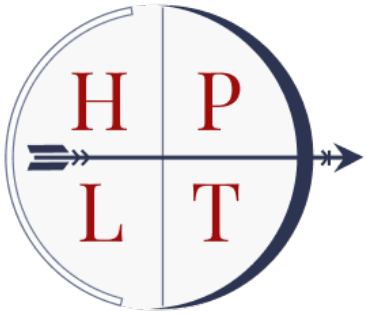
Co-funded by the European Union



HPLT: High Performance Language Technologies

First language models trained

Deliverable number: 4.1

Version 1.0

Horizon Europe

UK Research and Innovation

Funded by the European Union's Horizon Europe search and innovation programme under grant agreement No 101070350 and from UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10052546] programme

Project details

Project Acronym:	HPLT
Project Full Title:	HPLT: High Performance Language Technologies
Year of the Call:	2021
Type of Action:	HORIZON-IA (Innovation Action)
Grant Number:	101070350
Project URL:	https://hplt-project.org

Report details

First language models trained	
Lead author:	Sampo Pyysalo (UTURKU)
Contributing authors:	Risto Luukkonen (UTURKU) Andrey Kutuzov (UOSLO) David Samuel (UOSLO)
Internal reviewers:	Jörg Tiedemann (UHELS) Jaume Zaragoza (PROMPSIT)
Deliverable number:	4.1
Dissemination level:	Public (PU)
Contractual Delivery Date:	February 29, 2024
Actual Delivery Date:	February 29, 2024
Number of pages:	20

Document history

Version	Date	Changes
1.0	Feb 29, 2024	Original Submission

Abstract

This report provides a description of deliverable D4.1 – the initial release of language models created in the HPLT project. The release includes encoder-only models for 75 languages and and decoder-only models for languages selected for inclusion in initial model training. Evaluation of the models demonstrates advances over previously existing models for several languages and serves also to validate the quality of the HPLT data.

Contents

1 Executive summary

This deliverable, *First language models trained*, reports on the monolingual and multilingual language models trained during the first 18 months of the HPLT project. Translation models are described separately in deliverable D5.1 (*Translation models for select language pairs*). All models described herein build on the Transformer architecture (Vaswani et al., 2017), and the document is organized by the model class: encoder-only (BERT-like) models are presented in Section 3, and decoder-only (GPT-like) models in Section 4.

The EU-funded HPLT project applies high-performance computing to scale and advance language technologies. Taking advantage of recent advances in machine learning and astonishing storage capacities, it will create and process huge language data sets and produce language and translation models in a large number of languages. The resulting models will be tested from various angles to ensure smooth integration, high accuracy, and regulatory compliance concerning privacy, unwanted biases and ethical issues. The models and data sets will be a game changer in the language service market in the EU and beyond. The resulting models will be open, free and available from established language repositories for anyone interested in pursuing research or innovation projects.

The project, coordinated by the Charles University in Prague (CUNI), gathers partners from 5 different universities 2 HPC centers and a private NLP company from all around Europe.

CHARLES UNIVERSITY

UNIVERSITY OF OSLO

UNIVERSITY OF EDINBURGH

UNIVERSITY OF TURKU

UNIVERSITY OF HELSINKI

PROMPSIT

CESNET

SIGMA2

2 Introduction

The goals of this initial release of HPLT language models include completing and demonstrating the successful application of tools to create Transformer models, assess the quality of the first HPLT data release (Tiedeman et al., 2023) and the data cleaning processes (Ramírez-Sánchez, 2024), and provide open state-of-the-art models created using a replicable, transparently documented process.

This first release includes encoder-only and decoder-only models, the architectures and configurations of which follow current community practices, with the encoder-only models having approximately 100 million parameters, while the larger decoder-only models have more than 10 billion parameters. Due to the differences in scale, computational requirements, and evaluation practices, the models are trained using separate toolchains and evaluated using different protocols, described in Sections 3 (encoder-only models) and 4 (decoder-only models).

All models were trained on the LUMI supercomputer located in Finland (https://lumi-supercomputer.eu/). LUMI is currently the fastest supercomputer in Europe as well as one of the most greenest data centers in the world, with its energy consumption covered 100% with renewable electricity. The LUMI GPU partition used to create the models described herein consists of nearly 3000 nodes with four AMD MI250X GPUs each, with each of these devices drawing 500W.

The models are released through Hugging Face on the HPLT repository (https://huggingface.co/HPLT) and the repositories of project partners (https://huggingface.co/TurkuNLP, https://huggingface.co/norallm) and additionally made available via the HPLT website https://hplt-project.org/models .

3 Encoder-only models

In this section, we describe HPLT language models trained following the encoder-only Transformer architecture (Vaswani et al., 2017). Those are so called masked language models. In particular, we used the modification of the classic BERT model (Devlin et al., 2019) named LTG-BERT (Samuel et al., 2023). LTG-BERT is fully open, developed by the Language Technology Group at the University of Oslo (a member of the HPLT consortium) and is different from the original BERT in not a using next sentence prediction objective, swapping subword masking to span masking, and other minor architectural improvements.

