small Models, BIG Impact: Efficient Corpus and Graph-Based Adaptation of Small Multilingual Language Models for Low-Resource Languages

Daniil Gurgurov[1,3] **Ivan Vykopal**[2,4] **Josef van Genabith**[3] **Simon Ostermann**[3]
[1]Saarland University [2]Brno University of Technology
[3]German Research Center for Artificial Intelligence (DFKI)
[4]Kempelen Institute of Intelligent Technologies (KInIT)
{daniil.gurgurov, josef.van_genabith, simon.ostermann}@dfki.de, ivan.vykopal@kinit.sk

Abstract

Low-resource languages (LRLs) face significant challenges in natural language processing (NLP) due to limited data. While current state-of-the-art large language models (LLMs) still struggle with LRLs, smaller multilingual models (mLMs) such as mBERT and XLM-R offer greater promise due to a better fit of their capacity to low training data sizes. This study systematically investigates parameter-efficient adapter-based methods for adapting mLMs to LRLs, evaluating three architectures: Sequential Bottleneck, Invertible Bottleneck, and Low-Rank Adaptation. Using unstructured text from GlotCC and structured knowledge from ConceptNet, we show that small adaptation datasets (e.g., up to 1 GB of free-text or a few MB of knowledge graph data) yield gains in intrinsic (masked language modeling) and extrinsic tasks (topic classification, sentiment analysis, and named entity recognition). We find that Sequential Bottleneck adapters excel in language modeling, while Invertible Bottleneck adapters slightly outperform other methods on downstream tasks due to better embedding alignment and larger parameter counts. Adapter-based methods match or outperform full fine-tuning while using far fewer parameters, and smaller mLMs prove more effective for LRLs than massive LLMs like LLaMA-3, GPT-4, and DeepSeek-R1-based distilled models. While adaptation improves performance, pre-training data size remains the dominant factor, especially for languages with extensive pre-training coverage. The code for our experiments is available at https://github.com/d-gurgurov/Knowledge-Driven-Adaptation-LLMs.

1 Introduction

The need for effective natural language processing (NLP) tools for low-resource languages (LRLs) is pressing, as these languages lack sufficient data to train robust models (Joshi et al., 2020; Bird, 2022; Huang et al., 2023). While **massive state-of-the-art (SoTA) large language models (LLMs)** such as GPT-4 (OpenAI et al., 2024), LLaMA-2 (Touvron et al., 2023), Gemini (Team et al., 2023), BLOOM (Le Scao et al., 2023), and the DeepSeek model family (DeepSeek-AI et al., 2025) have demonstrated strong generalization capabilities across diverse tasks (Srivastava et al., 2022; Smith et al., 2022; Bang et al., 2023), they struggle to generalize effectively to LRLs (Cahyawijaya et al., 2023; Robinson et al., 2023; Hasan et al., 2024; Adelani et al., 2024a). **Smaller multilingual language models (mLMs)** like mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) often show greater promise for LRLs (Hu et al., 2020; Asai et al., 2023; Adelani et al., 2024b).

This work investigates parameter-efficient adaptation techniques (Houlsby et al., 2019) as an alternative to full fine-tuning, or continued pre-training, for adapting small mLMs to LRLs. We compare these approaches with the zero- and few-shot prompting and adapter-based adaptation of LLMs. Following Pfeiffer et al. (2020), Parović et al. (2023), and Gurgurov et al. (2024a), we integrate unstructured textual data and structured knowledge from knowledge graphs (KGs), exploring their complementary benefits. KGs, which encode cross-lingual semantic relationships, have been shown to be effective for various NLP tasks (Peters et al., 2019; Zhang et al., 2019; Wang et al., 2021), yet remain underexplored for LRLs. On the other hand, unstructured text provides rich contextual information and is widely used for adaptation (Neubig and Hu, 2018; Han and Eisenstein, 2019).

Our contributions are threefold:

- First, we show that **limited adaptation data yields significant gains**—up to 1 GB of free text or a few MB of KG data. We eval-

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 355–395
July 28-29, 2025

uate three adapter architectures: Sequential Bottleneck, Invertible Bottleneck, and Low-Rank Adaptation (Houlsby et al., 2019; Pfeiffer et al., 2020; Hou et al., 2022). Sequential Bottleneck excels in language modeling, while Invertible Bottleneck outperforms others on downstream tasks, likely due to differing parameterization. Adapter-based approaches match or outperform full fine-tuning while using fewer trainable parameters.

- Second, **we highlight the effectiveness of smaller mLMs, such as XLM-R, for LRLs, outperforming both few-shot prompting and adaptation of massive SoTA LLMs** such as GPT-3.5 (Ouyang et al., 2022b), LLaMA- 3 (Grattafiori et al., 2024), and DeepSeek-R1-based distilled models (DeepSeek-AI et al., 2025) on the tested tasks. This is in line with prior work suggesting that smaller models better align cross-lingual representations under constrained capacity (Wu et al., 2019; Dufter and Schütze, 2020; Yong et al., 2023) and shows that small LMs are often better suited for LRLs.

- Finally, analyzing 30 LRLs, we **show a direct relationship between pre-training and adaptation data size and performance**, with adaptation data providing diminishing returns for languages with larger pre-training data coverage. We also observe a moderate correlation between language modeling and downstream task performance, suggesting pseudo-perplexity as a useful proxy for evaluating adaptation quality.

2 Related Work

To improve multilingual models for LRLs without monolingual pre-training, researchers have explored full fine-tuning, adapter-based approaches, and other auxiliary methods.

2.1 Full Fine-Tuning Adaptation

Full fine-tuning has been widely used to enhance LRL performance. Neubig and Hu (2018) utilized similar-language post-training to reduce overfitting. Domain-adaptive fine-tuning (Han and Eisenstein, 2019) improved contextualized models like mBERT on specific domains (e.g. Middle English). Further, language-specific fine-tuning on monolingual corpora (Gururangan et al., 2020; Chau et al., 2020) and adaptation with transliterated data (Muller et al., 2021) boosted performance on diverse tasks, such as dependency parsing and tagging. Ebrahimi and Kann (2021) showed that fine-tuning on Bible corpora improved tagging and named entity recognition in languages unseen during pre-training.

2.2 Adapter-Based Adaptation

Adapters are parameter-efficient small modules that are inserted into model layers, avoiding catastrophic forgetting (French, 1999), reducing computational costs (Houlsby et al., 2019; Strubell et al., 2019), and requiring fewer training examples (Faisal and Anastasopoulos, 2022). Frameworks like MAD-X (Pfeiffer et al., 2020) introduced language and task adapters, improving named entity recognition. Extensions such as UDapter (Üstün et al., 2020) and MAD-G (Ansell et al., 2021) leveraged typological features for improved zero-shot inference. Hierarchical adapters based on language phylogeny (Faisal and Anastasopoulos, 2022), methods addressing resource imbalances with language combination (Lee et al., 2022a; Parović et al., 2022), and exposing task adapters to target languages during training to address training-inference mismatches (Parović et al., 2023) have further advanced adapter effectiveness. Recent work (Pfeiffer et al., 2022; Yong et al., 2023) emphasized the efficiency of adapter-based tuning over continued pre-training for LRLs, with performance tied to data quantity.

2.3 Knowledge Graph Integration

KGs improve the quality of static word embeddings (Faruqui et al., 2014; Speer et al., 2017; Gurgurov et al., 2024b) and, more recently, LMs by leveraging structured semantic relationships, predominantly for high-resoure languages (Miller, 1995; Navigli and Ponzetto, 2012; Speer et al., 2017). Approaches like KnowBERT (Peters et al., 2019) and ERNIE (Zhang et al., 2019) improve LMs through entity linkers and attention. LIBERT (Lauscher et al., 2020b) incorporates semantic constraints for better task performance. CN-ADAPT (Lauscher et al., 2020a) and K-Adapter (Wang et al., 2021) use bottleneck adapters (Houlsby et al., 2019) to inject structured knowledge into models, improving commonsense reasoning and relational tasks.

3 Methodology

This section describes our approaches to adapting mLMs for LRLs and the data resources used.

3.1 Model Adaptation

We adapt mBERT (Devlin et al., 2019) and XLM-R-base (Conneau et al., 2020) using three adapter architectures: Sequential Bottleneck (`Seq_bn`; Houlsby et al. (2019); Pfeiffer et al. (2020)), Sequential Bottleneck with Invertible Layers (`Seq_bn_inv`; Pfeiffer et al. (2020)), and Low-Rank Adaptation (`LoRA`; Hou et al. (2022)). Additionally, we adapt LLaMA-3-8B (Grattafiori et al., 2024), but exclusively with `Seq_bn_inv` adapters (due to computational constraints). Language adapters are pre-trained with a masked language modeling (MLM) objective (Devlin et al., 2019) for mBERT and XLM-R on structured data (ConceptNet; Speer et al. (2017)) and unstructured data (GlotCC; Kargaran et al. (2024)).[1] Further, we pre-train language adapters for LLaMA-3 with a causal language modeling (CLM) objective (Radford, 2018), only with unstructured data, leaving the exploration of graph knowledge injection into large-scale LMs for future work.

Task-specific adapters are trained on target language data using the `Seq_bn` architecture. These adapters are stacked on "frozen" LMs and language adapters, following prior work (Pfeiffer et al., 2020; Lee et al., 2022a; Parović et al., 2023). We also experiment with adapter fusion (Pfeiffer et al., 2021a), combining language adapters trained on different data types.

3.2 Data Sources

Structured Data. ConceptNet (Speer et al., 2017), a multilingual knowledge graph, provides common-sense knowledge across 304 languages. We preprocess the data by converting ConceptNet triples into natural language sentences, similar to Lauscher et al. (2020a) and Gurgurov et al. (2024a), using predefined predicates (Appendix A), and split it into train and validation sets.

Unstructured Data. GlotCC-V1 (Kargaran et al., 2024) is a large-scale multilingual corpus derived from CommonCrawl (Wenzek et al., 2020). It emphasizes LRLs, providing high-quality text in 1,000 languages. To simulate a low-resource environment for all languages, we limit each language to 1 GB (if it exceeds this limit), clean the data, and split it into training and validation sets.

[1]Full fine-tuning is performed only on the GlotCC data for mBERT and XLM-R due to ConceptNet's limited size.

4 Experimental Setup

This section details the experimental setup, including language selection, evaluation tasks, and adapter training procedures.

4.1 Languages

We selected 30 LRLs identified by Joshi et al. (2020) as low-resource—representing a diverse set that includes *Thai, Romanian, Bulgarian, Danish, Greek, Hebrew, Slovak, Slovenian, Latvian, Indonesian, Georgian, Bengali, Azerbaijani, Urdu, Macedonian, Telugu, Nepali, Marathi, Swahili, Welsh, Uzbek, Javanese, Sundanese, Sinhala, Amharic, Kurdish, Uyghur, Maltese, Tibetan, and Yoruba*—to evaluate adapter performance across underrepresented linguistic contexts. Table 5 (Appendix B) summarizes language-specific details.

4.2 Language Adapter Training

Language adapters were trained on mBERT and XLM-R for all languages using MLM with GlotCC and ConceptNet data. We evaluated `Seq_bn`, `Seq_bn_inv`, and `LoRA`, with the default hyperparameters (Appendix F). For LLaMA-3-8B, only GlotCC data was used with the `Seq_bn_inv` architecture and CLM objective for a subset of 5 languages due to computational constraints. Training consisted of up to 100,000 steps for GlotCC and 25,000 steps for ConceptNet, with a batch size of 16 and learning rate of 1e-4.

4.3 Task-Specific Training

Adapters were evaluated on four tasks. For *Masked Language Modeling* (MLM), we used the FLORES-200 devtest set (Team et al., 2022), comprising 1012 parallel sentences, and measured pseudo-perplexity (Salazar et al., 2019) as a proxy for linguistic acceptability. *Topic Classification* (TC) employed the 7-class SIB-200 dataset (Adelani et al., 2024a), training task adapters on predefined splits (701 train, 99 validation, 204 test examples) and fixed hyperparameters (Appendix F), with F1 scores computed on the test set (Sokolova et al., 2006). For *Sentiment Analysis* (SA), binary-class datasets from multiple sources (Table 6 in Appendix C) were used to train task adapters with similar hyperparameters, evaluating performance

Model	Configuration	TC (↑)		NER (↑)		SA (↑)		MLM (↓)	
		Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen
mBERT	Baseline	77.67	28.72	83.82	42.54	82.18	71.03	25.17	**124.67**
	+ LoRA (Glot)	78.74	36.65	84.2	44.51	82.75	73.27	10.44	7434.61
	+ Seq_bn (Glot)	79.28	41.42	84.46	45.04	82.99	73.3	**8.95**	12218.65
	+ Seq_bn_inv (Glot)	**79.35**	**42.4**	84.36	**45.64**	**83.64**	**73.91**	14.31	27170.23
	+ LoRA (ConceptNet)	77.87	24.88	84.38	41.32	82.59	70.79	37.37	126.44
	+ Seq_bn (ConceptNet)	78.39	25.87	84.35	41.2	81.9	70.48	41.22	139.25
	+ Seq_bn_inv (ConceptNet)	78.42	24.18	**84.7**	41.48	81.58	71.54	55.95	157.49
	+ Seq_bn (Glot+ConceptNet)	–	–	84.36	44.21	–	–	–	–
	+ Seq_bn_inv (Glot+ConceptNet)	–	–	84.36	44.93	–	–	–	–
	Full Fine-tune	81.73	43.65	–	–	84.07	73.97	9.25	81492.4
XLM-R	Baseline	81.14	34.52	77.33	54.45	87.45	60.72	15.65	203.96
	+ LoRA (Glot)	82.31	40.94	77.52	52.01	87.98	62.02	6.83	**97.99**
	+ Seq_bn (Glot)	83.63	49.72	78.57	54.4	**88.2**	**65.94**	**6.53**	122.08
	+ Seq_bn_inv (Glot)	**84.06**	**51.43**	78.17	55.64	**88.2**	65.88	10.56	713.65
	+ LoRA (ConceptNet)	80.71	29.08	78.38	52.71	87.48	60.00	20.29	902.31
	+ Seq_bn (ConceptNet)	80.82	33.19	77.64	49.39	87.09	58.64	20.01	482.22
	+ Seq_bn_inv (ConceptNet)	80.64	33.59	78.62	51.04	87.28	59.52	22.81	569.48
	+ Seq_bn (Glot+ConceptNet)	–	–	**80.83**	**61.83**	–	–	–	–
	+ Seq_bn_inv (Glot+ConceptNet)	–	–	80.68	60.31	–	–	–	–
	Full Fine-tune	85.61	57.3	–	–	88.56	68.19	10.57	206.68

Table 1: Results for mBERT and XLM-R across 4 tasks: Topic Classification (TC), Named Entity Recognition (NER), Sentiment Analysis (SA), Masked Language Modeling (MLM). All numbers are the averages for the 30 studied LRLs and provided separately for the languages included ("seen") and languages not included ("unseen") in the pre-training data of a model. The baselines are the models with a single task adapter for downstream tasks, or without adapters for MLM. The full results for each task are in the Appendix.

via F1 scores. Finally, *Named Entity Recognition* (NER) used the WikiANN dataset (Pan et al., 2017), with data distributions detailed in Table 7 (Appendix D), and was evaluated with the "seqeval" F1 score (Nakayama, 2018). The (Seq_bn) task adapter was trained with the default hyperparameters (Appendix F).

