On the Path to Make Ukrainian a High-Resource Language

Mykola Haltiuk
AGH University of Krakow
mhaltiuk@agh.edu.pl

Aleksander Smywiński-Pohl
AGH University of Krakow
apohllo@agh.edu.pl

Abstract

Recent advances in multilingual language modeling have highlighted the importance of high-quality, large-scale datasets in enabling robust performance across languages. However, many low- and mid-resource languages, including Ukrainian, remain significantly underrepresented in existing pretraining corpora. We present Kobza, a large-scale Ukrainian text corpus containing nearly 60 billion tokens, aimed at improving the quality and scale of Ukrainian data available for training multilingual language models. We constructed Kobza from diverse, high-quality sources and applied rigorous deduplication to maximize data utility. Using this dataset, we pre-trained Modern-LiBERTa, the first Ukrainian transformer encoder capable of handling long contexts (up to 8192 tokens). Modern-LiBERTa achieves competitive results on various standard Ukrainian NLP benchmarks, particularly benefiting tasks that require broader contextual understanding or background knowledge. Our goal is to support future efforts to develop robust Ukrainian language models and to encourage greater inclusion of Ukrainian data in multilingual NLP research.

1 Introduction

Recent progress in Large Language Models (LLMs) has been strongly driven by the scale and quality of pre-training data. While English enjoys massive, high-quality corpora, many other languages – including Ukrainian – remain significantly underrepresented in the datasets used for multilingual model training (Grattafiori et al., 2024, Nguyen et al., 2023, Penedo et al., 2024). As a result, it often receives less attention during training, leading to suboptimal performance on Ukrainian inputs in otherwise powerful multilingual models. To change this, we believe it is essential to make high-quality data widely available and to encourage its inclusion in future multilingual training pipelines.

To support this goal, we present *Kobza*, a new large-scale Ukrainian text corpus containing nearly 60 billion tokens. To our knowledge, this is the largest publicly available Ukrainian corpus to date. Kobza is designed to be easily integrated into multilingual data mixtures for LLM training, and we hope it will help raise the share of Ukrainian in such efforts.

Alongside the dataset, we pre-train Modern-LiBERTa, a long-context transformer encoder that supports input sequences of up to 8,192 tokens. The model builds on the ModernBERT Large (Warner et al., 2024) architecture, originally designed for efficient, high-throughput processing on modern hardware. Modern-LiBERTa is the first Ukrainian-language model capable of handling such long contexts, enabling improved performance on tasks that require document-level understanding.

Finally, we outline a broader initiative to support the development of Ukrainian Natural Language Processing. In the future work, we plan to expand Kobza to at least 100 billion tokens and to build lightweight tools for filtering and scoring document quality in Ukrainian. Although this part lies beyond the scope of this paper, it is a key component of our long-term vision: to elevate Ukrainian to a high-resource language in the era of large-scale language technologies.

Our main contributions can be defined as follows:

- We compile Kobza, the largest Ukrainian text corpus to date, comprising nearly 60B tokens, suitable for both monolingual and multilingual LLM training.
- We pre-train Modern-LiBERTa, the first Ukrainian encoder model with support for long sequences (up to 8,192 tokens).

Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025), pages 120–130

- By releasing Kobza[1] and Modern-LiBERTa[2] along with the source code[3], we contribute to ongoing efforts aimed at improving the quality and availability of Ukrainian data, with the long-term goal of making it a high-resource language for NLP.

2 Related Work

Recent best-performing language models share the same trait – the scale of their training data. Leading English and multilingual models, such as ModernBERT (trained on 2 trillion tokens, Warner et al., 2024), NeoBERT (600 billion, Breton et al., 2025), and EuroBERT (5 trillion, Boizard et al., 2025), exemplify this trend. These models benefit not only from the vast availability of high-quality English data (Soboleva et al., 2023), but also from a mature ecosystem of tools for data curation (Jennings et al., 2024) and synthesis (Gunasekar et al., 2023).

In contrast, many other languages, particularly mid- and low-resource ones, lag behind in terms of data availability and tooling. Ukrainian is a prime example. While it is spoken by tens of millions and supported by active linguistic and technological communities, the scale and quality of data available for large-scale pre-training still remains limited.

Large Multilingual Corpora The primary source of pre-training data for large models is the open web, typically accessed through initiatives such as Common Crawl (CC). CC data serves as the foundation for corpora like OSCAR (Ortiz Su'arez et al., 2020, Ortiz Su'arez et al., 2019), C4, mC4 (Raffel et al., 2019), CC100 (Wenzek et al., 2020), and Pile-CC (Gao et al., 2020). These datasets played a foundational role in early multilingual modeling efforts and continue to inform newer datasets with improved filtering and language coverage.

