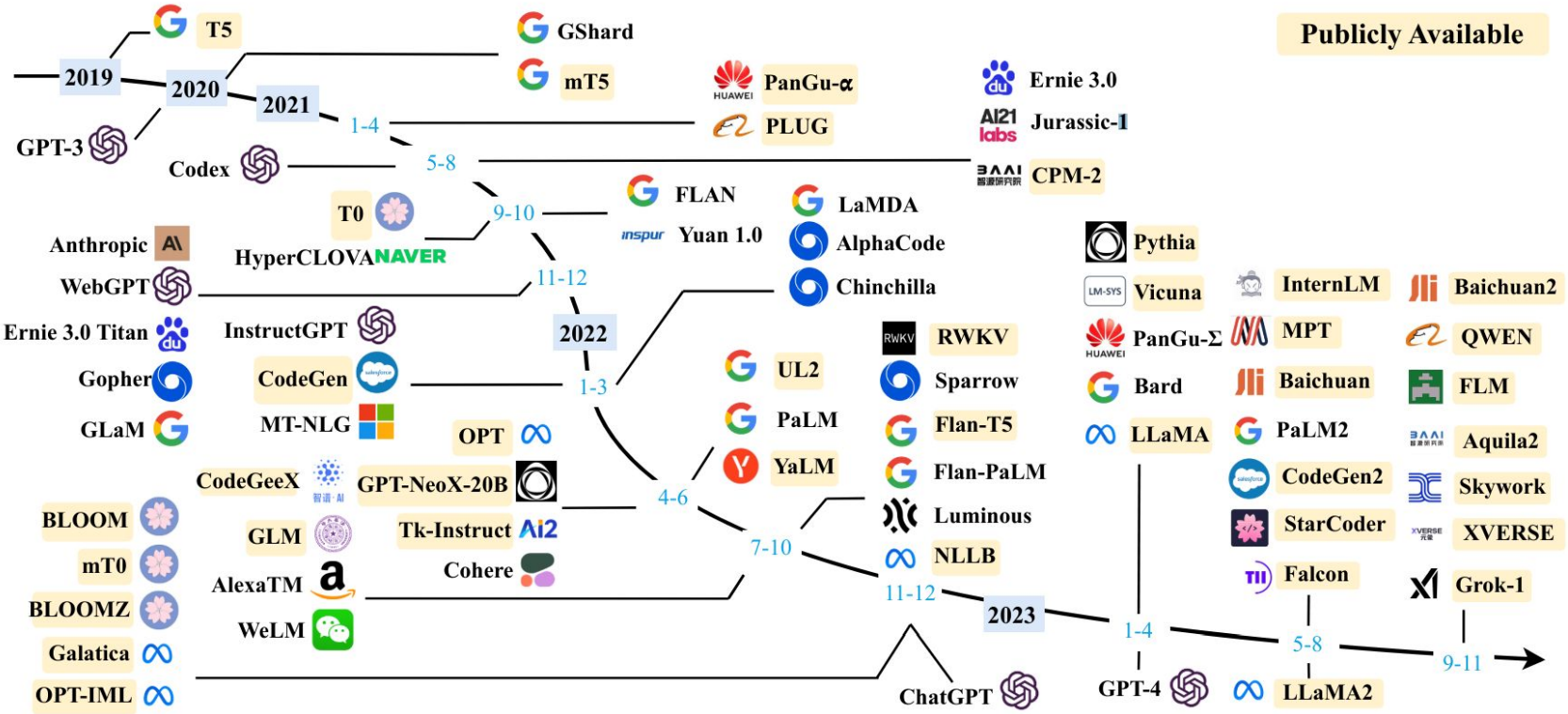
LiBERTa: Advancing Ukrainian Language Modeling through Pre-training from Scratch

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

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



A Survey of Large Language Models, Zhao et al. 2023

Daniel Han 
@danielhanchen

Maybe I found another Gemma bug? Can anyone from the #Gemma team confirm if Gemma uses approx gelu or exact gelu?
Keras=approx
Gemma_pytorch=exact
HF=exact
When comparing Keras to HF, torch.dist gets 4.7943, while tanh approx gets 0.0057 Maybe @suryabhupa? :)
pic.twitter.com/VJStwUzMih

