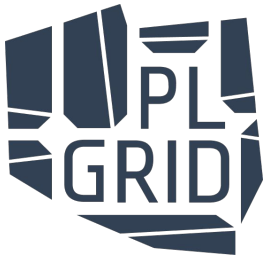
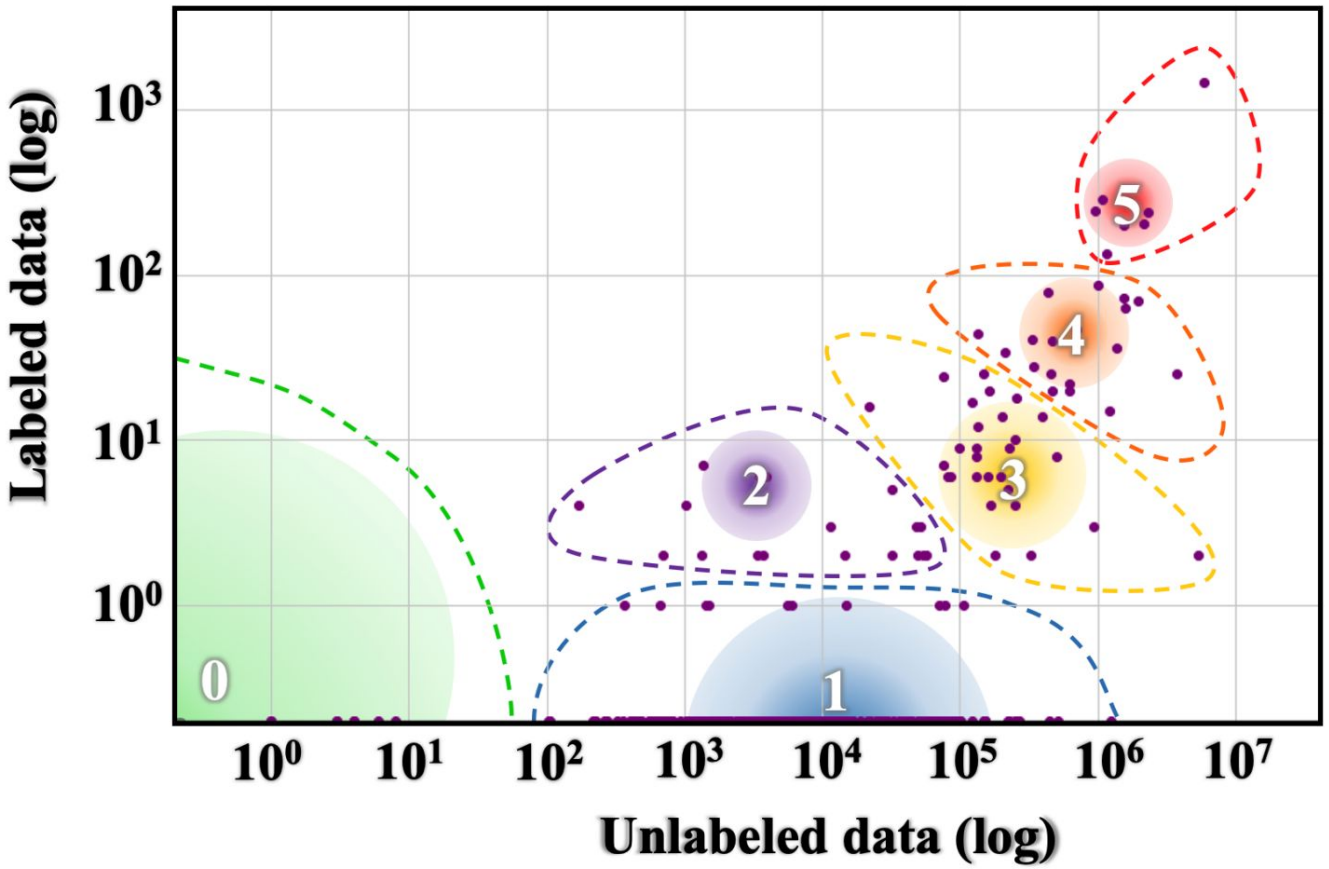
On the Path to Make Ukrainian a High-Resource Language