4.4 Baselines

For MLM, mBERT and XLM-R were evaluated without adapters; LLaMA-3 was not evaluated on this task due to its autoregressive nature. For TC, SA, and NER, baselines used a single Seq_bn task adapter, isolating the impact of language adapters and enabling direct comparisons with language adapter-enhanced models.

5 Results: Small mLMs

This section summarizes the outcomes of the mLM adaptation experiments. Tables 1 and 3 report the average results across 30 selected LRLs.

5.1 Masked Language Modeling

Glot-based adapters substantially improved pseudo-perplexity (Tables 10 and 11 Appendices G and H), particularly for mBERT. The Seq_bn adapter achieved the largest reduction, averaging a 65% improvement, followed by LoRA and Seq_bn_inv. For XLM-R, Seq_bn also excelled overall, while LoRA performed better for higher resourced languages. In contrast, ConceptNet-based adapters did not enhance MLM performance, likely due to the dataset's limited size and structured nature, but showed utility in downstream tasks (Section 5.2).

Full fine-tuning on GlotCC generally outperformed language adapters for mBERT (Table 10), while adapters applied to XLM-R often surpassed full fine-tuning (Table 11). Compared to larger models, Glot-based XLM-R adapters outperformed Glot500-m (Imani et al., 2023), despite the latter's larger vocabulary and more extensive training data. The performance of Glot500-m likely reflects its sampling strategy, which heavily prioritizes LRLs. Additionally, XLM-R-large without language adapters (Conneau et al., 2020) slightly surpassed XLM-R-base with adapters (Appendix J).

5.2 Downstream Tasks

We further fine-tuned task adapters stacked on language adapters and mLMs. The detailed results

are in Tables 15, 16, 18, 19, 21, and 22 (Appendices L, M, O, P, R, and S).[2]

5.2.1 Topic Classification

ConceptNet-based adapters showed marginal average improvements over the baseline. For mBERT, `Seq_bn_inv` primarily improved F1 scores for languages included in pre-training, but gains were inconsistent for others. Glot-based adapters demonstrated more substantial improvements, particularly for languages with less pre-training data. `Seq_bn_inv` achieved the best performance across both models, with mBERT showing an average 2-point F1 improvement for seen languages and 14 points for unseen ones, while XLM-R exhibited an average boost of 3 points for pre-trained languages and 17 points for excluded ones. Full fine-tuning provided better average results for both mBERT—4 points for seen and 15 points for unseen languages, and XLM-R—4 points and 23 points, respectively-with adapters being slightly behind. Additional experiments with `Seq_bn_inv` on LLaMA-3 showed an average 28-point improvement over single-task adapter setups.

5.2.2 Named Entity Recognition

For mBERT, ConceptNet adapters provided modest average improvements mostly for seen languages, with `Seq_bn_inv` achieving the highest gains of 1 F1 point on average. Glot-based adapters offered slightly lower gains for seen languages (0.5 points) but larger improvements for unseen ones, with `Seq_bn_inv` delivering an average gain of 3 points. XLM-R exhibited similar trends: ConceptNet adapters improved average scores by 1 point (`Seq_bn_inv`) for seen languages but showed decreases for unseen ones, while Glot-based adapters reached a 0.5-point improvement (`Seq_bn_inv`) for seen languages and 1 point for unseen ones. Meanwhile, LLaMA-3 with `Seq_bn_inv` failed to outperform its baseline.

Due to NER benefiting the most from ConceptNet adapters, we also experimented with the combination of ConceptNet and Glot adapters (`Seq_bn` and `Seq_bn_inv`) with adapter fusion (Pfeiffer et al., 2021a). This provided the greatest benefits for XLM-R, boosting F1 scores by up to 3 points for seen languages and 7 points for unseen ones, outperforming both individual adapters and the baselines. For mBERT, however, fusion did not produce additional improvements.

[2]Below, we report the average scores across languages for each configuration. Notably, numerous individual languages show improvements under each configuration.

Model	#Params (B)	TC (↑)	NER (↑)
mBERT+`Seq_bn_inv`	**0.177**	**71.92**	**85.28**
XLM-R+`Seq_bn_inv`	**0.279**	**80.79**	**85.42**
DeepSeek-R1-D-Llama	8	20.5	-
DeepSeek-R1-D-Qwen	14	41.88	-
DeepSeek-R1-D-Qwen	32	68.54	-
DeepSeek-R1-D-Llama	70	70.72	-
LLaMA-3	8	65.8	-
LLaMA-3.1	8	65.62	-
Gemma	7	60.21	-
Gemma-2	9	44.27	-
Qwen-1.5	7	40.41	-
Qwen2	7	56.82	-
GPT-3.5-turbo-0301	-	-	70.65
GPT-3.5-turbo-0613	-	45.02	-
GPT-4-0613	-	45.82	-
LLaMA-2	7	18.24	-
BLOOM	7	13.02	31.35
BLOOMz	7	17.51	20.92
mT0	13	-	17.48
Occiglot-eu5	7	28.56	
XGLM	7.5	29.98	-
Yayi	7	16.88	-
LLaMAX2 Alpaca	7	23.13	-
Mala-500-v2	10	5.74	-

Table 2: Average F1 scores on overlapping LRLs for LLMs and our Glot adapter-based mLMs on TC and NER. Prompting results are 3-shot, based on Ji et al. (2024) for TC and Asai et al. (2023) for NER. For NER, we report averages across eight overlapping languages, while the GPT-3.5 average is based on only two. TC results for GPT-3.5 and GPT-4 are zero-shot, as reported by Adelani et al. (2024a). DeepSeek results are zero-shot and were obtained in our evaluation. Per-language results are in Appendix U.

5.2.3 Sentiment Analysis

For mBERT, ConceptNet adapters showed limited average gains, with only `LoRA` surpassing the baseline for seen languages, with a 0.25-point improvement. Glot adapters consistently performed better across all architectures, with `Seq_bn_inv` achieving the highest F1 scores, with a 1.5-point improvement for seen and a 3-point gain for unseen languages. For XLM-R, ConceptNet adapters exhibited no average improvements, while Glot adapters consistently enhanced performance. `Seq_bn` and `Seq_bn_inv` achieved gains of up to 1 point for seen and 5 points for unseen languages. Full fine-tuning yielded similar results with a 2-point and 3-point boosts for mBERT, and 1-point and 8-point improvements respectively, for seen and unseen language groups.

Model	TC (↑)	SA (↑)	NER (↑)
mBERT+Seq_bn_inv	**71.92**	**73.68**	**59.32**
XLM-R+Seq_bn_inv	**80.79**	**83.35**	**69.26**
LLaMA-3 Baseline	31.93	58.83	45.18
LLaMA-3+Seq_bn_inv	60.26	68.68	45.12

Table 3: Average F1 scores over 5 selected LRLs for language adapter-tuned LLaMa-3-8B, mBERT, and XLM-R. Additionally, we present results for LLaMA3 with a single Seq_bn task adapter, similar to our baselines. Per-language results are in Appendix U.

Finally, Seq_bn_inv on LLaMA-3 resulted in a 10-point average improvement over its baseline.

6 Results: Small mLMs vs. SoTA LLMs

Compared to the zero-shot prompting of proprietary LLMs like GPT-3.5-Turbo (Ouyang et al., 2022a) and GPT-4 (OpenAI et al., 2024) on the SIB-200 TC task (Adelani et al., 2024a), our adapter-based models demonstrated superior performance across the 30 LRLs studied, as shown in Table 2. Further, our approach outperformed 3-shot results from LLaMA2-7B (Touvron et al., 2023), BLOOM-7B (Le Scao et al., 2023), instruction-tuned BLOOMZ-7B (Ji et al., 2024), XGLM (Lin et al., 2022), Occiglot-7B-eu5 (Barth et al., 2024), Yayi (Luo et al., 2023), LLaMaX2-7B-Alpaca (Lu et al., 2024), MaLA-500 (Lin et al., 2024), and recent models like LLaMA3-8B, LLaMA3.1-8B (Grattafiori et al., 2024), Gemma-7B, Gemma-2-9B (Team et al., 2024), Qwen-1.5-7B, and Qwen-2 (Yang et al., 2024). Additionally, our adapter-based approaches surpassed results reported by Asai et al. (2023) on the WikiAnn NER task for a subset of 8 overlapping LRLs. Their evaluation included zero- and few-shot prompting with GPT-3.5-Turbo, BLOOM-7B, and instruction-tuned BLOOMZ-7B and mT0-13B (Muennighoff et al., 2023). Distilled DeepSeek-R1 models (8B, 14B, 32B, and 70B) (DeepSeek-AI et al., 2025) failed to surpass smaller mLMs on TC.[3] Finally, Table 3 shows that although Seq_bn_inv language-adapter based LLaMA-3-8B improved performance over prompting and its single-task adapter baseline, it was still less effective than smaller mLMs like XLM-R for TC tasks.

[3]Results are zero-shot, with generated token output limited to 100.

7 General Findings and Discussion

This section highlights key insights gained from our experiments. We analyze performance trends of adapter-based and full fine-tuning approaches for small mLMs, compare their efficacy to LLMs, explore the relationship between language modeling and downstream task performance, and examine the impact of pre- and post-training data sizes on downstream task outcomes.

7.1 Performance Trends

For MLM, the Seq_bn adapter consistently achieved the best performance, likely due to its moderate parameter count (Table 9 Appendix F) aligning with the limited adaptation data. This partially confirms Mundra et al. (2024)'s findings that simple bottleneck adapters outperform other types, including Seq_bn_inv and LoRA. Conversely, LoRA, with even fewer parameters, excelled in languages with larger pre-training data in XLM-R, which may reflect that these languages require fewer parameters given their extensive pre-training coverage, considering the limited adaptation data (see Appendix I). Moreover, Pfeiffer et al. (2021a) noted that high-capacity adapters are less effective for XLM-R compared to mBERT.

For downstream tasks, Seq_bn_inv slightly outperformed other adapter configurations, with Seq_bn showing very similar performance in most cases, confirming findings by Pfeiffer et al. (2020) that invertible layers enhance adaptation by facilitating input and output embedding alignment. The advantage of Seq_bn_inv may also be attributable to its larger number of trainable parameters, which may benefit the task fine-tuning process. Yong et al. (2023) also report the superiority of using invertible layers for a subset of tested languages on the XNLI task (Conneau et al., 2018). Adapter fusion improved NER performance for XLM-R, likely due to the increased count of trainable parameters (compared to individual language adapters), as observed by Lee et al. (2022a). For mBERT, this improvement was not evident: Individual adapters likely provided sufficient capacity.

Adapter-based approaches outperformed full fine-tuning for XLM-R and matched mBERT's performance on MLM, while performing comparably on SA and slightly worse on TC, all with significantly fewer trainable parameters. This indicates that up to 1 GB

of adaptation data suffices for effective adapter training[4], but might be insufficient for fine-tuning larger models like XLM-R.

MLM performance (Tables 10 and 11 Appendices G and H) **was higher for languages supported by the model's vocabulary.** For unsupported languages in mBERT, such as Sinhala and Amharic, pseudo-perplexity was artificially low pre-adaptation due to overconfidence in predicting the UNK token. After adaptation, pseudo-perplexity scores increased, reflecting consistent predictions of non-language-specific tokens (e.g., punctuation). Languages with partial script support, such as Uyghur and Tibetan, showed minimal improvements. XLM-R's broader script coverage mitigated some issues but still struggled with Tibetan. This highlights the need for vocabulary extension when working with unseen languages (Zhang et al., 2020; Wang et al., 2020; Pfeiffer et al., 2021b).

7.2 Small vs. Large LMs for LRLs

Our findings emphasize the effectiveness of adapting smaller encoder-only mLMs with adapters over relying on prompting or adapting LLMs for LRLs. The superior performance of smaller mLMs compared to large-scale models has been explored in prior research. Wu et al. (2019) observed that limited capacity forces models to align semantically similar representations across languages rather than creating language-specific subspaces. Dufter and Schütze (2020) further showed that overparameterizing mBERT degrades its cross-lingual transfer ability and hypothesized that smaller models produce better language-independent representations by reusing parameters across languages, while larger models tend to partition capacity, limiting shared multilingual representations, later supported by Yong et al. (2023). Similarly, Shliazhko et al. (2023) found no performance improvements in mGPT when scaling from 1.3B to 13B parameters for classification and factual probing tasks, with mBERT and XLM-R outperforming larger models. Moreover, Pecher et al. (2024) noted that larger models do not consistently outperform smaller ones in fine-tuning or prompting settings. These findings, together with our results, collectively argue for prioritizing smaller mLMs over large-scale, resource-intensive models (Strubell et al., 2019) to advance performance on LRLs more efficiently and effectively.

7.3 Correlation Between Language Modeling and Downstream Task Performance

To investigate the relationship between language modeling and downstream task performance, we performed correlation analyses using Pearson (Cohen et al., 2009) and Spearman (Spearman, 1961) metrics. Results in Table 14 (Appendix K) show a moderate correlation between pseudo-perplexity and downstream task performance for XLM-R, both pre- and post-adaptation (using Glot data), but a less pronounced correlation for mBERT. **Lower pseudo-perplexity generally indicated better downstream performance for XLM-R and, to a lesser extent, for mBERT, suggesting its utility as a rough proxy for downstream task capabilities, particularly for larger mLMs**. These findings contrast with prior studies (Liang et al., 2022; Yong et al., 2023), which reported an unclear relationship between perplexity and task performance.[5] Post-adaptation, the correlation between pseudo-perplexity and downstream performance strengthened, particularly for tasks with consistent data quality (Figure 3). We conjecture that the stronger correlations observed for XLM-R likely arise from its optimized multilingual architecture and its extensive pre-training corpus.

7.4 Impact of Pre- and Post-Training Data Size on MLM and Downstream Tasks

We analyzed the relationship between pre- and post-adaptation data size and model performance. Before adaptation, pseudo-perplexity and downstream task performance were correlated with pre-training data size (Figure 1 and Table 12 Appendix I), as also found by Wu and Dredze (2020), Ahuja et al. (2023) and Bagheri Nezhad and Agrawal (2024). **Post-adaptation improvements primarily depended on pre-training and, surprisingly less so, on adaptation data volumes, with the latter providing only a marginal improvement**.[6] LRLs exhibited larger gains, while higher-resource languages faced diminishing returns or even reduced performance. The latter

[4]This is in line with Bapna et al. (2019), He et al. (2021), and Liu et al. (2022), who report that adapter-based tuning often surpasses full fine-tuning.

[5]Unlike these studies, we evaluate pseudo-perplexity across a diverse set of languages rather than models. This partially aligns with Xia et al. (2022), who observed a correlation between perplexity and few-shot learning results.

[6]Similarly, Kunz and Holmström (2024) show limited overall impact of adaptation data and language adapters.