Keras

```
x1 = self.gating_ffw(normalized_x)
x2 = self.gating_ffw_2(normalized_x)
x = keras.activations.gelu(x1, approximate=True) * x2
x = self.ffw_linear(x)
```

gemma_pytorch

```
def forward(self, x):
    gate = self.gate_proj(x)
    gate = F.gelu(gate)
    up = self.up_proj(x)
    fuse = gate * up
    outputs = self.down_proj(fuse)
    return outputs
```

exact

HF

```
def __init__(self, use_gelu_python: bool = False):
    super().__init__()
    if use_gelu_python:
        self.act = self._gelu_python
    else:
        self.act = nn.functional.gelu
```

exact

Daniel Han [@danielhanchen]. (2024, March 1). Twitter. https://twitter.com/danielhanchen/status/1763613620909580505

WECHSEL RoBERTa

	lang-uk NER (Micro F1)	WikiANN (Micro F1)	UD Ukrainian IU POS (Accuracy)
roberta-base-wechsel-ukrainian	90.81 (1.51)	92.98 (0.12)	98.57 (0.03)
roberta-large-wechsel-ukrainian	**91.24 (1.16)**	**93.22 (0.17)**	**98.74 (0.06)**
roberta-base-scratch-ukrainian*	89.57 (1.01)	92.05 (0.09)	98.31 (0.08)
roberta-large-scratch-ukrainian*	89.96 (0.89)	92.49 (0.15)	98.52 (0.04)
dbmdz/electra-base-ukrainian-cased-discriminator	90.43 (1.29)	92.99 (0.11)	98.59 (0.06)
xlm-roberta-base	90.86 (0.81)	92.27 (0.09)	98.45 (0.07)
xlm-roberta-large	90.16 (2.98)	92.92 (0.19)	98.71 (0.04)

Benjamin Minixhofer, benjamin/roberta-large-wechsel-ukrainian (2022), HuggingFace Hub

mC4

ασφαλιστικό πώς παίρνετε σύνταξη με λιγότερα ένσημα αρχειοθήκη ιστολογίου

головнаглавная страницамеблікорпусні меблітумби

підм сам озн н. в ж р одн підм себе звор р. в дод їм ос д в множ дод собі зв м. в дод якийсь неознач ч р одн зн в означ ній особ м. в ж р одн обст

гдз решебник по алгебре 7 класс мерзляк а.г. полонський в.б. рабінович ю.м. якір м.с. варіант 2 задание 160 2016 2017 онлайн

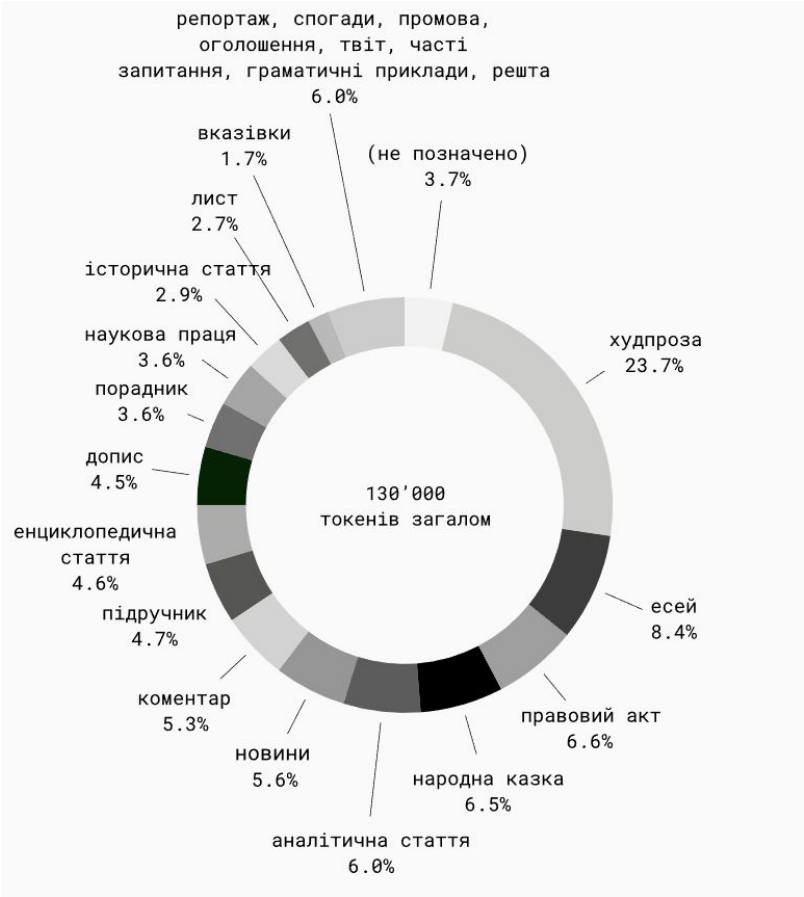
Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, Raffel et al. 2019