We trained a monolingual LTG-BERT model for every major language in the HPLT 1.2 data release (**75** models total). All the models are openly published on the HuggingFace Hub[1] and on the HPLT project website.[2] The training code is published on the HPLT GitHub repository.[3]

3.1 Training data

The tables 3.1 and 3.2 report the sizes of the training corpora for every language we trained a BERT model for.

[1] https://huggingface.co/HPLT
[2] https://hplt-project.org/deliverables
[3] https://github.com/hplt-project/HPLT-WP4

Language	MBytes	Words (millions)	Documents (thousands)
Afrikaans (af)	5,486	829	747
Arabic (ar)	322,934	31,848	26,800
Azerbaijani (az)	9,805	1,128	1,097
Belarusian (be)	4,925	394	356
Bulgarian (bg)	95,868	8,761	6,497
Bangla (bn)	44,056	2,766	2,875
Catalan (ca)	36,528	5,755	4,543
Czech (cs)	138,962	19,109	16,987
Welsh (cy)	781	124	111
Danish (da)	61,501	9,369	8,175
German (de)	816,024	110,983	101,414
Greek (el)	338,891	33,764	15,833
English (en)	16,492,853	2,314,978	1,021,383
Esperanto (eo)	675	101	67
Spanish (es)	1,157,696	181,226	129,291
Estonian (et)	13,338	1,740	1,475
Basque (eu)	2,509	324	343
Persian (fa)	438,144	47,581	30,897
Finnish (fi)	75,400	9,038	7,147
French (fr)	810,940	122,875	99,587
Irish (ga)	863	130	115
Galician (gl)	5,511	847	731
Gujarati (gu)	4,100	303	264
Serbo-Croatian (hbs)	72,923	10,026	8,680
Hebrew (he)	71,009	7,492	4,979
Hindi (hi)	88,618	7,542	5,774
Hungarian (hu)	114,968	14,391	11,708
Armenian (hy)	7,846	589	621
Indonesian (id)	294,223	42,077	31,420
Icelandic (is)	3,989	562	481
Italian (it)	519,502	74,452	53,525
Japanese (ja)	1,749,944	63,231	190,414
Georgian (ka)	10,868	573	533
Kazakh (kk)	6,641	471	406
Kannada (kn)	4,114	235	228
Korean (ko)	249,113	25,522	31,854
Kyrgyz (ky)	1,394	101	88
Latin (la)	2,037	294	301
Lithuanian (lt)	23,538	2,954	2,724
Latvian (lv)	12,597	1,594	1,537
Macedonian (mk)	8,244	736	734
Malayalam (ml)	10,748	517	469
Mongolian (mn)	9,452	803	594
Marathi (mr)	8,173	519	453
Malay (ms)	57,945	9,031	4,872
Maltese (mt)	790	102	111
Burmese (my)	8,639	357	239

Table 3.1: Data set sizes (uncompressed bytes and word tokens) for the HPLT monolingual release, part 1.

Language	MBytes	Words (millions)	Documents (thousands)
Norwegian Bokmål (nb)	53,632	8,301	6,115
Nepali (ne)	11,259	694	863
Dutch (nl)	219,649	33,301	31,745
Norwegian Nynorsk (nn)	1,927	298	228
Punjabi (pa)	2,194	184	152
Polish (pl)	327,648	44,170	39,378
Pashto (ps)	955	113	88
Portuguese (pt)	527,018	81,410	58,244
Romanian (ro)	129,709	19,491	14,468
Russian (ru)	3,554,44	284,583	224,196
Sinhala (si)	8,063	568	322
Slovak (sk)	36,473	4,982	4,617
Slovenian (sl)	17,091	2,512	2,196
Somali (so)	1,440	211	283
Albanian (sq)	8,834	1,341	1,241
Swedish (sv)	111,723	16,913	13,668
Swahili (sw)	4,471	668	698
Tamil (ta)	36,732	1,914	1,243
Telugu (te)	7,320	437	415
Thai (th)	110,351	4,332	8,192
Filipino (tl)	5,587	911	585
Turkish (tr)	320,747	42,650	27,051
Tatar (tt)	962	74	65
Ukrainian (uk)	136,393	10,574	9,306
Urdu (ur)	11,576	1,422	1,437
Uzbek (uz)	4,149	367	290
Vietnamese (vi)	299,851	49,362	31,504
Chinese (zh)	12,551,125	432,881	1,080,799
All	58,800,000	4,212,845	3,388,904

Table 3.2: Data set sizes (uncompressed bytes and word tokens) for the HPLT monolingual release, part 2.

3.2 Model description

All the HPLT encoder-only models share the same set of hyper-parameters, roughly following the BERT-base setup:

- hidden size: 768
- attention heads: 12
- layers: 12
- vocabulary size: 32768

Every model uses its own tokenizer trained on language-specific HPLT data. Note that we followed the prior HPLT decisions and trained one 'hbs' model on Serbian, Bosnian and Croatian data, due to these languages being particularly close to each other. When evaluating on Croatian ('hr') and Serbian ('sr') benchmark datasets in Section 3.3, this 'hbs' model was used.