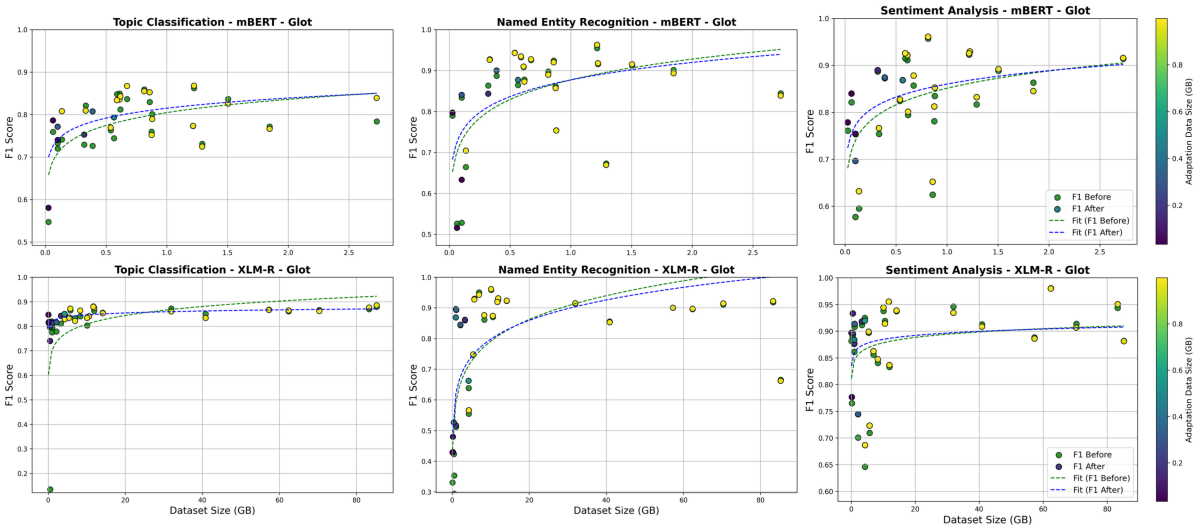

Figure 1: Correlation between the pre-training data sizes for mBERT and XLM-R and downstream task results for the pre-adaptation and post-adaptation results. The vertical bars indicate the amounts of adaptation data. The improvements in downstream performance for both models are primarily concentrated in languages with smaller pre-training data sizes, which are positioned on the left side of the plots (Section 7.4).

is likely due to the model encountering data already seen during pre-training (Lee et al., 2022b). Achieving further gains for well-represented languages may require increasing adaptation data and adapter capacity. A correlation analysis (Appendix I) demonstrates that adaptation data had a stronger impact on mBERT than XLM-R, likely because of its larger relative contribution as compared to pre-training data.

In downstream tasks, even small amounts of adaptation data (e.g., a few MB of graph-based data or a few hundred MB of free-text data) produced performance gains, consistent with Pfeiffer et al. (2020) and Yong et al. (2023). This was especially true for mBERT, where adaptation data constitutes a larger proportion relative to its overall training data. For XLM-R, adaptation data was more beneficial for LRLs, while its impact diminished for languages with pre-training data exceeding approximately 20 GB, as also observed by Adelani et al. (2024a). Diminishing returns suggest a threshold effect, where extensive pre-training coverage reduces the utility of adaptation data, indicating that larger adaptation datasets may be necessary for further gains. Figures 4, 5, and 6 demonstrate these trends, showing that underrepresented languages typically benefit more from even limited adaptation data, confirmed by correlation analyses (Appendices N, Q, and T).

The type of adaptation data influenced task-specific performance. ConceptNet-based adapters outperformed Glot-based adapters for NER in most languages, likely because ConceptNet contains straightforward NER information. This contrasts with the findings of Gurgurov et al. (2024a), who observed different trends when experimenting with a smaller subset of languages. Conversely, Glot-based adapters provided more consistent improvements across tasks, leveraging their larger adaptation data volumes (up to 1 GB for most languages). This emphasizes the important role of relative data size in determining the effectiveness of adaptation across tasks.

8 Conclusion

This study evaluated adapter-based adaptation of small mLMs to LRLs using structured and unstructured data, alongside continued pre-training and comparing them with SoTA LLMs. `Seq_bn` achieved the best results for MLM tasks, while `Seq_bn_inv` excelled in downstream tasks. Full fine-tuning offered limited advantages over adapters. Downstream performance was primarily influenced by pre-training data, with adaptation data providing incremental gains. Graph-based knowledge from ConceptNet, despite its small size, improved NER performance, while Glot data consistently delivered the largest gains across tasks. Our results generally suggest that smaller mLMs may be better suited for LRLs than LLMs, since mLMs efficiently align cross-lingual representations and generalize well under

data constraints.

Limitations

This study has three main limitations. First, adapters have specific hyperparameters that influence their behavior and capacity. Future work should systematically explore these hyperparameters and their effects on adapter performance. Second, the amount of adaptation data was limited to 1 GB per language due to computational constraints. Investigating the impact of larger datasets on model adaptation—e.g., utilizing the full GlotCC data without truncation—remains an open and promising direction. Increasing adapter capacity and adaptation data size and measuring adaptation effects as a function of both data volume and model capacity could provide valuable insights. Finally, some experiments were not conducted across all tasks due to resource constraints. For example, adapter fusion was applied only to named entity recognition, and full fine-tuning was only evaluated for small models on masked language modeling, topic classification, and sentiment analysis, but not on named entity recognition.

Acknowledgments

This work was supported by DisAI – Improving scientific excellence and creativity in combating disinformation with artificial intelligence and language technologies, a Horizon Europe-funded project under GA No. 101079164, and by the German Ministry of Education and Research (BMBF) as part of the project TRAILS (01IW24005).

References

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024a. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

David Ifeoluwa Adelani, A. Seza Doğruöz, André Coneglian, and Atul Kr. Ojha. 2024b. Comparing LLM prompting with cross-lingual transfer performance on indigenous and low-resource Brazilian languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 34–41, Mexico City, Mexico. Association for Computational Linguistics.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Adam Amram, Anat Ben David, and Reut Tsarfaty. 2018. Representations and architectures in neural sentiment analysis for morphologically rich languages: A case study from Modern Hebrew. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2242–2252, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2023. Buffet: Benchmarking large language models for few-shot cross-lingual transfer. *arXiv preprint arXiv:2305.14857*.

Soran Badawi, Arefeh Kazemi, and Vali Rezaie. 2024. Kurdisent: a corpus for kurdish sentiment analysis. *Language Resources and Evaluation*, pages 1–20.

Sina Bagheri Nezhad and Ameeta Agrawal. 2024. What drives performance in multilingual language models? In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 16–27, Mexico City, Mexico. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Ankur Bapna, Naveen Arivazhagan, and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation.

Fabio Barth, Manuel Brack, Maurice Kraus, Pedro Ortiz Suarez, Malte Ostendorf, Patrick Schramowski, and Georg Rehm. 2024. Occiglot euro llm leaderboard.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *60th Annual Meeting of the Association for Computational Linguistics, ACL 2022*, pages 7817–7829. Association for Computational Linguistics (ACL).

Jože Bučar, Martin Žnidaršič, and Janez Povh. 2018. Annotated news corpora and a lexicon for sentiment analysis in slovene. *Language Resources and Evaluation*, 52(3):895–919.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Dea Adhista, Emmanuel Dave, Sarah Oktavianti, Salsabil Akbar, Jhonson Lee, Nuur Shadieq, Tjeng Wawan Cenggoro, Hanung Linuwih, Bryan Wilie, Galih Muridan, Genta Winata, David Moeljadi, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung. 2023. NusaWrites: Constructing high-quality corpora for underrepresented and extremely low-resource languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–945, Nusa Dua, Bali. Association for Computational Linguistics.

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

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

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, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Keith Cortis and Brian Davis. 2019. A social opinion gold standard for the Malta government budget 2018. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 364–369, Hong Kong, China. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

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.

Alexiei Dingli and Nicole Sant. 2016. Sentiment analysis on maltese using machine learning. In *Proceedings of The Tenth International Conference on Advances in Semantic Processing (SEMAPRO 2016)*, pages 21–25.

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.

Abteen Ebrahimi and Katharina Kann. 2021. How to adapt your pretrained multilingual model to 1600 languages.

Luis Espinosa-Anke, Geraint Palmer, Padraig Corcoran, Maxim Filimonov, Irena Spasić, and Dawn Knight. 2021. English–welsh cross-lingual embeddings. *Applied Sciences*, 11(14):6541.

Fahim Faisal and Antonios Anastasopoulos. 2022. Phylogeny-inspired adaptation of multilingual models to new languages.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2014. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban

Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Daniil Gurgurov, Mareike Hartmann, and Simon Ostermann. 2024a. Adapting Multilingual LLMs to Low-Resource Languages with Knowledge Graphs via Adapters. *arXiv preprint arXiv:2407.01406*.

Daniil Gurgurov, Rishu Kumar, and Simon Ostermann. 2024b. Gremlin: A repository of green baseline embeddings for 87 low-resource languages injected with multilingual graph knowledge.

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

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling.

Md. Arid Hasan, Prerona Tarannum, Krishno Dey, Imran Razzak, and Usman Naseem. 2024. Do large language models speak all languages equally? a comparative study in low-resource settings.

Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jia-Wei Low, Lidong Bing, and Luo Si. 2021. On the effectiveness of adapter-based tuning for pretrained language model adaptation.

Yifan Hou, Wenxiang Jiao, Meizhen Liu, Carl Allen, Zhaopeng Tu, and Mrinmaya Sachan. 2022. Adapters for enhanced modeling of multilingual knowledge and text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3902–3917.

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 *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in

llms: Improving multilingual capability by cross-lingual-thought prompting.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Tim Isbister, Fredrik Carlsson, and Magnus Sahlgren. 2021. Should we stop training more monolingual models, and simply use machine translation instead? In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 385–390, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and Barry Haddow. 2024. Emma-500: Enhancing massively multilingual adaptation of large language models.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Dame Jovanoski, Veno Pachovski, and Preslav Nakov. 2015. Sentiment analysis in Twitter for Macedonian. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 249–257, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Georgios Kalamatianos, Dimitrios Mallis, Symeon Symeonidis, and Avi Arampatzis. 2015. Sentiment analysis of greek tweets and hashtags using a sentiment lexicon. In *Proceedings of the 19th panhellenic conference on informatics*, pages 63–68.

Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. 2024. GlotCC: An open broad-coverage commoncrawl corpus and pipeline for minority languages. *arXiv preprint*.

Muhammad Yaseen Khan, Shah Muhammad Emaduddin, and Khurum Nazir Junejo. 2017. Harnessing english sentiment lexicons for polarity detection in urdu tweets: A baseline approach. In *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, pages 242–249. IEEE.

Muhammad Yaseen Khan and Muhammad Suffian Nizami. 2020. Urdu sentiment corpus (v1.0): Linguistic exploration and visualization of labeled datasetfor urdu sentiment analysis. In *2020 IEEE 2nd International Conference On Information Science & Communication Technology (ICISCT)*. IEEE.

Jenny Kunz and Oskar Holmström. 2024. The impact of language adapters in cross-lingual transfer for nlu.

Elmurod Kuriyozov, Sanatbek Matlatipov, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2019. Construction and evaluation of sentiment datasets for low-resource languages: The case of uzbek. In *Human Language Technology. Challenges for Computer Science and Linguistics - 9th Language and Technology Conference, LTC 2019, Poznan, Poland, May 17-19, 2019, Revised Selected Papers*, volume 13212 of *Lecture Notes in Computer Science*, pages 232–243. Springer.

Anne Lauscher, Olga Majewska, Leonardo F. R. Ribeiro, Iryna Gurevych, Nikolai Rozanov, and Goran Glavaš. 2020a. Common sense or world knowledge? investigating adapter-based knowledge injection into pretrained transformers. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 43–49, Online. Association for Computational Linguistics.

Anne Lauscher, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2020b. Specializing unsupervised pretraining models for word-level semantic similarity.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model. *CoRR*.

Jaeseong Lee, Seung-won Hwang, and Taesup Kim. 2022a. FAD-X: Fusing adapters for cross-lingual transfer to low-resource languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 57–64, Online only. Association for Computational Linguistics.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022b. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Siyu Li, Kui Zhao, Jin Yang, Xinyun Jiang, Zhengji Li, and Zicheng Ma. 2022. Senti-exlm: Uyghur enhanced sentiment analysis model based on xlm. *Electronics Letters*, 58(13):517–519.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian

Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André F. T. Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual language models.

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

LocalDoc. 2024. Sentiment analysis datset for azerbaijani.

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. Llamax: Scaling linguistic horizons of llm by enhancing translation capabilities beyond 100 languages.

Yin Luo, Qingchao Kong, Nan Xu, Jia Cao, Bao Hao, Baoyu Qu, Bo Chen, Chao Zhu, Chenyang Zhao, Donglei Zhang, Fan Feng, Feifei Zhao, Hailong Sun, Hanxuan Yang, Haojun Pan, Hongyu Liu, Jianbin Guo, Jiangtao Du, Jingyi Wang, Junfeng Li, Lei Sun, Liduo Liu, Lifeng Dong, Lili Liu, Lin Wang, Liwen Zhang, Minzheng Wang, Pin Wang, Ping Yu, Qingxiao Li, Rui Yan, Rui Zou, Ruiqun Li, Taiwen Huang, Xiaodong Wang, Xiaofei Wu, Xin Peng, Xina Zhang, Xing Fang, Xinglin Xiao, Yanni Hao, Yao Dong, Yigang Wang, Ying Liu, Yongyu Jiang, Yungan Wang, Yuqi Wang, Zhangsheng Wang, Zhaoxin Yu, Zhen Luo, Wenji Mao, Lei Wang, and Dajun Zeng. 2023. Yayi 2: Multilingual open-source large language models.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Mounika Marreddy, Subba Reddy Oota, Lakshmi Sireesha Vakada, Venkata Charan Chinni, and Radhika Mamidi. 2022a. Am i a resource-poor language? data sets, embeddings, models and analysis for four different nlp tasks in telugu language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(1):1–34.

Mounika Marreddy, Subba Reddy Oota, Lakshmi Sireesha Vakada, Venkata Charan Chinni, and Radhika Mamidi. 2022b. Multi-task text classification using graph convolutional networks for large-scale low resource language. *arXiv preprint arXiv:2205.01204*.

Antonio Martínez-García, Toni Badia, and Jeremy Barnes. 2021. Evaluating morphological typology in zero-shot cross-lingual transfer. 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 3136–3153.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

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.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino D'ario M'ario Ant'onio Ali, Davis Davis, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Steven Arthur. 2023a. Afrisenti: A twitter sentiment analysis benchmark for african languages.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Sa'id Ahmad, Nedjma Ousidhoum, Abinew Ayele, Saif M Mohammad, and Meriem Beloucif. 2023b. Semeval-2023 task 12: Sentiment analysis for african languages (afrisenti-semeval). *arXiv preprint arXiv:2304.06845*.

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.

Nandini Mundra, Sumanth Doddapaneni, Raj Dabre, Anoop Kunchukuttan, Ratish Puduppully, and Mitesh M Khapra. 2024. A comprehensive analysis of adapter efficiency. In *Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD)*, pages 136–154.

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

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F Christiano, Jan Leike, and Ryan Lowe. 2022a. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022b. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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.

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

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

Samuel Pecar, Marian Simko, and Maria Bielikova. 2019. Improving sentiment classification in Slovak language. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 114–119, Florence, Italy. Association for Computational Linguistics.

Branislav Pecher, Ivan Srba, and Maria Bielikova. 2024. Comparing specialised small and general large language models on text classification: 100 labelled samples to achieve break-even performance.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. 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 43–54, Hong Kong, China. Association for Computational Linguistics.

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. 2021a. 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.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021b. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aabha Pingle, Aditya Vyawahare, Isha Joshi, Rahul Tangsali, and Raviraj Joshi. 2023. L3cube-mahasent-md: A multi-domain marathi sentiment analysis dataset and transformer models. *arXiv preprint arXiv:2306.13888*.