CulturaX (Nguyen et al., 2023) builds upon mC4 and OSCAR by applying more rigorous filtering. It re-labels languages using FastText (Bojanowski et al., 2017), discarding documents whose re-identified language mismatches the original. It then applies URL-based filtering to remove harmful or toxic domains, followed by basic quality metrics and a deduplication pass using MinHashLSH (Anand and Jeffrey David, 2011).

FineWeb 2 (Penedo et al., 2024) expands language coverage significantly, identifying documents using GlotLID (Kargaran et al., 2023), which supports many more languages than FastText. It applies per-language deduplication and filtering with language-specific parameters, including stopword lists. To balance frequency effects, the corpus is "rehydrated," meaning that documents are duplicated based on their frequency in the original crawl—though very frequent documents (appearing more than 1,000 times) are capped to a single instance, assuming lower quality.

HPLT 2.0 (Burchell et al., 2025) provides a complementary dataset by relying heavily on Internet Archive crawls rather than Common Crawl. Its pipeline includes OpenLID (Burchell et al., 2023) for language detection, followed by deduplication and filtering using the Web Document Scorer[4] (WDS), a quality estimation tool based on linguistic signals. Documents scoring below a quality threshold (e.g., WDS < 5) are discarded.

While these corpora make important strides toward better multilingual coverage, their treatment of Ukrainian often remains shallow. In many cases, filtering parameters and identification models are tuned for higher-resource languages, which can result in suboptimal data quality or volume for Ukrainian.

Ukrainian Corpora Several Ukrainian-focused corpora have also been developed over the last years: Zvidusil (Kotsyba et al., 2018), ukTenTen[5], Brown-UK (Starko and Rysin, 2023), etc. Among these, Malyuk[6], a compilation of UberText 2.0 (Chaplynskyi, 2023), the Ukrainian News dataset[7] and OSCAR, stands out as the largest and most linguistically rich. UberText 2.0, a core component of Malyuk, differs from multilingual corpora that rely heavily on large-scale web crawls by using custom web crawlers tailored specifically to Ukrainian-language sources. This results in high-quality documents, albeit potentially with reduced domain diversity. The dataset also includes multiple layers of linguistic annotation, such as tokenization, lemmatization, and part-of-speech tagging.

Model Architecture Another direction of improving language modeling has been the modifi-

[1] https://huggingface.co/datasets/Goader/kobza
[2] https://huggingface.co/Goader/modern-liberta-large
[3] https://github.com/Goader/ukr-lm
[4] https://github.com/pablop16n/web-docs-scorer/
[5] https://www.sketchengine.eu/uktenten-ukrainian-corpus/
[6] https://huggingface.co/datasets/lang-uk/malyuk
[7] https://huggingface.co/datasets/zeusfsx/ukrainian-news

cation of model architectures and pre-training procedures, addressing a range of goals: speeding up inference by making the model more compatible with modern GPU hardware, improving downstream performance across various tasks (Clark et al., 2020, He et al., 2021), and specifically optimizing for retrieval tasks, which have become increasingly prominent with the rise of Retrieval-Augmented Generation (RAG, Lewis et al., 2020). A further focus has been on extending the model's context window, allowing it to process significantly longer documents in a single pass.

Cross-Lingual Transfer Another important dimension of improvement in language model pre-training is cross-lingual transfer. It is now well established that initializing a model with weights from a related language model outperforms training from scratch, especially when the target language has limited data (Minixhofer et al., 2022).

Several methods have been proposed to bridge vocabularies and embedding spaces between languages. WECHSEL (Minixhofer et al., 2022) uses a bilingual dictionary to learn a linear transformation between embedding spaces. FOCUS (Dobler and de Melo, 2023) improves on this by leveraging overlapping subwords between source and target vocabularies. The most recent line of work, such as Trans-Tokenization (Remy et al., 2024), builds translation dictionaries from parallel corpora using FastAlign by Dyer et al. (2013) and applies additional alignment steps to handle multi-token mappings, increasing both accuracy and coverage.

These techniques have enabled the development of Ukrainian variants of RoBERTa (Liu et al., 2019), although so far these efforts have been limited to relatively small corpora and standard context windows.

3 Kobza

In this section, we describe the collection and preparation of the Kobza corpus – a large-scale Ukrainian text dataset.

