KyrText: A Multi-Domain Large-Scale Corpus for Kyrgyz Language

Tilek Chubakov
Independent Researcher
tchubakov@berkeley.edu

Abstract

Kyrgyz is a morphologically rich Turkic language that remains significantly underrepresented in modern multilingual language models. To address this resource gap, we introduce KyrText, a diverse, large-scale corpus containing 680.5 million words. Unlike existing web-crawled datasets which are often noisy or misidentified, KyrText aggregates high-quality news, Wikipedia entries, digitized literature, and extensive legal archives from the Supreme Court and Ministry of Justice of the Kyrgyz Republic. We leverage this corpus for the continual pre-training of mBERT, XLM-R, and DeBERTaV3, while also training RoBERTa architectures from scratch.

Evaluations across several benchmarks—including natural language inference (XNLI), question answering (BoolQ), sentiment analysis (SST-2), and paraphrase identification (PAWS-X)—demonstrate that targeted pre-training on KyrText yields substantial performance improvements over baseline multilingual models.

Our findings indicate that while base-sized models benefit immediately from this domain-specific data, larger architectures require more extensive training cycles to fully realize their potential. We release our corpus and suite of models to establish a new foundation for Kyrgyz Natural Language Processing.

1 Introduction

The development of robust language models relies on vast quantities of high-quality textual data. Although English and other high-resource languages are supported by large-scale web-crawled corpora, Kyrgyz continues to be a comparatively low-resource language. This work aims to bridge this gap by aggregating disparate sources into a unified corpus for Large Language Model (LLM) training.

In this work, we present:

1. **KyrText**: A diverse Kyrgyz corpus incorporating high-quality legal and literary texts.
2. **Kyrgyz-centric encoder models**: A suite of models optimized for Kyrgyz through continual pre-training and training from scratch.
3. **Benchmarking**: An evaluation of these models on downstream classification and natural language understanding tasks.

2 Related works

2.1 Multilingual Corpora

Kyrgyz language data is included in the widely used LLM training corpora: OSCAR (Abadji et al., 2022), mC4 (Xue et al., 2021), and CulturaX (Nguyen et al., 2024). However, these datasets were created using the web crawl data and contain noisy text and lack diversity. Even a cursory review reveals that a large ratio of the texts included in the Kyrgyz splits of these datasets is actually in other languages - predominantly in Kazakh. This is most probably caused by the poor performance of language identification libraries for Kyrgyz and Kazakh, which share a large common vocabulary.

2.2 Existing Kyrgyz Corpora

Earlier Kyrgyz resources include the Leipzig Corpora Collection (News and Community crawls from 2011 and 2017) (Leipzig Corpora Collection, a,b) and the Manas-UdS Kyrgyz Corpus (Kasieva et al., 2020). Recently, "The Cramer Project" (2025) released a Kyrgyz News Corpus (The Cramer Project, 2025) on Hugging Face. De-

Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026), pages 453–460

spite these, there is a lack of consolidated large-scale corpora that combine diverse text data.

2.3 Language models

To address the lack of pre-trained models for non-English languages, some works train models on a monolingual non-English language (Delobelle et al., 2020; Le et al., 2020; Carmo et al., 2020; Nguyen and Tuan Nguyen, 2020).
A broader approach is to pre-train multilingual models on many languages, such as mBERT (Devlin et al., 2019), mBART (Liu et al., 2020), and XLM-R (Conneau et al., 2020), which extend BERT, BART, and RoBERTa, respectively.
In this work we primarily focus on the continual pre-training of multilingual models. However, we also pre-train two models from scratch.

3 The text corpus

In order to achieve robustness and domain diversity, the corpus was designed to include both noisy web data and high-quality diverse texts from multiple domains. We compiled a multi-component dataset by aggregating the following sources:

3.1 Popular websites

We scraped the most popular news and entertainment websites in Kyrgyz as of late October 2025. Custom HTML parsers were utilized to ensure clean text extraction, followed by document-level deduplication. Language identification was carried out as a combination of website structure analysis, meta data checks, ensemble of language ID libraries, and high-frequency word checks.

3.2 Court Documents and Legislation

Legal texts were sourced from two primary government portals:

Digital Justice: Court documents downloaded from the Supreme Court of the Kyrgyz Republic portal (Supreme Court of the Kyrgyz Republic, 2025).