Following the principles of openness and transparency, we publish not only the final model checkpoints, but also intermediate ones: 10 checkpoints for every model, saved approximately every 3 000 steps. They can be used, e.g., to analyze how large language models are learning from data. We also publish the training statistics of all 75 runs online for full transparency.[4]

3.3 Model evaluation

Since the amount of available data varies a lot from one HPLT language to another, we do not position these models as aiming to achieve state-of-the-art performance across many NLP tasks. Instead, their role is to provide fully open and public lower boundary baselines for as many languages as possible; some of them never had a BERT model trained for them before. Still, we conduct substantial benchmarking efforts. In particular, we compare HPLT models to multilingual mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) models.

As downstream NLP tasks, we use part-of-speech tagging, lemmatization, dependency parsing and named entity recognition. For the first three tasks, we employ Universal Dependencies treebanks (de Marneffe et al., 2021) as the benchmark datasets; for named entity recognition, we employ WikiAnn datasets (Rahimi et al., 2019). Naturally, we were able to evaluate BERT models only on languages for which the corresponding benchmark datasets exist and are of sufficient size.

[4] https://api.wandb.ai/links/ltg/kduj7mjn

Model	af	ar	be	bg	ca	cs	cy	da	de	el	en	es	et	eu	fa	fi	fr	ga
HPLT	96.5	95.1	**95.5**	**97.8**	**97.4**	**98.3**	**89.2**	**97.8**	80.7	**96.1**	96.7	**96.0**	**97.1**	**92.3**	**96.4**	**96.8**	**98.1**	**88.7**
XLM-R	**96.6**	**95.2**	94.6	97.5	97.2	98.0	88.3	**97.8**	**89.4**	95.7	**96.8**	95.9	96.6	91.4	96.3	96.4	**98.1**	87.1
mBERT	95.8	94.3	94.1	97.0	97.1	97.8	87.2	96.7	88.8	94.6	96.1	95.7	96.0	91.0	95.9	95.1	97.8	86.5

Model	gl	he	hi	hr	hu	hy	id	is	it	ja	kk	ko	ky	la	lt	lv	mr	mt
HPLT	**97.1**	**96.5**	**93.6**	**96.4**	93.9	**92.7**	89.6	**88.6**	**98.1**	**97.8**	46.7	**89.9**	52.1	**97.1**	**92.0**	92.4	75.2	**97.0**
XLM-R	**97.1**	96.1	93.3	96.2	**94.4**	91.2	**89.8**	88.1	98.0	97.7	**47.3**	89.7	**53.0**	**97.1**	91.9	**92.8**	**75.7**	94.5
mBERT	96.9	95.6	92.4	95.5	92.8	88.7	89.5	87.7	98.0	97.5	41.6	88.6	49.8	96.9	87.7	91.6	74.0	94.7

Model	nb	nl	nn	pl	pt	ro	ru	sk	sl	sr	sv	ta	te	tr	uk	ur	vi	zh
HPLT	**97.6**	**97.1**	**97.7**	**96.9**	**94.1**	**97.7**	**94.5**	**91.2**	**98.1**	**96.2**	**97.4**	**82.8**	**96.3**	**91.5**	72.9	**80.5**	91.8	96.0
XLM-R	97.4	96.9	97.0	96.5	94.0	97.6	94.4	90.8	97.6	95.7	**97.4**	80.9	95.1	91.0	**94.7**	**80.5**	**92.1**	**96.3**
mBERT	97.0	96.2	96.6	95.6	93.6	97.3	93.8	89.1	96.7	95.2	96.5	79.6	94.7	90.4	93.1	80.3	89.8	96.2

Table 3.3: Results on part-of-speech tagging. We evaluate the `AllTags` performance – the exact match accuracy of the `UPOS`, `XPOS` and `UFeats` tags.

Model	af	ar	be	bg	ca	cs	cy	da	de	el	en	es	et	eu	fa	fi	fr	ga
HPLT	96.8	**95.2**	**93.8**	97.3	**99.4**	**99.4**	93.7	97.1	95.5	94.1	97.9	**99.4**	**95.2**	**96.0**	**99.4**	**91.6**	98.7	**96.1**
XLM-R	**97.9**	94.7	**93.8**	**97.7**	**99.4**	99.3	94.4	**97.6**	**97.7**	**94.7**	**98.0**	**99.4**	95.0	95.9	**99.4**	91.5	**98.8**	95.8
mBERT	97.8	94.5	93.2	97.5	**99.4**	99.3	**94.6**	97.2	97.6	94.6	97.8	**99.4**	94.8	95.7	99.1	90.6	98.6	95.5

