mGPT: Few-Shot Learners Go Multilingual

Oleh Shliazhko[1*], **Alena Fenogenova**[2], **Maria Tikhonova**[2,3],
Anastasia Kozlova[2], **Vladislav Mikhailov**[2*†], **Tatiana Shavrina**[2,4,5,6,*]
[1]Independent Researcher, The Netherlands, [2]SaluteDevices, Russia, [3]HSE University, Russia,
[4]AIRI, Russia, [5]AI Center, NUST MISiS, Russia, [6]Institute of Linguistics RAS, Russia
olehshliazhko@gmail.com, alenush93@gmail.com, mtihonova@hse.ru,
anastasi2510@gmail.com, vvmkhlvv@gmail.com, rybolos@gmail.com

Abstract

This paper introduces mGPT, a multilingual variant of GPT-3, pretrained on 61 languages from 25 linguistically diverse language families using Wikipedia and the C4 Corpus. We detail the design and pretraining procedure. The models undergo an intrinsic and extrinsic evaluation: language modeling in all languages, downstream evaluation on cross-lingual NLU datasets and benchmarks in 33 languages, and world knowledge probing in 23 languages. The in-context learning abilities are on par with the contemporaneous language models while covering a larger number of languages, including underrepresented and low-resource languages of the Commonwealth of Independent States and the indigenous peoples in Russia. The source code and the language models are publicly available under the MIT license.

1 Introduction

The advent of the Transformer architecture (Vaswani et al., 2017) has facilitated the development of various language models (LMs; Liu et al., 2020a). Although the well-established ''pretrain & finetune'' paradigm has led to rapid progress in NLP (Wang et al., 2019), it imposes several limitations. Finetuning relies on an extensive amount of labeled data. Collecting high-quality labeled data for new tasks and languages is expensive and resource-consuming (Wang et al., 2021). LMs can learn spurious correlations from finetuning data (Naik et al., 2018; Niven and Kao, 2019) and demonstrate inconsistent generalization, catastrophic forgetting, or brittleness to finetuning data order (McCoy et al., 2020; Dodge et al., 2020). Last but not least, finetuning requires additional computational resources and, therefore, aggravates the problem of a large carbon footprint (Bender et al., 2021).

The latest approaches address these limitations with zero-shot and few-shot learning, performing a task with LM scoring or conditioning on a few demonstration examples without parameter updates (Brown et al., 2020). Autoregressive LMs adopted via these paradigms have been widely applied in many NLP tasks (Schick and Schütze, 2021; Perez et al., 2021), notably in cross-lingual knowledge transfer (Winata et al., 2021) and low-resource language scenarios (Lin et al., 2022). However, model development for underrepresented typologically distant and low-resource languages (Wu and Dredze, 2020; Lauscher et al., 2020; Hedderich et al., 2021) and cross-lingual generalization abilities of autoregressive LMs (Erdem et al., 2022) have been left understudied.

This paper presents mGPT, a multilingual version of GPT-3 (Brown et al., 2020) available in 1.3B ($\text{mGPT}_{1.3B}$) and 13B (mGPT_{13B}) parameters. We aim (i) to develop a large-scale multilingual autoregressive LM that inherits the GPT-3's generalization benefits and (ii) to increase the linguistic diversity of multilingual LMs, making the first attempt to address languages of the Commonwealth of Independent States (CIS) and under-resourced languages of the indigenous peoples in Russia. We pretrain mGPT in 61 languages from 25 language families on Wikipedia and Colossal Clean Crawled Corpus (C4; Raffel et al., 2020). We analyze the mGPT's performance on various intrinsic and extrinsic tasks and compare it with the contemporaneous generative LMs.

Key Findings The analysis reveals that (i) mGPT_{13B} is comparable to $\text{XGLM}_{1.7B}$ (Lin et al., 2022) while having fewer weights and covering a larger number of languages, (ii) mGPT

[*]Work done while at SaluteDevices.
[†]Now at University of Oslo.

Transactions of the Association for Computational Linguistics, vol. 12, pp. 58–79, 2024. https://doi.org/10.1162/tacl_a_00633
Action Editor: Miguel Ballesteros. Submission batch: 8/2023; Revision batch: 6/2023; Published 1/2024.

shows confident performance on Austronesian, Austro-Asiatic, Japonic, Germanic, and Romance languages on multiple tasks and prominent language modeling abilities on the languages of the indigenous peoples in Russia, (iii) adding more demonstrations may result in performance degradation for both mGPT and XGLM, and (iv) hate speech detection is one of the most challenging tasks, receiving random guessing performance in the zero-shot and few-shot evaluation setups. External validation by the NLP community since the release[1] shows that $mGPT_{1.3B}$ can outperform large-scale LMs on SuperGLUE tasks and promote strong solutions for multilingual clause-level morphology tasks. We release the model evaluation code,[2] the $mGPT_{1.3B}$[3] and $mGPT_{13B}$[4] models. We hope to facilitate research on the applicability of autoregressive LMs in non-English languages and increase the linguistic inclusivity of the low-resource languages.

Language Family	Languages
Afro-Asiatic	Arabic (ar), Hebrew (he)
Austro-Asiatic	Vietnamese (vi)
Austronesian	Indonesian (id), Javanese (jv), Malay (ms) Tagalog (tl)
Baltic	Latvian (lv), Lithuanian (lt)
Basque	Basque (eu)
Dravidian	Malayalam (ml), Tamil (ta), Telugu (te)
Indo-European (Armenian)	Armenian (hy)
Indo-European (Indo-Aryan)	Bengali (bn), Marathi (mr), Hindi (hi), Urdu (ur)
Indo-European (Germanic)	Afrikaans (af), Danish (da), English (en), German (de), Swedish (sv)
Indo-European (Romance)	French (fr), Italian (it), Portuguese (pt), Romanian (ro), Spanish (es)
Indo-European (Greek)	Greek (el)
Indo-European (Iranian)	Ossetian (os), Tajik (tg), Persian (fa)
Japonic	Japanese (ja)
Kartvelian	Georgian (ka)
Koreanic	Korean (ko)
Kra-Dai	Thai (th)
Mongolic	Buryat (bxr), Kalmyk (xal), Mongolian (mn)
Niger-Congo	Swahili (sw), Yoruba (yo)
Slavic	Belarusian (be), Bulgarian (bg), Russian (ru), Ukrainian (uk), Polish (pl)
Sino-Tibetan	Burmese (my)
Turkic (Karluk)	Uzbek (uz)
Turkic (Kipchak)	Bashkir (ba), Kazakh (kk), Kyrgyz (ky), Tatar (tt)
Turkic (Oghuz)	Azerbaijani (az), Chuvash (cv), Turkish (tr), Turkmen (tk)
Turkic (Siberian)	Tuvan (tyv), Yakut (sax)
Uralic	Estonian (et), Finnish (fi), Hungarian (hu)

Table 1: A list of languages by the language family.

2 Related Work

Multilingual Transformers Recent years have featured the development of various monolingual and multilingual LMs initially designed for English. BERT (Devlin et al., 2019) has been replicated in other high-resource languages (Martin et al., 2020; Masala et al., 2020) and language families, e.g., Indian (Kakwani et al., 2020) and Balto-Slavic (Arkhipov et al., 2019). Massively multilingual LMs—mBERT, XLM-R (Conneau et al., 2020), RemBERT (Chung et al., 2021), mBART (Liu et al., 2020b) and mT5 (Xue et al., 2021)—have now pushed state-of-the-art results on various NLP tasks in multiple languages (Kalyan et al., 2021). Such models support more than 100 languages and vary in the architecture design and pretraining objectives. By contrast, our work presents one of the first multilingual *autoregressive* LMs covering more than 61 languages.