Legislation: Official normative legal acts from the Central Database of Legal Information of the Ministry of Justice (Ministry of Justice of the Kyrgyz Republic, 2025). This includes the Constitution, codes (Civil, Criminal, Labor, Tax), and acts from the Parliament (Jogorku Kenesh), President, and Cabinet of Ministers.

3.3 Wikipedia and Electronic Libraries

Wikipedia: A plain text version of Kyrgyz Wikipedia was extracted using a database dump from November 20, 2025.

Electronic Libraries: Documents from public, open-access libraries were scraped. For scanned documents, we employed the Deepseek OCR model to convert images into machine-readable text.

3.4 Data deduplication and language identification

The deduplication of the collected texts was done using a simple hash check after text normalization. Normalization included removing any non-alphanumeric characters and converting to lower case. 47% of the documents were removed from the *News* subset of the dataset. The removed documents were primarily non-content files. No duplicate documents were identified in the other subsets.

We estimated the LID precision by manually checking a stratified (by source) random sample of 5000 documents. No documents in languages other than Kyrgyz were found during the manual review, suggesting a high level of precision. All the documents in the *E-library* subset of the dataset were checked manually.

3.5 Corpus Statistics

Table 1 provides a detailed distribution across domains by the number of documents and words. The word counts were obtained by using a simple whitespace tokenizer.

4 Model pre-training

4.1 Tokenizer

We trained a Unigram LM tokenizer via SentencePiece (Kudo and Richardson, 2018) with a vocabulary size of 64,000. Given that Kyrgyz is morphologically rich, subword segmentation using Unigram LM provides better coverage and performance than standard BPE (Grönroos et al., 2020). The tokenizer was trained on a 10% sample of the KyrText corpus. Corpus tokenization statistics are

Table 1: Corpus statistics: Number of documents and words (mln) by domains.

Domain	Number of documents			Number of words (mln)				Upsampled for training		
	Train	Test	Total	Train	Test	Total	%	Factor	Words	%
E-library	597	32	629	22.3	1.1	23.5	3.45%	10	223.2	24.58%
Courts	171,044	1,728	172,772	160.8	1.6	162.4	23.86%	1	160.8	17.71%
Legal	155,450	1,571	157,021	66.8	0.7	67.5	9.91%	1	66.8	7.36%
News	1,895,059	19,143	1,914,202	413.7	4.3	418.0	61.43%	1	413.7	45.57%
Wikipedia	85,174	4,483	89,657	8.7	0.5	9.1	1.34%	5	43.4	4.78%
Total	**2,307,324**	**26,957**	**2,334,281**	**672.3**	**8.2**	**680.5**	**100%**	–	**907.9**	**100%**

provided in Table 2.
Even though vocabulary sizes can be an important factor for tokenizer performance in case of morphologically rich, agglutinative languages, we did not experiment with different vocabulary sizes due to computing budget constraints.

4.2 Training Modes and Architectures

Due to compute budget constraints, our primary strategy was continual pre-training. Additionally, we trained RoBERTa-Base and RoBERTa-Large architectures from scratch.

We used multilingual versions of the following encoder models:

Multilingual BERT (base version) (Devlin et al., 2019), *XLM-RoBERTa* (base and large versions) (Conneau et al., 2020), *Multilingual DeBERTa v3* (base version) (He et al., 2021), *RoBERTa* architecture (base and large versions) (Zhuang et al., 2021) was used for pre-training from scratch.

4.3 Training Procedure

Data Preparation: Wikipedia data was upsampled by a factor of 5, and the *E-library* subset by a factor of 10. The corpus was split into train and test sets with 99% and 1% ratios respectively. The training samples were chunked up to maximum sequence length of 512 across batches of size 1000.

Table 3: Token Counts per Model

Model	Tokens (Upsampled)
mbert-base	2689.8
mdeberta-base-v3	2365.5
roberta	1334.6
xlm-roberta	1920.3

Training configuration: We used a maximum sequence length of 512. Continual pre-training was carried out for 3 epochs using the Masked Language Modeling (MLM) objective. Pre-training from scratch was extended to 5 epochs.

