52.01	53.83
mmlu	58.07	20.64	33.90	48.66	52.84	57.35
Avg.	63.07	26.66	42.02	55.38	60.00	64.26

Table 5: Evaluation results comparing different fine-tuning configurations across various tasks for Estonian for LLAMA-3.1-8B. Best performing configurations are highlighted in dark green, second-best in light green.

Metric	Base	Full	MLP	LoRA	Random	Target
Monolingual Capabilities						
FLORES en→isl Latn	9.43	7.34	10.09	12.47	11.57	13.71
FLORES isl Latn→en	22.63	15.36	23.00	23.16	24.46	24.82
BELEBELE isl Latn	62.67	21.44	41.67	50.89	51.00	66.11
SIB200 isl Latn	41.18	2.45	24.51	25.98	28.43	42.16
OPUS en→isl Latn	11.57	7.37	10.68	13.30	12.74	13.45
OPUS isl Latn→en	11.71	13.52	16.40	17.55	17.14	16.23
Avg.	26.53	11.25	21.06	23.89	24.22	29.41
General Performance						
arc challenge	70.90	21.74	37.79	56.19	59.20	74.25
arc easy	85.79	23.33	55.96	74.39	76.84	85.61
hellaswag	40.38	8.29	25.24	32.65	27.75	44.67
piqa	71.87	26.55	54.79	69.48	61.48	77.31
winogrande	51.38	40.02	50.28	52.01	52.17	54.54
mmlu	58.07	17.18	35.86	46.77	45.40	58.92
Avg.	63.07	22.85	43.32	55.25	53.81	65.88

Table 6: Evaluation results comparing different fine-tuning configurations across various tasks for Icelandic for LLAMA-3.1-8B. Best performing configurations are highlighted in dark green, second-best in light green.

Metric	Base	Full	MLP	LoRA	Random	Target
Monolingual Capabilities						
FLORES en→kat Geor	3.65	1.98	3.52	3.18	3.52	3.42
FLORES kat Geor→en	18.87	1.12	6.61	18.72	18.02	19.01
BELEBELE kat Geor	59.67	0.00	12.67	60.33	49.56	58.89
SIB200 kat Geor	49.02	0.98	0.00	48.53	29.90	43.14
OPUS en→kat Geor	8.13	3.99	7.10	7.74	8.51	8.26
OPUS kat Geor→en	5.63	3.24	3.95	5.76	6.13	5.57
Avg.	24.16	1.88	5.64	24.04	19.27	23.05
General Performance						
arc challenge	70.90	9.03	10.70	71.57	63.55	69.90
arc easy	85.79	10.18	15.44	85.26	82.98	84.56
hellaswag	40.38	1.24	6.33	42.39	33.13	41.56
piqa	71.87	12.13	8.81	73.67	64.53	71.16
winogrande	51.38	3.95	37.57	53.12	50.91	51.38
mmlu	58.07	4.51	18.42	58.00	55.26	56.69
Avg.	63.07	6.84	16.21	64.00	58.39	62.54

Table 7: Evaluation results comparing different fine-tuning configurations across various tasks for Georgian for LLAMA-3.1-8B. Best performing configurations are highlighted in dark green, second-best in light green.

Metric	Base	Full	MLP	LoRA	Random	Target
Monolingual Capabilities						
FLORES en→lit Latn	10.81	4.09	8.51	11.95	12.39	12.96
FLORES lit Latn→en	22.67	10.75	20.50	23.28	24.56	24.72
BELEBELE lit Latn	66.44	34.33	46.33	53.00	64.56	67.00
SIB200 lit Latn	45.59	2.45	27.45	31.37	27.94	46.08
OPUS en→lit Latn	12.80	6.25	11.24	15.36	16.18	15.75
OPUS lit Latn→en	19.47	11.37	17.67	21.65	23.08	22.51
Avg.	29.63	11.54	21.95	26.10	28.12	31.50
General Performance						
arc challenge	70.90	20.74	37.79	54.52	67.89	70.23
arc easy	85.79	27.89	59.12	74.04	80.53	85.26
hellaswag	40.38	20.49	29.02	31.45	38.14	41.37
piqa	71.87	39.28	51.63	57.56	69.15	74.21
winogrande	51.38	51.07	49.96	50.43	50.36	53.35
mmlu	58.07	27.24	36.64	45.85	52.19	57.02
Avg.	63.07	31.12	44.03	52.31	59.71	63.58

Table 8: Evaluation results comparing different fine-tuning configurations across various tasks for Lithuanian for LLAMA-3.1-8B. Best performing configurations are highlighted in dark green, second-best in light green.

