How to Tune a Multilingual Encoder Model for Germanic Languages: A Study of PEFT, Full Fine-Tuning, and Language Adapters

Romina Oji and **Jenny Kunz**
Dept. of Computer and Information Science
Linköping University
romina.oji@liu.se and jenny.kunz@liu.se

Abstract

This paper investigates the optimal use of the multilingual encoder model mDeBERTa for tasks in three Germanic languages – German, Swedish, and Icelandic – representing varying levels of presence and likely data quality in mDeBERTas pre-training data. We compare full fine-tuning with the parameter-efficient fine-tuning (PEFT) methods LoRA and Pfeiffer bottleneck adapters, finding that PEFT is more effective for the higher-resource language, German. However, results for Swedish and Icelandic are less consistent. We also observe differences between tasks: While PEFT tends to work better for question answering, full fine-tuning is preferable for named entity recognition. Inspired by previous research on modular approaches that combine task and language adapters, we evaluate the impact of adding PEFT modules trained on unstructured text, finding that this approach is not beneficial.

1 Introduction

Massively multilingual encoder models like mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020) and mDeBERTa (He et al., 2021b) are a workhorse for NLP in many lower-resource languages. However, due to interference between languages (Conneau et al., 2020; Chang et al., 2023), these models can fall short of reaching their full potential for individual target languages: Monolingual models (Virtanen et al., 2019; Snæbjarnarson et al., 2022) and models with dedicated language modules (Pfeiffer et al., 2022; Blevins et al., 2024) frequently outperform them, raising the question for the best setups for different languages.

Parameter-efficient fine-tuning (PEFT) methods, such as bottleneck adapters (Houlsby et al., 2019), LoRA (Hu et al., 2022), and prefix tuning (Li and Liang, 2021), have emerged as an alternative to full fine-tuning of pre-trained language models. These methods preserve the model's representations and can lead to better generalisation (He et al., 2021c). This is especially relevant for multilingual models, trained on diverse data, of which the target language only constitutes a small fraction. Fully fine-tuning them on task-specific data risks overwriting some of the multilingual capabilities.

Language adapters – PEFT modules trained on unstructured text independently from task fine-tuning – have shown promise in cross-lingual transfer (Pfeiffer et al., 2020; Vidoni et al., 2020). We explore whether language adaptation modules are beneficial even in scenarios where cross-lingual transfer is *not* required, i.e., where we have in-language fine-tuning data. In addition, we use not only bottleneck (Pfeiffer) adapters but also LoRA (Hu et al., 2022), a method that has become popular for LLMs as its parameters can be merged with the model parameters, adding no inference overhead.

In this paper, we investigate strategies for adapting a multilingual encoder model to task data in three languages: German, Swedish, and Icelandic. For this, we use multilingual DeBERTa (He et al., 2021b), which is currently the best-performing model according to the ScandEval (Nielsen et al., 2024) leaderboard for Icelandic,[1] the lowest-resourced and thus the most challenging of the three languages.

Our findings indicate that the effectiveness of full fine-tuning versus PEFT varies by language. For German, a PEFT method consistently delivers the best results, although sometimes with marginal gains. For Swedish and Icelandic, the performance is task-dependent: PEFT is more beneficial for extractive question-answering (QA), while full fine-tuning works better for named entity recognition

[1] https://scandeval.com/icelandic-nlu/, as of 21/10/2024.

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 433–439
March 3-4, 2025

(NER). We hypothesise that in languages with quantitatively more limited or lower-quality representation in the pre-training data, there is less value in preserving the pre-existing representations and more value in increasing the learning capacity. In contrast, for higher-resource languages, capabilities from the pre-training phase are more impactful. Similarly, for extractive QA, pre-existing skills weigh higher, while the highly specific nature of NER benefits from full fine-tuning.

Language adapters do not provide consistent improvements in any of the tasks or languages tested. As the adaptation data we use has likely been used for pre-training the multilingual DeBERTa model, we conclude that the utilisation of this data at pre-training time has already been effective enough. Further adaptation, or specialisation, with this same data does not have a clear benefit.

