

On the Path to Make Ukrainian a High-Resource Language

Mykola Haltiuk

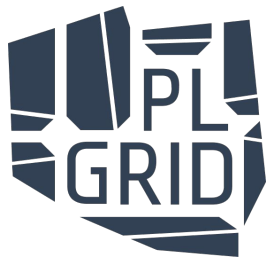
AGH University of Krakow

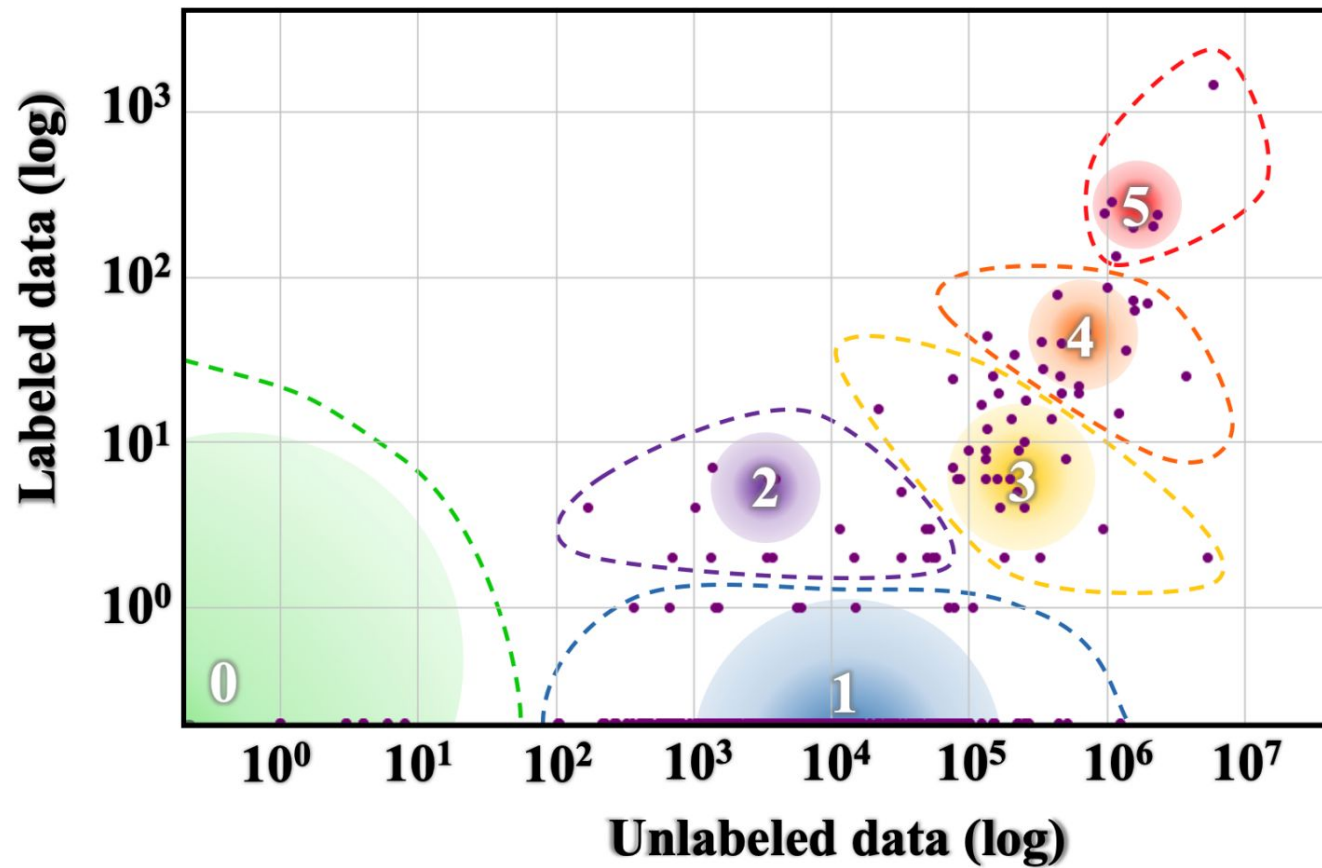
mhaltiuk@agh.edu.pl

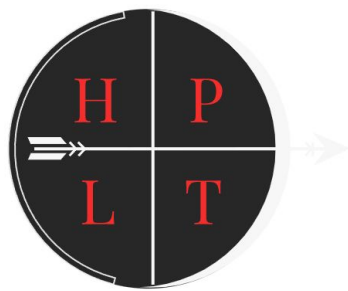
Aleksander Smywiński-Pohl

AGH University of Krakow

apohllo@agh.edu.pl



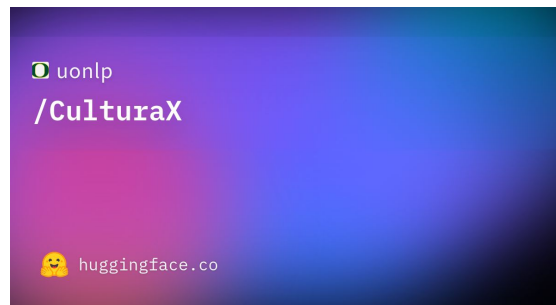




#21



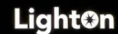
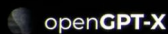
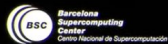
#24



#21