3.1 Sources

We rely on publicly available multilingual and monolingual corpora, prioritizing those that offer substantial Ukrainian coverage. Unlike some large-scale efforts that process raw Common Crawl data directly, we focus on merging curated datasets, reducing preprocessing overhead while preserving document diversity and quality.

CulturaX CulturaX (Nguyen et al., 2023) is a multilingual web corpus, where Ukrainian ranks 21st in terms of token count. The corpus contains 38 billion Ukrainian tokens, which represent 0.61% of its total volume, distributed across approximately 44 million documents.

FineWeb 2 FineWeb 2 (Penedo et al., 2024) includes Ukrainian as the 24th most represented language, with 23 billion words (0.86% of the total corpus) spread across 47 million documents.

HPLT 2.0 The HPLT 2.0 (Burchell et al., 2025) corpus offers 25 billion Ukrainian tokens, making it the 21st largest language in the collection. We use the cleaned version of this dataset, which includes 47 million documents.

Ukrainian News We incorporate the Ukrainian News dataset[8], which aggregates 16 million news articles from media outlets and over 6.5 million Telegram posts. This source adds both formal and informal texts and provides a high volume of short documents. We extract clean content using Trafilatura (Barbaresi, 2021), focusing on removing boilerplate and eliminating duplicate content.

UberText 2.0 UberText 2.0 (Chaplynskyi, 2023) is a monolingual Ukrainian corpus with approximately 2.5 billion tokens and 8.5 million documents. It comprises five domains: news, fiction, social, Wikipedia, and legal, offering a wide range of styles and document lengths.

3.2 Deduplication

Merging corpora from diverse sources inevitably introduces duplicate content. This issue is especially pronounced when datasets reuse similar web sources, such as Common Crawl. Duplicates may occur both as exact matches and near-duplicates due to differing preprocessing steps. To address this, we applied a two-stage deduplication process across the entire combined corpus.

Metadata-based In the first stage, we filter documents using metadata such as URLs and timestamps. This method captures many duplicates originating from processing the same documents from Common Crawl, even when the content differs. We validate this approach by calculating document similarity based on the normalized longest common subsequence (LCS):

[8] https://huggingface.co/datasets/zeusfsx/ukrainian-news

Subcorpora	Documents	Tokens
CulturaX	24,942,577	15,002,455,535
FineWeb 2	32,124,035	19,114,177,138
HPLT 2.0	26,244,485	20,709,322,905
UberText 2.0	6,431,848	2,904,208,874
Ukrainian News	7,175,971	1,852,049,111
Total	96,918,916	59,582,213,563

Table 1: Kobza token statistics

$$sim(a, b) = \frac{LCS(a, b)}{\min(|a|, |b|)}, \quad (1)$$

where a and b are document texts. This definition yields a 100% similarity if one text is a substring of the other. On a large sample of matched pairs, the average similarity was 92.9%, indicating that metadata-based deduplication effectively captures redundant documents. Overall, this step removes approximately 12% of the corpus.

MinHashLSH To identify near-duplicates not caught in the metadata phase, we apply MinHashLSH (Anand and Jeffrey David, 2011), a method that approximates Jaccard similarity over n-grams. We use 5-grams, a similarity threshold of 0.7, and implement the method using the `text-dedup`[9] package on Apache Spark for scalability. This stage removes an additional 33% of the documents.

3.3 Data Quality

While the included datasets have undergone quality filtering, either through heuristics (CulturaX, FineWeb 2, HPLT 2.0) or through source curation (Ukrainian News, UberText 2.0), these methods were not always optimized for Ukrainian. As a result, low-quality or noisy texts may still be present.

We highlight the need for a dedicated Ukrainian document quality scorer to improve future corpus construction. Developing such a tool remains an open direction for further research.

3.4 Statistics

The final Kobza corpus consists of nearly 60 billion tokens across about 97 million documents. It occupies 474GB of disk space in Parquet format with Snappy compression. Table 1 presents the number of tokens and documents per subcorpus.

As shown in Figure 1, a substantial share of the cumulative token distribution resides in longer documents. This makes the corpus especially suitable for training and evaluating models with extended context windows.

Each document in the Kobza corpus includes metadata such as the source, subsource, timestamp, and URL. This enables fine-grained data selection and filtering.

4 Modern-LiBERTa

This section outlines how we adapted ModernBERT (Warner et al., 2024), originally trained exclusively on English data, for use with Ukrainian. We describe the training corpus, model architecture, tokenizer, initialization approach, and training setup, including the extension to long-context sequences.