Metric	Base	Full	MLP	LoRA	Random	Target
Monolingual Capabilities						
FLORES en→lvs Latn	11.26	10.03	14.52	14.51	14.98	15.37
FLORES lvs Latn→en	24.47	7.43	24.24	25.04	27.36	26.80
BELEBELE lvs Latn	65.56	22.00	39.78	47.22	55.11	67.67
SIB200 lvs Latn	42.16	20.10	2.45	16.67	10.78	41.67
OPUS en→lvs Latn	12.88	9.78	13.82	16.69	16.70	17.08
OPUS lvs Latn→en	21.38	8.71	19.98	24.49	25.20	24.60
Avg.	29.62	13.01	19.13	24.10	25.02	32.20
General Performance						
arc challenge	70.90	20.07	25.42	53.51	68.23	69.23
arc easy	85.79	27.54	36.67	70.70	81.58	84.56
hellaswag	40.38	14.48	17.34	31.23	35.03	41.73
piqa	71.87	39.50	43.69	55.93	71.82	74.59
winogrande	51.62	47.43	46.80	50.04	50.43	55.17
mmlu	58.07	24.89	30.70	45.59	53.43	59.63
Avg.	63.10	28.98	33.44	51.17	60.09	64.15

Table 9: Evaluation results comparing different fine-tuning configurations across various tasks for Latvian for LLAMA-3.1-8B. Best performing configurations are highlighted in dark green, second-best in light green.

Metric	Base	Full	MLP	LoRA	Random	Target
Monolingual Capabilities						
FLORES en→mkd Cyrl	12.93	9.95	9.94	18.30	21.62	21.56
FLORES mkd Cyrl→en	33.37	6.36	27.85	33.89	34.32	33.45
BELEBELE mkd Cyrl	76.56	28.89	44.11	72.44	68.33	72.11
SIB200 mkd Cyrl	44.12	9.31	34.80	39.71	53.43	40.20
OPUS en→mkd Cyrl	10.72	11.89	12.21	15.56	18.76	18.38
OPUS mkd Cyrl→en	21.79	9.20	20.06	24.00	24.86	24.16
Avg.	33.25	12.60	24.83	33.98	36.89	34.98
General Performance						
arc challenge	70.90	31.10	43.48	69.90	66.22	69.23
arc easy	85.79	39.82	64.39	84.39	82.11	85.61
hellaswag	40.38	26.41	25.06	37.91	32.57	36.43
piqa	71.87	49.67	53.26	73.72	68.44	67.36
winogrande	51.62	49.41	49.25	52.96	51.78	52.09
mmlu	58.07	30.96	39.65	56.69	54.67	57.48
Avg.	63.10	37.90	45.85	62.60	59.30	61.37

Table 10: Evaluation results comparing different fine-tuning configurations across various tasks for Macedonian for LLAMA-3.1-8B. Best performing configurations are highlighted in dark green, second-best in light green.

Metric	Base	Full	MLP	LoRA	Random	Target
Monolingual Capabilities						
FLORES en→mlt Latn	9.19	16.81	24.25	20.96	17.71	20.92
FLORES mlt Latn→en	34.37	7.31	38.95	20.39	42.74	44.39
BELEBELE mlt Latn	57.67	25.11	40.00	50.89	58.11	66.44
SIB200 mlt Latn	37.75	10.29	24.02	27.94	33.82	38.73
OPUS en→mlt Latn	13.05	16.02	23.29	24.28	20.41	24.23
OPUS mlt Latn→en	29.27	10.67	30.24	19.60	36.79	39.36
Avg.	30.22	14.37	30.13	27.34	34.93	39.01
General Performance						
arc challenge	70.90	22.07	38.13	49.16	62.21	71.57
arc easy	85.79	21.23	56.84	72.46	82.11	83.86
hellaswag	40.38	9.70	26.72	31.45	35.07	39.00
piqa	71.87	42.11	50.87	61.97	72.85	73.34
winogrande	51.54	51.22	49.17	49.96	53.67	50.91
mmlu	58.07	26.26	32.66	45.00	52.58	57.87
Avg.	63.09	28.77	42.40	51.67	59.75	62.76

Table 11: Evaluation results comparing different fine-tuning configurations across various tasks for Maltese for LLAMA-3.1-8B. Best performing configurations are highlighted in dark green, second-best in light green.