The Rise of Sovereign LLMs

Model Builders Across Europe
Creating Sovereign LLMs with NVIDIA Nemotron

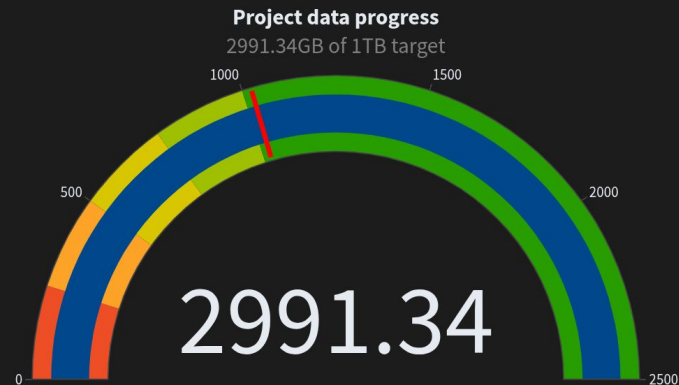
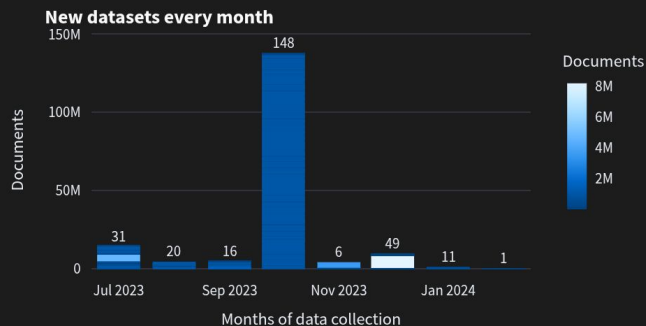




SpeakLeash a.k.a. Spichlerz

Datasets Dashboard

"SpeakLeash is an open collaboration project to build datasets for Language Modeling with a capacity of at least 1TB containing diverse texts in Polish. Our aim is to enable machine learning research and to train a Generative Pre-trained Transformer Model from collected data."