4.1 Training Data

Our training corpus combines Ukrainian and English text. The core of the Ukrainian data is the deduplicated version of the Kobza corpus, which contains approximately 60 billion tokens. We include the English Wikipedia[10], contributing roughly 6 billion tokens to support cross-lingual and knowledge-intensive tasks. This English portion accounts for about 10% of the total training mixture.

Inclusion of English Wikipedia is motivated by the frequent presence of English words and entities in Ukrainian texts – especially in news, technical, and academic domains – and its potential to improve performance on tasks such as Named Entity Recognition (NER) and Information Retrieval (IR).

4.2 Model Architecture

Modern-LiBERTa closely follows the ModernBERT architecture. It consists of 28 transformer layers with a hidden size of 1,024, totaling 410 million parameters. The design emphasizes efficiency, particularly for GPU acceleration, incorporating recent advances such as: Rotary Positional Embeddings (RoPE, Su et al., 2021) for effective long-sequence modeling, Flash Attention (Dao, 2023) for memory-efficient attention computation, alternating attention patterns that reduce the compute cost of scaling to long sequences without compromising model expressiveness.

[9] https://github.com/ChenghaoMou/text-dedup

[10] https://dumps.wikimedia.org/

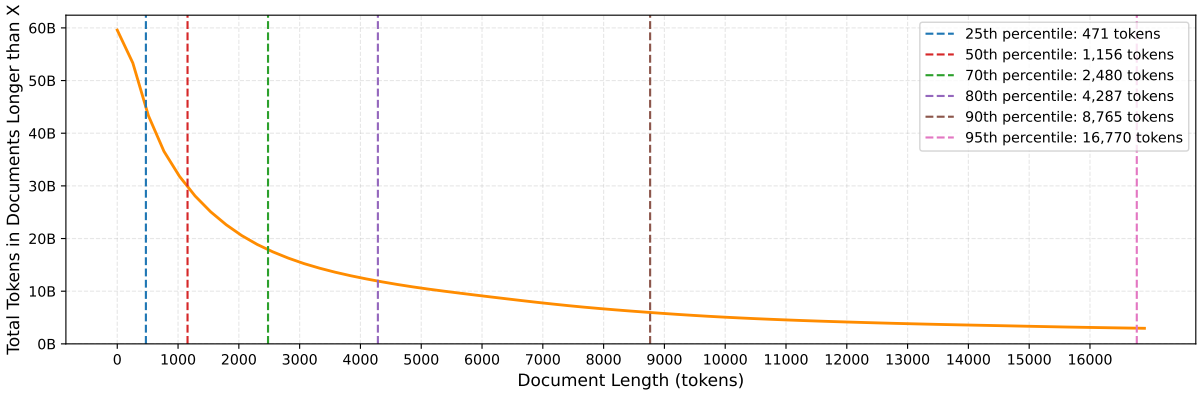

Figure 1: Cumulative token distribution with percentiles marked. y-axis indicates the total number of tokens originating from documents longer than x.

4.3 Tokenizer

We use the LiBERTa v2 (Haltiuk and Smywiński-Pohl, 2024) tokenizer with a vocabulary of 64,000 tokens. It was trained on Ukrainian text from the CC100 (Wenzek et al., 2020) corpus, with a small portion of English news articles included to improve coverage of named entities that may appear in English. This design choice aligns with our inclusion of English data in the pre-training mixture and helps ensure consistent tokenization of such entities during model training.

4.4 Weights Initialization

To accelerate convergence and transfer knowledge, we initialize the model using weights from the original English-trained ModernBERT Large. All layers are directly reused except for the input and output embeddings, which are replaced to match the new vocabulary.

For embedding initialization, we apply a TransTokenization procedure by Remy et al. (2024). Using parallel corpora, OpenSubtitles (Lison and Tiedemann, 2016) and NLLB (Costa-Jussà et al., 2022), we align Ukrainian and English tokens via FastAlign (Dyer et al., 2013). This allows us to map new tokens to semantically similar ones in the original vocabulary. We then construct each new embedding as a weighted linear combination of the corresponding English embeddings, using the official `transtokenizers`[11] toolkit.

4.5 Training Settings

The overall training was done in 2 phases: general pre-training phase with the sequence length of 1,024, which lasted for 140B tokens, and context extension phase, where it gets extended to 8,192, for 20B tokens. All the hyperparameters during each phase are presented in Table 2.

Objective Following MosaicBERT by Portes et al. (2023), we use the Masked Language Modeling (MLM) objective with the full-word masking rate of 30%.