Model	gl	he	hi	hr	hu	hy	id	is	it	ja	kk	ko	ky	la	lt	lv	mr	mt
HPLT	98.2	97.1	**99.0**	97.2	93.0	93.9	98.0	**96.5**	**98.8**	**98.3**	**66.4**	**94.4**	71.4	**99.3**	91.5	96.8	**91.7**	**100.0**
XLM-R	**98.3**	**97.2**	**99.0**	**97.4**	**94.3**	**94.9**	**98.3**	96.4	98.7	**98.3**	**66.4**	94.3	**73.8**	**99.3**	**91.6**	**97.5**	90.3	**100.0**
mBERT	**98.3**	97.0	98.9	97.2	93.0	94.4	98.2	96.2	98.6	**98.3**	65.1	94.0	71.0	99.2	90.2	96.9	88.8	**100.0**

Model	nb	nl	nn	pl	pt	ro	ru	sk	sl	sr	sv	ta	te	tr	uk	ur	vi	zh
HPLT	**98.8**	94.4	**98.5**	**98.2**	**98.3**	97.8	**98.6**	95.6	98.6	97.0	97.1	88.6	**100.0**	**91.9**	87.0	97.4	**99.9**	**99.9**
XLM-R	**98.8**	**94.7**	98.4	**98.2**	**98.3**	**97.9**	98.5	**96.1**	**98.7**	**97.3**	**97.6**	**89.7**	**100.0**	91.3	**97.2**	**97.5**	**99.9**	**99.9**
mBERT	98.5	94.1	98.2	97.8	98.1	97.7	98.3	95.7	98.5	97.2	97.3	87.9	**100.0**	91.1	96.9	**97.5**	**99.9**	**99.9**

Table 3.4: Results on lemmatization, the accuracy of predicting the correct lemmatized forms.

Model	af	ar	be	bg	ca	cs	cy	da	de	el	en	es	et	eu	fa	fi	fr	ga
HPLT	87.5	**86.1**	**91.1**	94.0	**94.4**	**94.4**	82.3	88.8	76.4	92.2	92.2	**93.1**	**90.8**	**88.1**	**93.9**	**93.3**	**94.5**	**83.4**
XLM-R	**88.0**	85.7	89.9	**94.4**	94.1	94.2	**82.8**	**89.1**	**87.1**	**93.5**	**92.6**	93.0	89.7	87.3	93.8	93.0	94.4	82.7
mBERT	87.2	84.7	88.1	92.7	93.6	93.5	80.8	86.7	84.6	91.7	91.3	92.3	88.1	85.3	92.7	90.2	93.8	81.3

Model	gl	he	hi	hr	hu	hy	id	is	it	ja	kk	ko	ky	la	lt	lv	mr	mt
HPLT	82.3	91.0	**93.5**	**91.3**	82.4	84.1	81.7	**86.9**	**94.6**	**94.6**	25.7	**89.4**	43.1	**92.0**	84.9	**90.9**	65.8	**83.2**
XLM-R	**82.6**	**91.6**	93.3	**91.3**	**86.7**	**85.3**	**82.7**	86.6	**94.6**	94.4	**27.5**	89.0	**54.3**	91.9	**85.7**	**90.9**	**67.7**	78.5
mBERT	82.3	89.8	92.6	90.2	84.3	80.4	82.4	85.2	94.1	94.1	26.1	88.0	52.1	90.9	79.3	88.8	65.5	78.2

Model	nb	nl	nn	pl	pt	ro	ru	sk	sl	sr	sv	ta	te	tr	uk	ur	vi	zh
HPLT	**94.5**	**93.8**	**94.6**	**95.3**	**84.9**	90.6	**93.6**	93.8	**94.8**	92.5	90.8	63.6	85.7	**73.6**	61.3	83.9	68.0	84.6
XLM-R	94.3	92.9	93.9	95.2	84.5	**91.0**	93.4	**94.4**	94.7	**93.0**	**92.1**	**64.9**	**87.9**	73.0	**91.8**	**84.2**	**70.3**	**86.9**
mBERT	93.2	91.6	92.9	93.7	83.4	89.5	92.6	92.9	93.4	92.3	89.4	62.9	85.6	70.9	89.4	82.8	66.5	86.1

Table 3.5: Results on dependency parsing. We measure the Labeled Attachment Score (LAS) performance.

Model	af	ar	be	ca	cs	cy	da	bg	es	et	eu	fa	fi	fr
HPLT	87.5	**88.2**	90.1	90.1	89.0	89.4	90.3	91.5	89.6	89.6	89.8	91.8	89.2	87.2
XLM-R	**90.1**	87.7	**90.3**	91.0	**91.2**	90.0	**91.6**	**92.2**	89.9	90.4	90.7	**92.9**	90.0	88.7
mBERT	90.0	86.9	91.7	**92.1**	**91.2**	**92.5**	91.2	**92.2**	**90.9**	**91.8**	**91.3**	92.0	**90.2**	**90.5**

