Train More Parameters But Mind Their Placement: Insights into Language Adaptation with PEFT

Jenny Kunz
Dept. of Computer and Information Science
Linköping University
jenny.kunz@liu.se

Abstract

Smaller LLMs still face significant challenges even in medium-resourced languages, particularly when it comes to language-specific knowledge – a problem not easily resolved with machine-translated data. In this case study on Icelandic, we aim to enhance the generation performance of an LLM by specialising it using unstructured text corpora. A key focus is on preventing interference with the models' capabilities of handling longer context during this adaptation. Through ablation studies using various parameter-efficient fine-tuning (PEFT) methods and setups, we find that increasing the number of trainable parameters leads to better and more robust language adaptation. LoRAs placed in the feed-forward layers and bottleneck adapters show promising results with sufficient parameters, while prefix tuning and $(IA)^3$ are not suitable. Although improvements are consistent in 0-shot summarisation, some adapted models struggle with longer context lengths, an issue that can be mitigated by adapting only the final layers.

1 Introduction

LLMs have strong multilingual capabilities and top the leaderboards even for less-represented languages (Nielsen et al., 2024). However, smaller LLMs still struggle with these languages, hampering fast and resource-efficient inference. Instruction tuning on machine-translated data can improve performance compared to English-only tuning (Muennighoff et al., 2023; Chen et al., 2024a) but models still fall short when evaluated on native benchmarks, likely due to missing language-specific knowledge (Chen et al., 2024b). While collecting large amounts of native instruction-tuning data could address this issue, this can be costly or infeasible. This makes techniques for adapting models using unstructured text data valuable.

In this paper, we perform ablations with parameter-efficient fine-tuning (PEFT) methods for language adaptation with unstructured text data *after* instruction alignment. This diverges from the standard setup for fine-tuning a model: Unlike typical fine-tuning, where the adaptation data closely matches the expected output format, the data we use is closer to the expected output in language but likely further from the target task format. Therefore, the setup risks interference with the original instruction-tuning objectives, possibly leading to *catastrophic forgetting* (Mccloskey and Cohen, 1989). In addition, hardware constraints made us choose a maximum context length smaller than the one used in pre-training, risking further performance degradation.

Therefore, we aim to identify setups that do not interfere with previously learned abilities. We attempt to avoid catastrophic forgetting with PEFT methods that leave the majority of or all model parameters unchanged: LoRA (Hu et al., 2022), IA^3 (Liu et al., 2022), bottleneck adapters (Houlsby et al., 2019) and prefix tuning (Li and Liang, 2021). We experiment with the number of learnable parameters, the placement of LoRA matrices in different Transformer modules and layers, as well as the training corpus used for adaptation.

We use the smallest instruction-tuned LLaMA 3.2 model (LlamaTeam, 2024) with 1B parameters and adapt it to Icelandic, evaluating performance on text summarisation. Our findings are that:

- LoRA and bottleneck adapters show improvements especially in 0-shot settings, though simply adding target-language task demonstrations also improves the performance substantially.
- A higher number of trainable parameters is better.
- LoRAs in the feed-forward layers are the best-performing setup, followed by bottleneck

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 323–330
March 3-4, 2025 ©2025 University of Tartu Library

adapters. LoRA in the attention layers works less well, particularly considering the number of trainable parameters. We therefore conclude that feed-forward modules are the most promising target in language adaptation.
- Prefix tuning hurts the model's capabilities.
- Some setups with few trainable parameters negatively impact 5-shot performance, possibly due to smaller context lengths at adaptation time compared to pre-training time. This can be resolved by restricting adapter placement to the top layers.

2 Experimental Setup

2.1 Models

We use *Llama-3.2-1B-Instruct*, the newest and smallest Llama model at the time of writing, with 1B, 16 layers, and a hidden size of 2048. This model has been tuned with instruction fine-tuning (Wei et al., 2022) and reinforcement learning with human feedback (Ouyang et al., 2024)[1].

