Intriguing Properties of Compression on Multilingual Models

Kelechi Ogueji* University of Waterloo kjogueji@uwaterloo.ca

Orevaoghene Ahia University of Washington oahia@cs.washington.edu

Gbemileke Onilude Cohere For AI Community lekeonilude@gmail.com

Sebastian Gehrmann Google Research gehrmann@google.com

Sara Hooker Cohere For AI sarahooker@cohere.com

Julia Kreutzer Google Research jkreutzer@google.com

Abstract

Multilingual models are often particularly dependent on scaling to generalize to a growing number of languages. Compression techniques are widely relied upon to reconcile the growth in model size with real world resource constraints, but compression can have a disparate effect on model performance for low-resource languages. It is thus crucial to understand the trade-offs between scale, multilingualism, and compression. In this work, we propose an experimental framework to characterize the impact of sparsifying multilingual pre-trained language models during fine-tuning. Applying this framework to mBERT named entity recognition models across 40 languages, we find that compression confers several intriguing and previously unknown generalization properties. In contrast to prior findings, we find that compression may improve model robustness over dense models. We additionally observe that under certain sparsification regimes compression may aid, rather than disproportionately impact the performance of low-resource languages.

1 Introduction

Scaling language models benefits multilingual settings, since it is difficult to maintain performance across a growing number of languages at a constant model size, a property also called the "curse of multilinguality" (Conneau and Lample, 2019; Conneau et al., 2020; Artetxe and Schwenk, 2019). However, the extent of growth in language model (LM) size (Radford et al., 2019; Brown et al., 2020; Zhang et al., 2022; Chowdhery et al., 2022) has made deployment to resource-constrained environments much more challenging (Warden and Situnayake, 2019; Samala et al., 2018; Treviso et al., 2022). To benefit from the performance gains conferred by scale, efficiency techniques that reduce model size while maintaining comparable aggregate performance are widely used, such as quantization (Shen et al., 2020), compression (Michel et al., 2019; Lagunas et al., 2021) and distillation (Tsai et al., 2019; Sanh et al., 2019; Pu et al., 2021).

While most compression techniques have minimal impact on aggregate performance numbers (Gale et al., 2019; Li et al., 2020; Hou et al., 2020; Chen et al., 2021; Bai et al., 2020; ab Tessera et al., 2021), the impact on individual sub-populations in the data, such as low-resource languages, can be far more severe (Hooker et al., 2019; Hooker et al., 2020; Ahia et al., 2021). Disparities in resource availability become more apparent at larger scale, both in terms of data and deployment resource availability. This makes compression all the more necessary, but also motivates a thorough consideration of the subsequent impact of compression on generalization.

In this work, we develop an experimental framework to investigate *the impact of compression during fine-tuning of pre-trained multilingual models* which we apply to Named Entity Recognition (NER) across 40 languages of the WikiAnn benchmark (Pan et al., 2017). We study the impact of compression on groups of languages across multiple dimensions—resourcedness, script, and language family—and evaluate the sensitivity of models to input perturbations along these groupings.

This leads us to discover the following *intriguing properties*: (1) Lower-performing languages disproportionately suffer under extreme levels of sparsity, as pruning amplifies disparities. However, low-resource languages present an intriguing *flip-flop* moment, where their performance may benefit from medium regimes of sparsity. (2) We find that dense models overfit to typical test cases, achieving a close-to-0 F1 score on slightly perturbed inputs, while compression can recover close to the original test performance. Our results stand in contrast to previous work that find that sparsity erodes robust-

*Work was done while at Google Research.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9092–9110

ness, suggesting more work is needed to understand the dynamics between compression and robustness. (3) The choice to prune model embeddings can completely negate the two benefits described in the previous observations, showing the importance of comparing the two cases in future analyses.

2 Related Work

The "curse-of-multilinguality" creates a trade-off between number of languages and size of a model (Conneau et al., 2020). However, training smaller models supporting fewer languages may not always be feasible (Abdaoui et al., 2020). Compressing large models has been shown to combat the curse, either by compressing the pre-trained model (Tsai et al., 2019; Sanh et al., 2019), or by compressing during finetuning, as in our case. While many studies investigate the impact of pruning on aggregate metrics in monolingual pre-trained LMs (Sanh et al., 2020; Goyal et al., 2020; Gordon et al., 2020; Budhraja et al., 2020; Sajjad et al., 2020; Lagunas et al., 2021; Xu et al., 2021; Du et al., 2021a; Ganesh et al., 2021), fewer works focus on multilingual settings (Mukherjee and Hassan Awadallah, 2020; Ansell et al., 2022). Yet, prior analyses find a disparate effect of removing attention heads or model layers on languages and language families distant from the training data in NER (Ma et al., 2021; Budhraja et al., 2021), demonstrating the importance of looking into sub-populations as we do in this study.

Studies that compare the robustness of compressed and dense models further find that compression may lead to erosion of performance on "challenging" samples and poor generalization (Ahia et al., 2021; Du et al., 2021a; Xu et al., 2021), a finding that we expand on and connect to language resourcedness. The technique we use to study robustness expands on studies that perturb training (Yaseen and Langer, 2021; Dai and Adel, 2020) or evaluation data (Dhole et al., 2021) in NER by introducing perturbations specific to languages, language families, and scripts.

3 Methodology

Data We conduct our experiments on WikiAnn (Pan et al., 2017), a multilingual NER dataset. WikiAnn was sourced from Wikipedia articles and automatically annotated with LOC (location), PER (person), and ORG (organisation) labels in the IOB2 format (Ramshaw and Marcus, 1995). It is considered a "silver standard" due to its automatic entity labels and noise (Lignos et al., 2022), but with its 176 languages it covers the most languages of any NER dataset. We focus our experiments on the 40 languages from the XTREME benchmark (Hu et al., 2020), with train-test splits defined by Rahimi et al. (2019). These training sets were built with stratified sampling to create a balance across entity types (Lignos et al., 2022), and are thus a subset of the total available data from the original WikiAnn. Table 1 lists language codes in ISO 639-1 and their available training data for fine-tuning.

# Sent.	Languages	Pretr. Token %
100	`jv`,`my`,`yo`	0.05
1000	`kk`,`sw`,`te`	0.19
5000	`af`,`hi`,`mr`	0.21
10000	`bn`,`eu`,`ka`,`ml`,`tl`	0.23
15000	`et`,`ta`	0.31
20000	`ar`,`bg`,`de`,`el`,`en`,`es`,`fa`,`fi` `fr`,`he`,`hu`,`id`,`it`,`ja`,`ko`,`ms` `nl`,`pt`,`ru`,`th`,`tr`,`ur`,`vi`,`zh`	2.93

Table 1: Data sizes and languages for WikiAnn and average representation for mBERT pre-training. The underlined languages are used for a comparison with monolingual fine-tuning.

Perturbations We test the robustness of compressed models by perturbing named entities in the test set. Previous work (Du et al., 2021a) show that sparse pretrained language models are less robust than their dense equivalents when evaluated on adversarial test sets, even when they perform similarly on in-distribution test sets. We adopt a data perturbation technique from Dai and Adel (2020) called *entity mention replacement*; an entity is randomly swapped with another entity of the same type (example sentences shown in App. D). We first perturb entities within same language for all the languages in our dataset (`in-language`). Secondly, we propose a new benchmark appropriate for testing the *cross-lingual robustness* of multilingual models on our downstream task. We perturb entities across different languages that share common linguistic properties. In particular, we group languages by family and script and perturb entities across languages within the same group (`in-script`, `in-family`).

Model We use the cased multilingual BERT (mBERT) (Devlin et al., 2019) for all our experi-

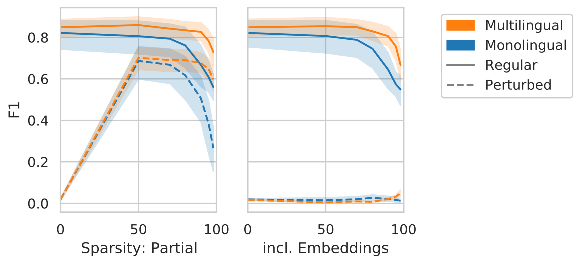

Figure 1: **Monolingual vs Multilingual:** F1 for monolingual and multilingual fine-tuning under regular and perturbed test conditions (`in-language`), averaged across languages (shaded areas: standard deviation).