Metric	Base	Full	MLP	LoRA	Random	Target
Monolingual Capabilities						
FLORES en→npi Deva	4.27	2.14	3.89	6.56	6.46	6.81
FLORES npi Deva→en	22.78	2.42	7.73	22.65	23.62	24.03
BELEBELE npi Deva	59.67	1.22	29.44	53.89	55.78	57.44
SIB200 npi Deva	36.27	4.41	4.41	27.45	33.33	57.84
OPUS en→npi Deva	12.36	5.75	10.96	17.61	17.50	18.90
OPUS npi Deva→en	25.97	12.53	21.17	31.08	28.00	30.87
Avg.	26.89	4.74	12.93	26.54	27.45	32.65
General Performance						
arc challenge	70.90	12.71	35.45	65.89	67.56	71.57
arc easy	85.79	19.47	49.12	81.75	82.98	84.39
hellaswag	40.38	16.38	14.32	37.46	36.68	36.31
piqa	71.87	36.83	51.25	62.35	73.83	71.11
winogrande	51.38	13.58	47.04	50.12	54.06	50.99
mmlu	58.07	17.50	33.18	54.08	54.08	57.74
Avg.	63.07	19.41	38.39	58.61	61.53	62.02

Table 12: Evaluation results comparing different fine-tuning configurations across various tasks for Nepali for LLAMA-3.1-8B. Best performing configurations are highlighted in dark green, second-best in light green.

Metric	Base	Full	MLP	LoRA	Random	Target
Monolingual Capabilities						
FLORES en→slk Latn	16.84	7.97	16.73	17.94	18.28	19.14
FLORES slk Latn→en	29.83	18.28	24.74	29.98	31.30	30.75
BELEBELE slk Latn	73.67	27.44	39.78	61.11	65.56	72.89
SIB200 slk Latn	40.20	20.59	39.22	44.61	33.82	46.08
OPUS en→slk Latn	17.21	9.33	16.08	17.34	18.55	18.24
OPUS slk Latn→en	23.13	17.56	22.42	25.59	27.12	26.27
Avg.	33.48	16.86	26.49	32.76	32.44	35.56
General Performance						
arc challenge	70.90	21.74	36.45	61.20	62.21	71.57
arc easy	85.79	23.86	51.05	78.25	78.60	85.44
hellaswag	40.38	15.04	27.87	40.04	36.74	43.68
piqa	71.87	43.25	54.30	59.25	70.29	71.76
winogrande	51.54	38.91	50.43	51.07	50.91	51.70
mmlu	58.07	21.36	35.99	52.25	51.53	57.67
Avg.	63.09	27.36	42.68	57.01	58.38	63.64

Table 13: Evaluation results comparing different fine-tuning configurations across various tasks for Slovak for LLAMA-3.1-8B. Best performing configurations are highlighted in dark green, second-best in light green.

Metric	Base	Full	MLP	LoRA	Random	Target
Monolingual Capabilities						
FLORES en→slv Latn	15.38	3.45	7.63	15.74	16.78	16.51
FLORES slv Latn→en	25.93	8.64	20.47	25.87	26.55	27.50
BELEBELE slv Latn	71.00	26.78	38.44	51.67	63.00	68.56
SIB200 slv Latn	41.67	23.04	16.18	20.59	53.92	47.55
OPUS en→slv Latn	14.46	2.74	10.53	14.26	16.00	16.38
OPUS slv Latn→en	16.82	3.26	16.63	19.86	20.24	20.89
Avg.	30.88	11.32	18.31	24.66	32.75	32.90
General Performance						
arc challenge	70.90	21.74	34.78	53.51	67.56	67.89
arc easy	85.79	28.60	36.49	71.23	82.46	83.51
hellaswag	40.38	8.79	24.92	34.08	39.72	36.48
piqa	71.87	45.65	56.15	70.13	72.03	75.08
winogrande	51.46	49.88	50.67	49.64	50.43	52.25
mmlu	58.07	27.43	29.13	43.31	55.06	56.69
Avg.	63.08	30.35	38.69	53.65	61.21	61.98

Table 14: Evaluation results comparing different fine-tuning configurations across various tasks for Slovenian for LLAMA-3.1-8B. Best performing configurations are highlighted in dark green, second-best in light green.

C.1 Mistral-Nemo-12B

Metric	Base	Random	Target
Monolingual Capabilities			
FLORES afr Latn→en	42.32	44.12	46.05
FLORES en→afr Latn	20.34	27.51	25.52
BELEBELE afr Latn	79.67	75.56	79.44
SIB200 afr Latn	31.86	43.14	31.86
OPUS afr Latn→en	19.69	30.04	30.54
OPUS en→afr Latn	16.77	19.24	17.99
Avg.	35.11	39.93	38.57
General Performance			
arc challenge	80.27	72.24	75.59
arc easy	89.82	83.33	87.72
hellaswag	40.86	38.31	42.32
piqa	80.03	77.15	77.20
winogrande	54.22	53.28	52.17
mmlu	61.66	55.85	61.33
Avg.	67.81	63.36	66.06

Table 15: Evaluation results comparing different fine-tuning configurations across various tasks for Afrikaans. Best performing configurations are highlighted in dark green, second-best in light green.