CulturaX

Code	Language	#Documents (M)					Filtering Rate (%)	#Tokens	
		Initial	URL Filtering	Metric Filtering	MinHash Dedup	URL Dedup		(B)	(%)
en	English	5783.24	5766.08	3586.85	3308.30	3241.07	43.96	2846.97	45.13
ru	Russian	1431.35	1429.05	922.34	845.64	799.31	44.16	737.20	11.69
es	Spanish	844.48	842.75	530.01	479.65	450.94	46.60	373.85	5.93
de	German	863.18	861.46	515.83	447.06	420.02	51.34	357.03	5.66
fr	French	711.64	709.48	439.69	387.37	363.75	48.89	319.33	5.06
				...					
ro	Romanian	89.37	89.25	45.99	42.8	40.33	54.87	39.65	0.63
sv	Swedish	103.04	102.76	58.67	52.09	49.71	51.76	38.49	0.61
uk	Ukrainian	81.50	81.44	50.95	47.12	44.74	45.10	38.23	0.61
fi	Finnish	59.85	59.80	36.69	32.15	30.47	49.09	28.93	0.46
ko	Korean	46.09	45.85	25.19	21.17	20.56	55.39	24.77	0.39
				...					

CulturaX: A Cleaned, Enormous, and Multilingual Dataset for Large Language Models in 167 Languages, Nguyen et al. 2023

Gold standard Universal Dependencies corpus for Ukrainian



Gold standard Universal Dependencies corpus for Ukrainian, Kotsyba et al. 2018

Tokenizer

Tokenizer	Size	Avg.	Hits
XLM-RoBERTa	250K	1.739	54.46%
Ukr-RoBERTa	52K	1.846	42.16%
WECHSEL-RoBERTa	50K	1.866	40.89%
Ukr-ELECTRA	32K	1.443	69.89%
LiBERTa	32K	**1.442**	**70.02%**

Pre-training Hyperparameters

Hyperparameter	Value
Peak Learning Rate	2e-4
Warm-up Steps	5K
Learning Rate Decay	Cosine
Effective Batch Size	1024
Batch Size per GPU	32
Gradient Accumulation Steps	4
Max Steps	85K
Weight Decay	0.01
Adam ϵ	1e-8
Adam β_1	0.9
Adam β_2	0.999
Gradient Clipping	1.0
Gradient Clipping Algorithm	L2

Evaluation Results

Model	NER-UK *micro-f1*	WikiANN *micro-f1*	UD POS *acc*	News *macro-f1*
Base Models				
XLM-R	90.86 (0.81)†	92.27 (0.09)†	98.45 (0.07)†	–
WECHSEL-RoBERTa	90.81 (1.51)†	92.98 (0.12)†	98.57 (0.03)†	–
Ukr-ELECTRA	90.43 (1.29)†	92.99 (0.11)†	98.59 (0.06)†	–
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)†	**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)

Fine-tuning Hyperparameters

Hyperparameter	Value
Peak Learning Rate	3e-5
Warm-up Ratio	0.05
Learning Rate Decay	Linear
Batch Size	16
Epochs	10
Weight Decay	0.05

What’s now?

LiBERTa V2

LiBERTa-V2 Evaluation

Model	**NER-UK** *micro-f1*	**WikiANN** *micro-f1*	**UD POS** *acc*	**News** *macro-f1*
Base Models				
XLM-R	90.86 (0.81)†	92.27 (0.09)†	98.45 (0.07)†	–
WECHSEL-RoBERTa	90.81 (1.51)†	92.98 (0.12)†	98.57 (0.03)†	–
Ukr-ELECTRA	90.43 (1.29)†	92.99 (0.11)†	98.59 (0.06)†	–
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)†	**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)**	**93.22 (0.14)**	**98.79 (0.06)**	95.67 (0.12)

Public Release

Goader/**ukr-lm**

Master's Thesis on the Ukrainian Language Model

1 Contributor

0 Issues

0 Stars

0 Forks

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

Thank you for your attention!