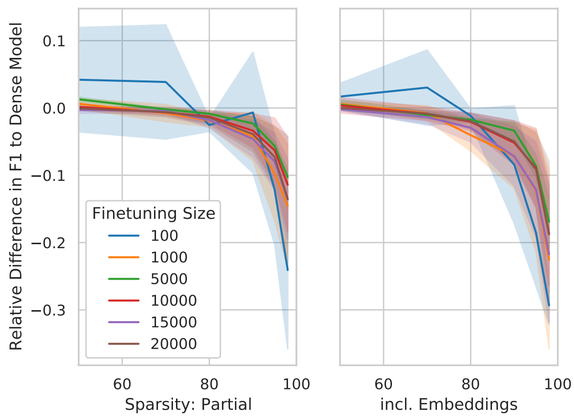

Figure 2: **Dense vs Sparse:** Mean relative difference in F1 for sparse multilingual models compared to the dense model. Results are averaged for languages grouped according to fine-tuning size.

ments because it is one of the most widely used and studied multilingual LMs (e.g., Pires et al., 2019; Rönnqvist et al., 2019; Wu and Dredze, 2019; Wang et al., 2020; Chi et al., 2020).[1] mBERT is trained on Wikipedia data from 104 languages, has approximately 177 million parameters, and a vocabulary size of around 120,000. We finetune mBERT by appending a linear classification layer to the model and updating all its parameters. Full hyperparameters are listed in App. A.1.

We evaluate the impact of sparsity in two settings: 1) In the `monolingual` setting we fine-tune on individual languages. For this setting, we select 10 languages with different available data size, language family and scripts (see Table 1). 2) In the `multilingual` setting we jointly fine-tune on all languages. We train all models with three random seeds, and evaluate F1 using `seqeval` (Nakayama, 2018) on the individual languages' evaluation data. We report mean results across runs after computing the micro-average F1 scores across entity classes.

Pruning We induce sparsity by applying Iterative Magnitude Pruning (IMP) (Han et al., 2015, 2016) during fine-tuning. IMP iteratively removes weights that are below a certain threshold until a desired target sparsity is reached. IMP is widely used and competitive with far more compute intensive approaches (Gale et al., 2019; Gordon et al., 2020; Du et al., 2021b; Ganesh et al., 2021), while allowing us to sparsify to an exact level. We compare two pruning strategies: 1) `partial` where we prune all dense layers except for embedding layers, 2) `incl. embeddings` where we prune all dense weights including embedding layers. Embeddings make up more than half (91M) of the 177M parameters in mBERT, while dense weights make up the rest. Hence, pruning embeddings allows us to significantly reduce the number of mBERT parameters. We consider five sparsity levels: 50%, 70%, 80%, 90%, 95% and 98%, corresponding to the percentage of weights pruned (hyperparameters in App. A.2. Preliminary experiments were conducted with lower sparsity levels (10%-40%) and yielded similar findings to those at moderate sparsity levels (50%-70%), motivating the sparsity intervals chosen. The chosen sparsity levels also align with general best practice in sparsity evaluation as presented in previous works. Moderate to high sparsity levels (50%+) are necessary for efficiency gains in the real-world and are usually studied in literature (Gale et al., 2019; Ahia et al., 2021; Ganesh et al., 2021).

4 Results and Discussion

4.1 Multilingual vs. Monolingual

Corroborating prior work on multilingual NER (Hu et al., 2020; Adelani et al., 2021), we find that the multilingual setting generally outperforms the monolingual one. Lower-resource languages tend to benefit more from crosslingual transfer.[2] We find that this finding holds under sparsity – multilingual models achieve higher F1 than monolingual models not only in the dense setting, but *across all sparsity levels*, as shown in Figure 1.[3] At high

[1] While XLM-R (Conneau et al., 2020) and others may perform better, the availability of fine-grained mBERT results through XTREME (Hu et al., 2020) allowed us to start from parameters that replicate the prior results.

[2] Kappa's $\tau = 0.39$ between multilingual gain and fine-tuning size for dense models.

[3] The three exceptions are Afrikaans (`af`) at 0% sparsity (−0.003), Hindi (`hi`) at 50% sparsity (−0.01) for partial prun-

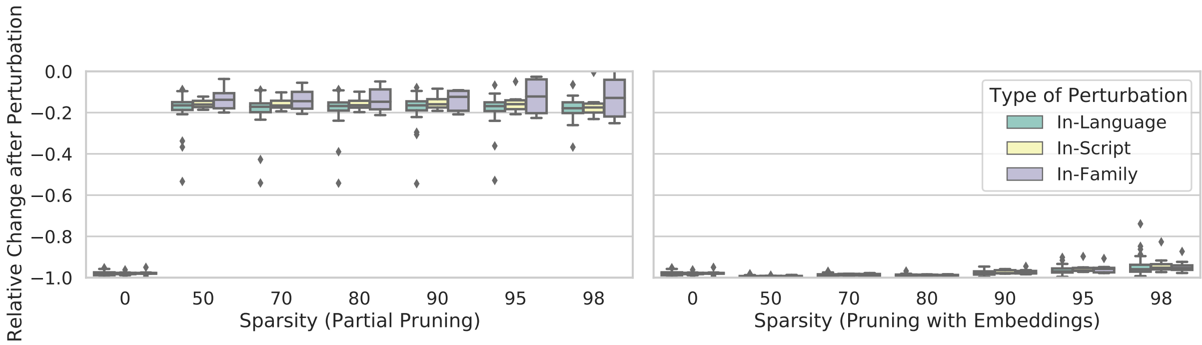

Figure 3: **Regular vs Perturbed:** We show the aggregated results across all languages after perturbation at different sparsity levels. Without pruning, the model performs poorly, which is overcome by partial pruning, but not pruning with embeddings. The relative performance drop is consistent across all pruning levels above 0.

sparsity levels, the loss in quality that is generally incurred is considerably lower for multilingual models. This suggests that when high levels of compression are necessary (e.g. for inference efficiency needs), *multilingual training should be preferred to monolingual training*, as it could help offset some of the erosion in the performance caused by the compression. Thus, we conclude that the benefits of *cross-lingual transfer are not inhibited by pruning*, and perhaps are even *more pronounced at a lower capacity* (Dufter and Schütze, 2020) for certain languages.

4.2 Impact of pruning across languages

Figure 2 displays the relative differences in F1 score between dense and sparse models across languages, grouped according to fine-tuning size.[4] At moderate sparsity levels (50%–70%), `partial` pruning surprisingly improves over the dense models, in particular those with less fine-tuning data. The majority of languages (26 out of 40) *benefit from moderate pruning* and yield slightly higher F1 with pruning than without. All three datasets with only 100 fine-tuning examples (`yo`, `my`, `jv`) benefit. This suggests that moderate pruning may benefit low-resource datasets when introduced during a finetuning regime. However, at high sparsity levels (70%–98%), the findings reverse. Those languages that have a lower frequency of representation in the finetuning dataset incur the highest absolute and relative loss in quality. We can observe the same trend when grouping languages according to their family or script, respectively (see Fig. 4 and 5 in App. B). The groups that start with the lowest average performance under the dense model, also suffer the most under extreme sparsity.

In conclusion, *moderate pruning levels should be explored for low-resource languages* since they may benefit such languages. This is especially important since models for low-resourced languages are often deployed in resource-constrained environments§ (Ahia and Ogueji, 2020; Nekoto et al., 2020; Ahia et al., 2021). Also, since *high sparsity levels reinforce existing disparities* (as measured by model performance and data availability) between languages and language groups, it is imperative that *practitioners pay attention to possible disparities when sparsifying models*.

4.3 How does pruning impact robustness?

Figure 3 shows the relative performance on the perturbed sets as a fraction of the corresponding unperturbed performance. Across all perturbation types, the dense model performs poorly, indicating that the model may have overfit to typical entities and the semantic context that appear in the training corpora.[5] Surprisingly, *partial pruning at any level* (shown left) improves upon the performance of the dense model. This finding disagrees with some prior works (Du et al., 2021a; Hooker et al., 2019; Sehwag et al., 2019) which find sparsity erodes different measures of robustness. However, the finding agrees with some other works. For example, Xu et al. (2021) found that pruning and post-training quantization improve BERT models' robustness to adversarial examples. Furthermore, Ahia et al.

ing, and Yoruba (yo) at 98% sparsity (−0.07) for pruning including the embeddings.