Metric	Base	Random	Target
Monolingual Capabilities			
FLORES cym Latn→en	36.21	43.11	38.94
FLORES en→cym Latn	9.09	21.73	25.54
SIB200 cym Latn	35.78	31.37	40.20
OPUS cym Latn→en	15.33	35.25	30.37
OPUS en→cym Latn	10.64	18.99	23.81
Avg.	21.41	30.09	31.77
General Performance			
arc challenge	80.27	63.88	71.57
arc easy	89.82	78.77	85.09
hellaswag	40.86	30.91	35.44
piqa	80.03	61.15	65.61
winogrande	54.22	51.07	51.93
mmlu	61.66	50.10	55.58
Avg.	67.81	55.98	60.87

Table 16: Evaluation results comparing different fine-tuning configurations across various tasks for Welsh. Best performing configurations are highlighted in dark green, second-best in light green.

Metric	Base	Random	Target
Monolingual Capabilities			
FLORES ekk Latn→en	23.86	25.62	25.54
FLORES en→ekk Latn	2.35	4.90	4.33
BELEBELE ekk Latn	73.00	62.44	71.00
SIB200 ekk Latn	32.35	29.90	36.76
OPUS ekk Latn→en	13.14	22.21	21.01
OPUS en→ekk Latn	6.37	14.12	11.60
Avg.	25.18	26.53	28.37
General Performance			
arc challenge	80.27	71.91	74.92
arc easy	89.82	83.68	88.42
hellaswag	40.86	32.99	40.67
piqa	80.03	68.61	77.75
winogrande	54.22	52.09	53.35
mmlu	61.66	54.93	59.83
Avg.	67.81	60.70	65.82

Table 17: Evaluation results comparing different fine-tuning configurations across various tasks for Estonian. Best performing configurations are highlighted in dark green, second-best in light green.

Metric	Base	Random	Target
Monolingual Capabilities			
FLORES en→isl Latn	6.05	6.37	11.14
FLORES isl Latn→en	22.24	5.40	25.24
BELEBELE isl Latn	67.00	60.00	70.67
SIB200 isl Latn	34.80	25.98	38.24
OPUS en→isl Latn	7.66	7.96	13.06
OPUS isl Latn→en	10.75	9.80	16.50
Avg.	24.75	19.25	29.14
General Performance			
arc challenge	80.27	64.55	76.59
arc easy	89.82	83.16	88.95
hellaswag	40.86	31.59	41.73
piqa	80.03	60.72	74.54
winogrande	54.22	51.70	53.67
mmlu	61.66	51.99	60.94
Avg.	67.81	57.28	66.07

Table 18: Evaluation results comparing different fine-tuning configurations across various tasks for Icelandic. Best performing configurations are highlighted in dark green, second-best in light green.

Metric	Base	Random	Target
Monolingual Capabilities			
FLORES en→kat Geor	2.40	7.97	3.83
FLORES kat Geor→en	19.32	20.11	19.58
BELEBELE kat Geor	71.89	68.56	72.33
SIB200 kat Geor	30.39	37.25	32.35
OPUS en→kat Geor	4.68	7.45	5.93
OPUS kat Geor→en	3.24	5.65	3.97
Avg.	21.99	24.50	23.00
General Performance			
arc challenge	80.27	69.23	77.26
arc easy	89.82	82.81	87.37
hellaswag	40.86	41.17	37.87
piqa	80.03	70.08	72.47
winogrande	54.22	52.33	53.99
mmlu	61.66	56.63	60.48
Avg.	67.81	62.04	64.91

Table 19: Evaluation results comparing different fine-tuning configurations across various tasks for Georgian. Best performing configurations are highlighted in dark green, second-best in light green.

Metric	Base	Random	Target
Monolingual Capabilities			
FLORES en→lit Latn	5.95	7.52	9.46
FLORES lit Latn→en	22.84	24.02	24.18
BELEBELE lit Latn	73.44	64.11	71.56
SIB200 lit Latn	33.82	25.98	34.80
OPUS en→lit Latn	6.97	10.58	11.70
OPUS lit Latn→en	16.29	20.54	19.94
Avg.	26.55	25.46	28.61
General Performance			
arc challenge	80.27	68.56	76.59
arc easy	89.82	86.32	88.77
hellaswag	40.85	36.47	41.65
piqa	80.03	67.25	72.20
winogrande	54.22	51.14	53.83
mmlu	61.66	52.25	60.16
Avg.	67.81	60.33	65.53

Table 20: Evaluation results comparing different fine-tuning configurations across various tasks for Lithuanian. Best performing configurations are highlighted in dark green, second-best in light green.