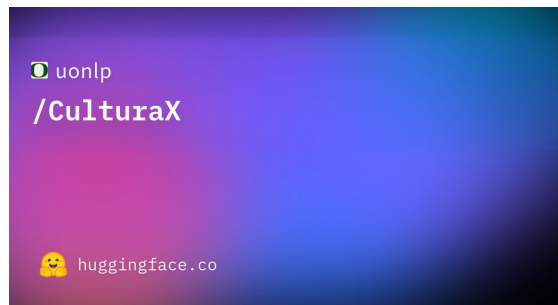
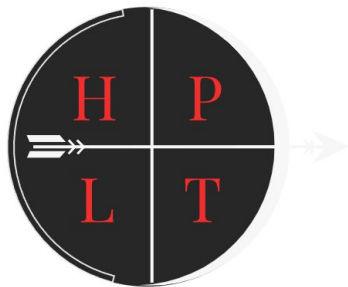


So far we managed to collect:

Total documents	Total characters	Total sentences	Total words	Total verbs	Total nouns	Total punctuations	Total symbols	Total stopwords
967 625 693	3 211 933 665 289	7 703 662 264	123 539 276 429	12 964 991 641	36 933 523 157	25 630 860 999	458 368 467	40 158 005 161

Kobza

Web Sources



Ukrainian News



UberText 2.0

Subcorpora	time span	# of sources	# of texts	# of tokens
<i>news</i>	2000-2023	38	7,208,299	2,172,526,177
<i>fiction</i>	n/a	2	23,796	253,321,894
<i>court</i>	2007-2021	1	111,658	285,252,442
<i>wikipedia</i>	2004-2023	1	2,819,395	499,603,082
<i>social</i>	2018-2022	264	885,314	63,472,353
total	-	-	11,048,462	3,274,175,948
total after filtering	-	-	8,592,389	2,489,454,148

Malyuk?



- **UberText 2.0** ✓
- **Ukrainian News** ✓
- **OSCAR** ✓ (included in CulturaX)

Metadata-Based Deduplication

```
{  
  'source': 'cultura-x',  
  'url': 'http://news.ua/148393',  
  'timestamp': '2017-01-24T17:19:39Z',  
  'text': 'Новини Одеси\п\пОдеський дельфінарій відкривається [...]'  
}
```

```
{  
  'source': 'fineweb-2',  
  'url': 'http://news.ua/148393',  
  'timestamp': '2017-01-24T17:19:39Z',  
  'text': 'Одеський дельфінарій відкривається в [...]'  
}
```

```
{  
  'source': 'hplt-2.0',  
  'url': 'http://odesa-news.ua/odeskyi-delfinariii',  
  'timestamp': '2016-12-14T08:43:12Z',  
  'text': 'Одеський дельфінарій відкривається у [...]'  
}
```

```
{  
  'source': 'fineweb-2',  
  'url': 'http://news.ua/148393',  
  'timestamp': '2017-01-24T17:19:39Z',  
  'text': 'Одеський дельфінарій відкривається в [...]'  
}
```

```
{  
  'source': 'hplt-2.0',  
  'url': 'http://odesa-news.ua/odeskyi-delfinariii',  
  'timestamp': '2016-12-14T08:43:12Z',  
  'text': 'Одеський дельфінарій відкривається у [...]'  
}
```

MinHashLSH Deduplication

```
{  
  'source': 'ukrainian-news',  
  'url': 'https://zaxid.net/domashniy_sorbet',  
  'timestamp': '2025-07-09T1:21:02Z',  
  'text': 'Домашнє морозиво з кавуна – це [...]'  
}
```

```
{  
  'source': 'ukrainian-news',  
  'url': 'https://zaxid.net/domashniy_sorbet',  
  'timestamp': '2025-07-09T1:21:02Z',  
  'text': 'Домашнє морозиво з кавуна – це [...]'  
}
```

```
{  
  'source': 'fineweb-2',  
  'url': 'http://news.ua/148393',  
  'timestamp': '2017-01-24T17:19:39Z',  
  'text': 'Одеський дельфінарій відкривається в [...]'  
}
```

```
{  
  'source': 'hplt-2.0',  
  'url': 'http://odesa-news.ua/odeskyi-delfinariii',  
  'timestamp': '2016-12-14T08:43:12Z',  
  'text': 'Одеський дельфінарій відкривається у [...]'  
}
```

```
{  
  'source': 'hplt-2.0',  
  'url': 'http://odesa-news.ua/odeskyi-delfinariii',  
  'timestamp': '2016-12-14T08:43:12Z',  
  'text': 'Одеський дельфінарій відкривається у [...]'  
}
```

