

Large-Scale Norwegian Language Data from the Web

Stephan Oepen, Language Technology Group, University of Oslo

Reflections from Two Ongoing EU Projects – December 11, 2025



1

Organizational Background



HEU & UKRI

2022-2025

8 Partners

1 Company

2 National HPC

Around 6 M€



DEP

2025-2028

22 Partners

5 Companies

5 National HPC

Around 37 M€

A Broad Alliance: Research, HPC, Innovation



Universities and Research Organizations



Companies



HPC Centres



MADLAD-400: A Multilingual And Document-Level Large Audited Dataset

Sneha Kudugunta[†] Isaac Caswell[◇] Biao Zhang[†] Xavier Garcia[†]
Christopher A. Choquette-Choo[†] Katherine Lee[†] Derrick Xin[†] Aditya Kusupati[◇]
Romi Stella[†] Ankur Bapna[†] Orhan Firat[†]
[†]Google DeepMind [◇]Google Research

<https://arxiv.org/abs/2309.04662>



A sparkling update with 1000s of languages

<https://huggingface.co/datasets/HuggingFaceFW/fineweb-2>



Nemotron-CC: Transforming Common Crawl into a Refined Long-Horizon Pretraining Dataset

**Dan Su^{*}, Kezhi Kong^{*}, Ying Lin^{*}, Joseph Jennings, Brandon Norick,
Markus Kliegl[†], Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro**

NVIDIA

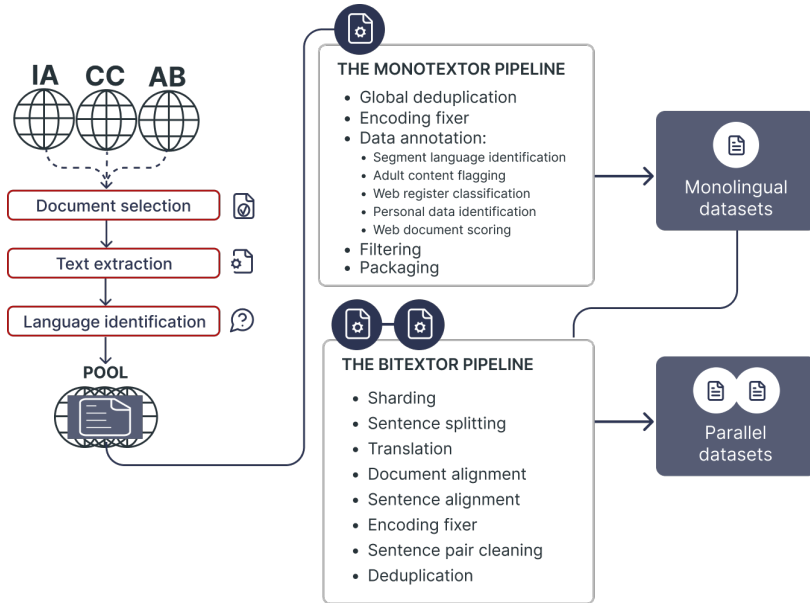
^{*}Equal contribution. [†]Correspondence to mkliegl@nvidia.com.