2 Related Work

PEFT methods not only reduce the number of trainable parameters and, consequently, memory usage in comparison to full fine-tuning, but there is also evidence suggesting that they provide better regularisation and help preserve pre-existing model capabilities. For example, He et al. (2021c) demonstrate that adapter-based fine-tuning outperforms full fine-tuning in cross-lingual transfer setups, likely by avoiding overfitting on the source language. Similarly, prefix tuning, another PEFT method, has been shown to surpass full fine-tuning in extrapolation scenarios (Li and Liang, 2021).

Other works have shown the effectiveness of bottleneck-style adapters in cross-lingual transfer as post-hoc trained language modules in encoder models. Pfeiffer et al. (2020) show that bottleneck language adapters in the Pfeiffer architecture improve performance in NER, commonsense classification, and extractive QA. Even Vidoni et al. (2020) report that language adapters are effective. Other research indicates that language adapters can aid in transferring knowledge to dialectal variants (Vamvas et al., 2024) and that sharing adapters across related languages can be beneficial (Faisal and Anastasopoulos, 2022; Chronopoulou et al., 2023). However, the success of language adapters may be task-specific and difficult to measure accurately when using machine-translated evaluation data (Kunz and Holmström, 2024). And notably, none of the works used multilingual DeBERTa models, which may explain divergences in results.

3 Experimental Setup

3.1 Model

We use the multilingual DeBERTa v3 model[2] as the base for our experiments. This model contains about 86 million parameters in its backbone, and the embedding layer, with a vocabulary of 250,000 tokens, adds another 190 million parameters, bringing the total to around 278 million parameters(He et al., 2021a). It was trained on 2.5 TB of the CC100 multilingual dataset (Wenzek et al., 2020; Conneau et al., 2020), which includes 100 languages, including Icelandic, Swedish, and German.

3.2 Tasks

We evaluate the fine-tuning and language adaptation methods on three tasks: extractive question answering (QA), named entity recognition (NER), and linguistic acceptability classification. This selection is inspired by coverage in the ScandEval benchmark (Nielsen et al., 2024) for all three languages while having structurally different tasks.

QA: For Icelandic, we use the *Natural Questions in Icelandic (NQiI)* dataset, which features questions from Icelandic texts written by Icelandic speakers.(Snæbjarnarson and Einarsson, 2022). For Swedish, we use the Swedish portion of ScandiQA, which was manually translated from English (Nielsen, 2023). For German, we use the human-labeled GermanQuAD dataset, which is natively German. (Möller et al., 2021).

NER: For Icelandic, we use the MIM-GOLD-NER dataset (Ingólfsdóttir et al., 2020), for Swedish, we use the Stockholm-Umeå Corpus (Kurtz and Öhman, 2022) and for German, we use GermanEval 2014 (Benikova et al., 2014).

Linguistic Acceptability: For all three languages, we use the respective portion of ScaLA (Nielsen, 2023), a binary classification dataset that judges the linguistic acceptability of sentences. Sentences are tagged as either grammatically correct or incorrect. This dataset is synthetically created by introducing corruptions based on the dependency trees of the sentences.

3.3 PEFT Methods

We use two different PEFT methods. **Pfeiffer adapters** (Pfeiffer et al., 2021, 2020) are a vari-

[2]loaded from `https://huggingface.co/microsoft/mdeberta-v3-base`

ation of bottleneck adapters (Houlsby et al., 2019), that is, small feed-forward layers that reduce the dimensionality of the input, process it, and then expand it back to the original size. They are inserted between the layers of the transformer model, and are the only parameters that are trained. **LoRA** (Hu et al., 2022) approximates the original weight updates as a low-rank decomposition by learning two low-rank matrices. Instead of updating the full set of model parameters, LoRA inserts trainable low-rank matrices into the self-attention of each layer of the model and updates only those.

3.4 PEFT Training

In the first step, we fine-tune individual *language adapters* for Icelandic, Swedish, and German, using the masked language modeling objective. We use 250,000 samples from the CC100 dataset and train a LoRA and a Pfeiffer language adapter for each language. Our language adapters are available at `https://huggingface.co/rominaoji`.