Optimizer We use StableAdamW (Wortsman et al., 2023) with a fully decoupled weight decay, implemented in the `optimi`[12] package. It ports Adafactor's update clipping (Shazeer and Stern, 2018) into AdamW (Loshchilov and Hutter, 2017) as a per-parameter learning rate modification, which has been shown to outperform regular gradient clipping.

Learning Rate Schedule Unlike ModernBERT, we stick to cosine decay with a peak learning rate of 5e-4 and decay to 5e-5 at the end of the first phase. The second phase follows the cosine decay schedule without any warm-up, starting at 5e-5 and decaying to 0.

Hardware Setup The training was conducted on the CYFRONET Helios Cluster on 4 nodes, each equipped with 4x GH200 96GB Superchips using Distributed Data Parallel (DDP, Li et al., 2020) strategy. We set the batch size per device to 16, gradient accumulation steps to 16, totaling an effective batch size of 4,096.

Context Length Extension For context length extension, we continued pre-training from the last checkpoint for an additional 20 billion tokens using a specially constructed data mixture with sequences up to 8,192 tokens long. It was developed following

[11] https://github.com/LAGoM-NLP/transtokenizer

[12] https://optimi.benjaminwarner.dev/

	Pretraining Phase	Context Length Extension
Training Tokens	140 billion	20 billion
Max Sequence Length	1,024	8,192
Batch Size	4,096	1,024
Batch Size per GPU	16	4
Gradient Accumulation	16	16
Learning Rate (Peak)	5e-4	5e-5
Schedule	Cosine	Cosine
Warmup (tokens)	5 billion	-
Decayed Learning Rate	5e-5	0
Weight Decay	1e-5	1e-6
Total Time (hours)	133	24
Optimizer	StableAdamW	
Betas	(0.90, 0.98)	
Epsilon	1e-6	
Training Hardware	16x GH200	
Training Strategy	Distributed DataParallel	

Table 2: Modern-LiBERTa training hyperparameters.

Fu et al. (2024), with the goal of preserving the original data distribution. The final mixture includes 8 billion tokens from documents with at least 4,096 tokens, another 8 billion from documents ranging between 1,024 and 4,096 tokens, and 4 billion from shorter documents under 1,024 tokens. This stratification was introduced to maintain the model's performance on shorter inputs, as prior work (Gao et al., 2024) has shown that the absence of short documents can significantly degrade performance on certain tasks. During the construction of the mixture, we also upsampled higher-quality sources, according to Gao et al. (2024).

5 Evaluation

In this section, we evaluate the performance of Modern-LiBERTa across a range of language understanding benchmarks for Ukrainian. We focus on two aspects: (1) intrinsic language modeling quality, measured via Masked Language Modeling (MLM) perplexity, and (2) performance on a set of standard downstream tasks, in comparison to existing Ukrainian and multilingual models.

5.1 Masked Language Modeling Perplexity

To assess the intrinsic modeling capabilities of Modern-LiBERTa, we report MLM perplexity and token-level accuracy. Since Modern-LiBERTa and LiBERTa v2 (Haltiuk and Smywiński-Pohl, 2024) use the same tokenizer, we are able to directly compare their results.

Definition We define perplexity over masked tokens as:

$$ppl(X) = \exp\left\{-\frac{1}{|M|}\sum_{x \in M} \log p_\theta(x \mid X - M)\right\} \tag{2}$$

where M denotes the set of masked tokens, $p_\theta(x \mid X - M)$ is the probability of a masked token x predicted by the model, given the unmasked context.

To align with common practice, we first mask 15% of words, then tokenize them using the target model's tokenizer. Each masked word is replaced with one or more `<mask>` tokens depending on how it is tokenized. The model predicts the probabilities for every input `<mask>` token, which are then used to compute perplexity as in Equation 2.

Datasets We report perplexity results on the following datasets, selected for their quality and diversity:

- **Ukrainian Universal Dependencies (UD)**: A curated corpus of well-formed Ukrainian documents with detailed linguistic annotations (Kotsyba et al., 2018). It contains over 100,000 tokens and serves as a standard benchmark for part-of-speech tagging.
- **Spivavtor (targets only)**: A collection of Ukrainian sentences derived from instruction-following tasks (Saini et al., 2024), including simplification, coherence, paraphrasing,

and fluency/grammatical error correction (including UA-GEC dataset by Syvokon et al., 2023). Only the fluency and grammatical error correction subset (approximately 44.5% of the data) is manually annotated in Ukrainian, while the rest is machine-translated from English. We use only the target outputs for evaluation, which vary in quality due to the mixed sources.