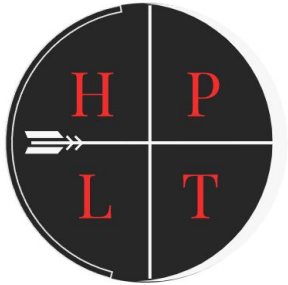
Mykola Haltiuk
AGH University of Krakow
mhaltiuk@agh.edu.pl

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

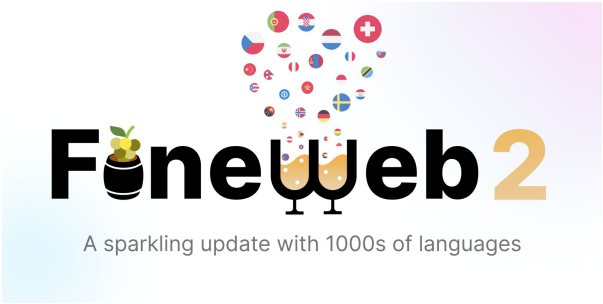




The State and Fate of Linguistic Diversity and Inclusion in the NLP World, Joshi et al. 2020


H P
L T

#21

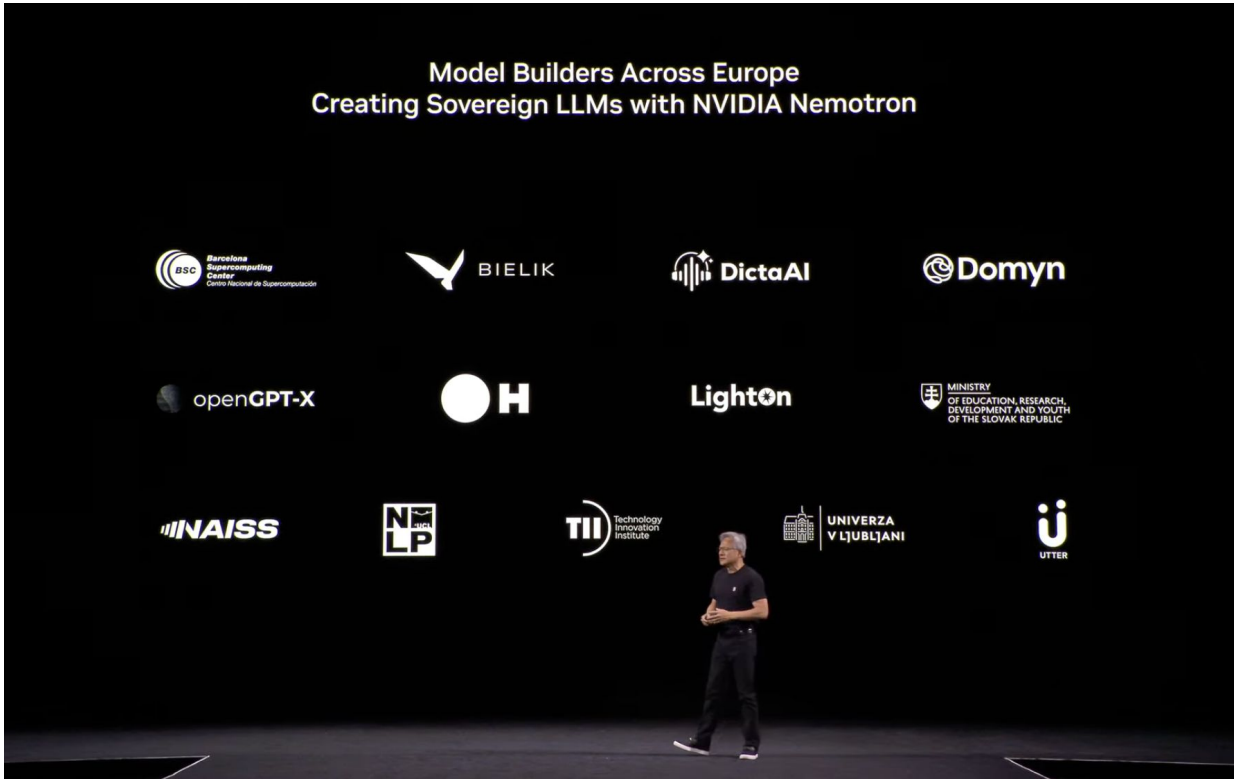

FineWeb 2
A sparkling update with 1000s of languages

#24


uonlp
/CulturaX
huggingface.co

#21

The Rise of Sovereign LLMs

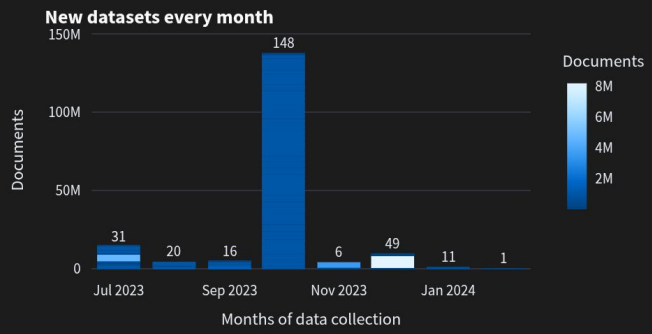


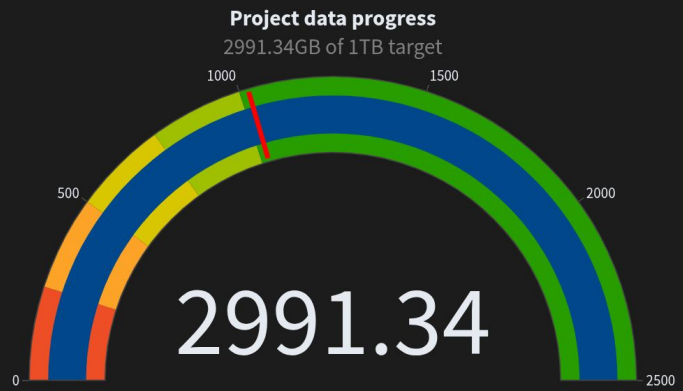
Bielik 7B v0. 1: A Polish Language Model--Development, Insights, and Evaluation, Ociepa et al. 2024

speakleash.org

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."