[4] A value of −0.1 means that this sparse model reaches 90% quality of the dense model, averaged across the languages within the same size bucket.

[5] Fig 9 shows that entity overlap between train and test set and model performance are correlated. This is particularly obvious for the highest (e.g., *(bn, ur, ms)*) and lowest performing languages (e.g., *(my, yo, jv)*). This may explain the poor performance of dense models on the perturbed test sets.

(2021) find that magnitude pruning improves model robustness to out-of-distribution shifts in machine translation. Despite the contradictions, our work represents an important step in understanding the impact of pruning on robustness, especially since we are one of the firsts to explore it multilingually. Interestingly, our findings are consistent across all perturbation types as their scope increases from languages (`in-language`) to scripts (`in-script`) and families (`in-family`). This suggests that sparsity can be explored as an avenue to improve robustness as has been explored in previous works (Xu et al., 2021; Ahia et al., 2021).

However, pruning the embeddings makes a crucial difference for the perturbed test cases. While pruning the embeddings does not matter for regular test set (see Figure 2), we observe the same severe drop in performance on the perturbed test-set as for the dense model. This suggests that including *model embeddings when pruning sharply erodes performance* on out-of-distribution rare artefacts, prompting a closer look into what is pruned in the embedding space and the *potential impact of sparsifying different parts of a model*.

5 Conclusion

This work investigates the effects of compression on multilingual pre-trained language models during fine-tuning. Our analysis revealed several intriguing properties of pruning that should inform future work in this direction: (1) Pruning dense layers up to $\sim 70\%$ may improve quality for low-frequency examples in the data and enhance model robustness. (2) The decision to prune embeddings may have critical impact on model robustness to out-of-distribution performance. (3) While low-performing languages benefit from moderate pruning, they are disproportionately harmed when pruning more aggressively. Based on these intriguing properties, we also make several recommendations to machine learning practitioners.

Limitations

We detail the following potential limitations of our work:

Noisy dataset: Lignos et al. (2022) shed light on several quality issues of the WikiAnn dataset that we are treating as a gold standard. Our results might thus not adequately reflect NER performance that can be achieved with cleaner and human-annotated datasets, such as the MasakhaNER (Adelani et al., 2021) or SADiLaR (Eiselen, 2016). Since the perturbations are based on the WikiAnn labels, we might be amplifying the existing label noise for the perturbed test sets and as a result underestimate model quality on clean perturbed examples. We try to combat the randomness by averaging results across three separate runs, but any issues intrinsic to WikiAnn will likely impact all three.

Other Multilingual Models and Downstream tasks: Multilingual pre-trained models such as XLM-R (Conneau et al., 2020) might yield a better performance or show slightly different trends across languages (Adelani et al., 2021). Other downstream tasks, especially generation tasks, might tolerate different levels of sparsity, and also show different crosslingual transfer capabilities (Wu and Dredze, 2019; Hu et al., 2020). However, since fine-grained prior results on the same WikiAnn splits were not available to us, we restricted the analysis to mBERT where we could verify that we can replicate the results reported by XTREME.

Evaluation metrics: We use F1 as the sole evaluation metric and trust it to reflect quality adequately across languages. Human evaluation and the use of qualitative evaluation metrics might reflect the quality for individual languages better.

Unknown factors influencing performance: The absolute performance for a given language can be influenced by many factors including size, family and script, relatedness to other languages, and the inherent difficulty of the NER task and the evaluation examples, as studied in related works (e.g., Pires et al., 2019; Wu and Dredze, 2020; Shaffer, 2021; Adelani et al., 2021; Muller et al., 2021; Deshpande et al., 2021). As a result, it is impossible to identify the exact cause for all our observations and we have to partially rely on correlational analyses.

References

Kale ab Tessera, Sara Hooker, and Benjamin Rosman. 2021. Keep the gradients flowing: Using gradient flow to study sparse network optimization.

Amine Abdaoui, Camille Pradel, and Grégoire Sigel. 2020. Load what you need: Smaller versions of mutililingual BERT. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 119–123, Online. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. 2021. The low-resource double bind: An empirical study of pruning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3316–3333, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Orevaoghene Ahia and Kelechi Ogueji. 2020. Towards supervised and unsupervised neural machine translation baselines for nigerian pidgin. *ArXiv*, abs/2003.12660.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, Dublin, Ireland. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Haoli Bai, Wei Zhang, Lu Hou, Lifeng Shang, Jing Jin, X. Jiang, Qun Liu, Michael R. Lyu, and Irwin King. 2020. Binarybert: Pushing the limit of bert quantization. *ArXiv*, abs/2012.15701.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Aakriti Budhraja, Madhura Pande, Pratyush Kumar, and Mitesh M. Khapra. 2021. On the prunability of attention heads in multilingual bert. *ArXiv*, abs/2109.12683.

Aakriti Budhraja, Madhura Pande, Preksha Nema, Pratyush Kumar, and Mitesh M. Khapra. 2020. On the weak link between importance and prunability of attention heads. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3230–3235, Online. Association for Computational Linguistics.

Xiao-Han Chen, Yu Cheng, Shuohang Wang, Zhe Gan, Zhangyang Wang, and Jing jing Liu. 2021. Earlybert: Efficient bert training via early-bird lottery tickets. *ArXiv*, abs/2101.00063.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek B Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Oliveira Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311.

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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2021. When is BERT multilingual? isolating crucial ingredients for cross-lingual transfer. *CoRR*, abs/2110.14782.

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.

Kaustubh D Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Srivastava, Samson Tan, et al. 2021. Nl-augmenter: A framework for task-sensitive natural language augmentation. *arXiv preprint arXiv:2112.02721*.

Mengnan Du, Subhabrata Mukherjee, Yu Cheng, Milad Shokouhi, Xia Hu, and Ahmed Hassan Awadallah. 2021a. What do compressed large language models forget? robustness challenges in model compression. *CoRR*, abs/2110.08419.

Mengnan Du, Subhabrata Mukherjee, Yu Cheng, Milad Shokouhi, Xia Hu, and Ahmed Hassan Awadallah. 2021b. What do compressed large language models forget? robustness challenges in model compression. *ArXiv*, abs/2110.08419.

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

Roald Eiselen. 2016. Government domain named entity recognition for South African languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3344–3348, Portorož, Slovenia. European Language Resources Association (ELRA).

Trevor Gale, Erich Elsen, and Sara Hooker. 2019. The state of sparsity in deep neural networks. *CoRR*, abs/1902.09574.

Trevor Gale, Erich Elsen, and Sara Hooker. 2019. The State of Sparsity in Deep Neural Networks. *arXiv e-prints*, page arXiv:1902.09574.

Prakhar Ganesh, Yao Chen, Xin Lou, Mohammad Ali Khan, Yin Yang, Hassan Sajjad, Preslav Nakov, Deming Chen, and Marianne Winslett. 2021. Compressing large-scale transformer-based models: A case study on BERT. *Transactions of the Association for Computational Linguistics*, 9:1061–1080.

Mitchell Gordon, Kevin Duh, and Nicholas Andrews. 2020. Compressing BERT: Studying the effects of weight pruning on transfer learning. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 143–155, Online. Association for Computational Linguistics.

Saurabh Goyal, Anamitra R. Choudhury, Saurabh Raje, Venkatesan T. Chakaravarthy, Yogish Sabharwal, and Ashish Verma. 2020. Power-bert: Accelerating bert inference via progressive word-vector elimination. In *ICML*.

Song Han, Huizi Mao, and William J. Dally. 2016. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding.

Song Han, Jeff Pool, John Tran, and William J. Dally. 2015. Learning both weights and connections for efficient neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NeurIPS'15, pages 1135–1143, Cambridge, MA, USA. MIT Press.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. 2019. What Do Compressed Deep Neural Networks Forget? *arXiv e-prints*, page arXiv:1911.05248.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2020. Characterising bias in compressed models.

Lu Hou, Lifeng Shang, X. Jiang, and Qun Liu. 2020. Dynabert: Dynamic bert with adaptive width and depth. *ArXiv*, abs/2004.04037.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

François Lagunas, Ella Charlaix, Victor Sanh, and Alexander Rush. 2021. Block pruning for faster transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10619–10629, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bei Li, Ziyang Wang, H. Liu, Quan Du, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2020. Learning light-weight translation models from deep transformer. *ArXiv*, abs/2012.13866.