- **UA-GEC (targets only)**: A high-quality, manually annotated grammatical error correction dataset. We report it separately from Spivavtor to target only carefully curated Ukrainian text.

- **Ukrainian Wikipedia**: A large and diverse corpus covering encyclopedic content[13]. It offers a complementary benchmark with longer and more knowledge-rich documents.

Results Results are presented in Table 3. Modern-LiBERTa consistently outperforms LiBERTa v2 across all datasets in both perplexity and token-level accuracy. All documents were truncated to 512 tokens to ensure a fair comparison, avoiding any advantage from ModernBERT's extended context window.

5.2 Tasks

Following LiBERTa, we evaluate Modern-LiBERTa on a set of Ukrainian NLU benchmarks. These include named entity recognition (NER), part-of-speech (POS) tagging, and text classification, enabling us to assess the model's ability to extract and generalize linguistic information.

- **NER-UK** and **NER-UK 2.0** (Chaplynskyi and Romanyshyn, 2024): Annotated corpora of Ukrainian named entities. NER-UK 2.0 includes additional entity types and more comprehensive annotations.

- **WikiANN** (Pan et al., 2017, Rahimi et al., 2019): A multilingual NER dataset, where examples are short and often require factual or encyclopedic knowledge.

- **UD POS Tagging** (Nivre et al., 2017): Based on the Universal Dependencies corpus, this task involves predicting POS tags for each token.

- **Ukrainian News Classification** (Panchenko et al., 2022): A news agency classification benchmark with class imbalance.

5.3 Results

We follow the same evaluation protocol as in WECHSEL-RoBERTa (Minixhofer et al., 2022) and LiBERTa, where each experiment is repeated 5 times with different random seeds, and both the average and standard deviation of the results are reported. This allows for a direct comparison of our metrics with those published for LiBERTa. The results for LiBERTa v2 are taken from the official conference presentation[14].

Modern-LiBERTa demonstrates competitive performance compared to current state-of-the-art models, such as WECHSEL-RoBERTa and LiBERTa v2, across most NLU tasks, as shown in Table 4. The most notable difference is on NER-UK 2.0, where Modern-LiBERTa underperforms the best model by over one percentage point.

On NER-UK, Modern-LiBERTa performs slightly worse than LiBERTa v2 in terms of absolute score, but shows much more consistent results across seeds. A similar pattern is observed on the Ukrainian News Classification task: while its performance is slightly behind WECHSEL-RoBERTa, it significantly outperforms LiBERTa v2. On the Universal Dependencies POS tagging benchmark, Modern-LiBERTa delivers nearly identical results to LiBERTa v2, with only a 0.01 percentage point difference.

On WikiANN, Modern-LiBERTa achieves the best results among all models, which may highlight the benefit of including English Wikipedia data during pretraining. Since WikiANN consists of very short, knowledge-dependent examples, where entity types often cannot be inferred from the local context, this improvement suggests that Modern-LiBERTa is effectively leveraging background knowledge acquired during pretraining.

It is important to note that most of these tasks involve short input sequences and, therefore, do not take advantage of Modern-LiBERTa's extended context window of up to 8,192 tokens.

As reported in the original ModernBERT paper, the base model did not achieve superior results on the GLUE benchmark (Wang et al., 2018) for NLU tasks either, but it showed strong performance on BEIR (Thakur et al., 2021), an information retrieval

[13] https://dumps.wikimedia.org/

[14] https://youtu.be/5qHkCZJNxJ0

Model	UD		Spivavtor		UA-GEC		Wikipedia	
	ppl ↓	*acc* ↑	*ppl* ↓	*acc* ↑	*ppl* ↓	*acc* ↑	*ppl* ↓	*acc* ↑
LiBERTa v2	15.51	52.81%	54.07	37.00%	76.00	33.77%	8.77	59.87%
Modern-LiBERTa	**8.96**	**58.82%**	**18.01**	**48.42%**	**22.22**	**44.71%**	**4.28**	**69.03%**

Table 3: MLM perplexity and token-level accuracy on selected high-quality Ukrainian datasets. Lower perplexity and higher accuracy indicate better modeling performance.