GPT-based Language Models Large-scale generative LMs (e.g., GPT-3; Brown et al., 2020) are triggering a shift from the ''pretrain & finetune'' paradigm to prompt-based learning (Liu et al., 2023a). The benefit of balancing the pretraining costs and performing standardized NLP tasks with a few demonstration examples has stimulated the development of open-source autoregressive LMs for English (e.g., Black et al., 2022; Biderman et al., 2023; Dey et al., 2023), Chinese (Zeng et al., 2021), and Russian (Zmitrovich et al., 2023). A few contemporaneous works extend the research on zero-shot and few-shot learning, evaluating the in-context abilities of GPT-based LMs in multilingual scenarios. Winata et al. (2021) report that English GPTs perform significantly better than random guessing with monolingual and multilingual prompts on typologically close languages, such as French, Spanish, and German. Lin et al. (2022) propose XGLM, a multilingual GPT-style LM in 30 languages, and empirically show that it can outperform its monolingual counterparts of the comparable number of parameters. We use XGLM as the main baseline in our experiments and analyze the results of comparing $mGPT_{1.3B}$ with other autoregressive LMs published after our release, such as BLOOM (Scao et al., 2023).

3 Method

3.1 Pretraining Data

Language Selection Table 1 summarizes the list of languages by their family. The pretraining

[1] As of the time of writing this paper, $mGPT_{1.3B}$ was publicly available. Note that $mGPT_{13B}$ is also now released.

[2] github.com/ai-forever/mgpt.

[3] hf.co/ai-forever/mGPT.

[4] hf.co/ai-forever/mGPT-13B.

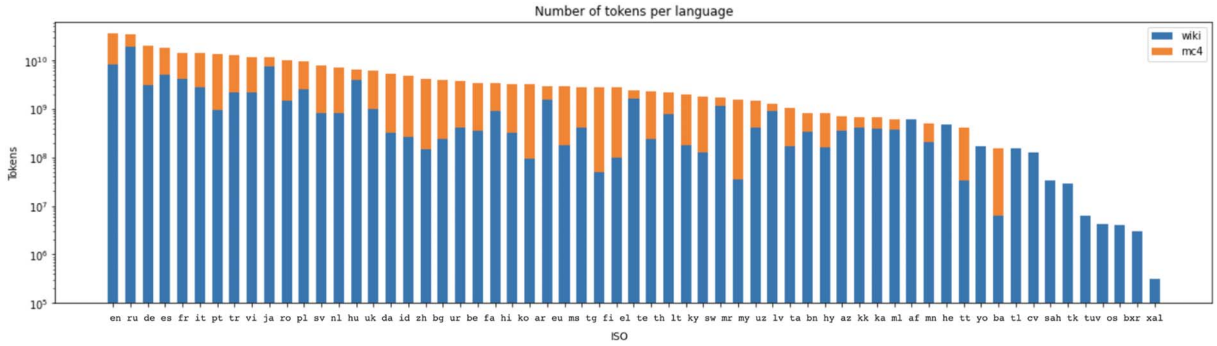

Figure 1: Number of tokens for each language in the pretraining corpus on a logarithmic scale.

corpus consists of a typologically weighted set of languages covered by cross-lingual benchmarks, such as XGLUE (Liang et al., 2020) and XTREME (Hu et al., 2020). The motivation behind the language choices is to narrow the gap between the high-resource and low-resource languages (Ducel et al., 2022). To this end, we include 20 languages from the tail of the C4 language list, the list of underrepresented languages of Russia, and the official and resource-lean CIS languages (Orekhov et al., 2016).

Data Preparation Pipeline Pretraining extensive LMs requires large volumes of high-quality data. Despite the explosive growth of web corpora resulting in the pretraining data volume of up to 6T tokens (Xue et al., 2021), the data quality is often unsatisfactory (Kreutzer et al., 2022). General approaches to maximizing the quality are based on manually curated heuristics (Yang et al., 2019b), the perplexity of LMs (Wenzek et al., 2020), and data quality classifiers (Brown et al., 2020). Our data preparation pipeline includes data collection, deduplication, and filtration.

Data Collection The pretraining corpus represents a collection of documents from Wikipedia and C4. The Wikipedia texts are extracted from the dumps (v. 20201101) with WikiExtractor (Attardi, 2015). The C4 data is downloaded using the Tensorflow datasets[5] (Paper, 2021).

Deduplication The text deduplication includes 64-bit hashing of each text in the pretraining corpus for keeping texts with a unique hash.

Filtration We follow Ortiz Suárez et al. (2019) on the C4 data filtration. We also filter the documents based on their text compression rate using `zlib`.[6] The most strongly and weakly compressing deduplicated texts are discarded. The compression range for an acceptable text is empirically defined as $\times 1.2$ to $\times 8$. The texts with an entropy of less than 1.2 contain code junk and entities, while those of more than 8 contain repetitive segments. The next step includes distinguishing between low and high-quality documents with a binary classifier. The classifier is trained with Vowpal Wabbit[7] on the Wikipedia documents as positive examples and the filtered C4 documents as negative ones. The remainder is cleaned by a set of language-agnostic heuristics. The size of the pretraining corpus is 46B (Wikipedia), and 442B UTF characters (C4), resulting in 600GB. Figure 1 shows the total number of tokens for each language, and the total number of documents in the pretraining corpus is presented in Figure 2.

3.2 Tokenization

The design of the tokenization method may have a significant impact on learning efficient representations, model memorization, and downstream performance (Mielke et al., 2021; Nogueira et al., 2021; Pfeiffer et al., 2021; Rust et al., 2021). We investigate the effect of the tokenization strategy on the model perplexity. We pretrain five strategy-specific versions of $\text{mGPT}_{163\text{M}}$ on a Wikipedia subset of the pretraining corpus. The tokenization strategy is selected based on their perplexity on a held-out Wikipedia sample (approx. 10.7MB), which is inferred as Equation 1.

$$PPL(t) = exp(-\frac{1}{|c|}\sum_{i=0}^{|t|} log_{p_\theta}(x_i|x_{<i})) \quad (1)$$

[5] `tensorflow.org/datasets/catalog/c4`.

[6] `docs.python.org/3/library/zlib`.

[7] `github.com/VowpalWabbit/vowpal_wabbit`.

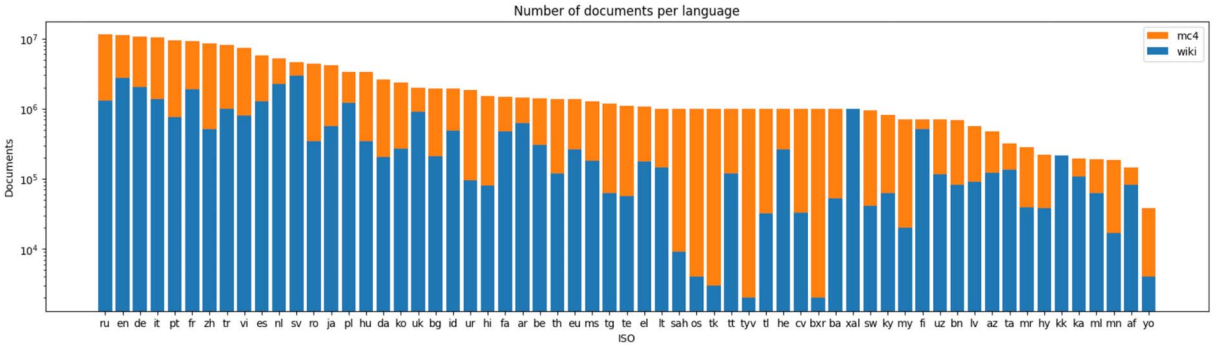

Figure 2: Number of documents for each language in the pretraining corpus on a logarithmic scale.

Strategy	Tokenization Example
DEFAULT	22, Birds, +, 3, birds, =, 25, birds
CASE	22, <case>, birds, +, 3, birds, …
ARITHMETIC	2, 2, <case>, birds, |, +, |, 3, …
COMBINED	2, 2, <case>, birds, |, +, |, 3, |, …
CHAR	2, 2, |, B, i, r, d, s, |, +, |, …

Table 2: Different tokenization strategies applied to the sentence ‘‘22 Birds + 3 birds = 25 birds’’. The resulting tokens are highlighted in the corresponding colors.

where t is an input text, $|t|$ is the length of the text in tokens, $|c|$ is the length of the text in characters. The perplexity is normalized over the number of characters since the tokenizers produce different numbers of tokens for t (Cotterell et al., 2018).