Task adapters are fine-tuned on target-language task data with the datasets described in Section 3.2.

For all adapters, we set the LoRA rank to 8 and the α to 16, while for the Pfeiffer method, the reduction factor is set to 16. For the implementation, we use the *adapters* library (Poth et al., 2023).

3.5 Setups

To find the optimal method to use mDeBERTa for the three languages, we fine-tune it using three setups: (1) **Full fine-tuning**, (2) tuning using only **task adapters**, and (3) using a **combination of language and task adapters** as in the MAD-X framework. In each setup, models are fine-tuned over five epochs.

As PEFT models require higher learning rates than full fine-tuning due to their lower number of trainable parameters, we determine a suitable rate for each setup by testing learning rates from 1-e4 to 9e-4 for PEFT and from 1e-5 to 9e-5 for full fine-tuning. This resulted in a learning rate of 3e-4 for both the language and task adaptation methods and 2e-5 for full fine-tuning. All experiments use a linear scheduler paired with the AdamW optimiser(Loshchilov, 2017). The code is available at `https://github.com/rominaoji/german-language-adapter`.

3.6 Evaluation

For the sake of simplicity, we only present F1 scores as the evaluation metric for all three tasks in this paper. While we have collected results on more metrics, we did not observe differences in the trends. The results are the mean of a five-fold cross-validation, with standard deviation.

4 Results and Discussion

All results are presented in Table 1. We discuss the effects of different task fine-tuning strategies on different languages and tasks (§4.1) and finally the effect of language adapters (§4.2).

4.1 Full Fine-Tuning Versus PEFT

Tasks: For the extractive QA tasks, we observe that PEFT methods generally outperform full fine-tuning. In German, there is a notable gap between full fine-tuning and both PEFT methods, with LoRA yielding the best results. For Icelandic, Pfeiffer adapters outperform both full fine-tuning and LoRA. For Swedish, the differences between setups are minimal. We hypothesise that for this task, the model benefits from the pre-trained representations and does not require the highest possible learning capacity to identify relevant text spans in these tasks.

In contrast, full fine-tuning is the best approach for NER tasks, outperforming the highest-performing PEFT method in Icelandic and Swedish, and performing on par with Pfeiffer adapters in German. This suggests that for this word-level task, a larger learning capacity is more crucial than preserving fine-grained capabilities from pre-training.

For ScaLA, the results are mixed. Full fine-tuning yields slightly higher scores for Icelandic, while Pfeiffer adapters perform marginally better for Swedish and German. Interpreting the performance on this task is challenging, as the dataset contains some corrupted instances that may be detectable with simple pattern-matching, while others require more fine-grained linguistic knowledge.

Languages: For German, Pfeiffer adapters consistently outperform full fine-tuning in QA and ScaLA tasks, and are either on par or slightly better for NER. LoRA performs best for QA but yields lower scores in the other two tasks. This suggests that German benefits from keeping the base model intact, likely due to its relatively large representation in the pre-training dataset.