<https://arxiv.org/abs/2412.02595>

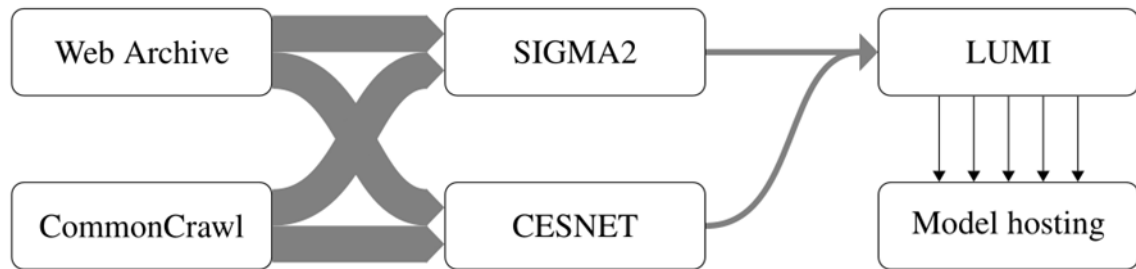
2

Web Data Processing

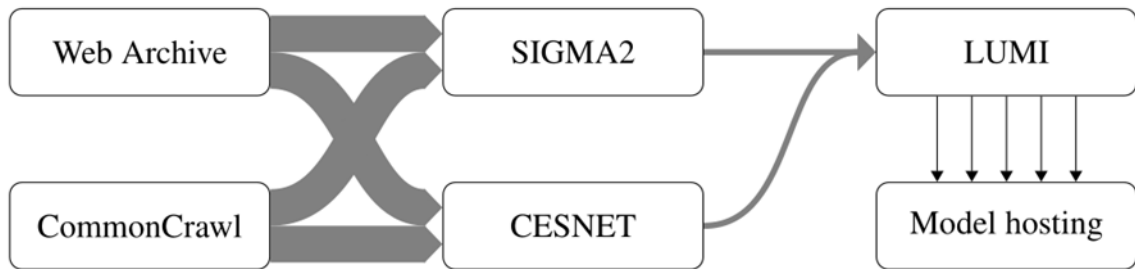
Pipeline for Text Extraction and Enrichment



The Big Picture: Large-Scale Training Data & (L)LMs



The Big Picture: Large-Scale Training Data & (L)LMs



- ▶ **Storage** allocations in CZ (CESNET), NO (Sigma2), and LUMI: **7 + 6 + 6 PiB**;
- ▶ Lightweight document extraction on **dedicated local nodes** at each storage site;
- ▶ Text extraction, data enrichment, cleaning, filtering, packaging, etc. on **LUMI**;
- ▶ Generous **national allocations** (CZ, FI, NO), about **30M cpu & 10M gpu hours**;
- ▶ **Public download** site (some **600 TiB**) hosted by Sigma2 in Norway (Lefdal Mine).

3

**Data
Statistics**

Some Contrastive Statistics



Language	HPLT 3.0			FineWeb 1 & 2			HPLT 2.0		
	T		%	T		%	T		%
English	16T	901	55	17T	695	78	3.9T	892	35
Multilingual	13T	1187	45	4.9T	976	22	7.2T	1178	65
Basque	3.2B	991	0.02	1.5B	951	0.03	2.0B	1030	0.03
Catalan	22B	853	0.17	12B	715	0.25	18B	976	0.25
Czech	126B	1171	0.93	67B	1015	1.37	95B	1266	1.32
Finnish	73B	1491	0.55	48B	1324	0.99	53B	1538	0.74
French	584B	968	4.32	292B	811	5.95	379B	943	5.24
Galician	3.1B	772	0.02	1.8B	695	0.04	2.7B	906	0.04
Norwegian	52B	1388	0.39	53B	1318	1.09	42B	1477	0.58
Spanish	658B	908	4.86	329B	746	6.71	471B	936	6.51
Ukrainian	81B	1014	0.60	49B	938	1.02	60B	1280	0.84
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Some Contrastive Statistics



Language	HPLT 3.0			FineWeb 1 & 2			HPLT 2.0		
	T		%	T		%	T		%
English	16T	901	55	17T	695	78	3.9T	892	35
Multilingual	13T	1187	45	4.9T	976	22	7.2T	1178	65
Basque	3.2B	991	0.02	1.5B	951	0.03	2.0B	1030	0.03
Catalan	22B	853	0.17	12B	715	0.25	18B	976	0.25
Czech	126B	1171	0.93	67B	1015	1.37	95B	1266	1.32
Finnish	73B	1491	0.55	48B	1324	0.99	53B	1538	0.74
French	584B	968	4.32	292B	811	5.95	379B	943	5.24
Galician	3.1B	772	0.02	1.8B	695	0.04	2.7B	906	0.04
Norwegian	52B	1388	0.39	53B	1318	1.09	42B	1477	0.58
Spanish	658B	908	4.86	329B	746	6.71	471B	936	6.51
Ukrainian	81B	1014	0.60	49B	938	1.02	60B	1280	0.84
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Some Contrastive Statistics