Model	NER-UK	NER-UK 2.0	WikiANN	UD POS	News
	micro-f1	*micro-f1*	*micro-f1*	*acc*	*macro-f1*
Large Models					
XLM-R	90.16 (2.98)†	–	92.92 (0.19)†	98.71 (0.04)†	95.13 (0.49)
WECHSEL-RoBERTa	91.24 (1.16)†	**85.72 (0.43)**	93.22 (0.17)†	98.74 (0.06)†	**96.48 (0.09)**
LiBERTa	91.27 (1.22)‡	–	92.50 (0.07)‡	98.62 (0.08)‡	95.44 (0.04)‡
LiBERTa-V2	**91.73 (1.81)**‡	85.47 (0.24)	93.22 (0.14)‡	**98.79 (0.06)**‡	95.67 (0.12)‡
Modern-LiBERTa	91.66 (0.57)	84.17 (0.18)	**93.37 (0.16)**	**98.78 (0.07)**	96.37 (0.07)

Table 4: Performance on NLU benchmarks for Ukrainian. Scores are averaged across 5 runs. Values in parentheses indicate standard deviation. $\cdot^{\dagger}$ indicates numbers provided by Minixhofer et al. (2022), $\cdot^{\ddagger}$ – by Haltiuk and Smywiński-Pohl (2024).

benchmark. Unfortunately, to the best of our knowledge, there is currently no comparable information retrieval dataset available for Ukrainian, which prevents us from evaluating Modern-LiBERTa's performance on this task. We believe that developing such a benchmark for Ukrainian, similar to efforts made for other languages (Poświata et al. (2024), Al Jallad and Ghneim (2023)), would have a significant impact on the progress of research on text embedding models for low-resource languages.

6 Conclusion

In this paper, we introduced Kobza, the largest publicly available Ukrainian text corpus, containing nearly 60 billion tokens collected from diverse, high-quality sources. Using this dataset, we trained Modern-LiBERTa, the first Ukrainian language model capable of processing long input sequences of up to 8,192 tokens. Our evaluation demonstrates that Modern-LiBERTa achieves competitive results on Ukrainian NLP benchmarks, especially benefiting tasks that rely on background knowledge.

We consider this work an important step toward elevating Ukrainian from its current status as an underrepresented language in multilingual models to a high-resource language. By releasing Kobza and Modern-LiBERTa, we aim to facilitate further advancements in Ukrainian NLP research and development. We encourage future multilingual modeling efforts to incorporate more Ukrainian data to enhance model performance and support richer linguistic diversity in NLP technologies.

Limitations

Despite careful selection and preprocessing, the Kobza corpus may contain content that is suboptimal for language modeling. The included datasets often reflect the biases of web-based sources, such as overrepresentation of sensationalist news, underrepresentation of marginalized voices, and an imbalance across genres and registers. Some documents may include misinformation, spam-like content, or machine-translated text, which can introduce noise or harmful patterns into trained models. These issues are particularly pronounced in multilingual corpora not specifically curated for Ukrainian, where language identification or filtering heuristics may fail.

Additionally, the lack of a dedicated quality scoring system for Ukrainian limits our ability to automatically filter out low-value or inappropriate content. As a result, the corpus may exhibit stylistic monotony, topical skew, or socio-linguistic gaps that affect downstream model robustness. Addressing these limitations requires future work on more principled corpus construction methods, with explicit attention to linguistic diversity, quality assurance, and social considerations.

These underlying biases and quality issues in the Kobza corpus may also be reflected in the models trained on it, including Modern-LiBERTa. As with many large-scale pretrained models, Modern-

LiBERTa may inherit stylistic, topical, or sociolinguistic imbalances present in the data, potentially affecting its fairness, generalizability, or performance across different use cases.

Acknowledgments

We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2024/017168.

The research presented in this paper was financed from the funds assigned by Polish Ministry of Science and Higher Education to AGH University of Krakow.

References

Khloud Al Jallad and Nada Ghneim. 2023. ARNLI: Arabic natural language inference entailment and contradiction detection. *Computer Science*, 24(2).

Rajaraman Anand and Ullman Jeffrey David. 2011. *Mining of massive datasets*. Cambridge university press.

Adrien Barbaresi. 2021. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131. Association for Computational Linguistics.

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M Alves, André Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, and 1 others. 2025. Eurobert: Scaling multilingual encoders for european languages. *arXiv preprint arXiv:2503.05500*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Lola Le Breton, Quentin Fournier, Mariam El Mezouar, and Sarath Chandar. 2025. Neobert: A next-generation bert. *arXiv preprint arXiv:2502.19587*.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. *arXiv preprint arXiv:2305.13820*.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Erik Henriksson, and 1 others. 2025. An expanded massive multilingual dataset for high-performance language technologies. *arXiv preprint arXiv:2503.10267*.