Ayu Purwarianti and Ida Ayu Putu Ari Crisdayanti. 2019. Improving bi-lstm performance for indonesian sentiment analysis using paragraph vector. In *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–5. IEEE.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Surangika Ranathunga and Isuru Udara Liyanage. 2021. Sentiment analysis of sinhala news comments. *Transactions on Asian and Low-Resource Language Information Processing*, 20(4):1–23.

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

Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2019. Masked language model scoring. *arXiv preprint arXiv:1910.14659*.

Salim Sazzed. 2020. Cross-lingual sentiment classification in low-resource Bengali language. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 50–60, Online. Association for Computational Linguistics.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2023. mgpt: Few-shot learners go multilingual.

Oyesh Mann Singh, Sandesh Timilsina, Bal Krishna Bal, and Anupam Joshi. 2020. Aspect based abusive sentiment detection in nepali social media texts. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 301–308.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. 2006. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*, pages 1015–1021. Springer.

Charles Spearman. 1961. The proof and measurement of association between two things.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Uga Sproģis and Matīss Rikters. 2020. What Can We Learn From Almost a Decade of Food Tweets. In *In Proceedings of the 9th Conference Human Language Technologies - The Baltic Perspective (Baltic HLT 2020)*, Kaunas, Lithuania.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Nicolas Stefanovitch, Jakub Piskorski, and Sopho Kharazi. 2022. Resources and experiments on sentiment classification for Georgian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1613–1621, Marseille, France. European Language Resources Association.

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

Arthit Suriyawongkul, Ekapol Chuangsuwanich, Pattarawat Chormai, and Charin Polpanumas. 2019. Pythainlp/wisesight-sentiment: First release.

Anca Tache, Gaman Mihaela, and Radu Tudor Ionescu. 2021. Clustering word embeddings with self-organizing maps. application on LaRoSeDa - a large Romanian sentiment data set. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 949–956, Online. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jar-

rett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Tarikwa Tesfa, Befikadu Belete, Samuel Abera, Sudhir Kumar Mohapatra, and Tapan Kumar Das. 2024. Aspect-based sentiment analysis on amharic text for evaluating ethio-telecom services. In *2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE)*, pages 1–6. IEEE.

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

Adam Tsakalidis, Symeon Papadopoulos, Rania Voskaki, Kyriaki Ioannidou, Christina Boididou, Alexandra I Cristea, Maria Liakata, and Yiannis Kompatsiaris. 2018. Building and evaluating resources for sentiment analysis in the greek language. *Language resources and evaluation*, 52:1021–1044.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418, Online. 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.

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.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2022. Nusax: Multilingual parallel sentiment dataset for 10 indonesian local languages.

Wilson Wongso, David Samuel Setiawan, and Derwin Suhartono. 2021. Causal and masked language modeling of javanese language using transformer-based architectures. In *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 1–7. IEEE.

Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Emerging cross-lingual structure in pretrained language models. *arXiv preprint arXiv:1911.01464*.

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.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Ves Stoyanov. 2022. Training trajectories of language models across scales. *arXiv preprint arXiv:2212.09803*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report.

Zheng-Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Indra Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023.

Bloom+1: Adding language support to bloom for zero-shot prompting.

Rong Zhang, Revanth Gangi Reddy, Md Arafat Sultan, Vittorio Castelli, Anthony Ferritto, Radu Florian, Efsun Sarioglu Kayi, Salim Roukos, Avi Sil, and Todd Ward. 2020. Multi-stage pre-training for low-resource domain adaptation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5461–5468, Online. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Yulei Zhu, Baima Luosai, Liyuan Zhou, Nuo Qun, and Tashi Nyima. 2023. Research on sentiment analysis of tibetan short text based on dual-channel hybrid neural network. In *2023 IEEE 4th International Conference on Pattern Recognition and Machine Learning (PRML)*, pages 377–384.

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

Appendix

A ConceptNet Tripple Conversion Mapping

ConceptNet Relationship	Natural Language Predicate
Antonym	is the opposite of
DerivedFrom	is derived from
EtymologicallyDerivedFrom	is etymologically derived from
EtymologicallyRelatedTo	is etymologically related to
FormOf	is a form of
PartOf	is a part of
HasA	belongs to
UsedFor	is used for
AtLocation	is a typical location for
Causes	causes
CausesDesire	makes someone want
MadeOf	is made of
ReceivesAction	receives action of
HasSubevent	is a subevent of
HasFirstSubevent	is an event that begins with subevent
HasLastSubevent	is an event that concludes with subevent
HasPrerequisite	has prerequisite of
HasProperty	can be described as
MotivatedByGoal	is a step toward accomplishing the goal
ObstructedBy	is an obstacle in the way of
Desires	is a conscious entity that typically wants
CreatedBy	is a process or agent that creates
CapableOf	is capable of
HasContext	is a word used in the context of
IsA	is a type of
RelatedTo	is related to
SimilarTo	is similar to
Synonym	is a synonym of
SymbolOf	symbolically represents
DefinedAs	is a more explanatory version of
DistinctFrom	is distinct from
MannerOf	is a specific way to do
LocatedNear	is typically found near

Table 4: ConceptNet relationships and their natural language predicates. This mapping is used for converting the ConceptNet KG data into natural language text.

B Language Details

Language	ISO	Language Family	CN (Sent-s)	CN (MB)	Glot (Doc-s)	Glot (MB)	mBERT?	XLM-R?	mBERT Data Size (GB)	XLM-R Data Size (GB)
Thai	th	Kra-Dai	123,859	6.95	2,391,253	977.68	✓	✓	1.29	85.24
Romanian	ro	Indo-European	70,236	2.47	8,657,002	1002.36	✓	✓	1.22	83.29
Bulgarian	bg	Indo-European	162,181	8.02	5,192,702	1014.73	✓	✓	1.50	70.37
Danish	da	Indo-European	66,109	2.27	8,743,767	1006.91	✓	✓	0.81	62.39
Greek	el	Indo-European	89,016	4.17	4,789,519	980.94	✓	✓	1.85	57.30
Hebrew	he	Afro-Asiatic	41,444	1.62	5,287,428	991.82	✓	✓	2.73	40.87
Slovak	sk	Indo-European	22,460	0.81	9,294,165	1006.96	✓	✓	0.61	31.96
Slovenian	sl	Indo-European	85,882	2.98	9,301,902	1007.91	✓	✓	0.67	14.16
Latvian	lv	Indo-European	66,408	2.4	8,301,651	988.21	✓	✓	0.33	11.94
Indonesian	ms	Austronesian	175,246	6.21	8,024,827	1022.01	✓	✓	0.59	11.73
Georgian	ka	Kartvelian	35,331	1.89	3,463,631	1014.24	✓	✓	0.88	10.55
Bengali	bn	Indo-European	8,782	0.46	2,940,197	993.44	✓	✓	1.22	10.10
Azerbaijani	az	Turkic	15,149	0.57	6,179,152	1016.68	✓	✓	0.62	8.33
Urdu	ur	Indo-European	13,315	0.51	4,220,566	1009.42	✓	✓	0.54	6.97
Macedonian	mk	Indo-European	38,116	1.54	5,037,552	1005.62	✓	✓	0.86	5.76
Telugu	te	Dravidian	33,476	1.72	3,162,535	1005.55	✓	✓	0.88	5.46
Nepali	ne	Indo-European	4,456	0.21	2,569,572	1012.63	✓	✓	0.14	4.32
Marathi	mr	Indo-European	7,232	0.37	402,575	157.3	✓	✓	0.32	3.33
Swahili	sw	Niger-Congo	12,380	0.39	2,450,753	323.27	✓	✓	0.10	2.15
Welsh	cy	Indo-European	18,313	0.61	3,174,686	360.24	✓	✓	0.39	1.07
Uzbek	uz	Turkic	4,362	0.16	4,018,172	481.49	✓	✓	0.57	0.95
Javanese	jv	Austronesian	3,448	0.13	367,795	43.56	✓	✓	0.10	0.20
Sundanese	su	Austronesian	1,880	0.07	323,610	43.55	✓	✓	0.06	0.08
Sinhala	si	Indo-European	1,782	0.1	1,655,641	586.21	✗	✓	✗	4.27
Amharic	am	Afro-Asiatic	1,814	0.07	667,881	203.65	✗	✓	✗	1.00
Kurdish	ku	Indo-European	12,246	0.44	376,260	134.7	✗	✓	✗	0.52
Uyghur	ug	Turkic	1,715	0.06	976,010	233.61	✗	✓	✗	0.43
Maltese	mt	Afro-Asiatic	3,895	0.14	1,389,527	182.17	✗	✗	✗	✗
Tibetan	bo	Sino-Tibetan	4,768	0.21	288,847	165.31	✗	✗	✗	✗
Yoruba	yo	Niger-Congo	1,044	0.05	278,003	34.51	✓	✗	0.03	✗

Table 5: Number of ConceptNet triples and GlotCC documents as well as corresponding data sizes per language, sorted by Glot (Doc-s) in descending order. The last four columns indicate the inclusion of the respective language in mBERT and XLM-R pre-training data, alongside the corresponding data sizes in GB. The sizes are approximated based on the openly available CC100 and WikiPedia datasets.

C Sentiment Analysis Data Details

Language	ISO code	Source	#pos	#neg	#train	#val	#test
Sundanese	su	Winata et al., 2022	378	383	381	76	304
Amharic	am	Tesfa et al., 2024	487	526	709	152	152
Swahili	sw	Muhammad et al., 2023a; Muhammad et al., 2023b	908	319	738	185	304
Georgian	ka	Stefanovitch et al., 2022	765	765	1080	120	330
Nepali	ne	Singh et al., 2020	680	1019	1189	255	255
Uyghur	ug	Li et al., 2022	2450	353	1962	311	530
Latvian	lv	Sproģis and Rikters, 2020	1796	1380	2408	268	500
Slovak	sk	Pecar et al., 2019	4393	731	3560	522	1042
Sinhala	si	Ranathunga and Liyanage, 2021	2487	2516	3502	750	751
Slovenian	sl	Bučar et al., 2018	1665	3337	3501	750	751
Uzbek	uz	Kuriyozov et al., 2019	3042	1634	3273	701	702
Bulgarian	bg	Martínez-García et al., 2021	6652	1271	5412	838	1673
Yoruba	yo	Muhammad et al., 2023a; Muhammad et al., 2023b	6344	3296	5414	1327	2899
Urdu	ur	Maas et al., 2011; Khan et al., 2017; Khan and Nizami, 2020	5562	5417	7356	1812	1812
Macedonian	mk	Jovanoski et al., 2015	3041	5184	6557	729	939
Danish	da	Isbister et al., 2021	5000	5000	7000	1500	1500
Marathi	mr	Pingle et al., 2023	5000	5000	8000	1000	1000
Bengali	bn	Sazzed, 2020	8500	3307	8264	1771	1772
Hebrew	he	Amram et al., 2018	8497	3911	8932	993	2483
Romanian	ro	Tache et al., 2021	7500	7500	10800	1200	3000
Telugu	te	Marreddy et al., 2022a; Marreddy et al., 2022b	9488	6746	11386	1634	3214
Welsh	cy	Espinosa-Anke et al., 2021	12500	12500	17500	3750	3750
Azerbaijani	az	LocalDoc, 2024	14000	14000	19600	4200	4200
Tibetan	bo	Zhu et al., 2023	5006	5000	7004	1501	1501
Kurdish	ku	Badawi et al., 2024	4065	3922	6000	993	994
Greek	el	Kalamatianos et al., 2015; Tsakalidis et al., 2018	5773	1313	5936	383	767
Javanese	jv	Wongso et al., 2021	12500	12500	17500	5025	2475
Maltese	mt	Dingli and Sant, 2016; Cortis and Davis, 2019	271	580	595	85	171
Thai	th	Suriyawongkul et al., 2019;	4778	6822	8103	1153	2344
Malay	ms	Purwarianti and Crisdayanti, 2019	7319	4005	7926	1132	2266

Table 6: Sentiment analysis data details.

D Named Entity Recognition Data Details

Language	ISO code	#train	#val	#test
Bulgarian	bg	20000	10000	10000
Indonesian	ms	20000	1000	1000
Maltese	mt	100	100	100
Nepali	ne	100	100	100
Javanese	jv	100	100	100
Uyghur	ug	100	100	100
Tibetan	bo	100	100	100
Sinhala	si	100	100	100
Sundanese	su	100	100	100
Amharic	am	100	100	100
Swahili	sw	1000	1000	1000
Georgian	ka	10000	10000	10000
Latvian	lv	10000	10000	10000
Slovak	sk	20000	10000	10000
Slovenian	sl	15000	10000	10000
Uzbek	uz	1000	1000	1000
Yoruba	yo	100	100	100
Urdu	ur	20000	1000	1000
Macedonian	mk	10000	1000	1000
Danish	da	20000	10000	10000
Marathi	mr	5000	1000	1000
Bengali	bn	10000	1000	1000
Hebrew	he	20000	10000	10000
Romanian	ro	20000	10000	10000
Telugu	te	1000	1000	1000
Welsh	cy	10000	1000	1000
Azerbaijani	az	10000	1000	1000
Greek	el	20000	10000	10000
Kurdish	ku	100	100	100
Thai	th	20000	10000	10000

Table 7: Named entity recognition data details.