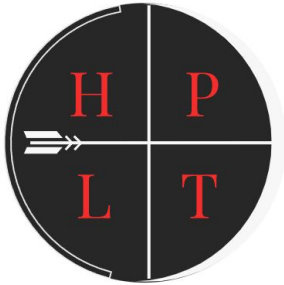
So far we managed to collect:

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

https://speakleash.org/

Kobza

Web Sources


H
P
L
T


FineWeb 2
A sparkling update with 1000s of languages


uonlp
/CulturaX
huggingface.co

Ukrainian News

https://huggingface.co/datasets/zeusfsx/ukrainian-news

UberText 2.0

Subcorpora	time span	# of sources	# of texts	# of tokens
news	2000-2023	38	7,208,299	2,172,526,177
fiction	n/a	2	23,796	253,321,894
court	2007-2021	1	111,658	285,252,442
wikipedia	2004-2023	1	2,819,395	499,603,082
social	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

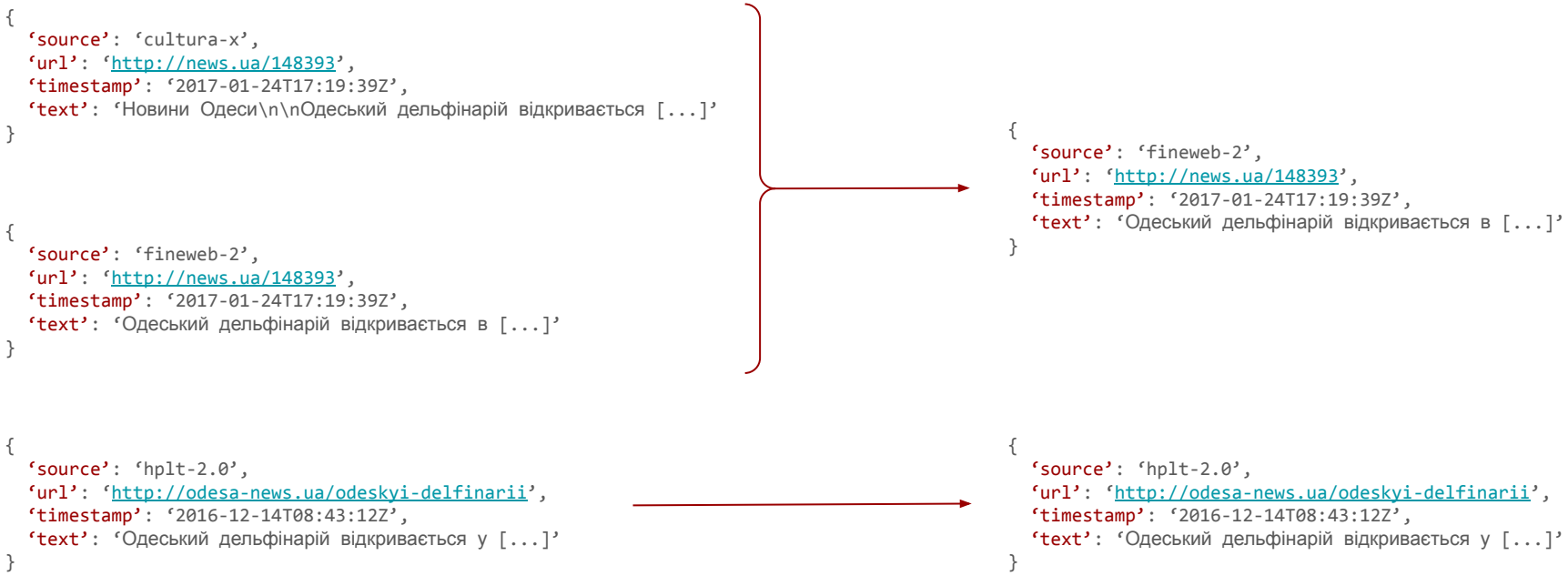
Introducing UberText 2.0: A Corpus of Modern Ukrainian at Scale, Chaplynskyi 2023