Dmytro Chaplynskyi. 2023. Introducing UberText 2.0: A corpus of Modern Ukrainian at scale. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Dmytro Chaplynskyi and Mariana Romanyshyn. 2024. Introducing NER-UK 2.0: A rich corpus of named entities for Ukrainian. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 23–29, Torino, Italia. ELRA and ICCL.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Konstantin Dobler and Gerard de Melo. 2023. FOCUS: Effective embedding initialization for monolingual specialization of multilingual models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13440–13454, Singapore. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 644–648.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. 2024. Data engineering for scaling language models to 128k context. *arXiv preprint arXiv:2402.10171*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, and 1 others. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. 2024. How to train long-context language models (effectively). *arXiv preprint arXiv:2410.02660*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio Cesar Teodoro Mendes, Allison Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero C. Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, S. Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuan-Fang Li. 2023. Textbooks are all you need. *ArXiv*, abs/2306.11644.

Mykola Haltiuk and Aleksander Smywiński-Pohl. 2024. LiBERTa: Advancing Ukrainian language modeling through pre-training from scratch. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 120–128, Torino, Italia. ELRA and ICCL.

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

Joseph Jennings, Mostofa Patwary, Sandeep Subramanian, Shrimai Prabhumoye, Ayush Dattagupta, Vibhu Jawa, Jiwei Liu, Ryan Wolf, Sarah Yurick, and Varun Singh. 2024. Nemo-curator: a toolkit for data curation. `https://github.com/NVIDIA/NeMo-Curator`. If you use this software, please cite it as below.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. GlotLID: Language identification for low-resource languages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Natalia Kotsyba, Bohdan Moskalevskyi, and Mykhailo Romanenko. 2018. Gold standard Universal Dependencies corpus for Ukrainian.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Damania, and Soumith Chintala. 2020. Pytorch distributed: Experiences on accelerating data parallel training. *CoRR*, abs/2006.15704.

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).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. Cite arxiv:1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *ArXiv*, abs/2309.09400.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Pedro Javier Ortiz Su'arez, Laurent Romary, and Benoit Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Pedro Javier Ortiz Su'arez, Benoit 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 f"ur Deutsche Sprache.

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

Dmytro Panchenko, Daniil Maksymenko, Olena Turuta, Mykyta Luzan, Stepan Tytarenko, and Oleksii Turuta. 2022. Ukrainian news corpus as text classification benchmark. In *ICTERI 2021 Workshops*, pages 550–559, Cham. Springer International Publishing.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2024. Fineweb2: A sparkling update with 1000s of languages.

Jacob Portes, Alexander Trott, Sam Havens, Daniel King, Abhinav Venigalla, Moin Nadeem, Nikhil Sardana, Daya Khudia, and Jonathan Frankle. 2023. Mosaicbert: A bidirectional encoder optimized for fast pretraining. *Advances in Neural Information Processing Systems*, 36:3106–3130.

Rafał Poświata, Sławomir Dadas, and Michał Perełkiewicz. 2024. Pl-mteb: Polish massive text embedding benchmark. *arXiv preprint arXiv:2405.10138*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

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

François Remy, Pieter Delobelle, Hayastan Avetisyan, Alfiya Khabibullina, Miryam de Lhoneux, and Thomas Demeester. 2024. Trans-tokenization and cross-lingual vocabulary transfers: Language adaptation of llms for low-resource nlp. *arXiv preprint arXiv:2408.04303*.

Aman Saini, Artem Chernodub, Vipul Raheja, and Vivek Kulkarni. 2024. Spivavtor: An instruction tuned Ukrainian text editing model. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 95–108, Torino, Italia. ELRA and ICCL.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. https://www.cerebras.net/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama.

Vasyl Starko and Andriy Rysin. 2023. Creating a POS gold standard corpus of Modern Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 91–95, Dubrovnik, Croatia. Association for Computational Linguistics.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *CoRR*, abs/2104.09864.

Oleksiy Syvokon, Olena Nahorna, Pavlo Kuchmiichuk, and Nastasiia Osidach. 2023. UA-GEC: Grammatical error correction and fluency corpus for the Ukrainian language. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 96–102, Dubrovnik, Croatia. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

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.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

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.

Mitchell Wortsman, Tim Dettmers, Luke Zettlemoyer, Ari Morcos, Ali Farhadi, and Ludwig Schmidt. 2023. Stable and low-precision training for large-scale vision-language models. *Advances in Neural Information Processing Systems*, 36:10271–10298.