Tokenization Strategies We considered five tokenization strategies incorporating specific representations of uppercase characters, numbers, punctuation marks, and whitespaces. Table 2 presents examples of the tokenization strategies.

- DEFAULT: BBPE (Wang et al., 2020);
- CASE: Each uppercase character is replaced with a special token <case> followed by the corresponding lowercase character;
- ARITHMETIC: The CASE strategy combined with representing numbers and arithmetic operations as individual tokens;
- COMBINED: The ARITHMETIC strategy combined with representing punctuation marks and whitespaces as individual tokens;
- CHAR: Character-level tokenization.

Pretraining Details The models are pretrained on 16 V100 GPUs for 600k training steps with a

Strategy	Avg. PPL
DEFAULT	**6.94**
CASE	8.13
ARITHMETIC	<u>7.99</u>
COMBINED	8.43
CHAR	9.47

Table 3: The average perplexity results. The best score is put in bold, the second best is underlined.

Model	Size	Layers	d_{model}
GPT-2	1.5B	48	1600
GPT-$3_{1.3B}$	1.3B	24	2048
GPT-3_{13B}	13B	40	5120

Table 4: Comparison of GPT-2 and GPT-3. The mGPT architecture replicates the parameters of GPT-$3_{1.3B}$ and GPT-3_{13B}, and uses sparse attention in alternating dense and sparse layers.

set of fixed hyperparameters: vocabulary size of 100k, context window of 2048, learning rate of $2e^{-4}$, and batch size of 4.

Results The experiment results are presented in Table 3. The DEFAULT model achieves the best results, outperforming the rest of the models by up to 2.5 of perplexity score. Based on this experiment, we select the DEFAULT strategy to pretrain the mGPT$_{1.3B}$ and mGPT$_{13B}$ models.

3.3 Model Architecture

The mGPT architecture is based on GPT-3. We use the architecture description by Brown et al., the GPT-2 code base (Radford et al., 2019) from HuggingFace (Wolf et al., 2020), and Megatron-LM (Shoeybi et al., 2020). Table 4 presents the

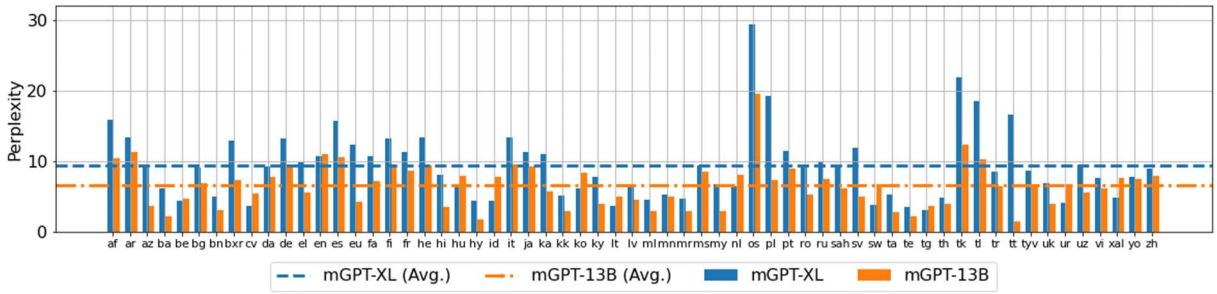

Figure 3: Language-wise perplexity results. Lower is better.

description of the GPT-2 and GPT-3 architectures of comparable sizes. With all the other hyperparameters equal, GPT-3 has fewer layers (*Layers*: 48 vs. 24) but a larger hidden size (d_{model}: 1600 vs. 2048) as opposed to GPT-2. GPT-3 also alternates the classic dense and sparse attention layers (Child et al., 2019).

3.4 Model Pretraining

The pretraining procedure mostly follows Brown et al. We utilize the DeepSpeed library (Rasley et al., 2020) and Megatron-LM (Shoeybi et al., 2020). We pretrain our LMs with a total batch size of 2048 and a context window of 512 tokens. The total number of the training steps is 600k, and the models have seen 400B tokens during pretraining. The pretraining took 14 days on a cluster of 256 V100 GPUs for $\text{mGPT}_{1.3\text{B}}$ and 22 days on 512 V100 GPUs for $\text{mGPT}_{13\text{B}}$. We report the computational, energy, and carbon costs in §7.2.

4 Experiments

4.1 Language Modeling

Method We estimate the language modeling performance on the held-out sets for each language. Here, perplexity is computed as described in §3.2, except that perplexity is normalized over the length of the input text t in tokens $|t|$. We also run statistical tests to analyze the effect of linguistic, dataset, and model configuration criteria:

- *Language script*: We divide the languages into two groups by their scrip—Latin and others (e.g., Cyrillic and Arabic)—and use the Mann-Whitney U test (Mann and Whitney, 1947) to analyze the perplexity distributions in the groups.
- *Pretraining corpus size*: We calculate the Pearson correlation coefficient (Pearson, 1895) to analyze the correlation between the language perplexity and the number of documents in this language in the pretraining corpus.
- *Model size*: We use the Mann-Whitney U test to analyze the effect of the model size.

Results by Language Figure 3 presents the perplexity scores for each language on the held-out sets. The $\text{mGPT}_{13\text{B}}$ model achieves the best perplexities within the 2-to-10 score range for the majority of languages, including Dravidian (Malayalam, Tamil, Telugu), Indo-Aryan (Bengali, Hindi, Marathi), Slavic (Belarusian, Ukrainian, Russian, Bulgarian), Sino-Tibetan (Burmese), Kipchak (Bashkir, Kazakh), and others. Higher perplexities up to 20 are for only seven languages from different families. The $\text{mGPT}_{1.3\text{B}}$ results have similar distribution but are consistently higher than $\text{mGPT}_{13\text{B}}$.

Results by Language Family Analyzing results by the language family (see Figure 4), we find that $\text{mGPT}_{13\text{B}}$ shows consistently lower perplexities as opposed to $\text{mGPT}_{1.3\text{B}}$. Specifically, $\text{mGPT}_{1.3\text{B}}$ underperforms $\text{mGPT}_{13\text{B}}$ on Basque, Greek, Kartvelian, and Turkic families.

Correlation Analysis We present the results in Table 5. We observe that the language modeling performance depends on the language script and model size. In particular, the non-Latin languages receive lower scores on average, while $\text{mGPT}_{13\text{B}}$ performs better than $\text{mGPT}_{1.3\text{B}}$ in this setting. However, the positive correlation between the pretraining corpus size and perplexity in particular languages can be attributed to the low diversity of the text domains in the pretraining monolingual corpora for the low-resource languages. Such corpora contain Wikipedia articles on a limited amount of general topics; therefore, the model

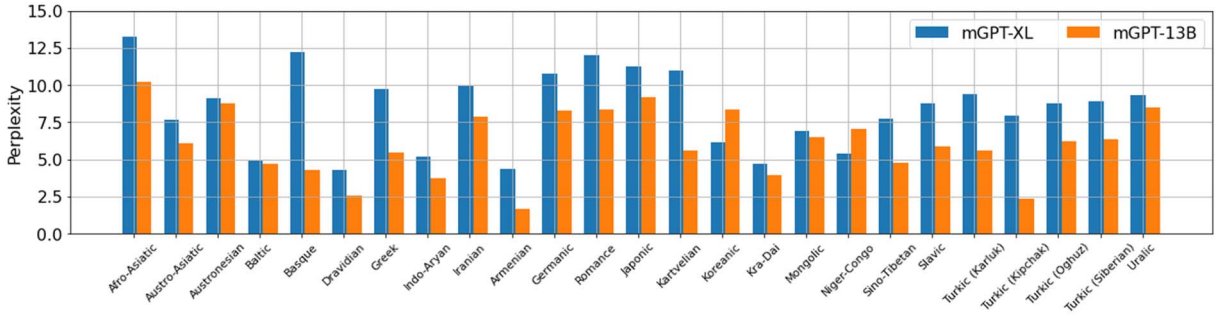

Figure 4: Family-wise perplexity results. The scores are averaged over the number of languages within each family.