Model	gl	hi	hr	ga	hy	is	it	id	kk	de	la	lt	lv	mr
HPLT	91.1	84.3	89.3	55.9	94.8	55.9	87.8	89.1	62.4	64.1	88.6	87.0	90.7	84.7
XLM-R	**93.3**	88.0	**91.6**	78.0	95.3	63.9	89.7	**91.6**	74.8	87.7	92.1	**89.3**	92.6	87.5
mBERT	92.5	**88.6**	91.5	**80.8**	**95.7**	**81.7**	**90.5**	91.3	**82.0**	**89.4**	**93.6**	89.1	**93.2**	**88.0**

Model	nb	nl	nn	pt	ro	ru	sk	el	sv	te	tr	ur	vi	uk
HPLT	91.1	88.6	93.2	88.0	91.2	85.6	91.2	90.2	93.5	61.7	90.8	93.8	89.2	77.5
XLM-R	**92.6**	90.4	93.6	90.3	93.6	86.9	92.9	**90.7**	**94.5**	70.9	92.0	**94.6**	90.6	91.7
mBERT	91.9	**91.7**	**95.8**	**91.2**	**94.5**	**88.0**	**93.2**	90.2	94.3	**77.2**	**92.2**	94.3	**91.9**	**92.0**

Table 3.6: Results on named entity recognition. We measure the `seqeval` balanced F1 score.

As shown in tables 3.3, 3.4 and 3.5, HPLT BERT models are on par with multilingual mBERT and XLM-R models or outperform them in the Universal Dependencies tasks. For named entity recognition (Table 3.6), this pilot generation of HPLT models falls slightly behind multilingual LLMs on all languages except Arabic, where HPLT BERT outperforms all other models.

Abbrev.	Name	Reference
Parsebank	Finnish Internet Parsebank	https://turkunlp.org/finnish_nlp.html
mC4	multilingual colossal, cleaned Common Crawl	https://huggingface.co/datasets/mc4
CC-Fi	Common Crawl Finnish	https://github.com/TurkuNLP/CC-Fi
Fiwiki	Finnish Wikipedia	https://fi.wikipedia.org/wiki
Lönnrot	Projekti Lönnrot	http://www.lonnrot.net
ePub	National library "epub" collection	https://kansalliskirjasto.finna.fi
Lehdet	National library "lehdet" collection	https://kansalliskirjasto.finna.fi
Suomi24	The Suomi 24 Corpus 2001-2020	http://urn.fi/urn:nbn:fi:lb-2021101527
Reddit-Fi	Reddit `r/Suomi` submissions and comments	https://www.reddit.com/r/Suomi
STT	Finnish News Agency Archive 1992-2018	http://urn.fi/urn:nbn:fi:lb-2019041501
Yle	Yle Finnish News Archive 2011-2018	http://urn.fi/urn:nbn:fi:lb-2017070501
	Yle Finnish News Archive 2019-2020	http://urn.fi/urn:nbn:fi:lb-2021050401
	Yle News Archive Easy-to-read Finnish 2011-2018	http://urn.fi/urn:nbn:fi:lb-2019050901
	Yle News Archive Easy-to-read Finnish 2019-2020	http://urn.fi/urn:nbn:fi:lb-2021050701
ROOTS	Responsible Open-science Open-collaboration Text Sources	https://huggingface.co/bigscience-data

Table 4.1: Data sources for Finnish models

4 Decoder-only models

Below, we describe the first generation of decoder-only auto-regressive language models trained within the HPLT project. Such LLMs are "generative" in the sense that they are primarily used to generate text or code. The first models are focused on Finnish and Norwegian languages, mostly because two of the HPLT partner institutions had already invested efforts into data and model preparation for these languages in the past, reducing the time required to create models for this initial release. Models for other languages as well as flagship multilingual models are currently being trained and will be released as they complete.

Note that we report performance scores of these models on specific downstream tasks, but not perplexity, since this metric is heavily dependent on the choice of held-out test set, the size of vocabulary, etc. At the same time, perplexity scores themselves reflect only how good the model is in predicting the next word given the previous ones, but not its performance in real-world tasks.

4.1 Finnish models

The FinGPT model family (Luukkonen et al., 2023) is a monolingual GPT-style model family for Finnish trained early on in HPLT. The models serve as a proof of concept and are used as a starting point and reference for the multilingual decoder-only models trained in the project. The models follow the BLOOM model architecture, having ALiBi positional embeddings and an additional normalization layer after the first embedding layer. The data sources compiled to train the FinGPT models are detailed in Table 4.1.

The FinGPT family consists of 7 purely monolingual models ranging from 186M to 13B parameters, each trained for 300B tokens (Table 4.2). An 8th model, dubbed BLUUMI, is a more experimental model created through continued pre-training of the well-known BLOOM model, an open, massively multilingual 176B-parameter model, which we trained for an additional 40B tokens with the Finnish corpus combined into its original pre-training dataset. Excepting for the BLUUMI model which uses the original BLOOM tokenizer, all of the models share the same BPE tokenizer with a vocabulary size

Model	Layers	Dim	Heads	Params
Small	12	768	12	186M
Medium	24	1024	16	437M
Large	24	1536	16	881M
XL	24	2064	24	1.5B
3B	32	2560	32	2.8B
8B	32	4096	32	7.5B
13B	40	5120	40	13.3B
BLUUMI	70	14336	112	176B

Table 4.2: Architectures of the FinGPT model family.