Language	HPLT 3.0			FineWeb 1 & 2			HPLT 2.0		
	T		%	T		%	T		%
English	16T	901	55	17T	695	78	3.9T	892	35
Multilingual	13T	1187	45	4.9T	976	22	7.2T	1178	65
Basque	3.2B	991	0.02	1.5B	951	0.03	2.0B	1030	0.03
Catalan	22B	853	0.17	12B	715	0.25	18B	976	0.25
Czech	126B	1171	0.93	67B	1015	1.37	95B	1266	1.32
Finnish	73B	1491	0.55	48B	1324	0.99	53B	1538	0.74
French	584B	968	4.32	292B	811	5.95	379B	943	5.24
Galician	3.1B	772	0.02	1.8B	695	0.04	2.7B	906	0.04
Norwegian	52B	1388	0.39	53B	1318	1.09	42B	1477	0.58
Spanish	658B	908	4.86	329B	746	6.71	471B	936	6.51
Ukrainian	81B	1014	0.60	49B	938	1.02	60B	1280	0.84
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Some Contrastive Statistics



Language	HPLT 3.0			FineWeb 1 & 2			HPLT 2.0		
	T		%	T		%	T		%
English	16T	901	55	17T	695	78	3.9T	892	35
Multilingual	13T	1187	45	4.9T	976	22	7.2T	1178	65
Basque	3.2B	991	0.02	1.5B	951	0.03	2.0B	1030	0.03
Catalan	22B	853	0.17	12B	715	0.25	18B	976	0.25
Czech	126B	1171	0.93	67B	1015	1.37	95B	1266	1.32
Finnish	73B	1491	0.55	48B	1324	0.99	53B	1538	0.74
French	584B	968	4.32	292B	811	5.95	379B	943	5.24
Galician	3.1B	772	0.02	1.8B	695	0.04	2.7B	906	0.04
Norwegian	52B	1388	0.39	53B	1318	1.09	42B	1477	0.58
Spanish	658B	908	4.86	329B	746	6.71	471B	936	6.51
Ukrainian	81B	1014	0.60	49B	938	1.02	60B	1280	0.84
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Some Contrastive Statistics



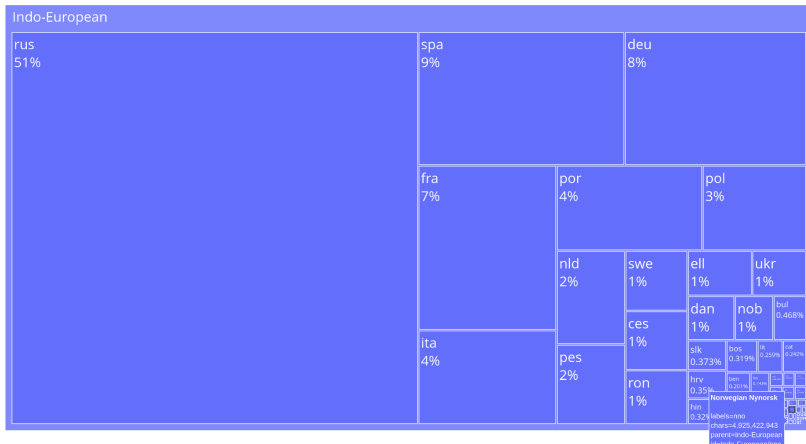
Language	HPLT 3.0			FineWeb 1 & 2			HPLT 2.0		
	T		%	T		%	T		%
English	16T	901	55	17T	695	78	3.9T	892	35
Multilingual	13T	1187	45	4.9T	976	22	7.2T	1178	65
Basque	3.2B	991	0.02	1.5B	951	0.03	2.0B	1030	0.03
Catalan	22B	853	0.17	12B	715	0.25	18B	976	0.25
Czech	126B	1171	0.93	67B	1015	1.37	95B	1266	1.32
Finnish	73B	1491	0.55	48B	1324	0.99	53B	1538	0.74
French	584B	968	4.32	292B	811	5.95	379B	943	5.24
Galician	3.1B	772	0.02	1.8B	695	0.04	2.7B	906	0.04
Norwegian	52B	1388	0.39	53B	1318	1.09	42B	1477	0.58
Spanish	658B	908	4.86	329B	746	6.71	471B	936	6.51
Ukrainian	81B	1014	0.60	49B	938	1.02	60B	1280	0.84
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Zooming in: Closer to Home