Constantine Lignos, Nolan Holley, Chester Palen-Michel, and Jonne Sälevä. 2022. Toward more meaningful resources for lower-resourced languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 523–532, Dublin, Ireland. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Weicheng Ma, Kai Zhang, Renze Lou, Lili Wang, and Soroush Vosoughi. 2021. Contributions of transformer attention heads in multi- and cross-lingual tasks. 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 1956–1966, Online. Association for Computational Linguistics.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems (NeurIPS)*, 32.

Subhabrata Mukherjee and Ahmed Hassan Awadallah. 2020. XtremeDistil: Multi-stage distillation for massive multilingual models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2221–2234, Online. 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.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

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

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is multilingual BERT fluent in language generation? In *Proceedings of the*

First NLPL Workshop on Deep Learning for Natural Language Processing, pages 29–36, Turku, Finland. Linköping University Electronic Press.

Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. 2020. Poor man's bert: Smaller and faster transformer models. *ArXiv*, abs/2004.03844.

Ravi K Samala, Heang-Ping Chan, Lubomir M Hadjiiski, Mark A Helvie, Caleb Richter, and Kenny Cha. 2018. Evolutionary pruning of transfer learned deep convolutional neural network for breast cancer diagnosis in digital breast tomosynthesis. *Physics in Medicine & Biology*, 63(9):095005.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Victor Sanh, Thomas Wolf, and Alexander Rush. 2020. Movement pruning: Adaptive sparsity by fine-tuning. In *Advances in Neural Information Processing Systems*, volume 33, pages 20378–20389. Curran Associates, Inc.

Vikash Sehwag, Shiqi Wang, Prateek Mittal, and Suman Jana. 2019. Towards compact and robust deep neural networks. *CoRR*, abs/1906.06110.

Kyle Shaffer. 2021. Language clustering for multilingual named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 40–45, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. 2020. Q-bert: Hessian based ultra low precision quantization of bert. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8815–8821.

Marcos Treviso, Tianchu Ji, Ji-Ung Lee, Betty van Aken, Qingqing Cao, Manuel R. Ciosici, Michael Hassid, Kenneth Heafield, Sara Hooker, Pedro H. Martins, André F. T. Martins, Peter Milder, Colin Raffel, Edwin Simpson, Noam Slonim, Niranjan Balasubramanian, Leon Derczynski, and Roy Schwartz. 2022. Efficient methods for natural language processing: A survey.

Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, Xin Li, and Amelia Archer. 2019. Small and practical BERT models for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3632–3636, Hong Kong, China. Association for Computational Linguistics.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

P. Warden and D. Situnayake. 2019. *TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers*. O'Reilly Media, Incorporated.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Ke Xu, Julian McAuley, and Furu Wei. 2021. Beyond preserved accuracy: Evaluating loyalty and robustness of BERT compression. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10653–10659, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Usama Yaseen and Stefan Langer. 2021. Data augmentation for low-resource named entity recognition using backtranslation. *ArXiv*, abs/2108.11703.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *ArXiv*, abs/2205.01068.

A Hyperparameters

A.1 Fine-tuning Hyperparameters

Train epochs: 60
Optimizer: AdamW (Loshchilov and Hutter, 2019)
Learning rate: 7e-5
Max sequence length: 512
Dropout: 0.1
Batch size:
Data size $\in \{100, 1000\}$: 8
Data size $\in \{5000\}$: 16
Data size $\in \{10000, 15000, 20000\}$: 16

A.2 Pruning Hyperparameters

Data size = 100:
pruning start step: 10
pruning end step: 60
pruning frequency: 10

Data size = 1000:
pruning start step: 100
pruning end step: 300
pruning frequency: 50

Data size $\in \{5000, 10000\}$:
pruning start step: 500
pruning end step: 1200
pruning frequency: 100

Data size = 15000:
pruning start step: 700
pruning end step: 1800
pruning frequency: 150

Data size = 20000:
pruning start step: 1000
pruning end step: 2400
pruning frequency: 200

B Additional Diagrams

Relative change for different groups of languages Figures 4 and 5 show the relative change in F1 compared to the dense model averaged across languages within the same family or with the same script, respectively, on the regular test set. Figures 6, 7 and 8 depict the corresponding results on the in-language perturbed test sets. Figure 9 shows the correlation between percentage entity overlap and F1 on dense multilingual models.

C Full Results

We present the results for individual languages on both the regular and perturbed test sets obtained via multilingual finetuning in tables 2, 3, 4 and 5

We present the results for individual languages on both the regular and perturbed test sets obtained via monolingual finetuning in tables 6, 7, 8 and 9

D Examples of Perturbed Test Sentences

We present examples of perturbed test sentences in the *in-language* setting for English (table 10) and Yoruba language table (11).

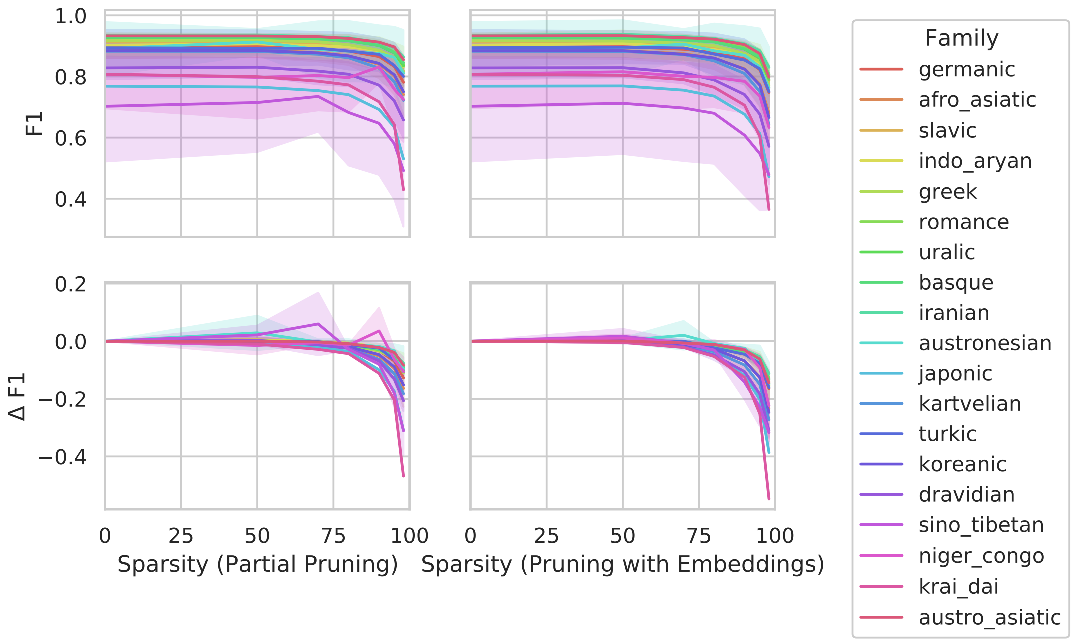

Figure 4: **Regular test:** Absolute F1 scores on top, relative differences in comparison to the dense model on the bottom. Results are averaged for languages grouped according to their *language families*. The shaded areas represent the standard deviation.

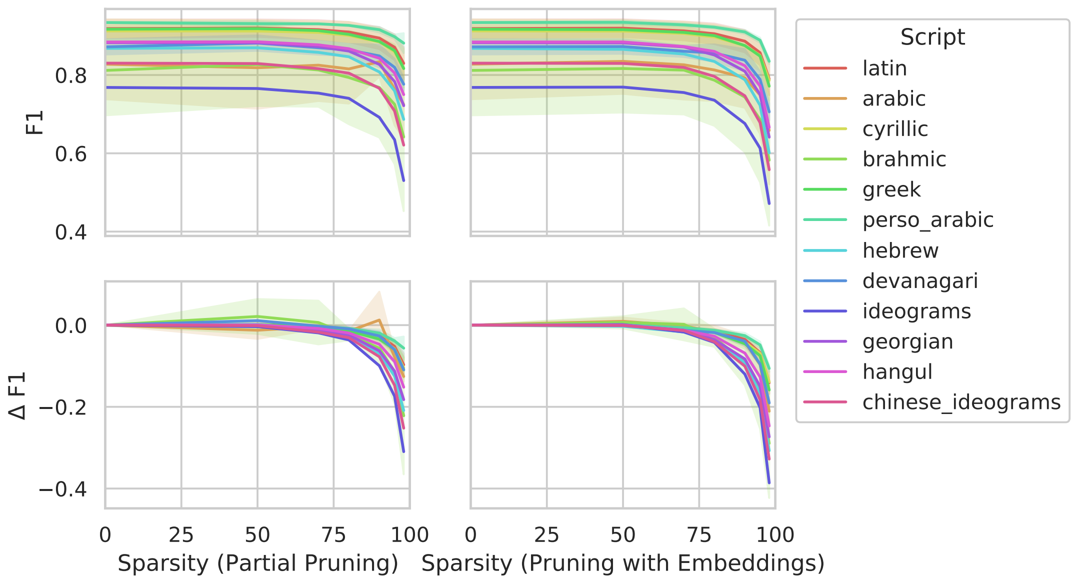

Figure 5: **Regular test:** Absolute F1 scores on top, relative differences in comparison to the dense model on the bottom. Results are averaged for languages grouped according to their *script*. The shaded areas represent the standard deviation.