Malyuk?

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

https://huggingface.co/datasets/lang-uk/malyuk

Metadata-Based Deduplication

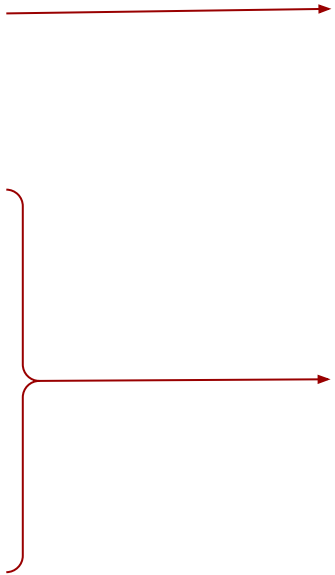


MinHashLSH Deduplication


```
{
  ‘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-delfinarii’,
  ‘timestamp’: ‘2016-12-14T08:43:12Z’,
  ‘text’: ‘Одеський дельфінарій відкривається у [...]’
}
```


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



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

Statistics

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

Document Length

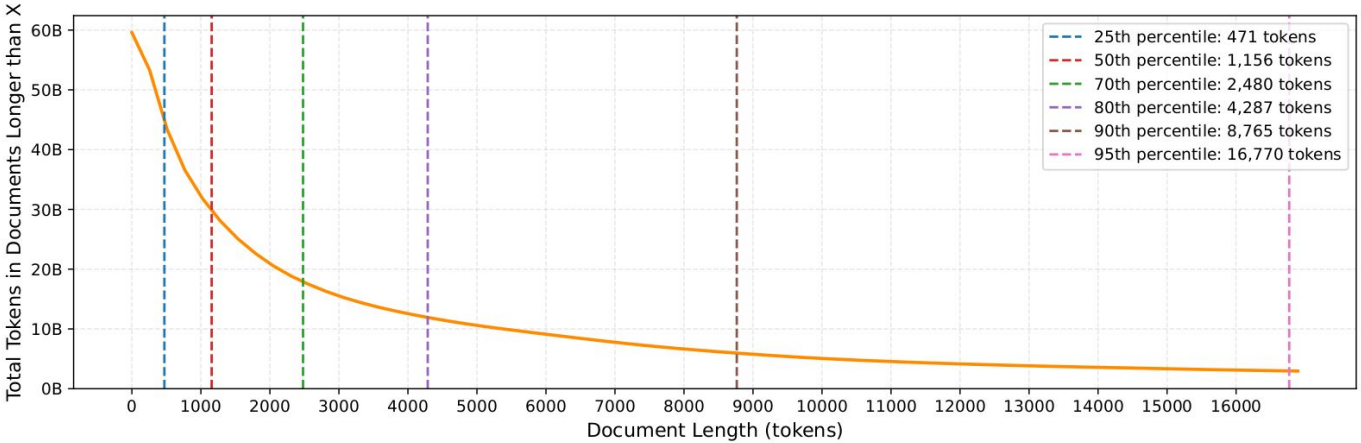


Figure 1: Cumulative token distribution with percentiles marked. y-axis indicates the total number of tokens originating from documents longer than x.