Model pre-training was conducted on NVIDIA RTX 4090 and RTX 5090 GPUs using single-GPU training. The total wall time was approximately 300 hours.

All models were trained using the AdamW optimizer. The learning rates was set $\mathbf{10^{-4}}$, $\boldsymbol{\beta_1}$ was set at 0.9, $\boldsymbol{\beta_2}$ at 0.98, $\boldsymbol{\epsilon}$ at $\mathbf{10^{-9}}$, weight decay at 0.01, and warmup ratio at 10%. FP16 precision was used for all models. The masking probability was set at 15%. The training batch sizes were adjusted based on the used GPU VRAM and model sizes.

Table 4: Batch sizes and training time

Model	Batch	GAS[1]	Time(h)
mbert-base-kgz	30	3	33
mdeberta-base-v3-kgz	10	4	66
roberta-base-kgz	40	4	25
roberta-large-kgz	48	3	41
xlm-roberta-base-kgz	8	9	60
xlm-roberta-large-kgz	13	5	70

4.4 MLM loss

We measured the MLM loss using the test set of the corpus. All multilingual models showed a decrease in MLM loss after the continual pre-training on the corpus. However, the RoBERTa-Large model trained from scratch exhibited a relatively high MLM loss. This is likely due to the model being undertrained (only 5 epochs) and the total training data volume being insufficient for a model of that scale. The decrease for the Multilingual DeBERTa v3 was the highest probably due to a very small size of the Kyrgyz split of the CC100 (Wenzek et al.,

[1]Gradient accumulation steps.

Table 2: Token Statistics: token counts and average tokens per word

	mBERT		Unigram		DeBERTa		XLM-RoBERTa	
Source	Tokens	Avg tpw	Tokens	Avg tpw	Tokens	Avg tpw	Tokens	Avg tpw
E-library	70.0	2.98	36.9	1.57	60.5	0.05	2.5	2.24
Courts	442.8	2.73	215.8	1.33	395.4	0.03	16.3	1.95
Legal	203.8	3.02	91.8	1.36	178.4	0.01	38.8	2.06
News	1184.6	2.83	567.6	1.36	1038.9	0.08	13.2	1.95
Wikipedia	25.5	2.79	14.6	1.59	23.9	0.02	0.4	2.24
Total	**1926.7**	**2.83**	**926.7**	**1.36**	**1697.1**	**0.01**	**1341.3**	**1.97**

2020) dataset. It should be noted that MLM loss comparisons are model-relative, not absolute.

Model	**MLM loss**
kyrgyz-bert	6.4
roberta-base-kgz	1.55
roberta-large-kgz	5.05
xlm-roberta-base	1.38
xlm-roberta-base-kgz	0.61
xlm-roberta-large	1.03
xlm-roberta-large-kgz	0.48
mbert-base	1.35
mbert-base-kgz	0.32
mdeberta-base-v3	19.27
mdeberta-base-v3-kgz	0.5

Table 5: MLM loss values on the test set.

5 Benchmarks

5.1 Tasks

News Topic Classification. We constructed a dataset for news topic classification. 4648 news articles were randomly selected across 5 categories: science (276), politics (1001), culture (886), economics (1579), and sport (906). The evaluation was conducted on the titles of news articles.

PAWS-X. The (Yang et al., 2019) PAWS-X dataset is a cross-lingual version of the (Zhang et al., 2019) PAWS dataset, designed for evaluating paraphrase identification across multiple languages: given a pair of sentences, the task is to predict whether they have the same meaning (paraphrase) or not. The PAWS-X pairs in English were translated to Kyrgyz using the Google Translation API with further manual review and post-editing.

BoolQ. BoolQ is a dataset for question answering (QA) where the answers are yes/no (Clark et al., 2019). The dataset is designed to test reading comprehension in a natural setting: given a passage and a question, predict whether the answer is "yes" or "no." We used the Kyrgyz translation of the dataset included in the KyrgyzLLM-Bench (Turatali et al., 2025).

XNLI. XNLI (Conneau et al., 2018) stands for Cross-lingual Natural Language Inference. The dataset is widely used for evaluating multilingual and cross-lingual understanding in NLP. The task is specified as given a premise and a hypothesis, predict the relationship: *Entailment:* the hypothesis logically follows from the premise;
Contradiction: the hypothesis contradicts the premise;
Neutral: neither entailment nor contradiction.
We translated a subset of the dataset with 7000 samples using the Google Translation API.