2.2 Adaptation Data

Our main dataset for adaptation is the Icelandic portion of CC100 (Conneau et al., 2020) that has been processed with CCNet filtering (Wenzek et al., 2020) to increase data quality. We randomly select 250,000 text chunks, with a maximum length of 1,024 tokens, resulting in 12.5M tokens. This data was likely seen during pre-training, i.e., the model is not exposed to new data but *primed* towards Icelandic. As web-crawled corpora are reportedly of lower quality for smaller languages (Kreutzer et al., 2022; Artetxe et al., 2022), we perform ablations with the curated Icelandic Gigaword Corpus (IGC) (Steingrímsson et al., 2018; Barkarson et al., 2022), using sections from its subsets *Books*, *Wiki*, *Social*, and *Journals*. Even here we use 250,000 chunks, resulting in 12M tokens. As the *Social* subset is by far the largest and we aim to have a large portion of highly curated text, we undersample it by using only 10%, resulting in a dataset composition of 9% *Books*, 17% *Wiki*, 22% *Journals*, and 52% *Social*.

2.3 Adaptation Methods and Setups

The code, prompt generator and adapters used for the experiments in this paper can be found at `github.com/jekunz/peft-la`. We use the Transformers (Wolf et al., 2020) and Adapters (Poth et al., 2023) libraries, a learning rate of 5e-5, a linear learning rate scheduler, and a batch size of 4.[2] All adapters are trained with a causal language modeling objective. We test the following methods and setups:

LoRA is a widespread adaptation technique for generative LLMs. In the most common setup, it adds low-rank decomposition matrices to the model's self-attention modules and trains only those. The matrices can be merged into the weights, removing the inference overhead. For LoRA in the attention module, we test ranks 1024, 256, 128, 32 and 8 and apply LoRA to the query and value matrices, which is reportedly the most stable setup (Fomenko et al., 2024). We also test LoRA in the feed-forward module and place LoRAs in all matrices using ranks 256, 128, 64, 32 and 8. For both module setups and all ranks, we use use $\alpha = 2r$.

IA^3 is the most parameter-efficient among the methods tested. It multiplies activations in the model's attention (key and value) and feed-forward matrices with learned vectors, adding hardly any overhead.

Bottleneck adapters add smaller intermediate layers with a down- and up-projection in between the model's layers. While popular for encoder model, bottleneck adapters are less common for generative LLMs as they increase the number of parameters and depth even during inference. We train Houlsby adapters with reduction factors of 64, 16 and 4.

Prefix tuning prepends a sequence of learnable prefix vectors to the input sequences, allowing the model to attend to the prefix vectors when generating the subsequent tokens. As the vectors add to the sequence length, even prefix tuning slows down inference. We use a prefix length of 30 tokens.

2.4 Evaluation

To assess generative performance, we evaluate abstractive text summarisation with the RÚV Radio News (RRN) dataset (Sverrisson and Einarsson, 2023) in the *main* $\rightarrow$ *intro* setup, i.e., generating the introduction from the main body of the article. We filter out articles missing one of these fields.

We evaluate the summaries using BERTScore (Zhang et al., 2020) (base model: *bert-base-multilingual-uncased*) to measure the representational similarity between the output and the reference, and ROUGE-L (Lin, 2004) for surface overlap, based on the longest common subsequence.

[1]Ablations with the base model *Llama-3.2-1B* showed inferior performance with and without adaptation.

[2]As the learning rate and scheduler are crucial in continued pre-training (Ibrahim et al., 2024), we also tested 1e-5 and 1e-4 and a cosine scheduler but did not observe large differences.

The models are evaluated in 0-shot, 1-shot and 5-shot setups with minimal prompts in Icelandic[3] that instruct the model to summarise the article in one paragraph and include markers for the start of both the article and the summary.

3 Results and Discussion

3.1 PEFT Methods

	0-shot	1-shot	5-shot
No Adapter	53.37 / 04.09	64.68 / 10.26	64.01 / 11.37
LoRA-qv-1024	63.61 / 08.57	66.53 / 11.70	65.50 / 12.06
LoRA-qv-256	63.27 / 08.32	65.55 / 11.05	62.97 / 10.56
LoRA-qv-128	62.55 / 07.63	64.51 / 10.78	62.23 / 10.54
LoRA-qv-32	61.06 / 06.62	62.68 / 08.98	55.42 / 05.60
LoRA-qv-8	60.45 / 05.21	61.53 / 08.23	56.62 / 06.42
LoRA-ff-256	**65.60 / 09.72**	**69.06 / 13.89**	**69.10 / 15.48**
LoRA-ff-128	64.67 / 08.87	**69.10 / 13.86**	68.36 / 14.55
LoRA-ff-64	63.72 / 07.86	67.72 / 12.60	67.46 / 13.65
LoRA-ff-32	62.94 / 07.19	67.61 / 12.18	67.42 / 13.76
LoRA-ff-8	61.69 / 06.36	64.85 / 10.39	62.66 / 10.09
$(IA)^3$	56.70 / 04.56	64.07 / 09.47	61.74 / 10.37
Bottlen.-4	63.78 / 08.15	66.75 / 11.74	66.74 / 13.21
Bottlen.-16	63.33 / 08.38	67.77 / 13.11	65.80 / 12.36
Bottlen.-64	60.66 / 05.16	64.79 / 09.96	61.32 / 08.59
Prefix	55.84 / 02.02	54.56 / 01.73	49.86 / 00.67

Table 1: Comparing adaptation methods. BERTScore F1 / ROUGE-L.