Metric	Base	Random	Target
Monolingual Capabilities			
FLORES en→lvs Latn	6.01	14.63	7.62
FLORES lvs Latn→en	25.29	26.96	26.65
BELEBELE lvs Latn	72.89	60.44	66.00
SIB200 lvs Latn	34.80	46.08	30.88
OPUS en→lvs Latn	6.86	15.63	10.67
OPUS lvs Latn→en	19.44	23.49	22.49
Avg.	27.55	31.21	27.39
General Performance			
arc challenge	80.27	64.21	72.91
arc easy	89.82	83.68	87.54
hellaswag	40.86	34.50	35.25
piqa	80.03	66.32	62.84
winogrande	54.22	51.22	51.93
mmlu	61.66	53.89	58.79
Avg.	67.81	58.97	61.54

Table 21: Evaluation results comparing different fine-tuning configurations across various tasks for Latvian. Best performing configurations are highlighted in dark green, second-best in light green.

Metric	Base	Random	Target
Monolingual Capabilities			
FLORES en→mkd Cyrl	8.85	16.78	8.43
FLORES mkd Cyrl→en	32.65	34.40	32.79
BELEBELE mkd Cyrl	76.11	73.56	76.67
SIB200 mkd Cyrl	31.37	31.86	44.12
OPUS en→mkd Cyrl	8.98	14.63	10.12
OPUS mkd Cyrl→en	20.96	24.28	22.71
Avg.	29.82	32.59	32.47
General Performance			
arc challenge	80.27	73.58	75.92
arc easy	89.82	87.02	87.89
hellaswag	40.86	37.81	44.30
piqa	80.03	74.97	76.17
winogrande	54.22	51.93	54.06
mmlu	61.66	57.28	59.57
Avg.	67.81	63.77	66.32

Table 22: Evaluation results comparing different fine-tuning configurations across various tasks for Macedonian. Best performing configurations are highlighted in dark green, second-best in light green.

Metric	Base	Random	Target
Monolingual Capabilities			
FLORES en→mlt Latn	3.20	19.27	19.41
FLORES mlt Latn→en	16.24	40.39	39.70
BELEBELE mlt Latn	41.33	64.67	72.44
SIB200 mlt Latn	30.88	44.61	33.33
OPUS en→mlt Latn	5.19	20.71	20.91
OPUS mlt Latn→en	16.52	31.51	33.82
Avg.	18.89	36.86	36.60
General Performance			
arc challenge	80.27	68.23	74.92
arc easy	89.82	80.88	86.67
hellaswag	40.86	35.46	39.45
piqa	80.03	72.74	65.23
winogrande	54.22	51.54	52.57
mmlu	61.66	52.45	59.24
Avg.	67.81	60.22	63.01

Table 23: Evaluation results comparing different fine-tuning configurations across various tasks for Maltese. Best performing configurations are highlighted in dark green, second-best in light green.

Metric	Base	Random	Target
Monolingual Capabilities			
FLORES en→npi Deva	1.58	3.21	2.90
FLORES npi Deva→en	18.36	18.36	19.39
BELEBELE npi Deva	64.11	51.56	60.67
SIB200 npi Deva	34.31	12.25	22.55
OPUS en→npi Deva	11.35	16.22	14.82
OPUS npi Deva→en	23.64	26.49	21.99
Avg.	25.56	21.35	23.72
General Performance			
arc challenge	80.27	51.51	72.58
arc easy	89.82	66.84	87.19
hellaswag	40.86	29.48	35.37
piqa	80.03	63.38	76.01
winogrande	54.22	52.49	53.12
mmlu	61.66	46.18	55.45
Avg.	67.81	51.65	63.29

Table 24: Evaluation results comparing different fine-tuning configurations across various tasks for Nepali. Best performing configurations are highlighted in dark green, second-best in light green.

Metric	Base	Random	Target
Monolingual Capabilities			
FLORES en→slk Latn	7.76	17.36	15.05
FLORES slk Latn→en	28.93	27.71	29.25
BELEBELE slk Latn	79.78	71.89	77.33
SIB200 slk Latn	34.80	44.12	28.43
OPUS en→slk Latn	8.73	17.68	15.89
OPUS slk Latn→en	20.26	25.70	26.11
Avg.	30.04	34.08	32.01
General Performance			
arc challenge	80.27	70.90	75.59
arc easy	89.82	85.09	88.77
hellaswag	40.86	36.49	40.12
piqa	80.03	70.95	78.62
winogrande	54.22	51.62	55.49
mmlu	61.66	53.36	60.42
Avg.	67.81	61.40	66.50

Table 25: Evaluation results comparing different fine-tuning configurations across various tasks for Slovak. Best performing configurations are highlighted in dark green, second-best in light green.