SST-2. The subset of the Stanford Sentiment Treebank (SST) (Socher et al., 2013) dataset, translated to Kyrgyz (Metinov et al., 2025) using machine translation tools. The dataset is designed as a binary sentiment classification task: classifying a sentence as positive or negative. Unlike fine-grained SST (5 labels), SST-2 collapses neutral and other categories into positive/negative, making it binary.

5.2 Fine-tuning

Models were fine-tuned for 3 epochs for the downstream tasks. We report the mean accuracy, used in GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) benchmarks, across three iterations.
We note that our results are based on three runs without formal significance testing due to computational constraints. No hyperparameter tuning was carried out and similar training parameters were used for all models to assure fair comparison. Train-

Table 6: Model Evaluation Results (Accuracy ± Standard Deviation)

Model	boolq	news	paws-x	sst	xnli
kyrgyz-bert	$\mathbf{0.60 \pm 0.0027}$	$\mathbf{0.83 \pm 0.0144}$	$\mathbf{0.58 \pm 0.0023}$	$\mathbf{0.82 \pm 0.0003}$	$\mathbf{0.47 \pm 0.0270}$
mbert-base	$\mathbf{0.65 \pm 0.0057}$	$\mathbf{0.87 \pm 0.0078}$	$\mathbf{0.77 \pm 0.0012}$	$\mathbf{0.84 \pm 0.0003}$	$\mathbf{0.60 \pm 0.0047}$
mbert-base-kgz	$\mathbf{0.66 \pm 0.0136}$	$\mathbf{0.90 \pm 0.0069}$	$\mathbf{0.70 \pm 0.0325}$	$\mathbf{0.85 \pm 0.0003}$	$\mathbf{0.64 \pm 0.0031}$
mdeberta-base-v3	$\mathbf{0.65 \pm 0.0352}$	$\mathbf{0.90 \pm 0.0045}$	$\mathbf{0.81 \pm 0.0078}$	$\mathbf{0.85 \pm 0.0008}$	$\mathbf{0.62 \pm 0.0080}$
mdeberta-base-v3-kgz	$\mathbf{0.67 \pm 0.0043}$	$\mathbf{0.91 \pm 0.0033}$	$\mathbf{0.75 \pm 0.0010}$	$\mathbf{0.87 \pm 0.0019}$	$\mathbf{0.61 \pm 0.0240}$
roberta-base-kgz	$\mathbf{0.64 \pm 0.0177}$	$\mathbf{0.91 \pm 0.0043}$	$\mathbf{0.60 \pm 0.0023}$	$\mathbf{0.83 \pm 0.0025}$	$\mathbf{0.58 \pm 0.0060}$
roberta-large-kgz	$\mathbf{0.63 \pm 0.0202}$	$\mathbf{0.92 \pm 0.0094}$	$\mathbf{0.58 \pm 0.0031}$	$\mathbf{0.82 \pm 0.0019}$	$\mathbf{0.53 \pm 0.0262}$
xlm-roberta-base	$\mathbf{0.66 \pm 0.0207}$	$\mathbf{0.90 \pm 0.0087}$	$\mathbf{0.81 \pm 0.0119}$	$\mathbf{0.87 \pm 0.0017}$	$\mathbf{0.62 \pm 0.0260}$
xlm-roberta-base-kgz	$\mathbf{0.70 \pm 0.0083}$	$\mathbf{0.92 \pm 0.0065}$	$\mathbf{0.71 \pm 0.0006}$	$\mathbf{0.88 \pm 0.0031}$	$\mathbf{0.66 \pm 0.0023}$
xlm-roberta-large	$\mathbf{0.72 \pm 0.1128}$	$\mathbf{0.90 \pm 0.0054}$	$\mathbf{0.66 \pm 0.1553}$	$\mathbf{0.90 \pm 0.0077}$	$\mathbf{0.55 \pm 0.1899}$
xlm-roberta-large-kgz	$\mathbf{0.79 \pm 0.0059}$	$\mathbf{0.92 \pm 0.0045}$	$\mathbf{0.85 \pm 0.0092}$	$\mathbf{0.89 \pm 0.0019}$	$\mathbf{0.76 \pm 0.0023}$

ing batch size of 8 was used for all tasks, except for SST-2, where 16 was used. The learning rate was set at 1e-5. The weight decay was set at 0.1. FP16 precision was used for all tasks.