E Language Adapters Evaluation Losses

ISO	ConceptNet						Glot					
	mBERT			XLM-R			mBERT			XLM-R		
	Seq_bn	LoRA	Seq_bn_inv	Seq_bn	LoRA	Seq_bn_inv	Seq_bn	LoRA	Seq_bn_inv	Seq_bn	LoRA	Seq_bn_inv
th	1.21	1.24	1.2	1.42	1.42	1.35	0.46	0.54	0.45	1.55	1.65	1.53
ro	1.41	1.46	1.34	1.43	1.43	1.33	1.37	1.52	1.34	1.27	1.3	1.26
bg	0.68	0.71	0.66	0.87	0.87	0.81	1.09	1.25	1.07	1.83	1.8	1.8
da	1.24	1.29	1.19	1.35	1.36	1.26	1.39	1.54	1.36	1.28	1.36	1.26
el	1.13	1.18	1.12	1.36	1.36	1.29	0.67	0.77	0.66	0.84	0.9	0.83
he	1.35	1.38	1.32	1.47	1.46	1.4	1.3	1.41	1.28	1.29	1.38	1.28
sk	1.22	1.28	1.16	1.39	1.39	1.28	1.09	1.19	1.06	1.16	1.19	1.14
sl	0.83	0.91	0.79	1.05	1.09	0.98	1.16	1.28	1.13	1.22	1.28	1.21
lv	1.32	1.4	1.25	1.47	1.51	1.37	1.11	1.29	1.07	1.28	1.37	1.25
ms	1.57	1.63	1.5	1.59	1.57	1.47	1.52	1.65	1.48	1.55	1.6	1.54
ka	1.15	1.19	1.14	1.38	1.35	1.3	0.79	0.91	0.77	1.12	1.18	1.11
bn	0.99	1.03	0.97	1.37	1.37	1.3	1.05	1.16	1.03	1.44	1.49	1.42
az	1.33	1.37	1.29	1.5	1.55	1.42	0.89	1.02	0.86	1.19	1.31	1.15
ur	1.43	1.48	1.4	1.62	1.61	1.51	1.15	1.31	1.12	1.38	1.44	1.36
mk	1.42	1.44	1.38	1.59	1.54	1.45	0.89	0.99	0.87	1.41	1.4	1.41
te	1.09	1.12	1.07	1.29	1.29	1.22	0.83	0.94	0.81	1.33	1.4	1.31
ne	1.26	1.31	1.21	1.53	1.52	1.42	0.77	0.9	0.75	1.38	1.45	1.35
mr	1.08	1.12	1.04	1.46	1.45	1.37	0.94	1.07	0.92	1.43	1.49	1.41
sw	1.54	1.63	1.51	1.64	1.73	1.56	0.94	1.13	0.9	1.13	1.22	1.1
cy	1.55	1.6	1.48	1.83	1.91	1.76	0.81	0.99	0.77	0.95	1.06	0.92
uz	1.22	1.3	1.18	1.55	1.62	1.45	0.85	1.01	0.82	1.06	1.17	1.03
jv	1.44	1.5	1.4	1.55	1.56	1.48	2.11	2.21	2.08	2.63	2.66	2.54
su	1.51	1.56	1.47	1.38	1.4	1.38	1.14	1.28	1.11	1.21	1.35	1.18
si	1.4	1.33	1.38	1.31	1.25	1.25	0.82	0.88	0.8	1.21	1.29	1.19
am	1.47	1.51	1.58	1.22	1.29	1.13	1.25	1.31	1.23	1.2	1.31	1.19
ku	1.64	1.73	1.61	1.91	2.04	1.86	0.93	1.05	0.9	0.76	1.02	0.71
ug	1.09	1.13	1.07	1.57	1.59	1.47	0.46	0.57	0.44	0.79	0.94	0.76
mt	1.41	1.44	1.39	1.53	1.68	1.5	0.84	1.08	0.8	0.93	1.2	0.87
bo	1.0	1.01	0.98	0.63	0.64	0.62	0.24	0.28	0.24	0.72	0.73	0.71
yo	1.12	1.27	1.1	1.77	1.79	1.76	0.87	1.04	0.84	0.83	1.03	0.78

Table 8: Evaluation losses for language adapters by model, architecture, and language.

Evaluation loss values were not predictive of MLM performance. Despite `Seq_bn_inv` achieving the lowest evaluation losses, it underperformed in MLM tasks, indicating that evaluation loss may be an unreliable training metric (suggested by Salazar et al. (2019)).

F Language Adapter Hyperparameters

Adapter Type	mBERT			XLM-R			LLaMA-3	
	`Seq_bn`	`Seq_bn_inv`	`LoRA`	`Seq_bn`	`Seq_bn_inv`	`LoRA`	`Seq_bn`	`Seq_bn_inv`
Trainable Params (No.)	894,528	1,190,592	294,912	894,528	1,190,592	294,912	67,248,128	75,642,880
Trainable Params (%)	0.505%	0.672%	0.166%	0.322%	0.429%	0.106%	0.896%	1.008%
Hyperparameters for LA	Batch Size: 16, Learning Rate: 1e-4, `Seq_bn` and `Seq_bn_inv`: Reduction Factor = 16, `LoRA`: $\alpha = 8$, $r = 8$						Batch Size: 1, Learning Rate: 1e-4	
Hyperparameters for TA	Batch Size: 32, Learning Rate: 1e-4, `Seq_bn`: Reduction Factor = 16, `LoRA`: $\alpha = 8$, $r = 8$						Batch Size for TC: 16; for SA and NER: 8, Learning Rate: 2e-5	

Table 9: Trainable parameters and hyperparameters for different adapter types in mBERT, XLM-R, and LLaMA-3. The rest of hyperparameters are as specified in the default adapter configurations in Adapterhub. LA - Langauge adapter, TA - Task adapter.

G Masked Language Modeling Pseudo-Perplexity - Part I

ISO	mBERT							
		ConceptNet			Glot			
	Base	Seq_bn	LoRA	Seq_bn_inv	Seq_bn	LoRA	Seq_bn_inv	FFT
he	18.36	19.71	18.29	19.85	**11.09**	12.31	12.51	8.78
el	4.69	6.17	5.55	6.92	**3.3**	3.54	3.49	2.71
bg	10.84	14.99	12.65	20.93	**5.4**	5.9	6.09	4.67
th	3.87	4.13	4.29	4.07	**2.94**	3.34	3.18	2.54
ro	11.49	13.47	12.67	22.39	**5.94**	6.59	8.67	6.75
bn	11.97	14.94	13.53	15.99	**9.11**	10.05	10.32	8.42
te	7.92	8.9	8.34	9.33	**6.09**	6.13	6.4	5.32
ka	6.52	6.3	6.0	6.54	**3.63**	4.06	3.91	2.6
mk	11.95	14.5	12.3	13.26	**5.83**	6.33	6.54	5.53
da	19.16	19.29	25.39	30.87	**11.13**	11.8	13.02	8.76
sl	13.57	18.09	14.32	26.86	**6.68**	7.26	8.58	4.91
az	12.47	15.2	13.48	24.26	**7.04**	7.89	7.9	5.83
sk	11.5	13.86	12.37	19.29	**5.98**	6.64	7.14	6.03
ms	36.26	53.66	50.17	128.6	**18.23**	20.01	22.71	16.95
uz	26.65	31.41	23.43	40.35	**5.84**	7.21	9.22	3.84
ur	22.59	23.02	21.74	26.4	**10.18**	12.0	12.89	7.16
cy	21.24	22.13	23.0	39.75	**6.08**	7.8	9.06	4.89
lv	14.14	18.31	16.21	33.14	**5.98**	7.13	7.48	4.58
mr	12.51	12.9	12.21	14.0	**5.84**	6.78	6.85	6.71
ne	12.72	14.19	13.08	15.36	**6.71**	7.21	8.68	4.88
jv	83.84	115.27	132.08	146.64	**19.4**	22.86	31.6	19.19
sw	42.53	57.57	52.21	79.5	**8.99**	12.48	16.09	7.19
su	102.16	177.27	183.04	227.87	**20.24**	23.2	34.29	34.93
yo	85.21	293.99	210.43	370.71	**23.14**	31.96	86.79	38.89
Avg.	25.17	41.22	37.37	55.95	**8.95**	10.44	14.31	9.25
mt†	531.59	432.99	456.64	457.43	**6.89**	9.87	15.02	5.95
ku†	**72.87**	119.29	101.13	149.74	1524.98	559.83	173.24	6381.75
ug†	112.63	96.52	86.31	121.15	**28.69**	67.26	75.53	313.64
si†	**16.29**	96.5	40.3	103.36	15640.68	8981.09	157397.73	443921.11
am†	**10.06**	31.41	26.93	23.47	56052.75	34924.59	4223.4	38289.93
bo†	**4.59**	58.78	47.33	89.81	57.94	65.03	1136.47	41.99
Avg.	**124.67**	139.25	126.44	157.49	12218.65	7434.61	27170.23	81492.4
Total	**45.07**	60.83	55.18	76.26	2450.89	1495.27	5445.49	16305.88

Table 10: Pseudo-perplexity scores comparison across different adapters for mBERT in ConceptNet and Glot. †Language not included in mBERT pre-training. FFT denotes full fine-tuning of a base model on the target-language Glot data. The underlined FFT scores indicate that FFT outperform the best performing adapter for a respective language.

H Masked Language Modeling Pseudo-Perplexity - Part II

ISO	XLM-R							
		ConceptNet			Glot			
	Base	Seq_bn	LoRA	Seq_bn_inv	Seq_bn	LoRA	Seq_bn_inv	FFT
th	**7.83**	8.67	8.86	10.11	8.78	7.97	9.39	22.16
ro	2.97	3.76	3.79	4.51	3.42	**2.96**	3.25	6.18
bg	**3.61**	4.88	5.51	5.4	3.63	3.7	3.64	6.12
da	4.29	5.56	5.94	6.21	6.69	**4.21**	4.58	7.9
el	**2.56**	3.17	3.1	3.46	2.97	2.63	2.87	3.81
he	**5.74**	6.17	6.36	6.74	5.8	5.84	5.99	10.95
sk	3.93	4.85	4.67	5.36	4.56	**3.68**	4.08	4.62
sl	4.79	7.31	7.41	8.68	4.35	**4.01**	4.95	5.3
lv	4.14	5.96	6.32	9.34	5.09	**3.92**	4.7	4.87
ms	10.79	15.02	15.82	17.26	8.97	**8.8**	9.65	12.55
ka	**3.88**	4.41	4.47	4.48	3.99	3.94	4.76	4.97
bn	6.5	7.22	7.17	7.6	**5.95**	6.28	8.0	6.69
az	7.52	11.21	11.45	15.95	8.27	**7.58**	9.7	14.11
ur	10.17	12.13	12.82	12.23	**9.53**	9.54	11.12	12.32
mk	5.19	6.74	7.51	7.28	4.82	**4.78**	**4.78**	8.14
te	6.76	8.12	8.11	8.31	**6.41**	6.66	9.92	7.6
ne	12.76	16.87	17.74	16.91	11.86	**11.82**	22.42	16.64
si	7.04	7.97	8.22	8.26	**5.74**	6.37	11.44	6.74
mr	10.25	11.83	12.12	12.67	9.11	**8.9**	16.42	21.99
sw	15.68	26.99	27.39	36.78	**7.76**	9.61	11.24	9.18
cy	9.37	13.94	16.05	17.51	**5.08**	5.88	8.11	4.7
am	10.87	14.77	15.4	15.15	**7.32**	8.44	17.0	10.49
uz	8.4	14.77	16.81	20.66	**5.46**	6.21	9.14	5.92
ku	159.39	72.75	84.04	69.25	**2.95**	4.34	19.27	3.88
ug†	6.87	13.97	12.48	16.76	**4.99**	5.97	12.48	16.13
jv	33.81	96.45	89.36	116.95	**12.49**	15.06	27.14	26.25
su	57.32	134.71	128.95	152.14	**10.41**	15.22	29.1	25.16
Avg.	15.65	20.01	20.29	22.81	**6.53**	6.83	10.56	10.57
mt‡	395.18	283.77	335.23	275.56	**3.19**	5.0	12.01	3.36
bo‡	**9.45**	937.1	2036.45	1209.39	353.49	274.66	1972.96	597.55
yo‡	207.26	225.8	335.24	223.49	**9.57**	14.31	155.99	19.12
Avg.	203.96	482.22	902.31	569.48	122.08	**97.99**	713.65	206.68
Total	34.48	66.23	108.49	77.48	18.09	**15.94**	80.87	30.18

Table 11: Pseudo-perplexity scores comparison for XLM-R across different adapters in ConceptNet and Glot. ‡Language not included in XLM-R pre-training. FFT denotes full fine-tuning of a base model on the target-language Glot data. The underlined FFT scores indicate that FFT outperform the best performing adapter for a respective language.

I Correlation Between Pseudo-Perplexity Pre- and Post-training Data Sizes

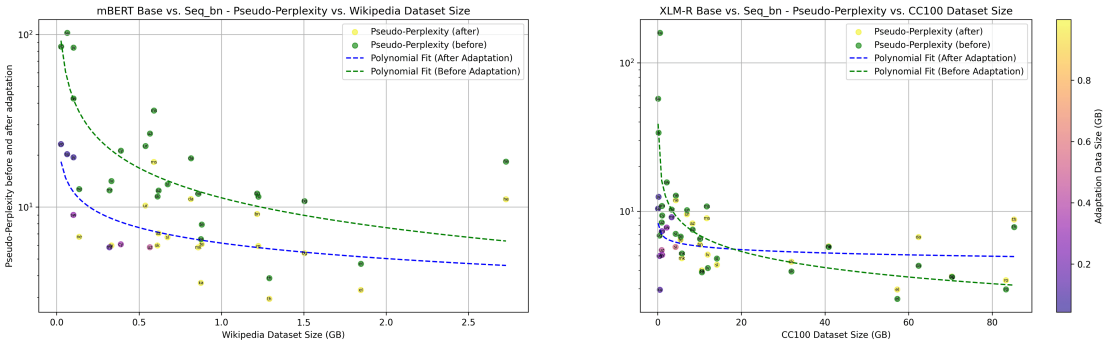

Figure 2: Correlation between the pre-training data sizes for mBERT and XLM-R and the pseudo-perplexities with the values fit in the log-space for the pre-adaptation and post-adaptation results.

	Model	Pearson (p-value)	Spearman (p-value)
Pre-adapt	mBERT	-0.37 (*0.07*)	-0.51 (*0.01*)
	XLM-R	-0.32 (*0.1*)	-0.39 (*0.04*)
Post-adapt	mBERT	-0.69 (*<0.001*)	-0.79 (*<0.001*)
	XLM-R	-0.27 (*0.16*)	-0.79 (*<0.001*)

Table 12: Pearson and Spearman Correlations for mBERT and XLM-R between pseudo-perplexity and amounts of pre-training and post-training data for the pre-adaptation and post-adaptation results. Post-adaptation results are based on the models with Seq_bn language adapters and denote the correlation between the sum of the pre-training and adaptation data sizes and pseudo-perplexity scores after the adaptation.

As illustrated in Figure 2, the improvements in pseudo-perplexity for both models are primarily concentrated in languages with smaller pre-training data sizes, which are positioned on the left side of the plots. These languages benefit the most from the adaptation process. Conversely, for languages with substantial representation in the pre-training data, the improvements are less pronounced or nonexistent. *This suggests that underrepresented languages in the pre-training data can achieve significant gains in pseudo-perplexity even with modest amounts of adaptation data and low-capacity adapters (smaller parameter counts). In contrast, further improvements for well-represented languages may require increasing the capacity of the adapters to better utilize their substantial pre-training representation.* The stagnation, or drops, in the performance on the languages with extensive pre-training data effects can also be attributed to the model seeing the same (duplicated) data that was seen during pre-training, which makes the "value" of data lower since the model sees the duplicates (Lee et al., 2022b).

J Comparison of XLM-R-base with Glot500 and XLM-R-large

ISO	XLM-R-base	Adapted XLM-R-base	XLM-R-large	Glot-500m
th	7.83	7.97	**4.92**	31.34
ro	2.97	2.96	**2.06**	13.29
bg	3.61	3.63	**2.53**	14.16
da	4.29	4.21	**2.78**	28.06
el	2.56	2.97	**1.87**	6.87
he	5.74	5.8	**3.19**	32.80
sk	3.93	3.68	**2.30**	26.36
sl	4.79	4.01	**2.60**	41.98
lv	4.14	3.92	**2.51**	14.55
ms	10.79	8.8	**6.71**	38.46
ka	3.88	3.94	**2.69**	10.77
bn	6.50	5.95	**3.99**	19.36
az	7.52	7.58	**4.40**	17.46
ur	10.17	9.53	**6.10**	25.60
mk	5.19	4.78	**3.23**	14.00
te	6.76	6.41	**4.31**	17.19
ne	12.76	11.82	**8.06**	23.19
mr	10.25	8.9	**5.77**	27.95
sw	15.68	**7.76**	8.90	44.82
cy	9.37	5.08	**4.35**	25.74
uz	8.40	5.46	**3.92**	15.33
jv	33.81	**12.49**	17.83	73.46
su	57.32	**10.41**	26.42	52.65
si	7.04	5.74	**4.50**	15.03
am	10.87	7.32	**6.73**	25.56
ku	159.39	**2.95**	126.40	23.35
ug	6.87	4.99	**3.80**	13.67
Avg.	15.65	**6.26**	10.11	25.66
mt	395.18	**3.19**	317.81	7.93
bo	9.45	274.66	**3.99**	26.74
yo	207.26	**9.57**	155.57	96.80
Avg.	203.96	95.81	159.12	**43.82**
Total	34.48	**15.22**	25.01	27.48

Table 13: Average pseudo-perplexity scores for 30 languages across three model configurations. For the adapted XLM-R-base, we pick the adapter with the best performance.