Statistics

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

Table 1: Kobza token statistics

Document Length

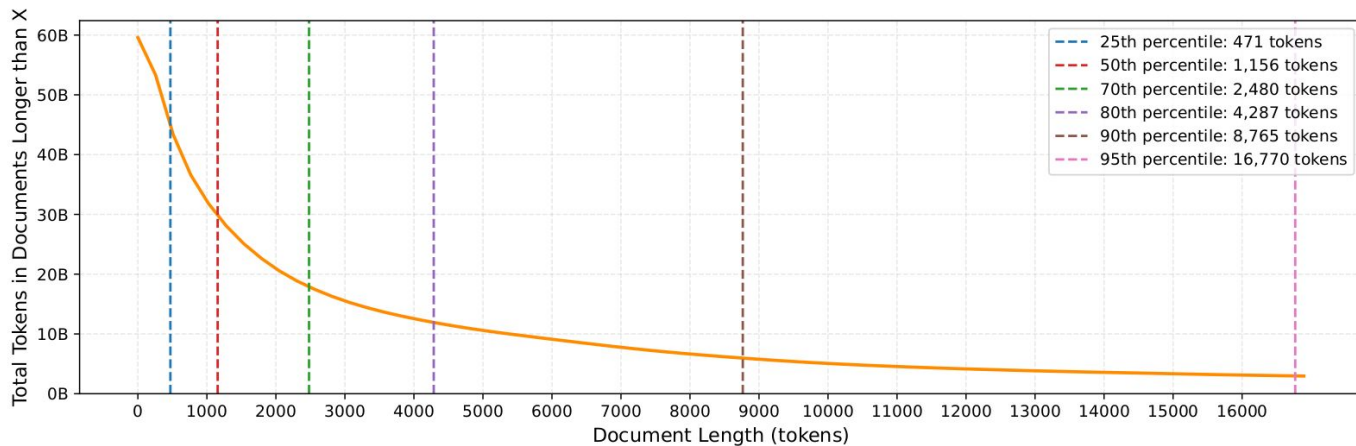


Figure 1: Cumulative token distribution with percentiles marked. y -axis indicates the total number of tokens originating from documents longer than x .

Modern LiBERTa

Pre-training Hyperparameters

	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.

Perplexity

Model	UD		Spivavtor		UA-GEC		Wikipedia	
	<i>ppl</i> ↓	<i>acc</i> ↑	<i>ppl</i> ↓	<i>acc</i> ↑	<i>ppl</i> ↓	<i>acc</i> ↑	<i>ppl</i> ↓	<i>acc</i> ↑
LiBERTa v2	15.51	52.81%	54.07	37.00%	76.00	33.77%	8.77	59.87%
<i>Modern-LiBERTa</i>	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.

Benchmarks

Model	NER-UK <i>micro-f1</i>	NER-UK 2.0 <i>micro-f1</i>	WikiANN <i>micro-f1</i>	UD POS <i>acc</i>	News <i>macro-f1</i>
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) [‡]
<i>Modern-LiBERTa</i>	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. [†] indicates numbers provided by [Minixhofer et al. \(2022\)](#), [‡] – by [Haltiuk and Smywiński-Pohl \(2024\)](#).

Public Release

Goadar/ukr-lm

Master's Thesis on the Ukrainian Language Model



1

Contributor



0

Issues



6

Stars



0

Forks



Datasets:

Goadar

kobza

like 5

Tasks:

Text Generation

Fill-Mask

Modalities:

Text

Formats:

parquet

Languages:

Ukrainian

Size:

10M - 100M

Libraries:

Datasets

Dask

Croissant

+1

License:

cc0-1.0

Dataset card

Data Studio

Files and versions

Community 1

Settings

Dataset Viewer (First 5GB)

Auto-converted to Parquet

API

Embed

Data Studio

Split (1)

train · ~48.6M rows (showing the first 1.03M)