Fine-tuning of models for benchmark tasks was carried out an NVIDIA RTX 3090 GPU.

5.3 Results

Table 6 presents the evaluation results across all five benchmark tasks.

Impact of Continual Pre-training: Continual pre-training on KyrText consistently improved model performance. Most notably, *xlm-roberta-large-kgz* achieved the strongest average performance across tasks, with 0.79 on BoolQ, 0.85 on PAWS-X, 0.76 on XNLI, 0.92 on News. Comparing base models with their *-kgz* counterparts reveals substantial gains: *xlm-roberta-large-kgz* improved over *xlm-roberta-large* by 7 percentage points on BoolQ (0.79 vs 0.72) and 19 percentage points on XNLI (0.76 vs 0.55), and 18 percentage points on PAWS-X (0.85 vs 0.66). However, it should be noted that *xlm-roberta-large* demonstrated very high variance across training experiments, whereas the Kyrgyz fine-tuned version was stable. The instability of the *xlm-roberta-large* model could be explained by relative size of the used datasets to the model size.

Task Difficulty: News topic classification proved the easiest task, with most models achieving 0.90+ accuracy. XNLI was the most challenging, with the best model reaching 0.76 accuracy. The *kyrgyz-bert* baseline showed particularly low performance on XNLI (0.47), highlighting the importance of multilingual pretraining foundations.

Model Comparison: Among continually pretrained models, *xlm-roberta-large-kgz* demonstrated the strongest overall performance, followed by *xlm-roberta-base-kgz* and *mdeberta-base-v3-kgz*. Interestingly, base-sized models with continual pretraining often outperformed larger models without it, suggesting the effectiveness of domain-specific adaptation.

Training from Scratch: Models trained from scratch (*roberta-base-kgz, roberta-large-kgz*) showed competitive performance on News and SST tasks but lagged on PAWS-X and XNLI, consistent with their higher MLM losses (Table 5) and suggesting these tasks benefit more from multilingual knowledge.

6 Conclusion

This research introduced KyrText, a comprehensive, multi-domain Kyrgyz corpus totaling 680.5 million words. By consolidating high-quality legal documents, news, Wikipedia entries, and digitized literary texts, this work addresses the critical scarcity of high-quality linguistic resources for the Kyrgyz language—a morphologically rich Turkic language historically classified as low-resource.
The efficacy of the KyrText corpus was demonstrated through the continual pre-training of multilingual encoder models and the development of RoBERTa-based architectures from scratch.

Key findings from our evaluation include:

Performance Gains: Targeted continual pre-training significantly enhanced model performance across downstream tasks. We also observed

substantial reduction in Masked Language Modeling (MLM) loss across all evaluated multilingual models.

Architectural Limitations: While continual pre-training yielded immediate benefits, results suggest that larger architectures, such as RoBERTa-Large, require more extensive data volumes and training epochs to fully realize their performance potential compared to base-sized models.

Task Versatility: The suite of Kyrgyz-centric models demonstrated competitive accuracy across diverse natural language understanding tasks, including sentiment analysis (SST-2), paraphrase identification (PAWS-X), question answering (BoolQ) and natural language inference (XNLI).

By providing both the KyrText corpus and a suite of optimized encoder models, this study establishes a robust foundation for advanced Kyrgyz Natural Language Processing (NLP). This contribution facilitates the development of more accurate and culturally nuanced language technologies, effectively bridging the gap between Kyrgyz and higher-resource languages in the era of Large Language Models.

Limitations

The development of the KyrText corpus and the subsequent evaluation of Kyrgyz-centric models are subject to several limitations that should be considered:

Hardware and Compute Constraints: Due to significant computing budget constraints, the study primarily utilized a continual pre-training strategy, only training a small number of architectures from scratch.

Model Undertraining: The RoBERTa-Large model trained from scratch exhibited a relatively high Masked Language Modeling (MLM) loss, likely indicating that the model was undertrained at only 5 epochs.