Criterion	Model	Test	p-value
Language script	$mGPT_{1.3B}$	M-W U test	0.012
	$mGPT_{13B}$		0.000
Pretraining corpus size	$mGPT_{1.3B}$	Pearson	0.137
	$mGPT_{13B}$		0.307
Model size	$mGPT_{1.3B}$	M-W U test	0.0007
	$mGPT_{13B}$		

Table 5: Correlation analysis results.

learns the distribution in the corpora without being able to generalize well. In general, the results align with Scao et al. (2023), who report that the considered criteria can affect the knowledge acquired by $BLOOM_{1B}$ and $BLOOM_{176B}$.

4.2 Downstream Evaluation

We conduct an extrinsic evaluation of mGPT and baselines on classification and sequence labeling tasks in zero-shot and few-shot settings. In the zero-shot setting, the model is shown a test example formatted as a prompt in natural language, while in the few-shot setting, the model is provided with k demonstrations from the training data specified via prompts. The prompt examples for each task are presented in Table 6.

4.2.1 Classification

Tasks The classification tasks include commonsense reasoning (XCOPA; Ponti et al., 2020), natural language inference (XNLI; Conneau et al. 2018), Winograd schema challenge (XWINO; Tikhonov and Ryabinin, 2021), paraphrase detection (PAWSX; Yang et al., 2019a), and hate speech detection (Davidson et al., 2017).

Method mGPT utilizes per-token cross-entropy loss, which is reduced to negative log probability due to one-hot encoding of the tokens. We select the target label associated with the prompt that results in the lowest sum of negative log probabilities for its tokens. The few-shot experiments are run five times with different random seeds, while the zero-shot experiments are run only once since the model loss is determined.

Baselines The $XGLM_{1.7B}$ and $XGLM_{7.5B}$ models are used as the baselines in the classification experiments. We reproduce the XGLM evaluation based on the methodology by Lin et al. (2022) and use the model weights and code available in the `fairseq`[8] library (Ott et al., 2019). We select prompts according to the templates reported by Lin et al. Prompts for non-English languages are automatically translated with Google Translate.

Results Table 7 presents the classification results averaged across languages. The ''✗'' tag marks k-shot settings not reported by Lin et al. We do not perform them for reproducibility purposes and fair comparison. The results by Lin et al. are reproduced in the zero-shot setup, and some scores are even slightly higher. However, not all results are reproduced, e.g., PAWSX and XNLI. We attribute this to potential differences in the translated prompts.

Overall, we observe that $mGPT_{1.3B}$ is comparable with $XGLM_{1.7B}$ while having fewer weights and is pretrained in twice as many languages. $mGPT_{13B}$ performs better than $XGLM_{7.5B}$ in zero-shot setting on all tasks except XNLI. At the same time, it lags behind in a few-shot setting being better than $XGLM_{7.5B}$ only in XNLI and PAWSX tasks. Comparing the performance across languages, we find that English receives the highest accuracy for all tasks. The

[8] `github.com/pytorch/fairseq/xglm`.

Task	Template	Output Candidates
XNLI	<s> {sentence 1}, right? {label} {sentence 2} </s>	Yes (Entailment); Also (Neutral) No (Contradiction)
PAWSX	<s> {sentence 1}, right? {label} {sentence 2} </s>	Yes; No
XWINO	<s> {sentence start} {candidate} {sentence end} </s>	✗
XCOPA	<s> {sentence} because {candidate answer} </s> <s> {sentence} so {candidate answer} </s>	✗
Hate Speech	<s> The sentence is {label}. {sentence} </s>	sexist, racist, offensive, abusive, hateful (Positive) normal, common, ok, usual, acceptable (Negative)
NER	<s>lang: {lang} \n Tagged sentence: {sentence with tags}	I-LOC, I-MISC, I-ORG, I-PER, O
POS	<s>lang: {lang} \n Tagged sentence: {sentence with tags}	ADJ, ADP, ADV, AUX, CCONJ, DET, INTJ, NOUN, NUM, PART, PRON, PROPN, PUNCT, SCONJ, SYM, VERB, X

Table 6: Prompt examples for each downstream task. The examples are in English for illustration purposes.

Model	k-shot	XWINO	PAWSX	XCOPA	XNLI	Hate Speech
$mGPT_{1.3B}$	0	56.2	53.1	55.5	40.6	50.0
	1	57.0	51.3	54.9	36.1	✗
	4	56.8	52.2	54.8	37.4	50.8
	16	54.5	52.2	54.8	37.9	✗
$mGPT_{13B}$	0	59.3	51.5	58.2	42.6	**53.1**
	1	61.0	50.6	57.9	37.5	✗
	4	61.8	51.6	58.3	41.4	51.5
	16	59.2	**55.1**	57.3	33.3	✗
$XGLM_{1.7B}$	0	54.2	50.3	55.5	42.6	50.1
	1	58.0	45.9	56.8	36.4	✗
	4	57.9	45.9	56.2	38.8	49.5
	16	✗	44.2	56.1	36.5	✗
$XGLM_{7.5B}$	0	59.2	50.1	55.5	**44.7**	50.1
	1	63.7	46.4	60.6	36.9	✗
	4	**64.2**	45.3	61.4	40.1	51.8
	16	✗	44.9	**62.5**	40.0	✗

Table 7: Accuracy scores (%) on classification tasks averaged across languages.

$mGPT_{1.3B}$ and $mGPT_{13B}$ models show high accuracy for the Austronesian, Dravidian, Japonic, Germanic, and Romance language families. Only the Afro-Asiatic family gets low accuracy. The mGPT models perform better than the XGLM counterparts for Austronesian, Koreanic, and Romance languages.

Our results on hate speech detection are consistent with Lin et al. The performance is slightly better across the five languages but still close to random guessing (see Table 8). The manual analysis shows that the behavior is sensitive to the input prompts, most notably for Polish. Increasing the number of demonstrations can lead to performance degradation on some classification tasks for both mGPT and XGLM.

Model	k-shot	en	es	pt	pl	it
$mGPT_{1.3B}$	0	55.1	52.1	42.3	50.0	50.2
	4	50.1	50.2	51.7	51.5	50.4
$mGPT_{13B}$	0	59.0	**55.2**	46.9	50.0	**54.6**
	4	52.2	50.0	50.8	**53.4**	51.0
$XGLM_{1.7B}$	0	54.8	51.8	52.3	50.0	54.5
	4	51.0	48.8	49.2	46.7	51.0
$XGLM_{7.5B}$	0	**61.7**	52.4	**52.3**	50.0	49.0
	4	51.8	51.3	51.5	51.4	52.9

Table 8: Accuracy scores (%) on hate speech detection by language. The best score is put in bold, the second best is underlined.

4.2.2 Sequence Labeling

Tasks The sequence labeling tasks include named entity recognition (NER) and part-of-speech tagging (POS) from the XGLUE benchmark (Liang et al., 2020). To address other medium-resource and resource-lean languages, we use the Universal Dependencies treebanks (UD; Nivre et al., 2016) to evaluate POS-tagging in Armenian, Belarusian, Buryat, Kazakh, Tatar, Ukrainian, and Yakut.