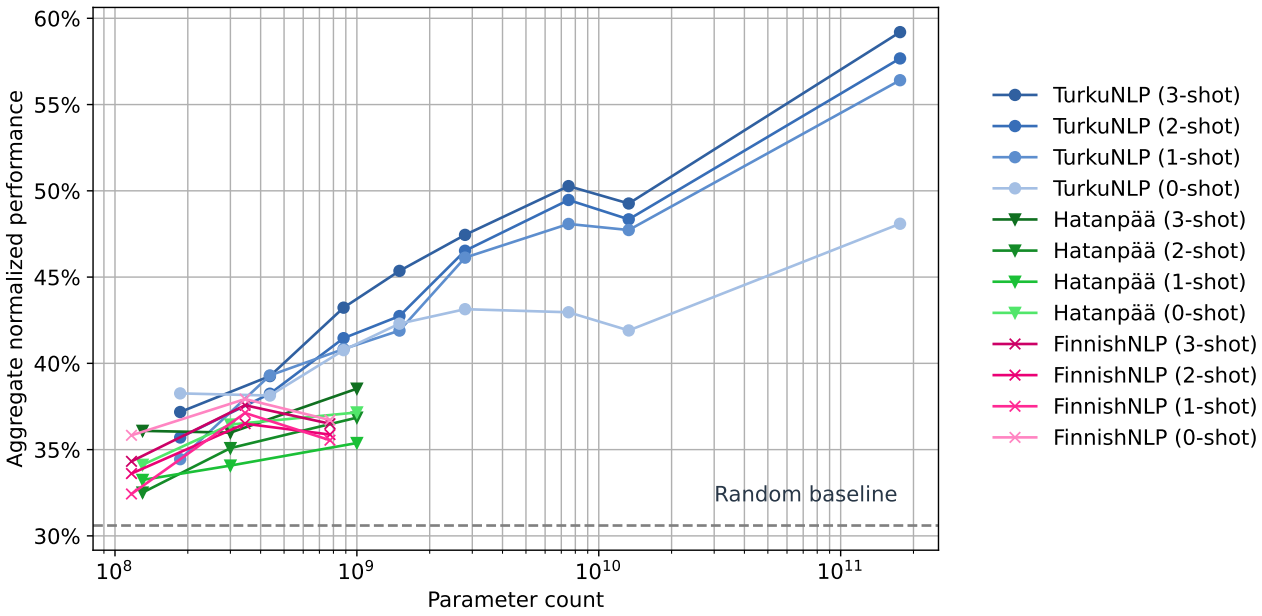


Figure 4.1: Overall FIN-bench evaluation results on FinGPT-models vs. previously released Finnish decoder-only models.

of 128k trained on a sample of the pretraining data. These models excel on the FIN-bench (Luukkonen et al., 2023) evaluation dataset, outperforming previously released Finnish GPT-like models (Fig 4.1). All of the models are available via Hugging Face at `https://huggingface.co/TurkuNLP`.

4.2 Norwegian models

Within the HPLT project, three new large auto-regressive decoder-only language models for Norwegian were trained (so called NoraLLM models, see Table 4.3):

1. **NorMistral-7b-warm**[1] – an LLM initialized from Mistral-7b-v0.1[2] and continuously pre-trained on Norwegian data;
2. **NorMistral-7b-scratch**[3] – a Mistral-based LLM pre-trained from scratch on Norwegian data;
3. **NorBLOOM-7b-scratch**[4] – a BLOOM-based LLM pre-trained from scratch on Norwegian data.

All the models are pre-trained on the same dataset and with the same byte-based BPE tokenizer of 32 768 tokens. The models are pre-trained exclusively on publicly available data, combining the resources from the public part of the NCC corpus,[5] from the cleaned HPLT corpus, and from CulturaX.[6] This resulted in over 34B subword tokens of Norwegian (Bokmål or Nynorsk) in total, which amounts to about 26.7B whitespace-separated tokens. The corpus was also augmented with programming code from Starcoder;[7] 20% of the 260B subword tokens are sampled from this code corpus. The natural language data is repeated six times to get the pre-training budget of 260B tokens, in accordance with findings from (Muennighoff et al., 2023).

Model	Training Data	Size	Context	Tokens	LR
NorMistral-7b-warm	NCC+HPLT+CulturaX+Starcoder	7B	2k	260B	1.0 x 10-4
NorMistral-7b-scratch	NCC+HPLT+CulturaX+Starcoder	7B	2k	260B	3.0 x 10-4
NorBLOOM-7b-scratch	NCC+HPLT+CulturaX+Starcoder	7B	2k	260B	1.2 x 10-4

Table 4.3: NoraLLM generative language models

The models were pre-trained using the Megatron-DeepSpeed library on LUMI. Pre-training one model took approximately 70k GPU hours of computation. We release our codebase at https://github.com/ltgoslo/norallm.