Data Volume for Large Architectures: The total training data volume within KyrText may be insufficient for larger-scale models like RoBERTa-Large to reach their maximum potential performance.

Tokenizer Hyperparameters: While subword segmentation was optimized for Kyrgyz's morphologically rich nature, experiments with different vocabulary sizes were not conducted due to budget limitations.

Evaluation Scope: No hyperparameter tuning was performed during the fine-tuning stage for downstream tasks to ensure a baseline for comparison, which may mean the reported accuracies do not represent the absolute ceiling for these models.

Translation Artifacts: Several benchmark datasets, including XNLI, SST-2, and PAWS-X, relied on machine translation (Google Translation API), which may introduce synthetic artifacts or translation errors despite manual review for some subsets. Lack of robust and diverse evaluation benchmarks for Kyrgyz is a major limitation. However, creation of such benchmarks is beyond the scope of this work.

Data Governance and Ethics

All datasets used in this work are derived from publicly accessible sources. The corpus consists of news articles, official documents (including court decisions), and materials hosted in publicly available electronic libraries. Under Article 8 of the Law of the Kyrgyz Republic "On Copyright and Related Rights," news articles and official documents are not considered protected works.

To mitigate potential copyright and redistribution concerns, the released dataset does not include the full text of documents obtained from electronic libraries. Instead, it contains URLs pointing to the original public sources, enabling users to independently access the materials under the terms and conditions set by the respective content providers. The dataset is released for research purposes only.

No personal or sensitive data were intentionally collected. Court decisions published by the (Supreme Court of the Kyrgyz Republic, 2025) are anonymized prior to publication, and the study relies exclusively on these already anonymized versions.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto A. Lotufo. 2020. PTT5: pre-training and validating the T5 model on brazilian portuguese data. *CoRR*, abs/2008.09144.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

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

Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. 2020. Morfessor EM+Prune: Improved subword segmentation with expectation maximization and pruning. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3944–3953, Marseille, France. European Language Resources Association.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

A. A. Kasieva, J. Knappen, S. Fischer, and E. Teich. 2020. A new kyrgyz corpus: Sampling, compilation, annotation. Poster session, Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Leipzig Corpora Collection. a. Kyrgyz community corpus based on material from 2017.

Leipzig Corpora Collection. b. Kyrgyz news corpus (crawled 2011).

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

Adilet Metinov, Gulida M. Kudakeeva, and Gulnara D. Kabaeva. 2025. Kyrgyzbert: A compact, efficient language model for kyrgyz nlp. *Preprint*, arXiv:2511.20182.

Ministry of Justice of the Kyrgyz Republic. 2025. Centralized database of legal information of the ministry of justice of the kyrgyz republic. `https://cbd.minjust.gov.kg/`.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources*

and Evaluation (LREC-COLING 2024), pages 4226–4237, Torino, Italia. ELRA and ICCL.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Supreme Court of the Kyrgyz Republic. 2025. Digital justice portal of the supreme court of the kyrgyz republic. `https://portal.sot.kg/`.

The Cramer Project. 2025. Kyrgyz news corpus. `https://huggingface.co/datasets/the-cramer-project/Kyrgyz_News_Corpus`. Hugging Face dataset, cc-by-nc-4.0 license.

Timur Turatali, Aida Turdubaeva, Islam Zhenishbekov, Zhoomart Suranbaev, Anton Alekseev, and Rustem Izmailov. 2025. Bridging the gap in less-resourced languages: Building a benchmark for kyrgyz language models. In *2025 10th International Conference on Computer Science and Engineering (UBMK)*, pages 1673–1677.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. *SuperGLUE: a stickier benchmark for general-purpose language understanding systems*. Curran Associates Inc., Red Hook, NY, USA.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. 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.

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.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. 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.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

7 Appendix A. Data and code availability

The KyrText corpus and tokenized datasets for pre-training are available at: https://huggingface.co/collections/tchubakov/kyrtext.

The source code for model pre-training and task-specific fine-tuning is provided at: https://github.com/tilekchubakov/kyrtext.

The released code reproduces the core experiments reported in this paper. Additional utilities, extensions, and model checkpoints will be added in future releases.