Metric	Base	Random	Target
Monolingual Capabilities			
FLORES en→slv Latn	6.47	10.02	13.19
FLORES slv Latn→en	24.99	20.81	17.12
BELEBELE slv Latn	77.56	68.33	71.89
SIB200 slv Latn	32.84	32.35	26.96
OPUS en→slv Latn	7.18	10.74	8.69
OPUS slv Latn→en	15.11	17.59	11.76
Avg.	27.36	26.64	24.93
General Performance			
arc challenge	80.27	68.56	72.91
arc easy	89.82	83.51	84.74
hellaswag	40.86	34.75	37.04
piqa	80.03	64.58	75.24
winogrande	54.22	53.04	54.14
mmlu	61.66	55.13	57.41
Avg.	67.81	59.93	63.58

Table 26: Evaluation results comparing different fine-tuning configurations across various tasks for Slovenian. Best performing configurations are highlighted in dark green, second-best in light green.

C.2 Generation Prompts

Task Prompts

We use standardized prompt templates for each task category to ensure consistency across models and datasets. Below we show the representative formats. Placeholders such as {source_text}, {passage}, or {question} are replaced with the actual content at runtime.

Translation.

```
Translate this sentence into
{target_name}: {source_text}
Translation:
```

Classification.

```
Classify the following text into
one of these topics:
A) Science/Technology
B) Travel
C) Politics
D) Sports
E) Health
F) Entertainment
G) Geography

Text: {text}

Answer:
```

Comprehension.

```
Read the following passage and
pick the right answer.

Passage: {passage}

Question: {question}

{options}

Answer:
```

General Capabilities. We adopt dataset-specific templates for MMLU and commonsense reasoning benchmarks:

- **MMLU:**

```
Answer the following multiple
choice question:

{question}

{options}

Answer:
```

- **HellaSwag:**

```
Complete the following
scenario by choosing the
most likely continuation:

{ctx}

{options}

Answer:
```

- **PIQA:**

```
Choose the most appropriate
solution for the following
goal:

Goal: {goal}

A) {sol1}
B) {sol2}

Answer:
```

- **Winogrande:**

```
Complete the sentence by
choosing the correct option:

{sentence}

A) {option1}
B) {option2}

Answer:
```

- **ARC (Easy/Challenge):**

```
Answer the following
question:

{question}

{options}

Answer:
```

C.3 Generation Settings

The generation is done using the following parameters:

```
SamplingParams(   temperature=0,
repetition_penalty=1.1,
max_tokens=128,
top_p=1.0,             top_k=-1,
stop_token_ids=[eos_token_id]
if eos_token_id is not None else
[],   stop=[
"\n", "\n\n",
"Question:", "Q:",
"Answer:", "A:",
"Translation:",             "The
translation", "Translate:",
"Summary:",    "The    summary",
"Summarize:",
"Passage:",    "The    passage",
"Text:",
"Options:", "The options", "The
choices", "Choices:",
"Context:", "The context", "The
problem", "Problem:",
"\n\nQuestion", "\n\nAnswer",
"</s>", "<|endoftext|>",
"Human:", "Assistant:"
],
skip_special_tokens=True )
```

D Cross-lingual Alignment

	ar	de	en	es	fr	it	ja	ru	zh
Baseline	85.55	92.88	90.63	91.47	91.21	91.45	86.55	88.21	87.46
Full	86.73	91.90	94.65	91.14	91.48	89.98	86.10	87.36	88.83
MLP	84.22	90.39	94.05	89.85	90.39	89.62	85.06	87.48	88.19
Adapter	84.79	95.11	92.10	93.23	93.21	93.07	86.58	88.59	87.47
Random	86.76	95.06	92.94	92.84	92.95	92.66	88.15	90.18	89.31
Target	87.75	96.10	95.03	94.24	94.52	93.89	90.24	91.25	91.46

Table 27: Layer-averaged Cosine similarity (expressed as percentages) between the target language (Afrikaans) and other languages for LLAMA-3.1-8B.

	ar	de	en	es	fr	it	ja	ru	zh
Baseline	83.22	87.82	85.92	87.72	87.58	88.11	83.09	84.76	83.24
Full	88.11	90.77	92.58	91.70	91.88	92.55	87.44	87.37	89.85
MLP	84.88	90.89	91.98	89.65	90.63	90.65	85.80	87.55	88.96
Adapter	88.13	93.40	90.40	93.30	93.29	94.08	86.11	89.41	86.25
Random	89.94	92.71	89.41	92.48	92.39	92.13	86.22	89.49	86.88
Target	94.35	95.49	93.20	94.90	95.57	95.62	91.70	92.71	91.57