Method We use a modified approach to the sequence labeling tasks compared to §4.2.1. Given a sentence of n words, we iteratively predict the label for each word x_i using the preceding words $x_{<i}$ and their predicted labels $l_{<i}$ as the context using a template "$x_{<i}l_{<i}\_$", where i is the current token index and "_" is a placeholder. The only exception is the first token x_i used as the

Model	de	en	es	nl	Avg.
Random	1.9	3.1	1.8	1.6	2.1
$mGPT_{1.3B}$	12.2	22.1	12.7	13.1	15.0
$mGPT_{13B}$	5.6	20.9	10.4	6.7	10.9
$M\text{-}BERT_{base}$	69.2	90.6	75.4	77.9	78.2
$XLM\text{-}R_{base}$	70.4	90.9	**75.2**	79.5	79.0
Unicoder	**71.8**	**91.1**	74.4	**81.6**	**79.7**

Table 9: F1-scores for NER by language. The mGPT models are evaluated in the 4-shot setting. The best score is put in bold, the second best is underlined.

context. The placeholder is filled with each possible target label $l \in L$ at each step. We select the label with the lowest sum of losses per token in the resulting string. The experiments are run in the zero-shot and 4-shot settings.[9]

Example Consider an example for the POS-tagging task "I [PRON] WANT [VERB] IT [PART] . [PUNCT]", which requires 4 procedure steps. First, we combine the placeholder in the string "I_" with each possible POS tag and select the most probable candidate. Next, we repeat the procedure for "I_l_i WANT_", and so on.

Baselines We use results reported in Liang et al. as the baselines: M-BERT, XLM-R, and Unicoder (Huang et al., 2019). Note that the baselines are *finetuned* on the corresponding training set. The performance is evaluated with the F1-score (NER) and the accuracy score (POS-tagging)[10] according to the XGLUE methodology.

NER Results Table 9 shows counterintuitively that $mGPT_{1.3B}$ outperforms $mGPT_{13B}$ on all languages. 4-shot falls behind finetuned models but significantly outperforms random guessing for both mGPT models. Per-language language analysis shows a large gap between English and other languages (for $mGPT_{13B}$ the F1-score on English is more than twice higher than for any of the other languages), while for German, both models perform the worst. This pattern coincides with the baseline results. In addition, it could be noted that while for $mGPT_{1.3B}$ the F1-score exceeds the 10 percent threshold for all languages, this is not the case for $mGPT_{13B}$.

POS-tagging Results POS-tagging results for the XGLUE benchmark and resource-lean languages are presented in Table 10. Similarly to the NER task, $mGPT_{1.3B}$ outperforms $mGPT_{13B}$ practically in all languages except for Italian. On average $mGPT_{1.3B}$ achieves accuracy score of 0.24 while $mGPT_{13B}$ only scores 0.21. These results are still far behind fine-tuned models; however, they are significantly higher than random guessing. Analyzing the results for the low-resource languages, it can be seen that $mGPT_{1.3B}$ performance is comparable with its performance on XGLUE, while the $mGPT_{13B}$ scores are lower.

4.3 Knowledge Probing

Method We probe our models for factual knowledge in 23 languages using the mLAMA dataset (Kassner et al., 2021). The task is to complete a knowledge triplet <*subject, relation, object*> converted to templates for querying LMs. Consider an example from the original LAMA (Petroni et al., 2019) for English, where <*Dante, born-in, X*> is converted to the template *''Dante was born in [MASK]''*. We follow Lin et al. to design the probing task. As each such query contains hundreds of negative candidates on average, we limit the number of candidates to three, i.e., one is the ground truth candidate and the other two candidates are randomly sampled from the provided knowledge source. The probing performance is evaluated with precision@1 averaged over all relations per language.

Results Figure 5 outlines the results for $mGPT_{1.3B}$ and $mGPT_{13B}$. The overall pattern is that the performance is equal to or above 0.6 for Germanic, Romance, Austro-Asiatic, Japonic, and Chinese languages. However, Uralic, Slavic, Koreanic, and Afro-Asiatic languages receive scores of lower than 0.5. We also find that scaling the number of model parameters usually boosts the performance for high-resource languages up to 5 points, while no significant improvements are observed in the other languages. Comparing our results with Lin et al., we conclude that our models achieve lower performance than $XGLM_{7.5B}$ almost in all languages and perform on par with $GPT3\text{-}Curie_{6.5B}$.

[9] We report the results only in the 4-shot setting since the manual analysis reveals that the models have failed to capture the task, giving constant predictions without any additional examples.

[10] We evaluate the sequence labeling tasks using the XGLUE code: `github.com/microsoft/XGLUE`.

Model	XGLUE																			CIS & Low-Resource UD						
	ar	bg	de	el	en	es	fr	hi	it	nl	pl	pt	ru	th	tr	ur	vi	zh	Avg.	be	bxr	hy	kk	sah	tt	uk
Random	6.5	6.5	6.0	5.2	4.4	5.7	5.5	6.7	6.6	6.6	5.9	4.7	6.0	6.4	6.8	1.2	7.0	7.1	5.8	1.3	5.7	5.9	2.6	9.6	8.7	4.8
$mGPT_{1.3B}$	16.5	24.5	30.6	20.9	40.0	24.3	27.0	16.2	25.4	28.8	28.3	24.6	29.4	12.9	30.4	15.0	25.6	19.5	24.4	**21.5**	**28.4**	**14.7**	**22.8**	**19.9**	**21.4**	**22.5**
$mGPT_{13B}$	11.7	21.8	26.8	16.1	36.0	22.2	25.0	12.3	26.5	26.5	24.2	21.8	21.8	9.5	26.8	12.7	21.5	12.5	20.9	10.6	7.7	7.3	9.4	11.8	9.2	10.9
$M\text{-}BERT_{base}$	52.4	85.0	88.7	81.5	95.6	86.8	87.6	58.4	91.3	88.0	81.8	88.3	78.8	43.3	69.2	53.8	54.3	58.3	74.7	✗	✗	✗	✗	✗	✗	✗
$XLM\text{-}R_{base}$	67.3	**88.8**	**92.2**	**88.2**	**96.2**	**89.0**	**89.9**	**74.5**	**92.6**	**88.5**	**85.4**	**89.7**	**86.9**	**57.9**	72.7	**62.1**	55.2	**60.4**	**79.8**	✗	✗	✗	✗	✗	✗	✗
Unicoder	**68.6**	88.5	92.0	88.3	96.1	89.1	89.4	69.9	92.5	88.9	83.6	89.8	86.7	57.6	**75.0**	59.8	**56.3**	60.2	**79.6**	✗	✗	✗	✗	✗	✗	✗

Table 10: Accuracy scores (%) for XGLUE and Universal Dependencies POS-tagging by language. mGPT models are evaluated in the 4-shot setting. The best score is put in bold, the second best is underlined.

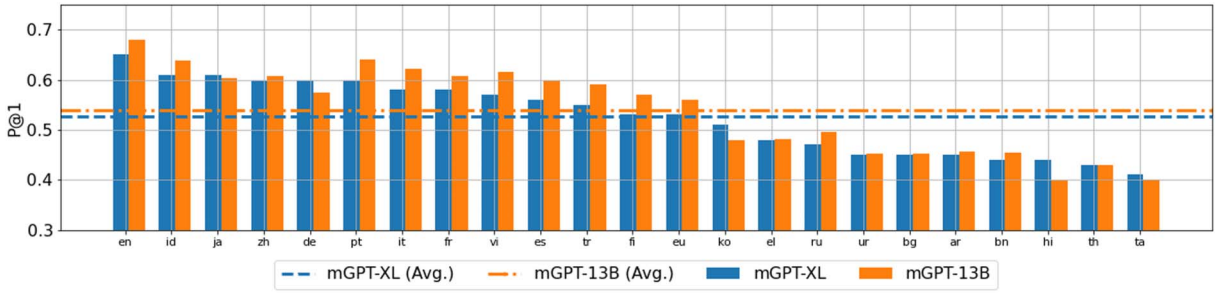

Figure 5: Knowledge probing results for 23 languages. The performance of a random baseline is 0.33.