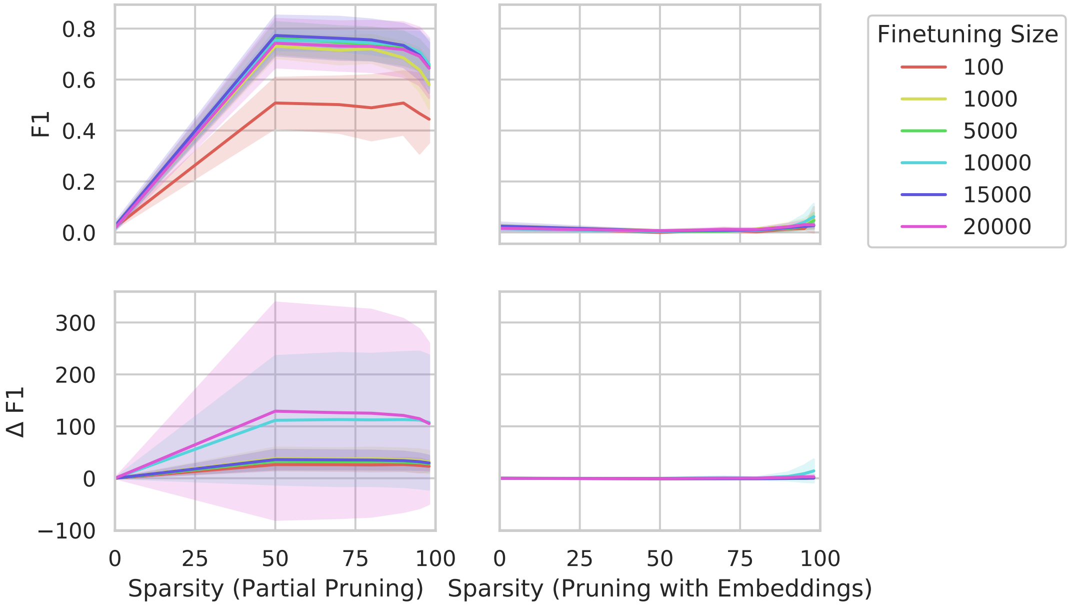

Figure 6: **In-language perturbation test:** Absolute F1 scores on top, relative differences in comparison to the dense model on the bottom. Results are averaged for languages grouped according to their *fine-tuning size*. The shaded areas represent the standard deviation.

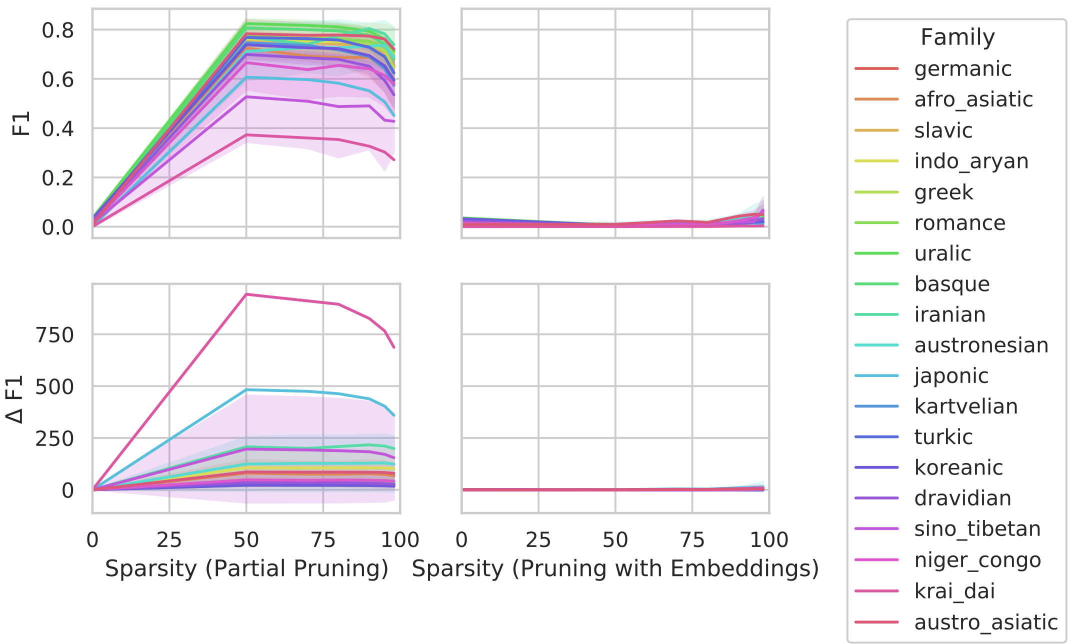

Figure 7: **In-language perturbation test:** Absolute F1 scores on top, relative differences in comparison to the dense model on the bottom. Results are averaged for languages grouped according to their *language families*. The shaded areas represent the standard deviation.

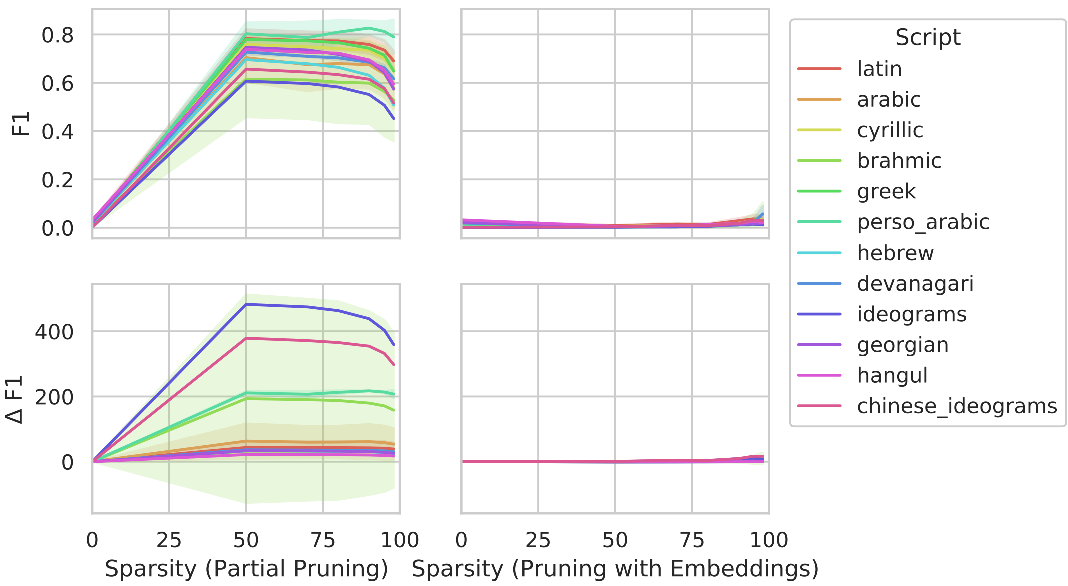

Figure 8: **In-language perturbation test:** Absolute F1 scores on top, relative differences in comparison to the dense model on the bottom. Results are averaged for languages grouped according to their *script*. The shaded areas represent the standard deviation.