As shown in Table 1, language adaptation consistently improves 0-shot summarisation scores. However, for 1-shot and 5-shot setups, the results are more mixed, and in some setups decrease compared to the baseline without adaptation. That the 1-shot setup without adaptation already shows comparable performance to many adaptation setups implies that in-context learning, where possible, can be an alternative to language adaptation for this model.

The best-performing method are LoRAs in the feed-forward layers. Even bottleneck adapters with a reduction factor of 16 or 4 consistently increase scores, although there is a noticeable difference in performance to feed-forward LoRA. As illustrated in Figure 1, feed-forward LoRA also results in the highest BERTScores relative to the number of parameters added, followed by bottleneck adapters. LoRA in the attention matrices requires substantially more parameters to reach a comparable performance. These results show that the placement of the PEFT modules in the Transformer architecture plays a crucial role even if the number of trainable parameters is the same.

Some setups interfere with the model's ability to operate on longer inputs as the performance especially in the 5-shot setup decreases. We hypothesise this is a result of limiting the context length to 1,024 tokens during the adaptation process. LoRA in the attention module is the most heavily affected setup, suggesting that the effectiveness of self-attention when processing longer contexts is harmed.

We observe that performance improves as the LoRA rank increases or the bottleneck reduction factor decreases, indicating that sufficient learning capacity is necessary for better results in language adaptation. This is in line with the underwhelming performance of $(IA)^3$, which introduces the fewest parameters. Designed as an alternative to in-context learning for task adaptation, $(IA)^3$ does not transfer well to language adaptation.

Prefix tuning with textual data decreases the performance substantially for the 1- and 5-shot setups. We assume that as prefixes have a direct impact on the generation, prefixes that diverge from the expected output format harm the model's abilities to match the latter. For this reason, prefix-tuning an instruction-tuned model on unlabelled text does not work, whereas prefix-tuning on specific tasks like summarisation, or instruction tuning in general, works well as shown by Zhang et al. (2024a).

3.2 Ablation 1: LoRA Modules

	0-shot	1-shot	5-shot
q,v	63.27 / 08.32	65.55 / 11.05	62.97 / 10.56
ff	65.60 / 09.72	69.06 / 13.89	69.10 / 15.48
ff + q,v	65.44 / 09.61	68.44 / 13.14	68.89 / 15.17

Table 2: Comparing LoRA module placement. BERTScore F1 / ROUGE-L; LoRA rank 256

We have a closer look at the module placement of LoRAs and compare LoRA in the self-attention module, LoRA in the feed-forward module, and LoRA both in the self-attention and the feed-forward module.

In the results given in Table 2, we see that for the same rank, LoRA in the feed-forward module is better than in the attention module. Moreover, it is slightly better than LoRA in both the attention and the feed-forward modules. We find this surprising given that the latter option has the most trainable parameters and conclude that having LoRA even

[3]We also tested English instructions, which led to slightly worse results, except for the *no adapters* model, where English instruction slightly improved the 0-shot performance.