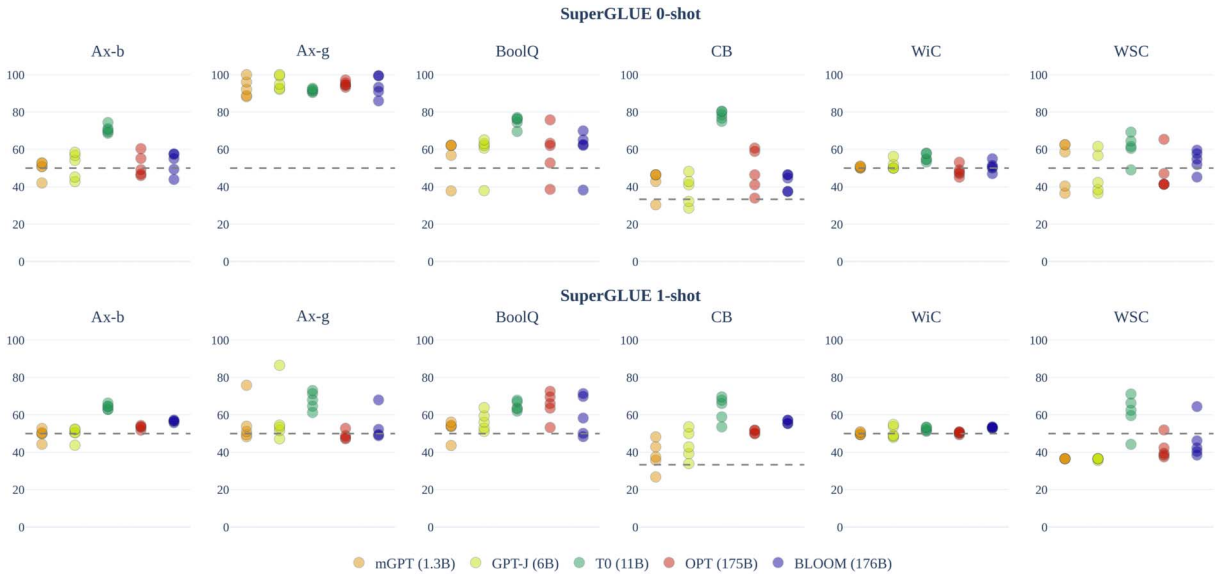

Figure 6: The SuperGLUE evaluation results in the zero-shot and one-shot settings (Scao et al., 2023).

4.4 External Evaluation

General Language Understanding Scao et al. (2023) compared the performance of $BLOOM_{176B}$, $mGPT_{1.3B}$, OPT_{175B} (Zhang et al., 2022), $GPT\text{-}J_{6B}$ (Wang and Komatsuzaki, 2021), and $T0_{11B}$ (Victor et al., 2022) on subset of tasks from the SuperGLUE benchmark (Wang et al., 2019) in the zero-shot and one-shot settings. The results of evaluating the models using five prompts are presented in Figure 6. The $mGPT_{1.3B}$ model has comparable performance despite having fewer weights. In the zero-shot setting, the performance of $mGPT_{1.3B}$, $BLOOM_{176B}$, OPT_{175B}, and $GPT\text{-}J_{6B}$ on the considered tasks is above random guessing. We also observe the strong performance of $mGPT_{1.3B}$ on the Winogender Schema Diagnostics (Ax-g). In the one-shot setting, $mGPT_{1.3B}$ performs on par with $GPT\text{-}J_{6B}$, and the resulting variability is significantly reduced across all prompts.

ISO	Avg. length	Distinct_1	Vocabulary size	Unique_1	Entropy_1	TTR	MSTTR
en	39.13 ± 22.61	0.071	387	103	6.175	0.097	0.228
fr	23.53 ± 17.92	0.128	486	181	6.875	0.159	0.346
de	30.85 ± 17.33	0.113	453	159	6.850	0.151	0.340
es	12.71 ± 15.54	0.102	413	124	6.818	0.148	0.315
zh	3.157 ± 2.39	0.492	188	124	7.055	0.525	0.526

Table 11: The results for lexical diversity of generated texts on the GEM story generation task.

Multilingual Clause-level Morphology The first shared task on Multilingual Clause-level Morphology (Goldman et al., 2022) covers nine languages and includes three sub-tasks: (i) inflection (generating a word form given a lexeme and a set of morphosyntactic features), (ii) reinflection (reinflect an input sentence according to a given set of morphosyntactic features), and (iii) detect a root and its features in an input sentence. Acikgoz et al. (2022) develop a first-place solution based on $\text{mGPT}_{1.3B}$ and prefix-tuning method, outperforming other solutions and baselines on the third task.

4.5 Generation Evaluation

Method We compute seven lexical diversity metrics from Gehrmann et al. (2021) using the mGPT outputs[11] on 100 test set samples from the story generation task in five languages: English, French, German, Spanish, and Chinese (Chen et al., 2022). The diversity metrics include the Shannon Entropy over unigrams (Entropy_1), the mean segmented type-token ratio over segment lengths of 100 (MSTTR), the ratio of distinct unigrams over the total number of unigrams (Distinct_1), and the counter of unigrams that appear once in the collection of generated outputs (Unique_1).

Results The results are presented in Table 11. The diversity metrics scores for Chinese are the highest, while the mean generated text length is the shortest. This is likely due to its logographic writing. The results for the Indo-European languages are similar (French, German, and Spanish), indicating that $\text{mGPT}_{1.3B}$ generates diverse texts in these languages. Surprisingly, the metrics are lower for English, with the average text length being longer. Our current natural language generation evaluation approach lacks downstream tasks, which we leave for future work.

[11] We use the generation hyperparameters: $temperature = 1$, $max\_length = 100$, $top\_k = 5$, $top\_p = 0.9$.

5 Discussion

Our key takeaways on pretraining and evaluating large-scale multilingual autoregressive LMs are summarized below.

5.1 Model Scaling

Empirical Results The language modeling results for $\text{mGPT}_{1.3B}$ and mGPT_{13B} suggest that the model scaling improves its generation abilities for all given languages (see §4.1). However, it does not improve performance on the downstream and probing tasks (see §4.2; §4.3). Overall, the language modeling performance depends on the model size and the pretraining corpus size in a language, and smaller models may better encode linguistic information than larger ones. These findings align with Scao et al. (2023).

Takeaways Our work had been conducted a year before the Chinchilla scaling laws were introduced (Hoffmann et al., 2022). According to the advanced methods of scaling LMs, our pretraining corpus can be sufficiently extended to improve the generalization abilities of the mGPT_{13B} model. At the same time, the pretraining corpus design can promote the model underfitting and overfitting on particular languages. We believe it can be accounted for by aggregating the language-specific cross-entropy loss and producing language weights similar to Xie et al. (2023).

5.2 Lack of Data

Empirical Results Another challenging factor is the lack of high-quality data for the low-resource languages. Although mGPT shows promising results on the language modeling and sequence labeling tasks for the underrepresented languages (see §4.1, §4.2), the low amount of evaluation

Language	HuggingFace URL	PPL
Armenian	hf.co/ai-forever/mGPT-1.3B-armenian	1.7
Azerbaijan	hf.co/ai-forever/mGPT-1.3B-azerbaijan	5.4
Bashkir	hf.co/ai-forever/mGPT-1.3B-bashkir	7.1
Belorussian	hf.co/ai-forever/mGPT-1.3B-belorussian	27.7
Bulgarian	hf.co/ai-forever/mGPT-1.3B-belorussian	15.2
Buryat	hf.co/ai-forever/mGPT-1.3B-buryat	17.6
Chuvash	hf.co/ai-forever/mGPT-1.3B-chuvash	28.8
Georgian	hf.co/ai-forever/mGPT-1.3B-georgian	16.9
Kalmyk	hf.co/ai-forever/mGPT-1.3B-kalmyk	14.0
Kazakh	hf.co/ai-forever/mGPT-1.3B-kazakh	3.4
Kirgiz	hf.co/ai-forever/mGPT-1.3B-kirgiz	8.2
Mari	hf.co/ai-forever/mGPT-1.3B-mari	21.2
Mongol	hf.co/ai-forever/mGPT-1.3B-mongol	4.4
Ossetian	hf.co/ai-forever/mGPT-1.3B-ossetian	18.7
Persian	hf.co/ai-forever/mGPT-1.3B-persian	33.4
Romanian	hf.co/ai-forever/mGPT-1.3B-romanian	3.4
Tajik	hf.co/ai-forever/mGPT-1.3B-tajik	6.5
Tatar	hf.co/ai-forever/mGPT-1.3B-tatar	3.7
Turkmen	hf.co/ai-forever/mGPT-1.3B-turkmen	28.5
Tuvan	hf.co/ai-forever/mGPT-1.3B-tuvan	40.8
Ukranian	hf.co/ai-forever/mGPT-1.3B-ukranian	7.1
Uzbek	hf.co/ai-forever/mGPT-1.3B-uzbek	6.8
Yakut	hf.co/ai-forever/mGPT-1.3B-yakut	10.6

Table 12: A list of the $\text{mGPT}_{1.3B}$ models continuously pretrained on monolingual corpora for 23 languages.

resources limits the scope of analyzing the model generalization abilities. The correlation between the model performance and the amount of pretraining data in a language (see §4.1, and, e.g., Lauscher et al., 2020; Ahuja et al., 2022) further highlights the need for creating text corpora in such languages.