Search this dataset

text	id	url
string · lengths	string · lengths	string · lengths
17	17	0
1.41M	107	6.66k
Наркоторговка з Донецька сприйняла арешт як розіграш до дня народження Жінка пригощала гостей...	hplt-2.0/d8983381ecbcee9ff6d05d5d4b0cea59	http://tsn.ua/tsikavinki/naik-spriinyala-aresht-yak-rozizgra
Магній – це мікроелемент, який називають незамінним для нашого організму, адже він присутній майже в...	hplt-2.0/222b56a3ce7f39335b9760c783e252c9	https://inly.ua/magnesium-ua
Чудовий настрій від теплої погоди пропонуєть зобразити за допомогою пензлів та фарб, передає...	fineweb-2/00baf8b2-612c-4059-b41e-83c211edbe59	https://pr.ua/news/mariupolqs-gerouem-kartin-molodi
Detmore House Detmore House пропонує ідеальне помешкання у Чептенхами. Готель знаходиться на...	hplt-2.0/b4d4fcddb4b1d36a746a82abe6358941	http://detmore-house-bed-brea-cheitenham.co.uk/uk/
Короткий огляд Виробник: Edis (Венгрія) № Реєстраційного посвідчення : UA/6382/01/03 Назва...	hplt-2.0/2e8044278c8ded1035e19b419ee02374	https://1sa.com.ua/emlodin-10
За шопінгом в Пасажі непомітно може пролетіти не тільки речі, але й ніч. Так... в бутику Chloe...	cultura-x/uk_part_00091.parquet/341524	

< Previous

1

2

3

...

10,280

Next >

Downloads last month

575

Use this dataset

Edit dataset card

Size of the auto-converted Parquet files (First 5GB):

2.66 GB

Number of rows (First 5GB):

1,028,000

Estimated number of rows:

48,635,739

Models trained or fine-tuned on Goadar/kobza

Goadar/modern-liberta-large

Fill-Mask · Updated 28 days ago · 18

Kobza

On the Path to Make Ukrainian a High-Resource Language

Kobza is the largest publicly available Ukrainian corpus to date, comprising nearly **60 billion tokens** across **97 million documents**. It is designed to support pretraining and fine-tuning of large language models (LLMs) in Ukrainian, as well as multilingual settings where Ukrainian is underrepresented.

https://huggingface.co/datasets/Goadar/kobza

Goadar/modern-liberta-large like 0

Fill-Mask Transformers PyTorch Goadar/kobza Ukrainian modernbert arxiv:2412.13663 License: cc-by-4.0

Model card Files and versions Community Settings

Train Deploy Use this model

Edit model card

Modern-LiBERTa

Modern-LiBERTa is a ModernBERT encoder model designed specifically for **Ukrainian**, with support for **long contexts up to 8,192 tokens**. It was introduced in the paper On the Path to Make Ukrainian a High-Resource Language presented at the [UNLP @ ACL 2025](#).

The model is pre-trained on **Kobza**, a large-scale Ukrainian corpus of nearly 60 billion tokens. Modern-LiBERTa builds on the [ModernBERT](#) architecture and is the first Ukrainian language model to support long-context encoding efficiently.

The goal of this work is to **make Ukrainian a first-class citizen in multilingual and monolingual NLP**, enabling robust performance on complex tasks that require broader context and knowledge access.

All training code and tokenizer tools are available in the [Goadar/ukr-lm](#) repository.

Evaluation

Downloads last month
18



Inference Providers

Fill-Mask

This model isn't deployed by any Inference Provider.

Ask for provider support

Dataset used to train Goadar/modern-liberta-large

Goadar/kobza

Viewer · Updated 28 days ago · 48.6M · 575 · 5

Collection including Goadar/modern-liberta-large

LiBERTa Collection

3 items · Updated 28 days ago

<https://huggingface.co/Goadar/modern-liberta-large>

Thank you for your attention!