We additionally compare XLM-R adapted with Glot language adapters against two larger models: XLM-R-large (Conneau et al., 2020) and Glot500-m (Imani et al., 2023) (Table 13). Both models provide distinct points of comparison. XLM-R-large shares the same architecture as XLM-R-base but with a significantly larger size (550M parameters). XLM-R-large outperformed smaller models with adapters on MLM, suggesting that adapter effectiveness might be inherently constrained by the base model's capacity. In contrast, Glot500-m, while only slightly larger than XLM-R-base (395M parameters), introduces an extended vocabulary to support new scripts from a 600GB multilingual corpus and fine-tunes the weights of XLM-R-base. Its training employs a sampling strategy with an alpha of 0.3, prioritizing low-resource languages over high-resource ones. While this approach improves its performance on many low-resource languages, it results in suboptimal outcomes for well-represented languages.

This comparison is particularly relevant as it evaluates whether fine-tuning XLM-R-base with Glot-based language adapters can surpass the performance of these larger models. Furthermore, Glot500-m provides a unique benchmark, as it was trained on the same multilingual corpus used for our adapters, albeit without the computational constraints that limited our data size for adaptation.

K Correlation Between Pseudo-Perplexity and Downstream Tasks

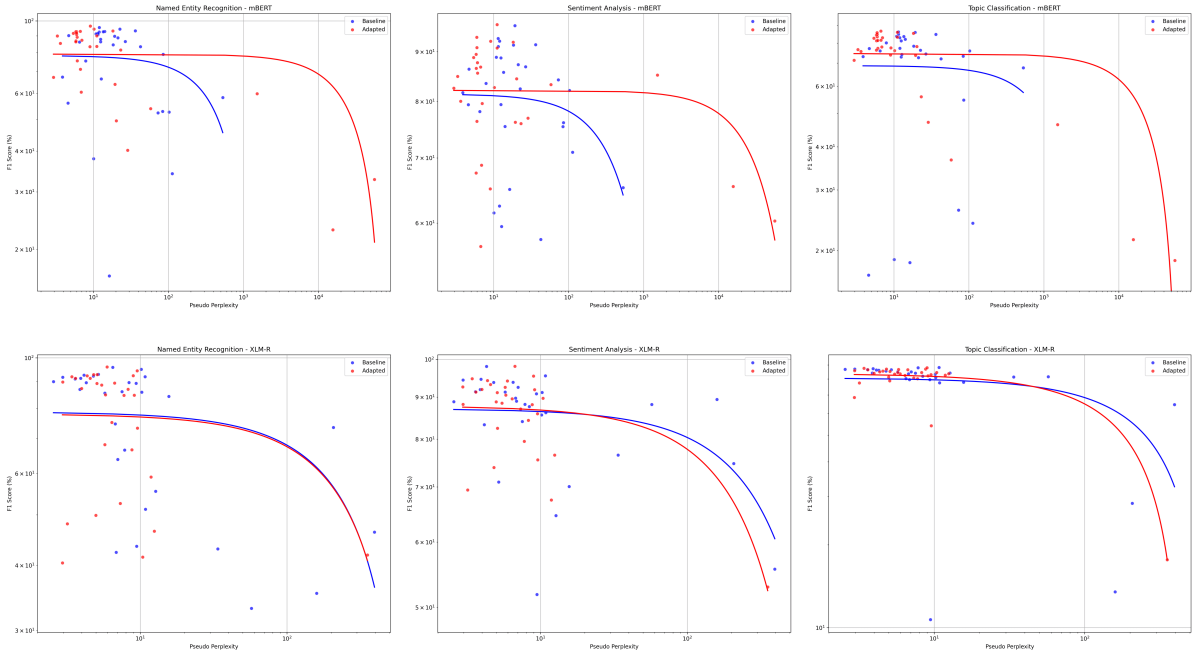

Figure 3: Correlation between the downstream performance for mBERT and XLM-R pre- and post-adaptation and the pseudo-perplexities.

Model	Task	Pre-Adapt		Post-Adapt	
		Pearson (p-value)	**Spearman (p-value)**	**Pearson (p-value)**	**Spearman (p-value)**
mBERT	TC	-0.09 (*0.62*)	-0.25 (*0.18*)	-0.66 (*<0.001*)	-0.42 (*0.02*)
	SA	-0.29 (*0.12*)	-0.15 (*0.42*)	-0.45 (*0.01*)	-0.23 (*0.23*)
	NER	-0.28 (*0.13*)	-0.22 (*0.24*)	-0.54 (*0.002*)	-0.49 (*0.006*)
XLM-R	TC	-0.48 (*0.007*)	-0.68 (*<0.001*)	-0.88 (*<0.001*)	-0.20 (*0.3*)
	SA	-0.47 (*0.009*)	-0.55 (*0.002*)	-0.64 (*<0.001*)	-0.38 (*0.04*)
	NER	-0.42 (*0.02*)	-0.62 (*<0.001*)	-0.35 (*0.06*)	-0.28 (*0.13*)

Table 14: Pearson and Spearman Correlations for mBERT and XLM-R (Pre-Adapt and Post-Adapt) between pseudo-perplexity and task performance. Post-Adapt is represented by the models adapted with the Seq_bn language adapters.

L Topic Classification Results - Part I

ISO	mBERT							
		ConceptNet			Glot			
	Base	Seq_bn	LoRA	Seq_bn_inv	Seq_bn	LoRA	Seq_bn_inv	FFT
he	79.79	**83.99**	82.87	82.11	83.26	83.43	83.91	83.24
el	**79.47**	77.95	79.14	78.12	76.65	77.92	76.64	84.81
bg	**84.39**	83.71	84.17	83.38	82.64	82.87	82.58	85.88
th	74.18	**74.66**	73.9	74.42	71.34	74.47	72.47	76.44
ro	86.95	87.86	86.45	**88.37**	85.8	86.63	86.8	89.06
bn	76.18	77.65	74.52	76.69	77.51	**78.09**	77.34	77.3
te	80.03	**82.35**	80.04	81.13	77.32	81.2	78.95	79.33
ka	76.28	73.26	74.26	74.07	75.68	**78.23**	75.19	79.82
mk	83.44	84.48	84.34	83.79	84.53	84.92	**85.25**	84.96
da	87.06	86.85	86.63	**87.72**	86.03	86.48	85.5	85.8
sl	83.6	85.07	83.75	86.22	86.71	85.39	**86.73**	86.43
az	81.09	83.72	82.53	83.38	82.93	82.55	**84.29**	82.01
sk	84.37	83.49	83.98	**85.4**	84.79	84.43	83.57	84.52
ms	84.31	84.65	84.1	82.94	**85.4**	84.59	83.39	84.38
uz	76.57	73.89	73.71	75.76	**81.32**	74.44	79.35	85.35
ur	76.7	73.7	74.85	74.76	76.06	75.26	**76.94**	78.18
cy	72.37	72.23	71.6	73.49	**81.47**	77.16	80.75	85.53
lv	82.28	**83.63**	82.42	82.45	83.48	82.56	80.94	85.02
mr	73.21	**77.29**	76.22	76.61	76.37	75.73	75.28	78.84
ne	73.72	77.55	74.62	76.02	**81.59**	75.21	80.8	79.11
jv	72.4	73.32	**75.12**	73.11	73.71	74.09	74.02	75.89
sw	69.17	70.53	69.89	70.21	73.93	69.05	**77.15**	85.89
su	76.15	77.42	77.62	77.0	78.21	**79.2**	78.63	79.97
yo	54.18	52.11	52.08	54.89	55.93	55.93	**58.05**	63.66
Avg.	77.67	78.39	77.87	78.42	79.28	78.74	**79.35**	81.73
mt†	69.86	69.83	69.85	68.79	78.0	78.09	**79.8**	83.32
ku†	28.76	23.78	15.71	19.93	46.41	40.22	**46.85**	52.82
ug†	23.4	22.21	20.9	22.17	47.18	31.68	**48.91**	56.26
si†	17.45	14.3	14.88	14.95	**21.53**	21.25	20.4	19.08
am†	17.75	14.01	18.47	12.94	18.74	**20.3**	18.07	16.88
bo†	12.59	11.08	9.48	6.33	36.67	28.36	**39.17**	33.53
Avg.	28.72	25.87	24.88	24.18	41.42	36.65	**42.2**	43.65
Total avg.	67.88	67.89	67.27	67.57	71.71	70.32	**71.92**	74.11

Table 15: F1 scores comparison across different adapters for mBERT in ConceptNet and Glot for topic classification. All results are averaged over 3 independent runs with different random seeds.

M Topic Classification Results - Part II

ISO	XLM-R							
		ConceptNet			Glot			
	Base	Seq_bn	LoRA	Seq_bn_inv	Seq_bn	LoRA	Seq_bn_inv	FFT
th	87.93	87.19	87.22	85.99	86.97	86.8	**88.5**	84.21
ro	86.94	87.0	86.85	**88.02**	87.47	86.95	87.6	88.03
bg	86.55	86.0	**87.81**	86.41	86.46	86.33	86.19	87.53
da	86.04	84.94	83.7	84.26	86.47	84.88	**86.41**	87.06
el	**86.74**	85.59	85.64	84.32	85.77	85.28	86.6	88.1
he	85.02	84.9	83.8	84.79	**86.62**	84.19	83.36	84.67
sk	**87.18**	85.53	84.81	85.2	85.46	86.52	86.03	85.59
sl	85.47	86.24	**86.95**	86.28	84.94	86.67	85.28	88.12
lv	86.25	87.83	86.93	**88.97**	85.22	86.38	87.41	87.52
ms	**88.12**	87.11	85.82	85.81	87.94	85.21	87.94	89.49
ka	84.08	**85.37**	83.79	83.18	83.92	85.0	83.95	82.27
bn	80.29	81.11	80.85	82.09	**83.56**	82.59	83.38	84.95
az	84.05	85.86	84.24	85.07	84.43	85.16	**86.39**	86.08
ur	**83.25**	81.04	80.29	82.35	82.97	81.98	82.17	83.97
mk	86.45	86.41	86.99	85.45	86.94	85.97	**87.15**	88.15
te	83.58	**83.64**	84.26	83.13	82.43	84.13	83.43	85.65
ne	84.14	83.98	83.92	83.77	82.65	**84.71**	82.85	84.2
si	84.92	84.54	84.86	82.23	84.49	83.37	**84.99**	84.53
mr	81.03	82.84	81.34	80.08	82.2	79.54	**84.23**	84.21
sw	77.83	75.58	76.23	77.97	80.23	78.73	**81.57**	85.95
cy	79.54	78.44	80.1	78.99	78.83	79.15	**81.37**	85.17
am	77.5	78.4	77.93	77.91	80.67	77.52	**81.51**	84.22
uz	81.93	78.73	78.43	76.97	**83.35**	81.13	80.68	86.37
ku	13.49	14.09	15.76	17.28	68.57	46.29	**73.97**	81.72
ug	79.56	79.11	78.67	78.86	81.29	**82.23**	80.14	84.95
jv	81.35	79.32	82.23	81.43	**83.59**	81.84	81.74	81.2
su	81.5	81.25	79.65	80.42	84.51	83.86	**84.66**	84.49
Avg.	81.14	80.82	80.71	80.64	83.63	82.31	**84.06**	85.61
mt‡	64.56	63.62	61.43	64.43	77.39	69.74	**77.92**	84.35
bo‡	10.69	9.89	9.73	11.74	17.65	**17.85**	16.93	20.41
yo‡	28.29	26.06	16.07	24.6	54.13	35.24	**59.44**	67.13
Avg.	34.52	33.19	29.08	33.59	49.72	40.94	**51.43**	57.3
Total avg.	76.48	76.05	75.54	75.93	80.24	78.17	**80.79**	82.77

Table 16: F1 scores comparison across different adapters for XLM-R in ConceptNet and Glot for topic classification. All results are averaged over 3 independent runs with different random seeds.

N Correlation Between Topic Classification and Pre- and Post-training Data

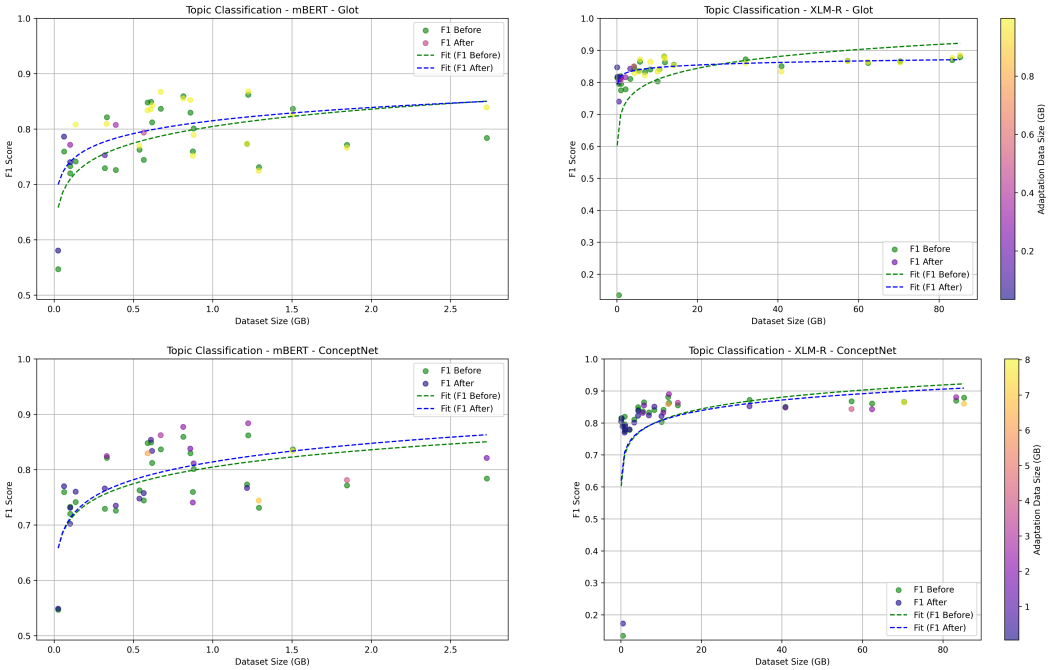

Figure 4: Correlation between the downstream performance for mBERT and XLM-R and the pre-training data and adaptation data.

Model	Task	Pre-Adapt		Post-Adapt (Glot)		Post-Adapt (CN)	
		P (p-value)	S (p-value)	P (p-value)	S (p-value)	P (p-value)	S (p-value)
mBERT	TC	0.35 (*0.1*)	0.53 (*0.008*)	0.45 (*0.03*)	0.32 (*0.13*)	0.38 (*0.06*)	0.55 (*0.006*)
XLM-R	TC	0.28 (*0.16*)	0.82 (*<0.005*)	0.55 (*0.002*)	0.75 (*<0.005*)	0.28 (*0.15*)	0.83 (*<0.005*)

Table 17: Pearson and Spearman Correlations for mBERT and XLM-R (Pre-Adapt and Post-Adapt) between task performance and data amounts. Post-Adapt is represented by the models adapted with the Seq_bn_inv language adapters and denote the correlation between the sum of the pre-training and adaptation data sizes and downstream task performance scores after the adaptation.