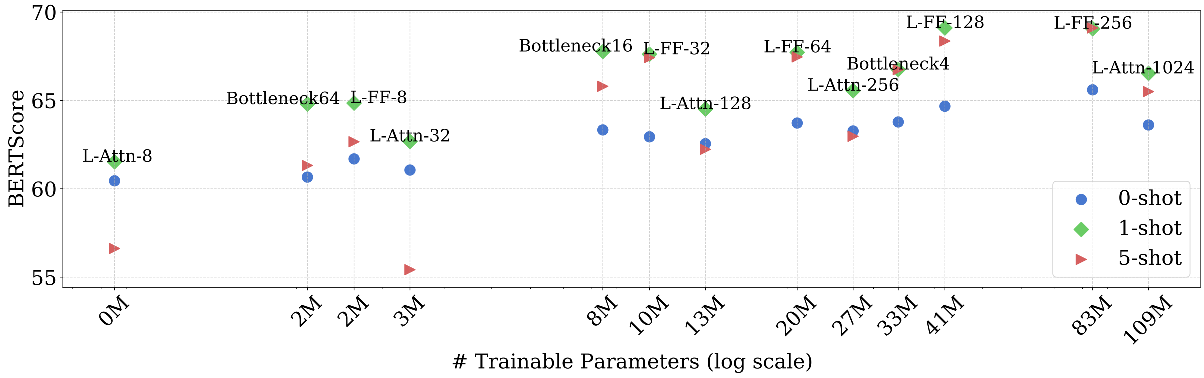

Figure 1: Number of trainable parameters plotted against BERTScores. Prefix tuning (34M parameters) and $(IA)^3$ (49K parameters) are excluded.

in the attention matrices is at best unnecessary.

3.3 Ablation 2: Layer Exclusion

	0-shot	1-shot	5-shot
No Adapter	53.37 / 04.09	**64.68 / 10.26**	64.01 / 11.37
All Layers	**61.06 / 06.62**	62.68 / 08.98	55.42 / 05.60
All but last 2	59.39 / 04.83	60.26 / 07.37	57.73 / 06.51
All but last 4	59.89 / 04.93	62.37 / 08.70	58.29 / 07.02
Only last 2	59.64 / 03.64	63.55 / 08.37	**65.40 / 12.42**
Only last 4	58.78 / 04.56	62.20 / 08.22	61.94 / 10.57

Table 3: Layer Exclusion experiments. BERTScore F1 / ROUGE-L; Self-attention (qv) LoRA rank 32.

Fine-tuning primarily affects the final layers of a model (Merchant et al., 2020; Mosbach et al., 2020; Zhou and Srikumar, 2022). We explore two strategies focusing on these layers: (1) *excluding* the final layers during adaptation to preserve the instruction-tuning capabilities while focusing on general language learning, which is likely stored in earlier layers, and (2) adapting *only* the final layers, as this may be sufficient and could maintain the model's robustness with respect to the limited context length used in our adaptation process (a key issue highlighted in Section 3.1).

We test the two hypotheses using self-attention LoRA with rank 32 as this configuration shows strong 0-shot performance but suffers in the 5-shot setup. The results in Table 3 show that the first hypothesis does not hold; excluding the last layers does not improve the performance and, in some cases, degrades it. The second hypothesis, however, appears plausible: restricting LoRA modules to the last two layers yields the best 5-shot results among all setups in Table 3, outperforming the baseline without adaptation. However, this comes at the expense of a slight decrease in 0-shot performance. We are hopeful that these insights can guide us in developing customised methods for language adaptation.

3.4 Ablation 3: Training Corpora

	0-shot	1-shot	5-shot
CCNet	63.27 / 08.32	65.55 / 11.05	62.97 / 10.56
IGC	60.80 / 05.75	61.02 / 06.48	58.31 / 06.17
CCNet	65.60 / 09.72	69.06 / 13.89	69.10 / 15.48
IGC	63.66 / 08.10	66.19 / 10.46	66.37 / 12.00
CCNet	63.78 / 08.15	66.75 / 11.74	66.74 / 13.21
IGC	61.39 / 05.58	64.95 / 09.79	65.24 / 11.54

Table 4: Comparing text corpora for adaptation. BERTScore F1 / ROUGE-L; LoRA-qv-256 (above), LoRA-ff-256 (middle) and bottleneck reduction factor 4 (below).

In Table 4, we do not observe a benefit of training on the IGC; on the contrary, the performance is consistently lower. While this is in line with previous research (Artetxe et al., 2022; van Noord et al., 2024), note that we do not test on any task where high-quality generation is important but on text summarisation, which can rely on copying chunks of text. We also note that CCNet is probably more diverse, and that different mixes from the IGC may lead to different results. We therefore believe that it is worthwhile to continue testing on curated data.

3.5 Future Work

In order to test whether our findings generalise, we plan to extend our approach to other languages, larger models and adapters trained on more data, and to explore the effect of training on longer contexts. Based on our experiments on the placement and training of adapters in Section 3.3, we hope to find a sweet spot for language adaptation where no relevant information is overwritten but generation performance is improved. Inspiration could be taken from methods that automatically detect, and assign more parameters to, layers of particular importance (Zhang et al., 2023; Yao et al., 2024).