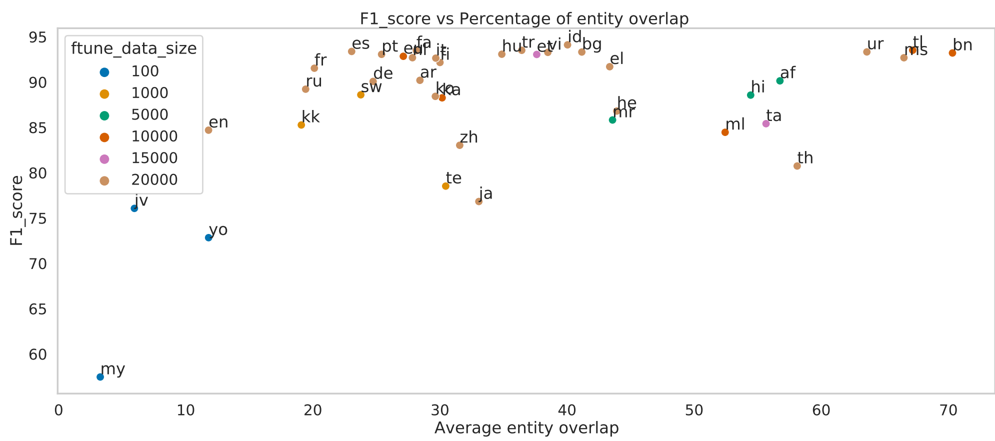

Figure 9: **Entity overlap:** Absolute F1 scores of dense multilingual model *vs* percentage overlap of entities between train and test set. The colors indicate the size of finetuning data per language.

languages	0	50	70	80	90	95	98
af	0.9014	0.9164	0.9002	0.8944	0.8874	0.8571	0.8204
ar	0.9020	0.9033	0.8983	0.8891	0.8719	0.8447	0.7943
bg	0.9332	0.9317	0.9287	0.9253	0.9090	0.8881	0.8524
bn	0.9321	0.9360	0.9326	0.9313	0.9191	0.9032	0.8760
de	0.9007	0.9021	0.8968	0.8886	0.8683	0.8375	0.7823
el	0.9170	0.9184	0.9133	0.9039	0.8856	0.8625	0.8164
en	0.8470	0.8481	0.8404	0.8305	0.8036	0.7654	0.6951
es	0.9338	0.9346	0.9295	0.9241	0.9139	0.8952	0.8612
et	0.9305	0.9311	0.9283	0.9212	0.9057	0.8844	0.8433
eu	0.9285	0.9271	0.9232	0.9165	0.9016	0.8788	0.8451
fa	0.9348	0.9356	0.9305	0.9270	0.9107	0.8938	0.8634
fi	0.9217	0.9210	0.9178	0.9103	0.8918	0.8666	0.8220
fr	0.9153	0.9154	0.9109	0.9069	0.8892	0.8618	0.8168
he	0.8681	0.8696	0.8578	0.8466	0.8078	0.7596	0.6866
hi	0.8857	0.8955	0.8823	0.8748	0.8692	0.8348	0.7880
hu	0.9308	0.9331	0.9290	0.9214	0.9057	0.8803	0.8396
id	0.9411	0.9401	0.9385	0.9335	0.9247	0.9094	0.8809
it	0.9265	0.9261	0.9224	0.9146	0.8992	0.8721	0.8280
ja	0.7683	0.7655	0.7537	0.7404	0.6918	0.6346	0.5303
jv	0.7607	0.8505	0.7509	0.7345	0.7375	0.7138	0.5987
ka	0.8827	0.8823	0.8727	0.8612	0.8270	0.7833	0.7216
kk	0.8527	0.8510	0.8544	0.8404	0.8343	0.7834	0.7522
ko	0.8843	0.8848	0.8771	0.8672	0.8428	0.8053	0.7499
ml	0.8446	0.8462	0.8365	0.8281	0.7972	0.7505	0.6805
mr	0.8582	0.8676	0.8572	0.8530	0.8284	0.8043	0.7650
ms	0.9269	0.9219	0.9361	0.9336	0.9099	0.8965	0.8679
my	0.5746	0.6003	0.6532	0.5594	0.5274	0.4516	0.3622
nl	0.9269	0.9264	0.9233	0.9188	0.9013	0.8709	0.8257
pt	0.9306	0.9335	0.9292	0.9252	0.9118	0.8918	0.8516
ru	0.8922	0.8930	0.8890	0.8770	0.8598	0.8317	0.7823
sw	0.8860	0.8924	0.8837	0.8751	0.8671	0.8530	0.8231
ta	0.8541	0.8486	0.8484	0.8319	0.7984	0.7607	0.7019
te	0.7853	0.7958	0.7678	0.7621	0.7192	0.6483	0.5907
th	0.8074	0.7993	0.7845	0.7724	0.7171	0.6424	0.4293
tl	0.9352	0.9389	0.9292	0.9289	0.9287	0.9300	0.8946
tr	0.9351	0.9338	0.9301	0.9256	0.9105	0.8887	0.8478
ur	0.9333	0.9269	0.9310	0.9266	0.9208	0.9018	0.8994
vi	0.9328	0.9326	0.9302	0.9247	0.9123	0.8966	0.8549
yo	0.7284	0.7015	0.7225	0.7172	0.7956	0.6635	0.6264
zh	0.8303	0.8293	0.8162	0.8048	0.7661	0.7080	0.6209
means	0.8795	0.8827	0.8764	0.8667	0.8492	0.8151	0.7622
medians	0.9017	0.9093	0.8993	0.8918	0.8788	0.8551	0.8166

Table 2: F1 scores for multilingual fine-tuning on the regular data for various levels of sparsity without pruning embedding layers.

languages	0	50	70	80	90	95	98
af	0.9014	0.9134	0.8960	0.8870	0.8810	0.8412	0.7878
ar	0.9020	0.9034	0.8955	0.8849	0.8624	0.8279	0.7593
bg	0.9332	0.9320	0.9270	0.9196	0.9018	0.8777	0.8222
bn	0.9321	0.9543	0.9359	0.9197	0.9029	0.8951	0.8078
de	0.9007	0.9006	0.8959	0.8854	0.8547	0.8227	0.7377
el	0.9170	0.9158	0.9089	0.9006	0.8752	0.8483	0.7714
en	0.8470	0.8491	0.8415	0.8283	0.7988	0.7604	0.6677
es	0.9338	0.9316	0.9275	0.9236	0.9078	0.8896	0.8377
et	0.9305	0.9305	0.9244	0.9172	0.8926	0.8642	0.7946
eu	0.9285	0.9260	0.9193	0.9128	0.8903	0.8640	0.8059
fa	0.9348	0.9379	0.9307	0.9244	0.9064	0.8843	0.8301
fi	0.9217	0.9202	0.9139	0.9067	0.8814	0.8505	0.7814
fr	0.9153	0.9147	0.9090	0.8983	0.8805	0.8525	0.7869
he	0.8681	0.8656	0.8537	0.8346	0.7880	0.7226	0.6012
hi	0.8857	0.8858	0.8709	0.8718	0.8578	0.8028	0.7212
hu	0.9308	0.9302	0.9257	0.9189	0.8943	0.8628	0.7952
id	0.9411	0.9400	0.9385	0.9342	0.9204	0.9014	0.8521
it	0.9265	0.9253	0.9214	0.9136	0.8941	0.8602	0.7886
ja	0.7683	0.7691	0.7552	0.7357	0.6761	0.6129	0.4716
jv	0.7607	0.7576	0.8329	0.7503	0.7273	0.6433	0.5623
ka	0.8827	0.8821	0.8718	0.8511	0.8096	0.7502	0.6412
kk	0.8527	0.8585	0.8567	0.8258	0.8053	0.7821	0.7140
ko	0.8843	0.8844	0.8727	0.8602	0.8238	0.7720	0.6660
ml	0.8446	0.8425	0.8261	0.8172	0.7695	0.7139	0.6184
mr	0.8582	0.8597	0.8504	0.8406	0.8178	0.7745	0.6905
ms	0.9269	0.9402	0.9198	0.9200	0.9091	0.8757	0.8229
my	0.5746	0.5948	0.5741	0.5627	0.4686	0.4160	0.3978
nl	0.9269	0.9266	0.9226	0.9151	0.8949	0.8648	0.7951
pt	0.9306	0.9318	0.9273	0.9216	0.9069	0.8831	0.8182
ru	0.8922	0.8923	0.8854	0.8750	0.8489	0.8220	0.7504
sw	0.8860	0.8880	0.8753	0.8659	0.8571	0.8332	0.7648
ta	0.8541	0.8512	0.8365	0.8161	0.7662	0.7078	0.6072
te	0.7853	0.7923	0.7725	0.7326	0.6867	0.6071	0.4890
th	0.8074	0.8039	0.7896	0.7646	0.7059	0.6036	0.3651
tl	0.9352	0.9360	0.9324	0.9310	0.9217	0.9041	0.8123
tr	0.9351	0.9330	0.9301	0.9208	0.9017	0.8677	0.7862
ur	0.9333	0.9313	0.9256	0.9200	0.9137	0.8934	0.8398
vi	0.9328	0.9334	0.9270	0.9222	0.9042	0.8745	0.7975
yo	0.7284	0.7426	0.7261	0.7279	0.7109	0.6368	0.5016
zh	0.8303	0.8296	0.8194	0.7964	0.7469	0.6782	0.5581
means	0.8795	0.8814	0.8741	0.8614	0.8341	0.7936	0.7105
medians	0.9017	0.9084	0.8960	0.8862	0.8688	0.8372	0.7681