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

	UD		Spivavtor		UA-GEC		Wikipedia	
Model	*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.

Benchmarks

Model	NER-UK *micro-f1*	NER-UK 2.0 *micro-f1*	WikiANN *micro-f1*	UD POS *acc*	News *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).

Public Release

Goader/**ukr-lm**

Master's Thesis on the Ukrainian Language Model

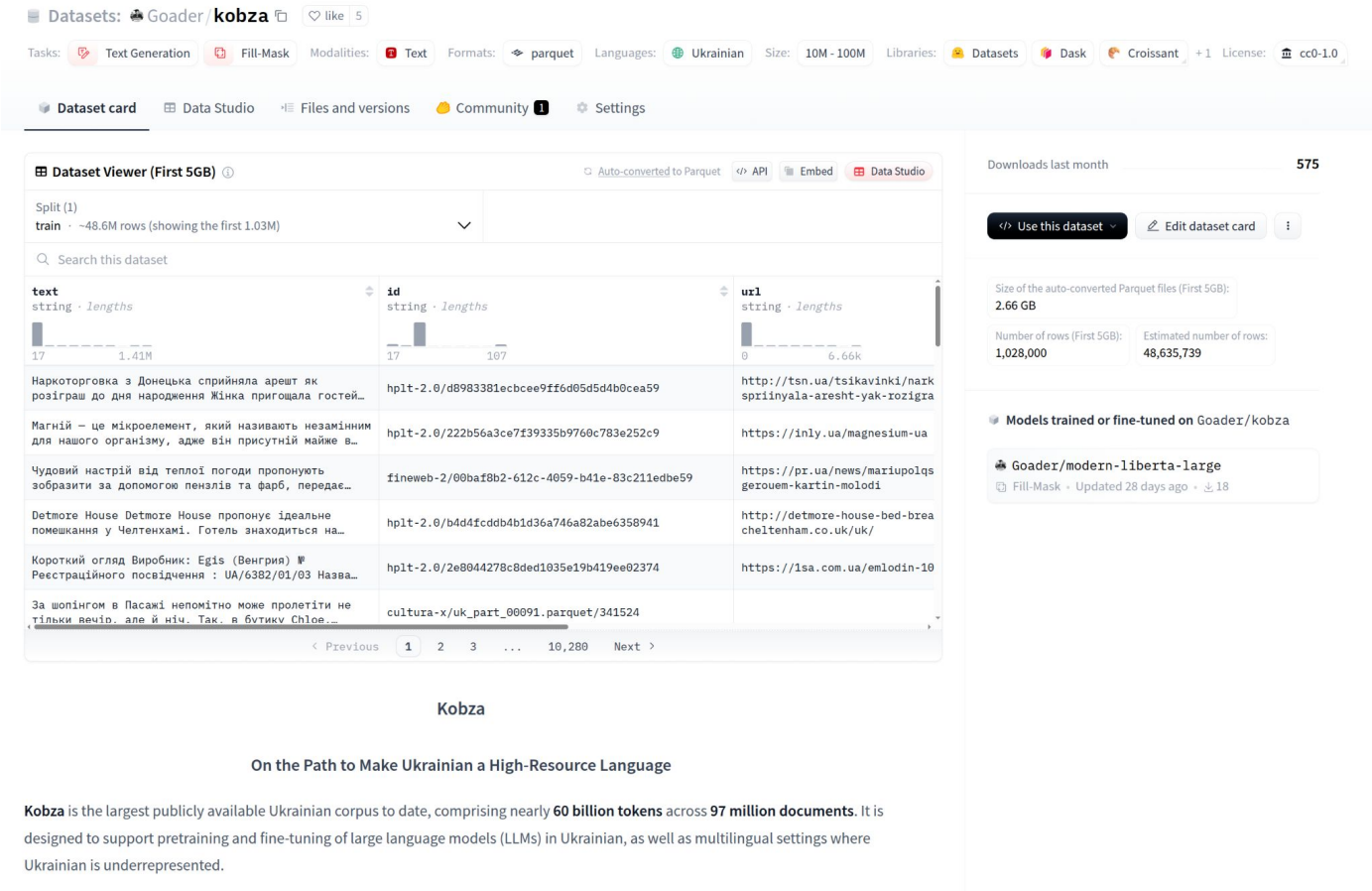
1 Contributor

0 Issues

6 Stars

0 Forks

https://github.com/Goader/ukr-lm

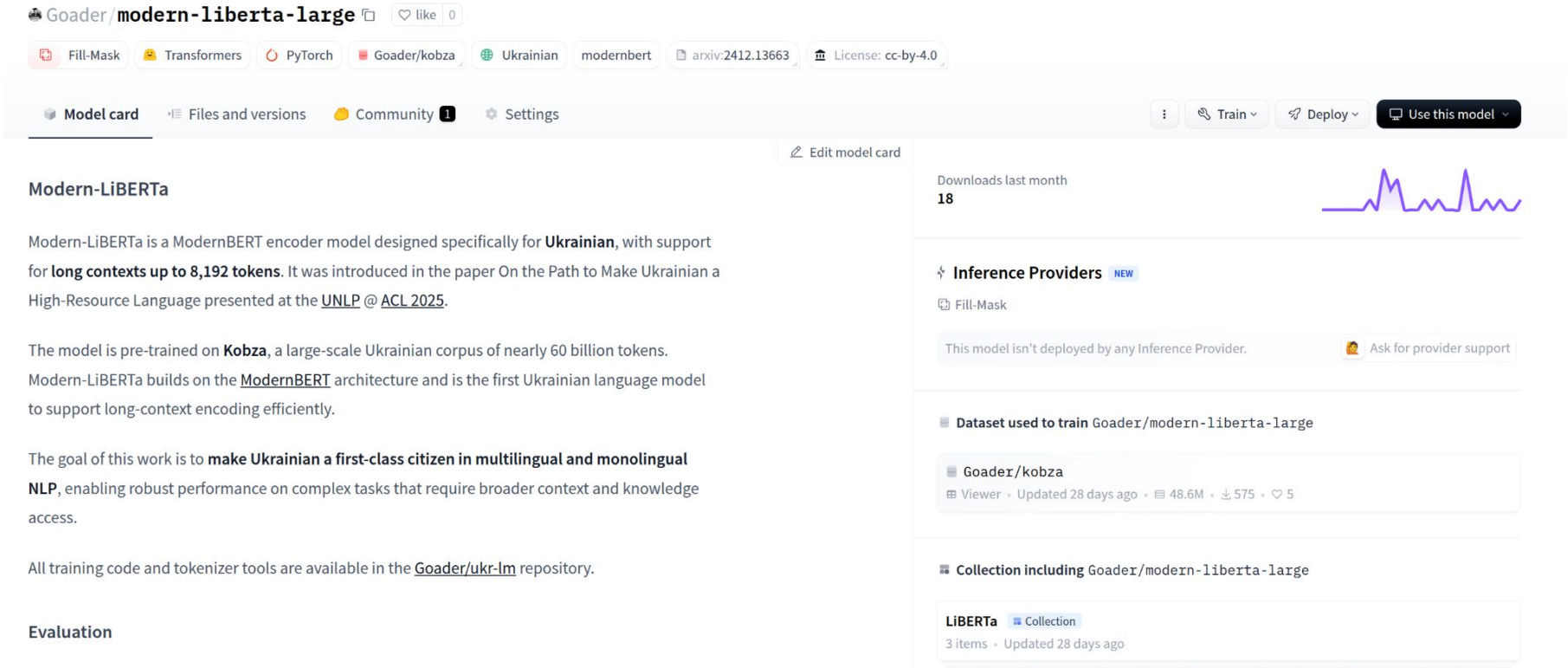


Kobza

On the Path to Make Ukrainian a High-Resource Language

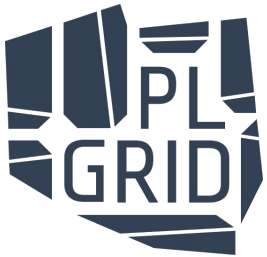
Kobza is the largest publicly available Ukrainian corpus to date, comprising nearly **60 billion tokens** across **97 million documents**. It is designed to support pretraining and fine-tuning of large language models (LLMs) in Ukrainian, as well as multilingual settings where Ukrainian is underrepresented.

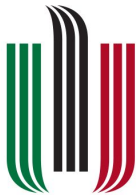
https://huggingface.co/datasets/Goader/kobza



https://huggingface.co/Goader/modern-liberta-large

Thank you for your attention!





AGH University of Krakow