Takeaways The question of addressing the discrepancy in data distribution across the world's languages remains unresolved. Our data collection and filtration approach is equivalent for all considered languages. Extending the language-agnostic heuristics is restrained due to the lack of linguistic expertise. However, we assume that experimenting with the training data for the text quality classifiers can improve the resulting quality of the corpora for the low-resource languages (e.g., training the classifiers on different mixtures of data in the medium and high-resource languages).

As the follow-up work, we release 23 versions of the $\text{mGPT}_{1.3B}$ model continuously pretrained with language modeling objective on monolingual corpora for medium-resource and low-resource languages collected through collaboration with the NLP community. Table 12 summarizes the models by language and the language modeling performance on the held-out monolingual test sets. Examples of the corpora include Eastern Armenian National Corpus (Khurshudyan et al., 2022), OpenSubtitles (Lison and Tiedemann, 2016), and TED talks. Continued pretraining on additional data improves the language modeling performance.

5.3 Language Selection

Empirical Results Results of $\text{mGPT}_{1.3B}$ on most of the classification tasks are on par or better than the results of the $\text{XGLM}_{1.7B}$ given that mGPT covers twice as many languages (see §4.2). However, mGPT underperforms the baselines on several multi-class classification and probing tasks.

Takeaways We find that balancing the pretraining corpus by the language family helps improve the language modeling abilities for underrepresented languages due to their typological similarity with the medium and high-resource languages (see §4.1). However, increasing language diversity can lead to performance degradation because of the curse of multilinguality and a limited model capacity (Conneau et al., 2020).

5.4 Tokenization

Empirical Results We conduct an ablation study to analyze the impact of the tokenization strategy on language modeling performance. We find that the considered strategies do not improve the model's perplexity. However, the main drawback of the perplexity-based evaluation is that it only partially assesses the model generalization abilities.

Takeaways The optimal tokenization method and vocabulary size remain an open question, particularly in the multilingual setup (Mielke et al., 2021). There are no established methods for defining the vocabulary size based on the amount of textual data in different languages. Our experiments are limited to a fixed vocabulary size, and we leave further investigation of the tokenization strategies and their configurations for future work.

5.5 Zero-shot and Few-shot Performance

Empirical Results

- Increasing the number of demonstrations does not always lead to improvements but decreases the performance on some downstream tasks (see §4.2.1; §4.2.2). This observation aligns with Lin et al. (2022) and Brown et al. (2020).
- The zero-shot and few-shot performance may not exceed the random guessing on particular tasks, which points to the failure of a model

to follow the guidance in the demonstration examples (see §4.2.1; §4.2.2).

- The prompting approach is unstable and hardly universal across languages, as indicated by the model sensitivity to the prompts.
- The mGPT models can assign higher probabilities to the most frequent tag in the input for the sequence labeling tasks (see §4.2.2).

Takeaways

- The stability of the models with respect to the prompts may be improved using prompt-tuning (Liu et al., 2023b) and contextual calibration (Zhao et al., 2021) as shown in §4.4.
- The generalization capabilities of the autoregressive LMs in sequence labeling tasks is an underexplored area. While our LMs achieve results higher than random guessing, the low performance can be attributed to the probability distribution shifts between the pretraining corpora and the prompts. We leave the investigation of the alternative prompt design (Liu et al., 2023a) and structured prediction methods (Liu et al., 2022) for future work.

6 Conclusion

We introduce the $\text{mGPT}_{1.3B}$ and mGPT_{13B} models, which cover 61 languages from linguistically diverse 25 language families. Our model is one of the first autoregressive LMs for economically endangered and underrepresented CIS and low-resource languages. The architecture design choices are based on the preliminary tokenization experiments and their perplexity-based evaluation. The model evaluation experiments include language modeling, standardized cross-lingual NLU datasets and benchmarks, world knowledge probing, and social bias tasks. We evaluate the in-context learning abilities in zero and few-shot settings with a negative log-likelihood probability. We present a detailed analysis of the model performance, limitations, and ethical considerations. Despite the space for further quality growth and solving the highlighted limitations, the model shows significant potential and can become the basis for developing generative pipelines for languages other than English, especially the low-resource ones. This initiative has been developed for 23 diverse languages through collaboration with the NLP community. We hope to benefit cross-lingual knowledge transfer, annotation projection, and other potential applications for economically challenged and underrepresented languages and diversify the research field by shifting from the Anglo-centric paradigm.

7 Ethical Statement and Social Impacts

7.1 Low-resource Languages

NLP for resource-lean scenarios is one of the leading research directions nowadays. The topic's relevance has led to proactive research on low-resource languages. Our work falls under this scope, introducing the first autoregressive LM for 61 languages. To the best of our knowledge, we present one of the first attempts to address this problem for 20 languages of the Commonwealth of Independent States and the indigenous peoples in Russia.

7.2 Energy Efficiency and Usage

Pretraining large-scale LMs requires many computational resources, which is energy-intensive and expensive. To address this issue, we used the sparse attention approach suggested by Brown et al. (2020) and reduced the computational resources required to achieve the desired performance. The CO2 emission of pretraining the mGPT models is computed as Equation 2 (Strubell et al., 2019):

$$CO2 = \frac{PUE * kWh * I^{CO2}}{1000} \qquad (2)$$

The power usage effectiveness (PUE) of our data centers is not more than 1.3, the spent power is 30.6k kWh ($\text{mGPT}_{1.3B}$) and 91.3 kWh (mGPT_{13B}), and the CO2 energy intensity (I^{CO2}) in the region is 400 grams per kWh. The resulting CO2 emission is 15.9k kg ($\text{mGPT}_{1.3B}$) and 47.5k kg (mGPT_{13B}). The emission is comparable with a single medium-range flight of a modern aircraft, which usually releases about 12k kg of CO2 per 1k km. Despite the costs, mGPT can be efficiently adapted to the user needs via few-shot learning, bringing down potential budget costs in the scope of applications in multiple languages, such as generating the content, augmenting labeled data, or summarizing news. The multilingual pretraining saves on data annotation and energy consumption,

alleviating the carbon footprint. Model compression techniques, e.g., pruning and distillation, can reduce inference costs.

7.3 Social Risks of Harm

Stereotypes and unjust discrimination present in pretraining corpora lead to representation biases in LMs. LMs can reflect historical prejudices against disadvantaged social groups and reproduce harmful stereotypes about gender, race, religion, or sexual orientation (Weidinger et al., 2022). We have analyzed mGPT's limitations on social risks of harm involving hate speech on the hate speech detection task. Our results are similar to Lin et al. (2022) in that the performance is close to random guessing. This may indicate a significant bias in the pretraining corpus, a mutual influence of languages during training, or methodological problems in the test set. We do not claim that our evaluation setup is exhaustive, and we assume that other biases can be revealed through a direct model application or an extended evaluation.

7.4 Potential Misuse

The misuse potential of LMs increases with their ability to generate high-quality texts. Malicious users can perform a socially harmful activity that involves generating texts, e.g., spreading propaganda and other targeted manipulation (Jawahar et al., 2020). We recognize that our models can be misused in all supported languages. However, adversarial defense and artificial text detection models can mitigate ethical and social risks of harm. Our primary purpose is to propose multilingual GPT-style LMs for **research and development** needs, and we hope to work on the misuse problem with other developers and experts in mitigation research in the future.