	HPLT 3.0			FineWeb 1 & 2			HPLT 2.0		
Bokmål	51T	1402	97.0	53T	1336	98.3	40T	1489	95.8
Nynorsk	1.6T	1052	3.0	0.9T	735	1.7	1.8T	1254	4.2

	HPLT 3.0			FineWeb 1 & 2			HPLT 2.0		
Bokmål	51T	1402	97.0	53T	1336	98.3	40T	1489	95.8
Nynorsk	1.6T	1052	3.0	0.9T	735	1.7	1.8T	1254	4.2



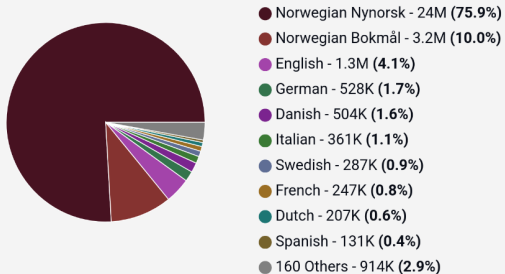
Tokens	Languages
$\geq 1,000,000,000,000$ (one trillion)	eng
$\geq 100,000,000,000$ (hundred billion)	deu, fra, ell, ita, nld, pol, por, spa
$\geq 10,000,000,000$ (ten billion)	bos, bul, cat, ces, dan, est, fin, hrv, hun, lav, lit, nor, ron, slk, slv, swe, tur, ukr
$\geq 1,000,000,000$ (one billion)	eus, glg, isl, kat, mkd, sqi, srp
$< 1,000,000,000$ (one billion)	gle, mlt

Table 2.2: Tiers of available textual data for OpenEuroLLM target languages.

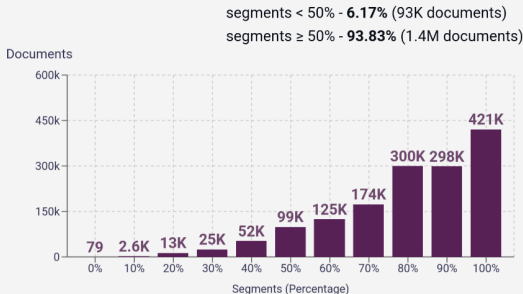
Language Identification: Use with a Bag of Salt



Number of segments in the Norwegian Nynorsk corpus ⓘ



Percentage of segments in Norwegian Nynorsk inside documents ⓘ



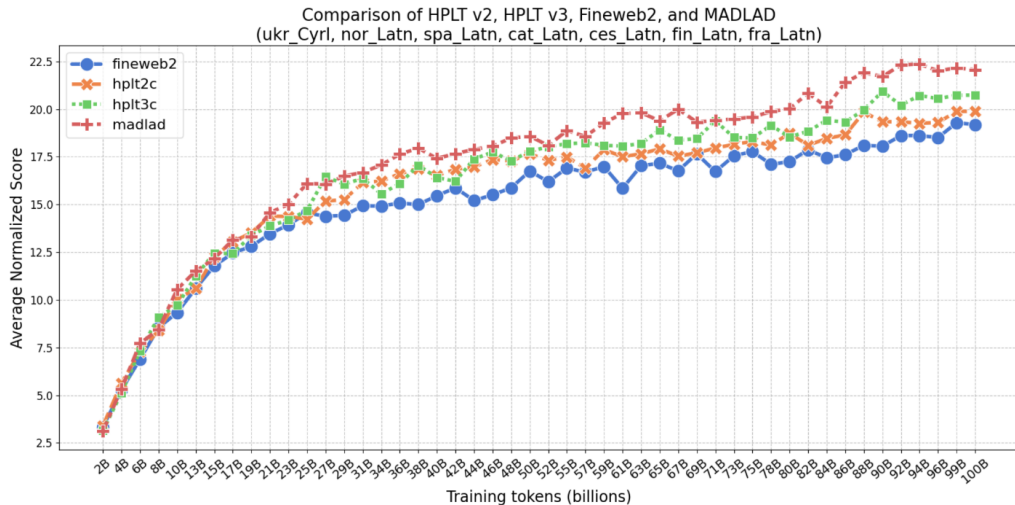
https://analytics.hplt-project.org/viewer/HPLT-v3-nob_Latn.yaml

https://analytics.hplt-project.org/viewer/HPLT-v3-nno_Latn.yaml

4

Multilingual Evaluation

HPLT-E: Seven European Languages (2B–100BT)