A common approach to mitigate interference is episodic memories – mixing in examples from previous tasks (Chaudhry et al., 2019), in our case, instruction-tuning data. This has shown promise in other works (Jiang et al., 2024; Parmar et al., 2024), making it worthwhile to incorporate.

One challenge in evaluating language adaptation methods is that automatic metrics for generative performance provide limited and potentially misleading insights. While running extensive human evaluations for all ablations in this paper is impractical, a human study of model outputs for the most promising setups, across a diverse set of prompts, should be included in future evaluations.

4 Related Work

Razumovskaia et al. (2024) find that LoRA language adaptation with unstructured text data improves the linguistic quality of generated texts in human ratings but usefulness and performance on a (translated) natural language inference benchmark remain low. Their study indicates that benchmark evaluation could underestimate the usefulness of language adaptation in chat and generation setups.

Work on testing other PEFT architectures than LoRA for language adaptation of LLMs has been sparse. While bottleneck-style language adapters trained on text corpora are a common setup for cross-lingual transfer with encoder models (Pfeiffer et al., 2020; He et al., 2021; Faisal and Anastasopoulos, 2022), they have been largely overlooked for generative models, likely due to the inference overhead that can be avoided with LoRA, as the latter works equally well for task fine-tuning. Our experiments show that similar findings hold for language adapters: Bottleneck adapters perform well but there are LoRA setups that reach the same performance or are better while avoiding the overhead.

Recent language adaptation works have focused on target-language instruction fine-tuning, often with machine-translated data (Muennighoff et al., 2023; Chen et al., 2024a; Holmström and Doostmohammadi, 2023). In cross-lingual transfer, multilingual instruction tuning has shown promise, particularly for generative tasks (Kew et al., 2023) and for larger models (Chen et al., 2024a). However, models trained on machine-translated data may perform well on translated evaluation sets but struggle on native benchmarks (Chen et al., 2024b).

5 Conclusion

We tested a range of PEFT methods for language adaptation using unstructured text corpora, finding that LoRA in the feed-forward modules yielded the most promising results, followed by bottleneck adapters. LoRA in the attention modules performed less well, was less robust to larger context lengths and needed more parameters for a comparable performance. Combining LoRAs in both the attention and feed-forward modules did not improve over feed-forward LoRAs only, and may even lead to slightly decreased performance. Prefix tuning and $(IA)^3$ were not suitable at all.

Our results show that across architectures, more trainable parameters lead to better scores, showing, perhaps unsurprisingly, that sufficient learning capacity is crucial for language adaptation.

Some adaptation setups led to a decline in performance as contexts get longer; possibly a result of restricted context lengths during adaptation. However, this issue can be mitigated by training only the last layers. Notably, we did not observe any positive effects from using higher-quality pre-training data sourced from narrower domains.

Moving forward, with a higher resource investment, we see the potential that more training data, possibly with instruction data in the mix, and longer context lengths improve the performance further. However, to truly assess the potential of these methods, we need more diverse, language-native evaluation data, as well as fine-grained human evaluations that assess various aspects of generated language quality and content.

Limitations

The meaningfulness of automated text summarisation metrics when using news text summaries as references has been questioned and is highly dependent on the dataset (Zhang et al., 2024b). While our search for effective setups yielded conclusive results with BERTScore and ROUGE-L, moving forward, it will be crucial to incorporate human evaluations and more diverse tasks to accurately assess performance across a broader and better-interpretable range of criteria.

As we have discussed in Section 5, we see a critical need for more language-native evaluation data, in particular datasets that incorporate significant language-specific knowledge (Chen et al., 2024b). Testing on a limited set of language-native tasks most of which are classification tasks, or on machine-translated data, may give a limited picture of the effect of language adaptation.

Due to computational constraints, we were unable to include larger models or more than one language in this study. As a result, it remains unclear whether our findings apply to other languages, especially those that are typologically more different from or closer to English.

Acknowledgments

I thank my colleagues Kevin Glocker, Kätriin Kukk, Julian Schlenker, Marcel Bollmann, Noah-Manuel Michael and Romina Oji for valuable discussions at all stages of this project and feedback on earlier drafts, and the anonymous reviewers for their constructive feedback and insightful suggestions.