References

Emre Can Acikgoz, Tilek Chubakov, Muge Kural, Gözde Şahin, and Deniz Yuret. 2022. Transformers on multilingual clause-level morphology. In *Proceedings of the The 2nd Workshop on Multi-lingual Representation Learning (MRL)*, pages 100–105, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.mrl-1.10

Kabir Ahuja, Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. 2022. Multi task learning for zero shot performance prediction of multilingual models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5454–5467, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.374

Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3712

Giuseppe Attardi. 2015. WikiExtractor. https://github.com/attardi/wikiextractor

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. https://doi.org/10.1145/3442188.3445922

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136,

virtual+Dublin. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.bigscience-1.9

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yiran Chen, Zhenqiao Song, Xianze Wu, Danqing Wang, Jingjing Xu, Jiaze Chen, Hao Zhou, and Lei Li. 2022. MTG: A benchmark suite for multilingual text generation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2508–2527, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-naacl.192

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*.

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. https://doi.org/10.18653/v1/2020.acl-main.747

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1269

Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2085

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 512–515. https://doi.org/10.1609/icwsm.v11i1.14955

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.

Nolan Dey, Gurpreet Gosal, Zhiming, Chen, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, and Joel Hestness. 2023. Cerebras-GPT: Open compute-optimal language models trained on the cerebras wafer-scale cluster. https://doi.org/10.48550/arXiv.2304.03208

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. https://doi.org/10.48550/arXiv.2002.06305

Fanny Ducel, Karën Fort, Gaël Lejeune, and Yves Lepage. 2022. Do we name the languages we study? The #BenderRule in LREC and ACL articles. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 564–573, Marseille, France. European Language Resources Association.

Erkut Erdem, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara Plank, Andrii Babii, Oleksii Turuta, Aykut Erdem, Iacer Calixto, et al. 2022. Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning. *Journal of Artificial Intelligence Research*, 73:1131–1207. https://doi.org/10.1613/jair.1.12918

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.gem-1.10

Omer Goldman, Francesco Tinner, Hila Gonen, Benjamin Muller, Victoria Basmov, Shadrack Kirimi, Lydia Nishimwe, Benoît Sagot, Djamé Seddah, Reut Tsarfaty, and Duygu Ataman. 2022. The MRL 2022 shared task on multilingual clause-level morphology. In *Proceedings of the The 2nd Workshop on Multi-lingual Representation Learning (MRL)*, pages 134–146, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.mrl-1.14

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.201

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. https://doi.org/10.48550/arXiv.2203.15556

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

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. 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 2485–2494, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1252

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V. S. 2020. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.208

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N. C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.445

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. AMMUS: A survey of transformer-based pretrained models in natural language processing. https://doi.org/10.48550/arXiv.2108.05542

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.284

Victoria Khurshudyan, Timofey Arkhangelskiy, Misha Daniel, Vladimir Plungian, Dmitri Levonian, Alex Polyakov, and Sergei Rubakov. 2022. Eastern Armenian national corpus: State of the art and perspectives. In *Proceedings of the Workshop on Processing Language Variation: Digital Armenian (DigitAm) within the 13th Language Resources and Evaluation Conference*, pages 28–37, Marseille, France. European Language Resources Association.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72. https://doi.org/10.1162/tacl_a_00447

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.363

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.484

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. https://doi.org/10.48550/arXiv.2112.10668

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pre-Train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35. https://doi.org/10.1145/3560815

Qi Liu, Matt J. Kusner, and Phil Blunsom. 2020a. A Survey on Contextual Embeddings. https://doi.org/10.48550/arXiv.2003.07278

Tianyu Liu, Yuchen Eleanor Jiang, Nicholas Monath, Ryan Cotterell, and Mrinmaya Sachan. 2022. Autoregressive structured prediction with language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 993–1005, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.70

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023b. GPT understands, too. *AI Open*. https://doi.org/10.1016/j.aiopen.2023.08.012

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020b. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742. https://doi.org/10.1162/tacl_a_00343

H. Mann and D. Whitney. 1947. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Annals of Mathematical Statistics*, 18(1):50–60. https://doi.org/10.1214/aoms/1177730491

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: A tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.645

Mihai Masala, Stefan Ruseti, and Mihai Dascalu. 2020. RoBERT – a Romanian BERT model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6626–6637, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.581

R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2020. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.blackboxnlp-1.21

Sabrina J. Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, and Samson Tan. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. https://doi.org/10.48550/arXiv.2112.10508

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the*

57th Annual Meeting of the Association for Computational Linguistics, pages 4658–4664, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1459

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2021. Investigating the limitations of transformers with simple arithmetic tasks. https://doi.org/10.48550/arXiv.2102.13019

Boris Orekhov, I. Krylova, I. Popov, E. Stepanova, and L. Zaydelman. 2016. Russian minority languages on the web: Descriptive statistics. In *Vladimir Selegey (chief ed.), Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference ''Dialogue''*, pages 498–508.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, pages 9–16, Mannheim. Leibniz-Institut für Deutsche Sprache. https://doi.org/10.14618/ids-pub-9021

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-4009

David Paper. 2021. TensorFlow datasets. *State-of-the-Art Deep Learning Models in TensorFlow: Modern Machine Learning in the Google Colab Ecosystem*, pages 65–91. https://doi.org/10.1007/978-1-4842-7341-8_3

Karl Pearson. 1895. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352):240–242. https://doi.org/10.1098/rspl.1895.0041

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *Advances in Neural Information Processing Systems*, 34.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1250

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. 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. https://doi.org/10.18653/v1/2021.emnlp-main.800

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.185

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu.

2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506. https://doi.org/10.1145/3394486.3406703

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? On the monolingual performance of multilingual language models. 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 3118–3135, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.243

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Alshaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles

Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A. Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S. Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. BLOOM: A 176B-parameter open-access multilingual language model. https://doi.org/10.48550/arXiv.2211.05100

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.185

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2020. Megatron-LM: Training multi-billion parameter language models using model parallelism. https://doi.org/10.48550/arXiv.1909.08053

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. https://doi.org/10.18653/v1/P19-1355

Alexey Tikhonov and Max Ryabinin. 2021. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning. In *Findings of the*

Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3534–3546, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.310

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Sanh Victor, Webson Albert, Raffel Colin, Bach Stephen, Sutawika Lintang, Alyafeai Zaid, Chaffin Antoine, Stiegler Arnaud, Raja Arun, Dey Manan, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, 32.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 billion parameter autoregressive language model.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020. Neural machine translation with byte-level subwords. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9154–9160. https://doi.org/10.1609/aaai.v34i05.6451

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.354

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229. https://doi.org/10.1145/3531146.3533088

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 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.mrl-1.1

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

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. https://doi.org/10.18653/v1v.repl4nlp-1.16

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. 2023. DoReMi: Optimizing data mixtures speeds up language model pretraining.

https://doi.org/10.48550/arXiv.2305.10429

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.41

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019a. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1382

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019b. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32.

Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, Chen Li, Ziyan Gong, Yifan Yao, Xinjing Huang, Jun Wang, Jianfeng Yu, Qi Guo, Yue Yu, Yan Zhang, Jin Wang, Hengtao Tao, Dasen Yan, Zexuan Yi, Fang Peng, Fangqing Jiang, Han Zhang, Lingfeng Deng, Yehong Zhang, Zhe Lin, Chao Zhang, Shaojie Zhang, Mingyue Guo, Shanzhi Gu, Gaojun Fan, Yaowei Wang, Xuefeng Jin, Qun Liu, and Yonghong Tian. 2021. PanGu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. https://doi.org/10.48550/arXiv.2104.12369

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open pre-trained transformer language models. https://doi.org/10.48550/arXiv.2205.01068

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergey Markov, Vladislav Mikhailov, and Alena Fenogenova. 2023. A family of pretrained transformer language models for Russian. https://doi.org/10.48550/arXiv.2309.10931