Table 3: F1 scores for multilingual fine-tuning on the regular data for various levels of sparsity with pruning embedding layers.

languages	0	50	70	80	90	95	98
af	0.0314	0.8349	0.8193	0.8142	0.7988	0.7648	0.7240
ar	0.0060	0.7543	0.7046	0.7091	0.7426	0.7133	0.6623
bg	0.0237	0.7829	0.7712	0.7711	0.7702	0.7400	0.6911
bn	0.0055	0.7619	0.7489	0.7568	0.7620	0.7641	0.7289
de	0.0257	0.8019	0.7946	0.7849	0.7562	0.7187	0.6690
el	0.0230	0.7792	0.7737	0.7659	0.7429	0.7139	0.6481
en	0.0143	0.6843	0.6781	0.6645	0.6407	0.6128	0.5552
es	0.0119	0.7803	0.7666	0.7767	0.7790	0.7630	0.7207
et	0.0350	0.8283	0.8216	0.8125	0.7939	0.7635	0.7161
eu	0.0207	0.8065	0.7996	0.7945	0.7773	0.7403	0.6806
fa	0.0037	0.7696	0.7405	0.7744	0.8042	0.7827	0.7385
fi	0.0390	0.8398	0.8337	0.8264	0.8070	0.7722	0.7246
fr	0.0221	0.7632	0.7551	0.7528	0.7405	0.7157	0.6780
he	0.0201	0.6957	0.6788	0.6638	0.6310	0.5774	0.5076
hi	0.0196	0.7199	0.7073	0.7102	0.6765	0.6526	0.6200
hu	0.0316	0.8044	0.7952	0.7933	0.7793	0.7471	0.6948
id	0.0118	0.8038	0.7921	0.7979	0.7916	0.7801	0.7375
it	0.0226	0.7867	0.7756	0.7723	0.7511	0.7259	0.6807
ja	0.0013	0.6068	0.5967	0.5824	0.5513	0.5067	0.4518
jv	0.0161	0.5384	0.5771	0.5588	0.5972	0.5862	0.5102
ka	0.0216	0.7465	0.7356	0.7174	0.6901	0.6415	0.5734
kk	0.0242	0.7693	0.7667	0.7620	0.7208	0.6703	0.5889
ko	0.0324	0.7384	0.7259	0.7227	0.6946	0.6520	0.5940
ml	0.0215	0.6995	0.6962	0.6806	0.6653	0.6103	0.5482
mr	0.0192	0.7342	0.7113	0.6959	0.6931	0.6709	0.6129
ms	0.0094	0.7493	0.7642	0.7757	0.7597	0.7902	0.7403
my	0.0276	0.3975	0.3742	0.3414	0.3658	0.2884	0.3389
nl	0.0233	0.7759	0.7663	0.7628	0.7503	0.7149	0.6662
pt	0.0170	0.7586	0.7453	0.7452	0.7394	0.7138	0.6908
ru	0.0188	0.7349	0.7264	0.7116	0.6993	0.6621	0.6095
sw	0.0118	0.7434	0.7217	0.7415	0.7210	0.7015	0.6716
ta	0.0142	0.7174	0.7021	0.6987	0.6740	0.6276	0.5759
te	0.0304	0.6803	0.6564	0.6581	0.6143	0.5424	0.4819
th	0.0004	0.3727	0.3600	0.3537	0.3266	0.3028	0.2716
tl	0.0024	0.7526	0.7777	0.7707	0.7826	0.7873	0.7679
tr	0.0254	0.7667	0.7596	0.7530	0.7354	0.7095	0.6588
ur	0.0039	0.8362	0.8343	0.8449	0.8486	0.8417	0.8412
vi	0.0090	0.7831	0.7768	0.7779	0.7734	0.7612	0.7208
yo	0.0172	0.5882	0.5532	0.5675	0.5609	0.5259	0.4841
zh	0.0017	0.6567	0.6440	0.6336	0.6146	0.5758	0.5165
means	0.0179	0.7286	0.7182	0.7149	0.7031	0.6733	0.6273
medians	0.0194	0.7564	0.7471	0.7529	0.7399	0.7135	0.6642

Table 4: F1 scores for multilingual fine-tuning on the perturbed data for various levels of sparsity without pruning embedding layers.

languages	0	50	70	80	90	95	98
af	0.0314	0.0058	0.0076	0.0049	0.0243	0.0239	0.0228
ar	0.0060	0.0058	0.0076	0.0104	0.0134	0.0168	0.0202
bg	0.0237	0.0048	0.0070	0.0081	0.0129	0.0219	0.0239
bn	0.0055	0.0008	0.0028	0.0013	0.0113	0.0494	0.1111
de	0.0257	0.0055	0.0145	0.0113	0.0251	0.0305	0.0234
el	0.0230	0.0035	0.0082	0.0094	0.0120	0.0152	0.0192
en	0.0143	0.0082	0.0214	0.0128	0.0398	0.0490	0.0439
es	0.0119	0.0069	0.0149	0.0146	0.0274	0.0395	0.0410
et	0.0350	0.0072	0.0110	0.0123	0.0202	0.0251	0.0233
eu	0.0207	0.0061	0.0106	0.0097	0.0224	0.0266	0.0303
fa	0.0037	0.0038	0.0037	0.0075	0.0094	0.0280	0.0357
fi	0.0390	0.0051	0.0112	0.0116	0.0190	0.0219	0.0213
fr	0.0221	0.0102	0.0190	0.0131	0.0366	0.0457	0.0436
he	0.0201	0.0029	0.0073	0.0067	0.0141	0.0212	0.0242
hi	0.0196	0.0026	0.0024	0.0096	0.0155	0.0326	0.0815
hu	0.0316	0.0058	0.0087	0.0118	0.0166	0.0182	0.0177
id	0.0118	0.0112	0.0137	0.0082	0.0149	0.0247	0.0227
it	0.0226	0.0098	0.0164	0.0147	0.0317	0.0352	0.0332
ja	0.0013	0.0021	0.0059	0.0054	0.0130	0.0144	0.0115
jv	0.0161	0.0000	0.0156	0.0000	0.0098	0.0162	0.0042
ka	0.0216	0.0037	0.0073	0.0069	0.0119	0.0172	0.0190
kk	0.0242	0.0061	0.0048	0.0148	0.0137	0.0184	0.0219
ko	0.0324	0.0058	0.0075	0.0128	0.0178	0.0261	0.0210
ml	0.0215	0.0014	0.0028	0.0034	0.0063	0.0176	0.0255
mr	0.0192	0.0022	0.0033	0.0159	0.0066	0.0141	0.0332
ms	0.0094	0.0178	0.0286	0.0295	0.0489	0.0738	0.0586
my	0.0276	0.0000	0.0130	0.0078	0.0222	0.0104	0.1038
nl	0.0233	0.0074	0.0157	0.0131	0.0284	0.0313	0.0291
pt	0.0170	0.0106	0.0181	0.0158	0.0379	0.0515	0.0532
ru	0.0188	0.0072	0.0122	0.0103	0.0249	0.0374	0.0440
sw	0.0118	0.0112	0.0137	0.0141	0.0405	0.0555	0.0798
ta	0.0142	0.0048	0.0060	0.0065	0.0171	0.0224	0.0308
te	0.0304	0.0038	0.0083	0.0117	0.0154	0.0208	0.0437
th	0.0004	0.0003	0.0010	0.0009	0.0025	0.0026	0.0034
tl	0.0024	0.0075	0.0179	0.0118	0.0437	0.0892	0.1235
tr	0.0254	0.0043	0.0063	0.0090	0.0143	0.0174	0.0167
ur	0.0039	0.0018	0.0047	0.0028	0.0137	0.0269	0.0246
vi	0.0090	0.0095	0.0234	0.0175	0.0424	0.0504	0.0491
yo	0.0172	0.0000	0.0000	0.0000	0.0083	0.0187	0.0401
zh	0.0017	0.0031	0.0098	0.0083	0.0179	0.0309	0.0304
means	0.0179	0.0054	0.0103	0.0099	0.0206	0.0297	0.0377
medians	0.0194	0.0053	0.0085	0.0100	0.0168	0.0249	0.0297