<https://github.com/hplt-project/hplt-e>

5

**Wrap-Up:
Looking Ahead**

The LLM Dilemma: Data Paucity & Unclear Terms of Use



Top 10 domains

Domain	Docs	% of total
wikipedia.org	91K	6.05%
blogspot.com	77K	5.10%
nrk.no	66K	4.40%
docplayer.me	26K	1.74%
framtida.no	24K	1.57%
wordpress.com	21K	1.42%
blogg.no	17K	1.13%
midtsiden.no	17K	1.11%
blogspot.no	17K	1.11%
ndla.no	16K	1.05%

- ▶ Majority of **web content** is published under **copyright**;
- ▶ like CC & IA: redistribute **without claim of ownership**;
- ▶ best effort to respect **common practices for 'opt-out'**;
- ▶ take-down mechanism (removal) by legitimate request;
- ▶ all metadata and **collection 'per se'** licensed as CC0.

<https://hplt-project.org/datasets/v3.0>

The LLM Dilemma: Data Paucity & Unclear Terms of Use



Top 10 domains

Domain	Docs	% of total
wikipedia.org	91K	6.05%
blogspot.com	77K	5.10%
nrk.no	66K	4.40%
docplayer.me	26K	1.74%
framtida.no	24K	1.57%
wordpress.com	21K	1.42%
blogg.no	17K	1.13%
midtsiden.no	17K	1.11%
blogspot.no	17K	1.11%
ndla.no	16K	1.05%

- ▶ Majority of **web content** is published under **copyright**;
- ▶ like CC & IA: redistribute **without claim of ownership**;
- ▶ best effort to respect **common practices for 'opt-out'**;
- ▶ take-down mechanism (removal) by legitimate request;
- ▶ all metadata and **collection 'per se'** licensed as CC0.

<https://hplt-project.org/datasets/v3.0>



INCREASING COORDINATION AMONG EU INITIATIVES,
E.G. OPENWEBSEARCH, LDS, TRUSTLLM, ETC.

HPLT 3.0: Very Large-Scale Multilingual Resources for LLM and MT Mono- and Bi-lingual Data, Multilingual Evaluation, and Pre-Trained Models

Stephan Oepen♣, Nikolay Arefev♣, Mikko Aulamo♣, Marta Bañón♡, Maja Buljan♣,
Laurie Burchell◇, Lucas Charpentier♣, Pinzhen Chen*, Mariia Fedorova♣, Ona de Gibert♣,
Barry Haddow*, Jan Hajič°, Jindřich Helcl♣, Andrey Kutuzov♣, Veronika Laippala*, Zihao Li♣,
Risto Luukkonen*, Bhavitvya Malik*, Vladislav Mikhailov♣, Amanda Myntti*,
Dayyán O'Brien*, Lucie Poláková°, Sampo Pyysalo*, Gema Ramírez Sánchez♡,
Janine Siewert♣, Pavel Stepachev*, Jörg Tiedemann♣, Teemu Vahtola♣,
Dušan Variš°, Fedor Vitiugin*, Tea Vojtěchová°, Jaume Zaragoza♡

♣ University of Oslo, Department of Informatics

♣ University of Helsinki, Department of Digital Humanities

♡ Prompsit Language Engineering

◇ The Common Crawl Foundation

• Edinburgh University, School of Informatics

° Charles University, Prague, Institute of Formal and Applied Linguistics

* TurkuNLP, University of Turku, Department of Computing

oe@ifi.uio.no

<https://arxiv.org/abs/2511.01066>

Stephan Oepen

Language Technology Group

`oe@ifi.uio.no`