The NoraLLM model evaluation is an ongoing effort. We provide our initial evaluation results on standard natural language understanding and generation tasks, and our evaluation design will be gradually extended. We recommend that a user should perform additional evaluations for their particular model application scenario, including safety and bias evaluations.

Our initial downstream evaluation is conducted on reading comprehension and sentiment analysis tasks using open-source peer-reviewed datasets and benchmarks in native Norwegian. We compare against other pre-trained generative language models that officially support Norwegian: NB-GPT-J, GPT-Sw3 6.7B, GPT-Sw3 6.7B v2, and Falcon-7B; we also include Mistral-7b-v0.1. Tables 4.4 and 4.5 report

[1] https://huggingface.co/norallm/normistral-7b-warm
[2] https://huggingface.co/mistralai/Mistral-7B-v0.1
[3] https://huggingface.co/norallm/normistral-7b-scratch
[4] https://huggingface.co/norallm/NorBLOOM-7b-scratch
[5] https://huggingface.co/datasets/NbAiLab/NCC
[6] https://huggingface.co/datasets/uonlp/CulturaX
[7] https://huggingface.co/datasets/vikp/starcoder_filtered

the evaluation results, showing that the HPLT models (especially *NorMistral-7b-warm*) outperform previously existing LLMs for Norwegian in most cases.

Model	0-shot	1-shot	16-shot
NorMistral-7b-warm	60.6	**77.8**	**87.3**
NorMistral-7b-scratch	47.3	62.2	80.1
NorBLOOM-7b-scratch	**75.7**	73.8	65.5
NB-GPT-J	48.4	56.5	65.2
GPT-Sw3-6.7B	61.5	72.2	76.5
GPT-Sw3-6.7B-v2	42.4	69.1	83.4
Falcon-7B	53.3	61.6	74.9
Mistral-7B-v0.1	70.2	72.9	84.8

Table 4.4: Norwegian auto-regressive models performance in sentence-level sentiment analysis on the NoRec dataset (Øvrelid et al., 2020) (macro F1). HPLT model names are given *in italics*.

Model	0-shot	1-shot	2-shot
NorMistral-7b-warm	**48.6**/**24.8**	63.6/40.0	66.5/43.8
NorMistral-7b-scratch	34.0/15.7	46.5/25.8	48.5/27.8
NorBLOOM-7b	35.0/13.3	47.7/28.0	49.3/30.1
NB-GPT-J	24.4/6.8	32.8/11.6	35.0/12.3
GPT-Sw3-6.7B	46.5/22.0	55.9/32.0	58.1/34.3
GPT-Sw3-6.7B-v2	46.9/22.5	61.1/38.9	66.0/44.5
Falcon-7B	15.8/7.0	27.3/13.9	27.4/13.1
Mistral-7B-v0.1	46.4/22.4	**64.9**/**41.1**	**71.7**/**49.4**

Table 4.5: Norwegian auto-regressive models performance in reading comprehension on the NorQuad dataset (Ivanova et al., 2023). Scores are F1 and EM (exact match). HPLT model names are given *in italics*.

Hyperparam	Count
Layers	54
Dim	7168
Heads	56

Table 4.6: Poro 34B hyperparameters

4.3 Poro model

We are building on the basis of the results and experience gained from training the monolingual models to create increasingly multilingual models, building towards complete coverage of all EU languages. The Poro model is the first of our multilingual generative models.

Poro is 34B parameter GPT-like model, like FinGPT-models, trained on 3 text modalities: English[8] (54%), Code[9] (32%) and Finnish (13%; Table 4.1) and a comparatively small portion of translations to facilitate crosslingual transfer. It is trained for the total of 1T tokens with a 2048 token context window. Poro uses a GPT2-style BPE-tokenizer with a vocabulary size of 128k tokens trained on an even distribution of English, Finnish and Code sampled from its training data.

At the time of writing, Poro is the best-performing openly released foundation model for Finnish (Fig 4.2). We have released 10 checkpoints throughout the training with a pacing of 100B tokens at https://huggingface.co/LumiOpen/Poro-34B.

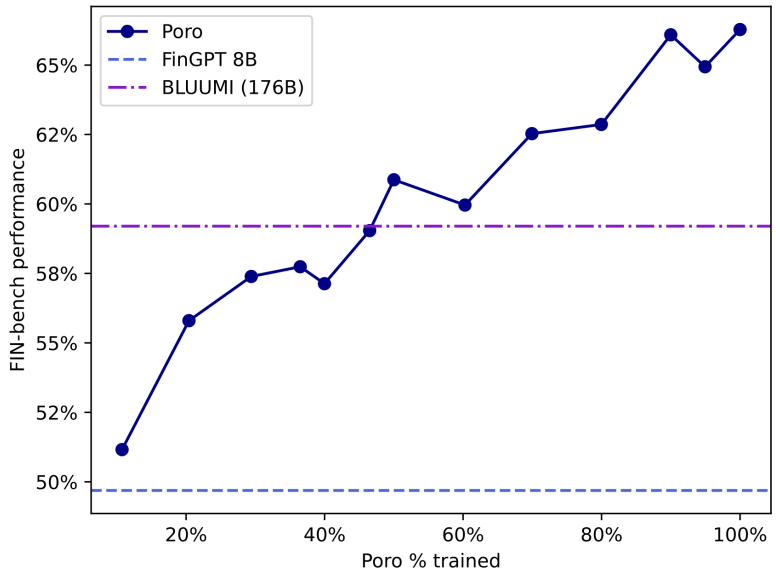


Figure 4.2: Comparison of Poro loss progression to the best monolingual FinGPT-model and the massive scale model BLUUMI.