		QA			NER			ScaLA		
TA	LA	Icelandic	Swedish	German	Icelandic	Swedish	German	Icelandic	Swedish	German
Full FT	-	57.52 ± 1.50	**35.08 ± 0.77**	73.56 ± 0.78	***92.35 ± 0.31***	***87.47 ± 0.41***	84.83 ± 0.33	**76.17 ± 1.32**	84.23 ± 1.07	83.90 ± 0.82
Pfeiffer	-	**59.31 ± 1.14**	***35.15 ± 1.00***	75.84 ± 0.92	91.37 ± 0.23	86.64 ± 0.51	***85.14 ± 0.22***	***76.35 ± 0.56***	**84.94 ± 1.04**	***84.53 ± 0.64***
LoRA	-	57.65 ± 2.11	34.76 ± 0.82	***77.17 ± 0.74***	89.69 ± 0.49	85.16 ± 0.32	84.12 ± 0.31	70.64 ± 2.78	82.32 ± 1.80	78.75 ± 2.34
Pfeiffer	Pfeiffer	***60.02 ± 1.46***	**35.07 ± 0.78**	76.66 ± 0.55	91.41 ± 0.23	**86.72 ± 0.28**	84.77 ± 0.38	75.38 ± 1.21	84.68 ± 0.76	**84.02 ± 0.64**
LoRA	Pfeiffer	57.44 ± 1.61	34.77 ± 0.60	**77.13 ± 0.28**	89.95 ± 0.50	85.11 ± 0.30	84.05 ± 0.35	71.31 ± 2.30	82.58 ± 2.01	78.85 ± 2.28
Pfeiffer	LoRA	59.24 ± 0.60	34.97 ± 0.82	76.31 ± 0.63	**91.49 ± 0.29**	86.50 ± 0.60	**85.08 ± 0.22**	75.06 ± 1.41	***84.98 ± 1.23***	83.86 ± 0.30
LoRA	LoRA	57.05 ± 1.74	34.40 ± 0.40	77.02 ± 0.35	89.64 ± 0.43	85.11 ± 0.30	84.08 ± 0.20	71.01 ± 3.00	82.97 ± 1.78	78.94 ± 2.37

Table 1: Mean F1 scores over five runs with standard deviation for all tasks and languages. The first column specifies the task adaptation method (TA), and the second one the language adaptation method (LA). The respectively highest score is highlighted in bold blue italics, the runner-up in bold black.

For Swedish, the performance of full fine-tuning and Pfeiffer adapters is similar across all three tasks, showing little variation.

For Icelandic, Pfeiffer adapters achieve higher scores in QA, while full fine-tuning performs better for NER. For ScaLA, both approaches produce comparable results. Icelandic's low representation in the CC-100 dataset used to train mDeBERTa might explain why it benefits less from the model's pre-training than German. While Swedish even has a slightly larger quantitative representation than German in open CC100 dumps,[3] it is unclear if the quality of the Swedish data matches that of the German data. For lesser-resourced languages, the quality of common-crawl corpora is often lower (Kreutzer et al., 2022; Artetxe et al., 2022), which may diminish the usefulness of pre-training for Swedish compared to German. Swedish has 13M speakers (10M L1),[4] whereas German has 175M speakers (95M L1),[5] which probably makes German higher-resource than Swedish, and may lead to higher-quality representation of German.

PEFT Methods: Except for German QA, Pfeiffer adapters outperform LoRA across all tasks. This may be due to architectural differences, though it is worth noting that Pfeiffer adapters in our setup have a higher learning capacity, with 896K trainable parameters compared to LoRA's 296K. Additionally, LoRA may require more extensive hyperparameter tuning than Pfeiffer adapters, as previous studies have shown its behavior to be unstable under certain conditions (Liu et al., 2024). A deeper exploration of how to improve LoRA's adaptation is left for future work.

[3] See e.g. `https://huggingface.co/datasets/statmt/cc100` as of 23/10/2024.

[4] `https://en.wikipedia.org/wiki/Swedish_language` as of 23/10/2024.

[5] `https://en.wikipedia.org/wiki/German_language` as of 23/10/2024.

4.2 Language Adaptation

Language adapters do not provide any significant benefits. When using Pfeiffer task adapters, performance remains similar whether language adapters are included or not. The only exception is Icelandic QA, where the combination of a Pfeiffer language adapter and a Pfeiffer task adapter achieves a slightly higher score compared to the best setup without language adapters. However, the difference is small and possibly due to result variability, as it falls within a standard deviation.

With LoRA task adapters, language adaptation methods sometimes result in a noticeable performance drop, suggesting potential interference. While prior work, such as Pfeiffer et al. (2020), reported improvements in similar tasks, their study focused on cross-lingual transfer, where no task data from the target language was available. In contrast, our setups use task data from the target language, and all the languages are present in the model's pre-training data. In addition, we use mDeBERTa-v3, which reportedly performs better for the languages in question than the XLM-R (Conneau et al., 2020) and multilingual BERT (Devlin et al., 2019) models that most other papers including Pfeiffer et al. (2020) use. These factors likely contribute to the fact that language adapters are unnecessary in our setup.