O Named Entity Recognition Results - Part I

ISO	mBERT								
		ConceptNet			Glot			Fusion	
	Base	Seq_bn	LoRA	Seq_bn_inv	Seq_bn	LoRA	Seq_bn_inv	Seq_bn	Seq_bn_inv
he	84.46	84.1	84.24	84.59	83.57	84.22	83.89	**84.84**	84.53
el	90.16	90.11	90.45	90.27	89.9	**90.5**	89.35	90.3	90.0
bg	91.25	91.64	91.64	91.48	91.64	91.59	91.56	**91.78**	91.76
th	67.34	65.65	66.79	66.68	67.22	**67.8**	66.95	67.36	67.57
ro	91.61	91.88	91.85	91.89	91.74	91.65	91.79	91.69	**92.17**
bn	95.46	96.07	95.82	**96.49**	96.42	96.03	96.3	95.86	96.1
te	75.41	76.17	76.94	75.29	75.51	74.69	75.37	76.53	**77.02**
ka	**86.17**	86.07	86.11	86.05	85.32	85.89	85.71	86.05	86.07
mk	92.43	92.09	92.3	92.2	**92.62**	92.2	92.02	91.61	91.98
da	89.76	90.08	**90.33**	89.74	90.02	89.72	88.99	89.41	89.48
sl	92.61	92.85	92.82	92.78	**92.93**	92.62	92.77	92.71	92.56
az	87.81	87.54	87.8	**88.23**	87.27	87.3	87.3	86.46	87.22
sk	90.87	90.88	90.89	90.96	90.83	**91.3**	90.99	**91.04**	90.84
ms	93.26	93.0	92.98	92.95	93.16	**93.93**	93.47	92.65	92.59
uz	86.48	86.69	86.58	86.33	86.87	86.46	87.73	87.5	**88.45**
ur	94.37	94.2	93.93	94.23	94.4	94.26	94.29	94.25	**94.85**
cy	88.72	89.34	89.35	89.05	89.18	89.36	**90.02**	88.95	88.71
lv	92.78	92.82	93.25	93.16	92.7	92.94	92.64	**93.34**	92.66
mr	**86.34**	86.19	85.97	86.29	86.32	86.07	84.35	86.24	86.22
ne	66.45	61.96	61.75	64.56	**71.12**	69.37	70.46	70.18	66.89
jv	52.87	62.83	61.76	**65.3**	63.97	58.73	63.34	57.21	58.67
sw	83.41	83.44	83.99	83.54	83.4	83.79	**84.07**	81.68	81.96
su	52.62	55.88	53.72	57.53	49.48	50.79	51.6	57.12	**57.74**
yo	79.0	83.02	**83.87**	83.1	81.48	79.58	79.74	79.81	78.54
Avg.	83.82	84.35	84.38	**84.7**	84.46	84.2	84.36	84.36	84.36
mt†	58.3	49.01	51.58	50.46	60.55	61.41	**64.93**	60.32	62.93
ku†	52.34	60.41	**59.92**	59.39	59.9	52.93	51.51	52.33	52.4
ug†	34.1	35.33	33.07	34.56	40.2	36.24	37.62	42.93	**44.05**
si†	16.59	13.41	14.06	13.94	22.97	14.58	19.94	20.7	**24.24**
am†	37.88	33.02	33.7	35.23	32.72	46.46	**46.49**	36.94	32.45
bo†	**56.04**	56.02	55.57	55.29	53.92	55.45	53.38	52.03	53.53
Avg.	42.54	41.2	41.32	41.48	45.04	44.51	**45.64**	44.21	44.93
Total avg.	75.56	75.72	75.77	76.05	76.58	76.26	**76.62**	76.33	76.47

Table 18: F1 scores comparison for mBERT in ConceptNet and Glot for named entity recognition. All results are averaged over 3 independent runs with different random seeds.

P Named Entity Recognition Results - Part II

ISO	XLM-R								
		ConceptNet			Glot			Fusion	
	Base	Seq_bn	LoRA	Seq_bn_inv	Seq_bn	LoRA	Seq_bn_inv	Seq_bn	Seq_bn_inv
th	66.55	66.4	**66.85**	66.76	66.63	65.29	66.2	65.89	66.82
ro	91.78	91.79	91.78	91.92	92.0	91.87	**92.18**	92.02	92.05
bg	91.09	91.22	91.36	**91.48**	91.34	91.4	91.43	90.91	91.43
da	89.58	89.57	89.54	89.45	89.44	89.85	89.72	89.85	**89.89**
el	90.03	90.32	89.88	90.14	89.89	90.02	90.02	90.18	**90.5**
he	85.56	85.48	85.45	84.99	84.92	**85.69**	85.28	85.35	85.4
sk	91.36	91.19	91.21	91.26	91.32	91.45	**91.49**	91.4	91.24
sl	92.28	**92.58**	92.16	92.41	92.36	92.05	92.33	92.21	92.12
lv	92.64	92.73	92.65	92.95	92.84	92.88	**93.1**	92.99	92.93
ms	92.0	92.36	91.65	92.28	92.4	92.06	91.9	**92.67**	91.82
ka	86.96	86.77	86.88	**87.73**	87.31	87.66	87.37	86.59	87.33
bn	95.87	95.66	95.9	96.06	96.07	96.13	96.09	95.57	**96.23**
az	86.13	85.34	86.47	86.53	87.03	86.63	**87.59**	86.23	86.38
ur	95.02	94.57	**95.04**	94.86	94.43	94.89	94.27	94.4	94.56
mk	92.97	92.47	**93.26**	92.28	92.83	92.68	92.72	92.32	92.46
te	74.67	73.64	**76.07**	74.27	75.18	74.38	74.82	72.92	73.91
ne	55.47	53.0	60.02	60.0	59.08	54.99	56.61	**67.84**	67.34
si	63.85	58.43	63.83	57.43	68.15	60.34	66.2	71.94	**73.66**
mr	85.92	85.86	85.5	85.77	84.75	85.25	**86.1**	85.8	85.52
sw	84.34	83.31	84.37	84.26	**84.72**	84.4	84.47	84.56	83.5
cy	89.33	88.9	88.88	88.97	89.3	**89.72**	89.41	89.4	89.36
am	51.22	49.9	49.29	48.18	52.57	47.17	51.67	**55.0**	52.55
uz	89.64	88.66	87.51	87.89	88.64	**89.97**	86.86	89.05	87.64
ku	35.34	39.53	42.99	43.83	40.41	31.43	29.4	**58.02**	56.93
ug	42.36	52.63	50.67	51.98	49.88	50.5	52.63	53.12	**58.5**
jv	42.99	45.64	44.7	50.87	46.51	44.7	47.96	**63.53**	58.81
su	33.07	38.4	42.26	48.32	41.47	39.76	42.89	**52.53**	49.61
Avg.	77.33	77.64	78.38	78.62	78.57	77.52	78.17	**80.83**	80.68
mt‡	46.31	32.69	40.11	32.13	48.03	41.54	53.57	**64.31**	57.57
bo‡	43.51	44.29	44.55	46.41	41.86	39.64	38.27	**48.15**	47.55
yo‡	73.54	71.2	73.46	74.59	73.3	74.87	75.09	73.04	**75.8**
Avg.	54.45	49.39	52.71	51.04	54.4	52.01	55.64	**61.83**	60.31
Total avg.	75.05	74.82	75.81	75.87	76.16	74.97	75.92	**78.93**	78.65

Table 19: F1 scores for XLM-R across ConceptNet and Glot for named entity recognition. All results are averaged over 3 independent runs with different random seeds.

Q Correlation Between Named Entity Recognition and Pre- and Post-training data

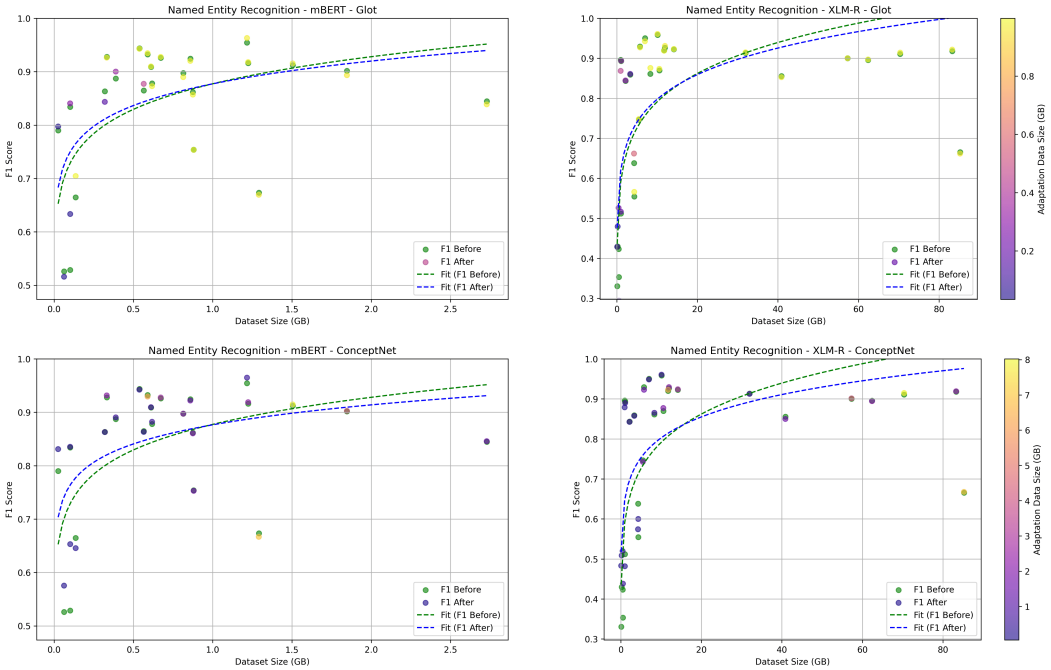

Figure 5: Correlation between the downstream performance for mBERT and XLM-R and the pre-training data and adaptation data.

Model	Task	Pre-Adapt		Post-Adapt (Glot)		Post-Adapt (CN)	
		P (p-value)	S (p-value)	P (p-value)	S (p-value)	P (p-value)	S (p-value)
mBERT	NER	0.32 (*0.1*)	0.32 (*0.1*)	0.42 (*0.04*)	0.29 (*0.2*)	0.20 (*0.3*)	0.44 (*0.03*)
XLM-R	NER	0.31 (*0.1*)	0.58 (*0.002*)	0.31 (*0.1*)	0.61 (*<0.005*)	0.32 (*0.1*)	0.60 (*<0.005*)

Table 20: Pearson and Spearman Correlations for mBERT and XLM-R (Pre-Adapt and Post-Adapt) between task performance and data amounts. Post-Adapt is represented by the models adapted with the Seq_bn_inv language adapters and denote the correlation between the sum of the pre-training and adaptation data sizes and downstream task performance scores after the adaptation.

R Sentiment Analysis Results - Part I

ISO	mBERT							
		ConceptNet			Glot			
	Base	Seq_bn	LoRA	Seq_bn_inv	Seq_bn	LoRA	Seq_bn_inv	FFT
he	91.42	91.55	90.44	90.81	90.79	90.87	**91.58**	90.6
el	**86.35**	86.27	86.05	86.22	84.88	84.95	84.52	86.38
bg	88.82	89.41	89.17	**89.54**	88.76	88.65	89.2	89.99
th	81.68	81.97	81.92	82.45	82.57	82.0	**83.23**	83.19
ro	92.87	92.67	92.62	92.64	93.13	**92.98**	92.96	93.7
bn	92.28	92.16	92.6	91.88	92.26	92.56	**92.57**	92.48
te	83.49	83.29	84.17	85.01	**85.55**	84.45	85.26	88.41
ka	78.12	78.1	76.68	76.05	80.03	80.23	**81.24**	86.97
mk	62.47	**69.01**	66.4	62.07	67.54	65.06	65.21	68.98
da	95.71	95.33	95.77	95.33	95.95	**96.15**	96.09	96.84
sl	85.71	86.46	86.28	85.81	86.79	86.4	**87.83**	88.66
az	79.42	79.59	79.72	80.03	79.62	**80.15**	80.13	81.44
sk	91.11	88.86	89.9	89.73	90.87	91.16	**92.18**	91.08
ms	91.5	92.03	91.87	91.99	92.06	91.7	**92.57**	93.83
uz	86.84	85.67	86.76	85.85	86.52	86.36	**86.85**	88.33
ur	82.43	81.89	82.01	82.13	82.69	82.66	**82.72**	83.81
cy	87.28	86.99	87.82	86.15	87.71	**87.76**	87.42	88.53
lv	75.41	75.66	73.99	74.71	76.32	75.41	**76.65**	79.24
mr	88.7	88.76	89.0	88.67	**89.43**	89.13	88.97	90.43
ne	59.51	51.46	**67.17**	55.31	56.77	59.35	63.19	63.47
jv	75.38	74.24	74.75	73.94	**76.16**	75.7	75.43	75.44
sw	57.71	54.25	57.24	52.9	65.05	62.21	**69.64**	54.6
su	82.13	84.25	84.62	83.33	84.42	**84.75**	83.99	84.06
yo	76.1	75.66	75.24	75.35	75.93	75.43	**77.85**	77.32
Avg.	82.18	81.9	82.59	81.58	82.99	82.75	**83.64**	84.07
mt†	65.24	65.68	62.82	66.88	68.79	**73.87**	65.34	74.11
ku†	84.2	82.82	83.97	83.37	85.14	84.46	**86.14**	85.55
ug†	70.94	68.35	72.67	72.19	76.91	71.35	**80.4**	76.63
si†	64.97	64.89	65.01	64.67	65.42	**66.02**	65.62	66.26
am†	61.45	62.02	60.87	61.45	60.3	61.62	**63.81**	59.48
bo†	79.4	79.12	79.38	80.67	**83.27**	82.33	82.14	81.77
Avg.	71.03	70.48	70.79	71.54	73.3	73.27	**73.91**	73.97
Total avg.	79.95	79.61	80.23	79.57	81.05	80.86	**81.69**	82.05

Table 21: F1 scores comparison across different adapters for mBERT in ConceptNet and Glot for sentiment analysis. All results are averaged over 3 independent runs with different random seeds.