[8] https://huggingface.co/datasets/cerebras/SlimPajama-627B
[9] https://huggingface.co/datasets/bigcode/starcoderdata

Model	Layers	Dim	FFN	Heads	Params
7B	32	4096	11008	32	7.5B
13B	40	5120	13824	40	13.3B
34B	54	7168	20480	56	34.5B

Table 4.7: Architectures of the Nordic model family.

	Tokens	Ratio
English	660.5B	33.0%
Swedish	161.1B	8.1%
Finnish	147.4B	7.4%
Danish	103.1B	5.2%
Norwegian	83.8B	4.2%
Icelandic	8.1B	0.4%
Cross-lingual	167.3B	8.4%
Code	661.0B	33.1%

Table 4.8: Training data distribution for the Nordic model family.

4.4 Nordic models

We are currently in the process of training a multilingual Nordic model family of 3 different model sizes: 7B, 13B and 33B parameters. It uses same the architecture as LLaMA, differing from Poro mainly in using Rotary positional encodings instead of ALiBI, not having the aforementioned extra layer normalization, and by using SwiGLU activation function instead of GeLU. In addition, the 33B parameter model uses group query attention with 8 key-value groups. The context window of all of the models is 4096 tokens.

The models are trained on a multilingual corpus consisting of English, Finnish, Swedish, Norwegian, Danish, Icelandic, and code, with the total dataset size of 2T tokens. The dataset is shared with Poro (Section 4.3) for Finnish English, and code, and with NoraLLM (Section 4.2) for Norwegian. For the other Scandinavian languages, the training data was drawn from mC4 (https://huggingface.co/datasets/mc4) as HPLT data was not yet available at the start of training. Following results from (Muennighoff et al., 2023), we upsample smaller languages up to 4x, and following the Poro process, we also include a cross-lingual signal from translatio pairs. The training data distribution is shown in Table 4.8. The tokenizer for the models is a GPT2-like BPE tokenizer with a vocabulary size of 128k tokens trained on an even distribution of English, Finnish, Swedish, Norwegian, Icelandic, Danish and code.

As the training process is long and still ongoing, we're providing a pre-release of model checkpoints at https://huggingface.co/HPLT for this deliverable.

4.5 European models

We have started the process to train a family of massively multilingual European models with a dataset covering all official EU languages.

Our flagship model is a 71B LLaMA-like model that is trained on 3T tokens, and it aims to be the state-of-the art model of multilingual European LLMs. The training process of this model is in its earliest phases due in part to recent changes in the hardware availability (LUMI getting divided into two partitions and max node count per job reduced from 1024 to 512), which required us to revise and adapt our plans for the training run. The context window of these models is 4096 tokens, and the tokenizer follows the design of the tokenizers for Poro and the Nordic models and is trained with a uniform distribution of all the languages.

In addition to the large flagship model, we will also train selected smaller models, but these are scheduled to be trained later since we want to keep priority on the flagship model and thus want to prevent our jobs competing with each other in the LUMI job queue. Model checkpoints will be made available through the HPLT website as they become available.

Bibliography

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.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

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.

Sardana Ivanova, Fredrik Andreassen, Matias Jentoft, Sondre Wold, and Lilja Øvrelid. 2023. NorQuAD: Norwegian question answering dataset. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 159–168, Tórshavn, Faroe Islands. University of Tartu Library.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Le Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, Samuel Antao, and Sampo Pyysalo. 2023. Fingpt: Large generative models for a small language. In *Conference on Empirical Methods in Natural Language Processing*.

Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling data-constrained language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Lilja Øvrelid, Petter Mæhlum, Jeremy Barnes, and Erik Velldal. 2020. A fine-grained sentiment dataset for Norwegian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5025–5033, Marseille, France. European Language Resources Association.

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.

Gema Ramírez-Sánchez. 2024. D3.1: software for cleaning data sets. HPLT deliverable.

David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. Trained on 100 million words and still in shape: BERT meets British National Corpus. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1954–1974, Dubrovnik, Croatia. Association for Computational Linguistics.

Jörg Tiedeman, Nikolay Arefev, Andrey Kutuzov, Stephan Oepen, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Ona de Gibert Bonet, and Pavel Straňák. 2023. D2.1: initial release of monolingual and parallel data sets. HPLT deliverable.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.