5 Conclusion

We compared the performance of the multilingual encoder model mDeBERTa across three task adaptation setups: full fine-tuning, bottleneck (Pfeiffer) adapters, and LoRA. Based on our evaluations across three tasks and three languages, we found that the choice of the best method is both task- and language-dependent. Specifically, extractive QA tasks benefit from PEFT methods, while NER gets better results with full fine-tuning. For Ger-

man, a higher-resourced language, PEFT consistently achieves higher scores. This suggests that the model benefits from fine-grained information learned during pre-training if coverage and (or) quality of the language data in the pre-training corpus are sufficiently high. In contrast, for lower-resourced languages, the increased learning capacity of full fine-tuning proves more advantageous.

We also tested language adaptation with Pfeiffer adapters and LoRA on unstructured text data before task adaptation. However, language adapters did not show any benefit. Access to target-language task data appears to dispense with the need for them, at least in our experiments where all languages are included in the pre-training data.

In future work, we aim to further explore the conditions under which PEFT methods versus full fine-tuning are most effective. We plan to investigate additional PEFT methods and tasks and optimise the LoRA setup, which may not have reached its full potential in our experiments.

Acknowledgments

We thank our colleagues Kevin Glocker, Kätriin Kukk, Julian Schlenker, Marcel Bollmann and Noah-Manuel Michael 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 research was supported by TrustLLM funded by Horizon Europe GA 101135671 and the National Graduate School of Computer Science in Sweden (CUGS). It was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The computations were enabled by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.

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.

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. Nosta-d named entity annotation for german: Guidelines and dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2524–2531, Reykjavik, Iceland. European Language Resources Association (ELRA).

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A. Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models.

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

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2023. Language-family adapters for low-resource multilingual neural machine translation. In *Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 59–72, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. 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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention.

Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jiawei Low, Lidong Bing, and Luo Si. 2021c. 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.

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*.

Svanhvít Lilja Ingólfsdóttir, Ásmundur Alma Guðjónsson, and Hrafn Loftsson. 2020. MIM-GOLD-NER – named entity recognition corpus (20.06). CLARIN-IS.

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.

Jenny Kunz and Oskar Holmström. 2024. The impact of language adapters in cross-lingual transfer for NLU. In *Proceedings of the 1st Workshop on Modular and Open Multilingual NLP (MOOMIN 2024)*, pages 24–43, St Julians, Malta. Association for Computational Linguistics.

Robin Kurtz and Joey Öhman. 2022. The kblab blog: Sucx 3.0 - ner.

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.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Timo Möller, Julian Risch, and Malte Pietsch. 2021. Germanquad and germandpr: Improving non-english question answering and passage retrieval. *arXiv preprint arXiv:2104.12741*.

Dan Saattrup Nielsen. 2023. Scandeval: A benchmark for scandinavian natural language processing. *arXiv preprint arXiv:2304.00906*.

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.

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

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

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.

Vésteinn Snæbjarnarson and Hafsteinn Einarsson. 2022. Natural questions in Icelandic. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4488–4496, Marseille, France. European Language Resources Association.

Vésteinn Snæbjarnarson, Haukur Barri Símonarson, Pétur Orri Ragnarsson, Svanhvít Lilja Ingólfsdóttir, Haukur Jónsson, Vilhjalmur Thorsteinsson, and Hafsteinn Einarsson. 2022. A warm start and a clean crawled corpus - a recipe for good language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4356–4366, Marseille, France. European Language Resources Association.

Jannis Vamvas, Noëmi Aepli, and Rico Sennrich. 2024. Modular adaptation of multilingual encoders to written Swiss German dialect. In *Proceedings of the 1st Workshop on Modular and Open Multilingual NLP (MOOMIN 2024)*, pages 16–23, St Julians, Malta. Association for Computational Linguistics.

Marko Vidoni, Ivan Vulić, and Goran Glavaš. 2020. Orthogonal language and task adapters in zero-shot cross-lingual transfer.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish.

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.