S Sentiment Analysis Results - Part II

ISO	XLM-R							
		ConceptNet			Glot			
	Base	Seq_bn	LoRA	Seq_bn_inv	Seq_bn	LoRA	Seq_bn_inv	FFT
th	88.18	88.26	88.43	**88.46**	88.11	88.31	88.13	86.39
ro	94.37	94.84	95.03	**95.04**	94.74	94.67	95.03	94.55
bg	91.36	90.66	**91.43**	91.41	91.26	90.93	90.65	90.79
da	98.04	97.84	**98.13**	98.02	98.09	98.04	97.98	97.82
el	88.82	88.92	**88.98**	88.73	88.19	88.25	88.61	88.75
he	91.26	89.66	**91.81**	91.25	90.48	90.27	90.85	90.2
sk	**94.6**	93.86	93.87	93.43	93.22	93.72	93.44	94.03
sl	93.75	93.46	**94.32**	92.68	94.23	93.57	93.86	92.73
lv	83.3	**83.78**	83.36	83.83	82.47	83.12	83.65	82.97
ms	95.51	95.27	**95.66**	95.57	95.44	95.29	95.53	95.26
ka	**91.92**	91.51	90.8	91.21	**91.92**	91.11	91.41	93.33
bn	93.78	94.14	94.3	**94.46**	94.13	94.1	94.43	94.41
az	84.05	84.05	84.05	83.98	84.32	84.2	**84.74**	85.19
ur	85.6	85.99	85.67	85.85	85.89	**86.7**	86.25	87.27
mk	70.96	69.22	67.05	69.45	**73.9**	70.74	72.31	71.68
te	89.72	89.15	89.59	89.22	89.56	89.72	**89.9**	90.92
ne	64.6	**69.37**	64.06	63.02	67.49	68.38	68.65	65.46
si	92.49	92.59	92.18	**93.21**	92.49	91.78	91.96	92.85
mr	91.17	91.8	91.9	91.8	91.87	**92.36**	91.8	92.43
sw	70.08	65.37	77.11	75.3	**79.52**	77.24	74.45	83.84
cy	90.83	91.01	90.57	90.65	91.12	90.88	**91.36**	91.01
am	86.15	83.77	84.2	82.88	87.04	**87.9**	87.7	87.49
uz	87.63	88.24	88.37	88.13	**88.47**	87.98	88.39	90.08
ku	89.39	89.73	89.08	89.78	92.57	89.09	**93.31**	95.31
ug	88.97	88.88	89.91	87.64	88.81	**90.01**	89.65	91.72
jv	76.51	77.34	77.01	77.14	76.51	76.79	**77.65**	75.53
su	88.15	82.66	85.17	84.41	89.69	**90.34**	89.69	89.03
Avg.	87.45	87.09	87.48	87.28	**88.2**	87.98	**88.2**	88.56
mt‡	55.63	55.19	55.32	54.13	**69.4**	63.15	69.31	70.38
bo‡	51.81	47.33	51.07	49.34	**52.92**	50.9	50.69	55.19
yo‡	74.73	73.4	73.6	75.09	75.5	72.0	**77.65**	78.99
Avg.	60.72	58.64	60.00	59.52	**65.94**	62.02	65.88	68.19
Total avg.	84.78	84.24	84.73	84.50	**85.98**	85.38	85.97	86.52

Table 22: F1 scores comparison across different adapters for XLM-R in ConceptNet and Glot for sentiment analysis. All results are averaged over 3 independent runs with different random seeds.

T Correlation Between Sentiment Analysis and Pre- and Post-training data

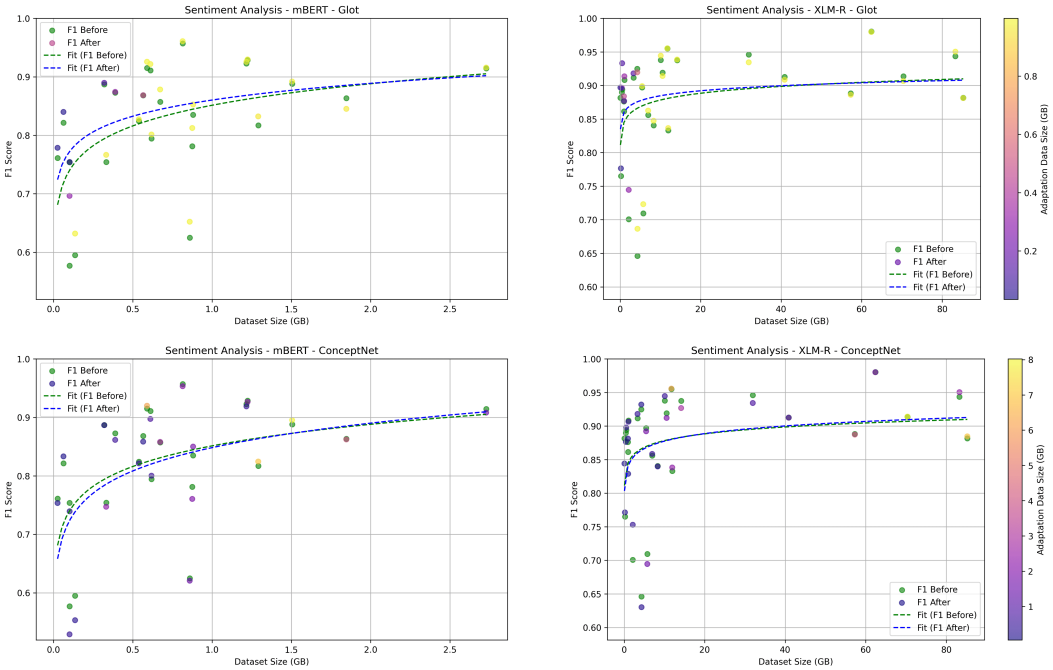

Figure 6: Correlation between the downstream performance for mBERT and XLM-R and the pre-training data and adaptation data.

Model	Task	Pre-Adapt		Post-Adapt (Glot)		Post-Adapt (CN)	
		P (p-value)	S (p-value)	P (p-value)	S (p-value)	P (p-value)	S (p-value)
mBERT	SA	0.45 (*0.03*)	0.50 (*0.01*)	0.38 (*0.07*)	0.41 (*0.05*)	0.39 (*0.06*)	0.52 (*0.009*)
XLM-R	SA	0.36 (*0.07*)	0.47 (*0.01*)	0.32 (*0.1*)	0.33 (*0.1*)	0.38 (*0.05*)	0.52 (*0.005*)

Table 23: Pearson and Spearman Correlations for mBERT and XLM-R (Pre-Adapt and Post-Adapt) between task performance and data amounts. Post-Adapt is represented by the models adapted with the Seq_bn_inv language adapters and denote the correlation between the sum of the pre-training and adaptation data sizes and downstream task performance scores after the adaptation.

U Results for Large-Scale Models for TC and NER

	GPT-3.5-turbo-0613	GPT-4-0613	LLaMAX2-7B-Alpaca	Llama-2-7b-chat-hf	Meta-Llama-3-8B	Meta-Llama-3.1-8B	Qwen1.5-7B	Qwen2-7B	bloom-7b1	bloomz-7b1	gemma-2-9b	gemma-7b	mala-500-10b-v1	mala-500-10b-v2	occiglot-7b-eu5	xglm-7.5B	yayi-7b-llama2
am	24.14	38.74	7.64	5.41	38.03	40.43	13.57	23.68	7.32	8.27	41.19	43.02	5.71	9.03	6.85	7.86	3.59
az	52.17	44.27	30.81	20.54	73.78	71.97	51.86	65.86	10.08	16.68	57.95	68.79	5.71	5.71	31.37	26.56	17.55
bn	54.29	50.55	23.79	9.35	65.89	63.43	42.62	66.08	10.75	20.93	51.22	66.91	5.71	5.69	22.42	28.25	12.99
bo	2.90	1.94	3.69	4.63	40.80	48.83	10.15	12.41	6.44	10.56	12.12	20.23	5.71	3.63	11.65	7.06	6.61
bg	54.80	58.33	31.47	29.92	64.95	63.53	55.15	77.06	20.17	16.58	51.85	63.26	5.71	5.23	44.70	41.81	24.77
ku	38.74	38.10	19.71	7.67	68.26	65.47	21.71	33.20	10.26	8.63	33.49	44.59	5.71	6.86	14.07	9.31	7.81
cy	43.08	42.47	26.76	18.08	68.75	68.69	37.47	49.93	10.45	18.09	50.38	56.87	5.71	5.71	26.21	17.57	19.37
da	52.71	52.17	33.17	34.03	73.03	73.73	57.73	75.95	17.85	21.90	45.05	71.14	5.71	5.39	49.20	56.88	32.02
el	54.29	60.27	21.84	21.69	70.22	73.70	46.99	63.73	11.97	11.90	39.08	67.20	5.71	5.71	31.48	55.80	20.84
he	56.84	51.09	24.39	17.55	69.01	69.80	46.93	70.07	10.87	8.53	44.35	64.03	5.71	4.76	22.82	10.66	9.51
jv	21.05	21.05	28.49	21.31	66.73	69.39	49.99	50.76	17.90	25.19	59.48	57.33	5.71	2.20	34.05	44.85	19.65
ka	47.19	43.68	18.37	15.25	68.58	63.50	32.76	52.02	3.50	14.76	58.73	69.17	5.71	8.13	25.17	9.35	13.24
lv	54.29	53.76	31.62	23.85	69.79	70.63	55.05	67.69	12.70	17.38	45.97	69.24	8.21	3.13	34.20	23.25	23.91
mr	52.71	51.09	19.90	14.04	64.84	63.07	39.41	56.66	26.78	29.62	27.30	56.58	5.71	5.83	19.63	23.24	9.59
mk	52.71	60.75	28.98	26.75	66.66	68.33	55.99	75.87	12.91	16.69	55.62	64.88	3.97	3.49	40.23	40.43	22.65
mt	44.27	50.55	29.18	23.07	63.25	67.22	44.26	56.10	11.45	20.18	43.93	62.54	5.71	5.71	34.33	28.45	24.09
ne	55.83	52.71	21.49	18.42	62.32	62.69	42.96	54.99	10.12	19.45	15.71	62.31	5.62	4.07	25.79	31.47	18.61
ro	51.64	54.80	34.88	31.49	70.19	72.20	56.43	74.64	20.10	20.76	52.51	69.08	5.71	5.71	47.32	43.15	30.92
si	23.38	62.63	8.66	4.81	60.25	57.45	12.49	29.29	5.98	9.38	46.12	65.92	6.60	2.20	10.82	5.48	5.71
sk	52.17	52.71	28.65	29.75	70.57	72.77	55.40	74.63	20.27	17.58	35.52	68.94	5.71	8.49	43.66	39.12	27.70
sl	53.76	47.76	33.60	31.05	75.67	70.18	55.53	63.56	11.10	17.18	48.42	67.87	9.22	3.30	40.19	30.21	28.09
su	26.38	20.26	28.22	23.89	63.50	67.46	46.31	58.94	17.55	21.68	60.68	65.78	5.71	7.69	32.18	44.52	21.59
sw	55.83	46.62	28.24	14.01	68.95	68.37	40.51	51.05	12.91	22.41	48.61	58.78	5.71	6.70	29.03	45.91	11.88
te	57.84	50.00	5.92	5.78	64.72	62.36	27.29	55.69	16.89	20.13	47.24	68.93	5.71	5.17	12.91	49.59	4.73
th	53.24	49.45	16.94	20.88	77.50	75.40	46.57	67.38	6.25	16.62	45.24	58.64	5.71	7.82	35.27	50.01	21.98
ug	44.27	46.04	6.53	6.90	66.23	62.22	12.37	54.64	9.29	11.72	33.74	45.77	7.54	3.66	16.20	7.76	7.13
ur	53.24	65.79	22.87	15.07	67.80	67.53	39.13	61.90	23.50	23.33	29.04	56.48	5.71	6.61	29.62	41.90	12.23
uz	44.87	34.82	29.50	13.49	69.53	68.35	33.53	54.55	10.05	13.89	56.50	65.44	5.71	11.34	26.86	15.93	10.11
yo	22.61	16.22	18.17	11.26	50.05	46.74	25.36	30.44	14.08	21.90	35.16	37.11	8.75	7.65	18.71	16.97	10.23
ms	49.45	55.83	30.46	27.35	74.10	73.28	56.74	76.05	11.13	23.31	55.96	69.52	5.71	5.71	40.15	46.19	27.33
Total avg.	45.02	45.82	23.13	18.24	65.80	65.62	40.41	56.83	13.02	17.51	44.27	60.21	6.04	5.74	28.57	29.98	16.88

Table 24: F1 Scores for All Large-Scale Models on TC. The results are based on 3-shot prompting, as reported by Ji et al. (2024). GPT-3.5 and GPT-4 results are zero-shot, obtained from Adelani et al. (2024a).

	Bloom	Bloomz	mT0	GPT-3.5-turbo-0301
th	1.0	0.2	1.4	-
el	19.7	13.0	12.8	69.3
ur	71.7	47.3	47.1	-
te	5.3	3.8	3.3	-
sw	58.8	26.8	24.3	-
bg	29.6	19.7	14.7	72.0
mr	27.9	20.4	12.3	-
bn	36.8	36.2	23.9	-
Total avg.	31.35	20.92	17.48	70.65

Table 25: Three-shot NER results across eight overlapping languages from BUFFET (Asai et al., 2023). The scores for GPT-3.5 are only provided for two languages.

	Qwen 1.5B	Qwen 7B	Llama 8B	Qwen 14B	Llama 70B
am	7.03	13.99	9.18	31.76	43.41
az	9.60	18.48	12.27	53.05	73.19
be	6.59	31.51	20.25	68.20	78.17
bo	2.38	8.17	9.67	18.92	62.63
bg	7.93	26.47	24.31	46.81	78.65
ku	6.77	18.48	20.10	17.98	77.52
cy	8.93	18.49	20.32	26.68	61.55
da	13.04	25.62	17.90	41.18	78.29
el	3.91	10.26	16.41	58.39	77.90
he	5.50	23.03	20.77	46.25	76.66
jv	10.51	19.45	19.53	28.04	66.43
ka	4.35	20.46	24.99	45.74	77.60
lv	11.14	14.29	17.60	44.14	74.09
mr	6.17	22.67	22.31	49.54	68.77
mk	4.91	24.16	22.44	44.38	77.66
mt	11.76	18.01	18.24	49.23	66.83
ne	4.70	23.59	26.36	55.34	69.25
ro	9.50	21.93	24.67	57.25	77.72
si	12.47	14.28	14.96	29.43	70.69
sk	6.66	15.61	21.37	45.38	75.80
sl	13.34	22.71	18.89	43.22	65.42
su	9.44	22.41	21.98	34.95	65.53
sw	10.38	11.15	15.45	18.60	67.94
te	9.19	17.90	27.21	38.99	75.35
th	8.49	40.22	20.80	73.49	74.23
ug	7.02	17.72	19.67	28.83	71.21
ur	3.71	27.47	24.23	47.75	80.07
uz	11.76	21.45	17.02	38.32	70.58
yo	6.70	13.49	15.20	18.57	45.55
ms	10.58	21.73	27.82	56.00	73.02
Total avg.	8.15	20.17	19.73	41.88	70.72

Table 26: F1 Scores for DeepSeek-R1 distilled models of various sizes for TC. The results are based on zero-shot prompting and were obtained in our evaluation.

Language	**TC**		**NER**		**SA**	
	LLaMA-3 (Baseline)	**LLaMA-3 +Seq_bn_inv**	**LLaMA-3 (Baseline)**	**LLaMA-3 +Seq_bn_inv**	**LLaMA-3 (Baseline)**	**LLaMA-3 +Seq_bn_inv**
cy	33.64	72.50	76.36	77.03	58.36	88.43
si	16.67	39.11	30.84	30.08	80.42	83.8
sw	29.05	60.21	67.08	67.33	45.47	51.22
ug	19.37	52.32	26.88	28.23	52.12	63.89
mt	60.93	77.14	24.72	22.94	57.77	56.06
Total avg.	31.93	60.26	45.18	45.12	58.83	68.68

Table 27: Comparison of F1 Scores for LLaMA-3 Baseline (fine-tuned with a task adapter) and LLaMA-3+Seq_bn_inv on TC, NER, and SA. All results are averaged over 3 independent runs with different random seeds.