Table 28: Layer-averaged Cosine similarity (expressed as percentages) between the target language (Welsh) and other languages for LLAMA-3.1-8B.

	ar	de	en	es	fr	it	ja	ru	zh
Baseline	85.45	91.29	89.18	90.46	90.19	90.67	86.09	88.40	86.75
Full	83.70	91.91	93.61	91.33	91.97	93.14	85.91	87.06	89.27
MLP	80.59	90.15	92.08	87.14	88.69	89.03	82.08	86.85	85.91
Adapter	86.19	93.87	90.52	92.56	92.51	92.83	86.99	89.58	87.47
Random	84.61	93.27	89.25	91.13	91.01	91.38	84.89	89.54	85.50
Target	90.07	95.43	93.87	94.13	93.91	93.73	90.77	92.25	91.84

Table 29: Layer-averaged Cosine similarity (expressed as percentages) between the target language (Estonian) and other languages for LLAMA-3.1-8B.

	ar	de	en	es	fr	it	ja	ru	zh
Baseline	82.82	88.54	85.85	87.65	87.45	87.89	83.22	85.24	83.55
Full	89.21	90.85	92.80	90.94	90.99	88.86	88.24	88.52	90.31
MLP	80.94	90.20	90.23	88.49	88.76	89.58	78.67	84.65	82.93
Adapter	85.49	93.79	89.50	92.61	92.24	92.57	85.22	88.89	86.21
Random	87.74	91.14	89.89	91.65	90.48	90.17	85.29	91.03	87.21
Target	91.14	93.44	93.11	94.09	94.10	94.25	90.11	91.98	90.65

Table 30: Layer-averaged Cosine similarity (expressed as percentages) between the target language (Icelandic) and other languages for LLAMA-3.1-8B.

	ar	de	en	es	fr	it	ja	ru	zh
Baseline	66.64	68.23	61.90	66.87	67.40	68.37	66.58	70.69	64.28
Full	64.74	64.17	57.36	64.37	63.03	64.32	63.36	66.32	61.19
MLP	67.84	69.05	61.96	67.05	67.74	68.95	67.34	69.48	65.94
Adapter	65.52	67.76	61.55	66.33	66.81	67.67	65.41	70.47	63.65
Random	67.02	69.26	62.24	67.80	68.14	69.37	66.86	71.79	65.19
Target	69.54	70.67	64.23	69.01	69.66	70.64	69.20	73.41	67.32

Table 31: Layer-averaged Cosine similarity (expressed as percentages) between the target language (Georgian) and other languages for LLAMA-3.1-8B.

	ar	de	en	es	fr	it	ja	ru	zh
Baseline	83.79	89.55	86.95	89.12	88.88	89.56	84.84	87.59	85.20
Full	86.30	92.68	92.10	91.05	91.04	91.40	88.21	87.89	89.71
MLP	82.28	91.34	91.10	89.48	89.88	90.35	85.49	86.40	87.62
Adapter	83.45	91.75	88.31	91.17	90.83	91.67	84.16	88.73	84.43
Random	83.26	91.59	88.84	91.37	90.73	91.36	84.83	88.83	85.68
Target	86.99	93.23	91.85	92.78	92.50	92.47	88.56	91.26	90.07

Table 32: Layer-averaged Cosine similarity (expressed as percentages) between the target language (Lithuanian) and other languages for LLAMA-3.1-8B.

	ar	de	en	es	fr	it	ja	ru	zh
Baseline	82.73	88.36	85.37	87.88	87.58	88.25	83.37	86.37	83.76
Full	86.70	89.06	90.55	90.20	89.57	88.98	87.45	84.64	89.21
MLP	80.97	88.03	88.93	86.05	86.14	86.56	82.48	84.67	86.19
Adapter	84.62	92.14	88.45	91.85	91.36	92.08	84.51	89.09	86.13
Random	82.07	91.14	87.65	90.65	90.18	90.39	83.74	87.63	84.94
Target	86.60	93.48	91.73	93.19	93.04	93.14	88.61	90.79	89.61

Table 33: Layer-averaged Cosine similarity (expressed as percentages) between the target language (Latvian) and other languages for LLAMA-3.1-8B.

	ar	de	en	es	fr	it	ja	ru	zh
Baseline	87.36	87.33	86.16	87.00	86.85	87.38	86.00	92.75	86.12
Full	85.37	89.55	90.10	86.95	89.35	87.79	84.21	92.15	87.69
MLP	84.71	88.09	90.33	87.73	87.95	88.06	86.25	90.71	88.38
Adapter	87.53	86.66	85.75	86.49	86.15	86.97	84.81	97.54	85.33
Random	94.88	89.60	87.52	89.51	89.00	90.22	88.70	98.18	88.11
Target	92.74	92.92	92.01	92.99	92.63	93.00	91.52	98.72	92.00