Table 5: F1 scores for multilingual fine-tuning on the perturbed data for various levels of sparsity with pruning embedding layers.

languages	0	50	70	80	90	95	98
en	0.8468	0.8421	0.8283	0.7987	0.7049	0.5618	0.5592
zh	0.8299	0.8262	0.8057	0.7726	0.6490	0.4759	0.4159
bn	0.9284	0.9319	0.9205	0.9130	0.8619	0.7773	0.6028
eu	0.9236	0.9179	0.9084	0.8904	0.8264	0.7209	0.6641
af	0.9044	0.8970	0.8927	0.8878	0.7944	0.6800	0.6740
hi	0.8827	0.9083	0.8643	0.8357	0.7579	0.6267	0.5863
sw	0.8617	0.8541	0.8553	0.8496	0.7554	0.7017	0.6900
te	0.7687	0.7481	0.7383	0.6859	0.4619	0.4864	0.4667
jv	0.5478	0.5044	0.4976	0.3387	0.3210	0.3883	0.4025
yo	0.7207	0.6266	0.6246	0.6387	0.5439	0.6567	0.5374
means	0.8215	0.8057	0.7936	0.7611	0.6677	0.6076	0.5599
medians	0.8543	0.8481	0.8418	0.8172	0.7302	0.6417	0.5728

Table 6: F1 scores for monolingual fine-tuning on the regular data for various levels of sparsity without pruning embedding layers.

languages	0	50	70	80	90	95	98
en	0.8468	0.8382	0.8157	0.7828	0.6560	0.5730	0.5593
zh	0.8299	0.8255	0.7974	0.7488	0.6073	0.4329	0.4075
bn	0.9284	0.9421	0.9179	0.9033	0.8288	0.6957	0.5463
eu	0.9236	0.9189	0.9051	0.8822	0.7885	0.6631	0.6583
af	0.9044	0.8981	0.8807	0.8512	0.7464	0.6460	0.6632
hi	0.8827	0.8670	0.8612	0.8224	0.7198	0.5636	0.5977
sw	0.8617	0.8751	0.8508	0.8148	0.7158	0.6870	0.6980
te	0.7687	0.7592	0.7188	0.6253	0.4601	0.4951	0.4857
jv	0.5478	0.5123	0.5055	0.3758	0.3494	0.4369	0.2939
yo	0.7207	0.6292	0.6271	0.6480	0.5786	0.5428	0.5779
means	0.8215	0.8066	0.7880	0.7454	0.6451	0.5736	0.5488
medians	0.8543	0.8526	0.8332	0.7988	0.6859	0.5683	0.5686

Table 7: F1 scores for monolingual fine-tuning on the regular data for various levels of sparsity with pruning embedding layers.

languages	0	50	70	80	90	95	98
en	0.0230	0.7032	0.6841	0.6570	0.5543	0.4206	0.3337
zh	0.0055	0.6551	0.6492	0.6245	0.5477	0.4262	0.2884
bn	0.0138	0.8090	0.8000	0.7783	0.7106	0.6376	0.4981
eu	0.0180	0.7938	0.7782	0.7470	0.6502	0.5274	0.3788
af	0.0271	0.8260	0.8185	0.7960	0.6921	0.5562	0.4475
hi	0.0166	0.7289	0.7094	0.6852	0.5889	0.4672	0.2965
sw	0.0214	0.7326	0.7490	0.6785	0.5173	0.4733	0.3058
te	0.0229	0.6581	0.6095	0.5602	0.3482	0.2932	0.1049
jv	0.0223	0.4165	0.3449	0.2146	0.1439	0.0000	0.0000
yo	0.0187	0.5396	0.5371	0.4288	0.3087	0.0168	0.0000
means	0.0189	0.6863	0.6680	0.6170	0.5062	0.3818	0.2654
medians	0.0201	0.7160	0.6967	0.6678	0.5510	0.4467	0.3011

Table 8: F1 scores for monolingual fine-tuning on the perturbed data for various levels of sparsity without pruning embedding layers.

languages	0	50	70	80	90	95	98
en	0.0230	0.0634	0.0465	0.0418	0.0401	0.0351	0.0231
zh	0.0055	0.0115	0.0160	0.0209	0.0316	0.0213	0.0101
bn	0.0138	0.0000	0.0124	0.0124	0.0009	0.0055	0.0020
eu	0.0180	0.0060	0.0116	0.0146	0.0202	0.0185	0.0249
af	0.0271	0.0013	0.0084	0.0190	0.0132	0.0197	0.0262
hi	0.0166	0.0337	0.0382	0.0104	0.0007	0.0041	0.0184
sw	0.0214	0.0092	0.0541	0.0526	0.0469	0.0576	0.0210
te	0.0229	0.0007	0.0029	0.0070	0.0016	0.0000	0.0000
jv	0.0223	0.0212	0.0074	0.0114	0.0034	0.0000	0.0000
yo	0.0187	0.0000	0.0000	0.0782	0.0526	0.0000	0.0000
means	0.0189	0.0147	0.0198	0.0268	0.0211	0.0162	0.0126
medians	0.0201	0.0076	0.0120	0.0168	0.0167	0.0120	0.0143

Table 9: F1 scores for monolingual fine-tuning on the perturbed data for various levels of sparsity with pruning embedding layers.

	Example english test sentences
Original	Much construction was undertaken during this period , such as the building of Palermo Cathedral .
Perturbed	Much construction was undertaken during this period , such as the building of Knott 's Soak City .
Original	It is found in Peru .
Perturbed	It is found in Carbon Cliff , Illinois .
Original	Alberto Mancini won in the final 7–5 , 2–6 , 7–6 , 7–5 against Boris Becker.
Perturbed	John Jones (footballer , born 1895) won in the final 7–5 , 2–6 , 7–6 , 7–5 against Sultan Ahmad Shah .
Original	It flows from Ägerisee through Lake Zug into the Reuss .
Perturbed	It flows from New Orleans through Humboldt County , Nevada into the Crow Agency , Montana .
Original	The album 's lead single “ Better Believe It ” featuring Young Jeezy and Webbie , was released on July 14 , 2009 .
Perturbed	The album 's lead single “ Better Believe It ” featuring W. S. Merwin and Empress Maria Theresa , was released on July 14 , 2009 .

Table 10: Example of test sentences for English language using the *entity mention replacement* (Dai and Adel, 2020) technique where an entity is randomly swapped with another entity of the same type.

	Example yoruba test sentences
Original	Egbé Olóèlúaráìlú àwn Aráàlù (Nàìjíríà)
Perturbed	Ilé-ìgbìm Aòfin Oníbínibí il Nàìjíríà)
Original	Agbègbè Ìjba Ìbíl Wudil
Perturbed	Agbègbè Ìjba Ìbíl Gúúsù-Ìwòrùn Èkìtì Wudil
Original	Àgbáj àwn Oríl-èdè Aòkan
Perturbed	Àkój àwn olórí ìjba il Bùrkínà Fasò Aòka .
Original	Àsìá il Tufalu.
Perturbed	Abdulsalami Abubakar Tufalu
Original	’ ” ’ ” j Fáráò ni gíptì Ayéijun
Perturbed	’ ” ’ ” j Yousaf Raza Gillani ni Nàìjíríà .

Table 11: Example of test sentences for Yoruba language using the *entity mention replacement* (Dai and Adel, 2020) technique where an entity is randomly swapped with another entity of the same type.