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. It was supported by TrustLLM funded by Horizon Europe GA 101135671. The computations were enabled by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre and by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

References

Mikel Artetxe, Itziar Aldabe, Rodrigo Agerri, Olatz Perez-de Viñaspre, and Aitor Soroa. 2022. Does corpus quality really matter for low-resource languages? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7383–7390, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Starkaður Barkarson, Steinþór Steingrímsson, and Hildur Hafsteinsdóttir. 2022. Evolving large text corpora: Four versions of the Icelandic Gigaword corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2371–2381, Marseille, France. European Language Resources Association.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K. Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. 2019. On tiny episodic memories in continual learning.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024a. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1347–1356, St. Julian's, Malta. Association for Computational Linguistics.

Pinzhen Chen, Simon Yu, Zhicheng Guo, and Barry Haddow. 2024b. Is it good data for multilingual instruction tuning or just bad multilingual evaluation for large language models?

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

Fahim Faisal and Antonios Anastasopoulos. 2022. Phylogeny-inspired adaptation of multilingual models to new 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 1: Long Papers)*, pages 434–452, Online only. Association for Computational Linguistics.

Vlad Fomenko, Han Yu, Jongho Lee, Stanley Hsieh, and Weizhu Chen. 2024. A note on lora.

Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jiawei Low, Lidong Bing, and Luo Si. 2021. On the effectiveness of adapter-based tuning for pretrained language model adaptation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2208–2222, Online. Association for Computational Linguistics.

Oskar Holmström and Ehsan Doostmohammadi. 2023. Making instruction finetuning accessible to non-English languages: A case study on Swedish models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 634–642, Tórshavn, Faroe Islands. University of Tartu Library.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference* *on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats Leon Richter, Quentin Gregory Anthony, Eugene Belilovsky, Timothée Lesort, and Irina Rish. 2024. Simple and scalable strategies to continually pre-train large language models. *Transactions on Machine Learning Research*.

Zhengbao Jiang, Zhiqing Sun, Weijia Shi, Pedro Rodriguez, Chunting Zhou, Graham Neubig, Xi Lin, Wen-tau Yih, and Srini Iyer. 2024. Instruction-tuned language models are better knowledge learners. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5421–5434, Bangkok, Thailand. Association for Computational Linguistics.

Tannon Kew, Florian Schottmann, and Rico Sennrich. 2023. Turning english-centric llms into polyglots: How much multilinguality is needed?

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haokun Liu, Derek Tam, Muqeeth Mohammed, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*.

LlamaTeam. 2024. The llama 3 herd of models.

Michael Mccloskey and Neil J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *The Psychology of Learning and Motivation*, 24:104–169.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online. Association for Computational Linguistics.

Marius Mosbach, Anna Khokhlova, Michael A. Hedderich, and Dietrich Klakow. 2020. On the interplay between fine-tuning and sentence-level probing for linguistic knowledge in pre-trained transformers. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 68–82, Online. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Dan Saattrup Nielsen, Kenneth Enevoldsen, and Peter Schneider-Kamp. 2024. Encoder vs decoder: Comparative analysis of encoder and decoder language models on multilingual nlu tasks.

Rik van Noord, Taja Kuzman, Peter Rupnik, Nikola Ljubešić, Miquel Esplà-Gomis, Gema Ramírez-Sánchez, and Antonio Toral. 2024. Do language models care about text quality? evaluating web-crawled corpora across 11 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5221–5234, Torino, Italia. ELRA and ICCL.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2024. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing*

Systems, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Jupinder Parmar, Sanjev Satheesh, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Reuse, don't retrain: A recipe for continued pretraining of language models.

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.

Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. 2023. Adapters: A unified library for parameter-efficient and modular transfer learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 149–160, Singapore. Association for Computational Linguistics.

Evgeniia Razumovskaia, Ivan Vulić, and Anna Korhonen. 2024. Analyzing and adapting large language models for few-shot multilingual nlu: Are we there yet?

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A very large Icelandic text corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Þór Sverrisson and Hafsteinn Einarsson. 2023. Abstractive text summarization for Icelandic. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 17–31, Tórshavn, Faroe Islands. University of Tartu Library.

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

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Kai Yao, Penlei Gao, Lichun Li, Yuan Zhao, Xiaofeng Wang, Wei Wang, and Jianke Zhu. 2024. Layer-wise importance matters: Less memory for better performance in parameter-efficient fine-tuning of large language models.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*.

Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. 2024a. LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024b. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Yichu Zhou and Vivek Srikumar. 2022. A closer look at how fine-tuning changes BERT. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1046–1061, Dublin, Ireland. Association for Computational Linguistics.