Table 34: Layer-averaged Cosine similarity (expressed as percentages) between the target language (Macedonian) and other languages for LLAMA-3.1-8B.

	ar	de	en	es	fr	it	ja	ru	zh
Baseline	84.92	87.62	85.09	87.37	87.32	89.50	82.94	84.95	83.17
Full	86.03	90.98	91.14	89.55	89.71	90.60	85.43	88.60	85.49
MLP	86.29	90.72	90.80	90.55	89.50	90.07	86.65	88.56	87.35
Adapter	88.20	93.48	89.95	93.85	93.66	96.17	85.00	88.92	85.74
Random	89.54	91.84	90.38	91.49	91.17	93.10	86.12	89.70	86.50
Target	92.23	93.53	92.42	92.79	92.63	94.90	89.06	90.80	89.52

Table 35: Layer-averaged Cosine similarity (expressed as percentages) between the target language (Maltese) and other languages for LLAMA-3.1-8B.

	ar	de	en	es	fr	it	ja	ru	zh
Baseline	85.43	84.39	82.60	83.91	83.85	84.81	85.77	86.33	84.49
Full	88.65	88.61	89.58	89.29	89.63	89.65	88.12	89.33	89.39
MLP	83.47	83.90	86.27	84.17	84.67	83.77	85.09	85.00	86.68
Adapter	89.31	86.91	84.48	86.07	85.95	86.98	88.93	89.12	86.84
Random	92.35	87.82	84.56	87.06	86.96	87.95	90.76	91.25	88.24
Target	94.93	92.81	89.43	91.40	91.43	92.10	94.87	94.62	92.49

Table 36: Layer-averaged Cosine similarity (expressed as percentages) between the target language (Nepali) and other languages for LLAMA-3.1-8B.

	ar	de	en	es	fr	it	ja	ru	zh
Baseline	87.04	91.90	90.14	91.61	91.24	91.61	87.26	90.57	87.85
Full	87.16	92.87	93.77	91.80	91.93	91.83	88.86	88.68	89.80
MLP	84.25	91.34	93.15	90.85	90.58	90.91	84.58	88.15	87.52
Adapter	85.71	92.33	90.44	92.23	91.83	92.12	86.20	90.18	87.07
Random	87.06	92.57	91.00	92.11	91.82	91.69	87.97	91.88	88.57
Target	87.56	93.88	93.35	93.52	93.25	93.33	89.86	92.60	90.45

Table 37: Layer-averaged Cosine similarity (expressed as percentages) between the target language (Slovak) and other languages for LLAMA-3.1-8B.

	ar	de	en	es	fr	it	ja	ru	zh
Baseline	85.88	91.34	89.69	91.13	90.40	91.31	86.13	89.15	86.98
Full	86.41	89.81	93.60	90.56	89.94	90.11	86.23	89.84	88.16
MLP	84.26	89.02	92.18	89.38	89.64	89.61	84.05	87.33	87.56
Adapter	86.40	92.67	91.10	92.64	92.03	92.60	86.11	90.97	87.25
Random	84.48	92.45	90.52	91.82	91.10	92.14	84.45	91.15	85.49
Target	88.57	94.18	93.79	94.25	93.47	94.17	88.72	93.27	90.21

Table 38: Layer-averaged Cosine similarity (expressed as percentages) between the target language (Slovenian) and other languages for LLAMA-3.1-8B.

E Computational Budget

All experiments were conducted on a single NVIDIA H100 GPU, except for the full fine-tuning setup, which required two H100 GPUs due to its higher memory and compute demands. Each model was trained for approximately 100 million tokens, corresponding to about 6-8 GPU hours per run for parameter-efficient methods and roughly 10 GPU hours for full fine-tuning.

F Licenses and Attribution

All experiments were conducted using publicly available models and datasets released under permissive or research-oriented licenses. The LLAMA-3.1-8B model is distributed under the Meta Llama 3 Community License Agreement, which permits research use and redistribution with attribution. MISTRAL-NEMO-12B is released under the Apache License 2.0, allowing open use and modification. Language-specific token data were obtained from the GlotCC corpus (CC0 for metadata, Common Crawl terms for text content). Evaluation datasets including FLORES, BELEBELE, SIB200, and OPUS were used in accordance with their respective research licenses, and English benchmarks (ARC, HELLASWAG, PIQA, WINOGRANDE, MMLU) were accessed under standard academic usage terms.