

HPLT & OpenEuroLLM: European Foundation Models

Nikolay Arefev, Mariia Fedorova, Jindřich Helcl, Andrey Kutuzov,
Vladislav Mikhailov, Stephan Oepen, David Samuel

UiO Language Technology Seminar – May 13, 2025



An Expanded Massive Multilingual Dataset for High-Performance Language Technologies

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, Farrokh Mehryary, Vladislav Mikhailov, Nikita Moghe, Amanda Myntti, Dayyán O'Brien, Stephan Oepen, Proyag Pal, Jousia Piha, Sampo Pyysalo, Gema Ramírez-Sánchez, David Samuel, Pavel Stepachev, Jörg Tiedemann, Dušan Variš, Tereza Vojtěchová, Jaume Zaragoza-Bernabeu

Training state-of-the-art large language models requires vast amounts of clean and diverse textual data. However, building suitable multilingual datasets remains a challenge. In this work, we present HPLT v2, a collection of high-quality multilingual monolingual and parallel corpora. The monolingual portion of the data contains 8T tokens covering 193 languages, while the parallel data contains 380M sentence pairs covering 51 languages. We document the entire data pipeline and release the code to reproduce it. We provide extensive analysis of the quality and characteristics of our data. Finally, we evaluate the performance of language models and machine translation systems trained on HPLT v2, demonstrating its value.

<https://arxiv.org/abs/2503.10267>

1

Organizational Background

The HPLT Consortium (2022-2025)



High-Performance Language Technologies (HPLT)



HPLT gathers together academic partners, an NLP company and HPC centers from all around Europe.



SIGMA2



UNIVERSITY OF TURKU



UNIVERSITY OF EDINBURGH



PROMPSIT L.E.



CESNET



UNIVERSITY OF OSLO



UNIVERSITY OF HELSINKI

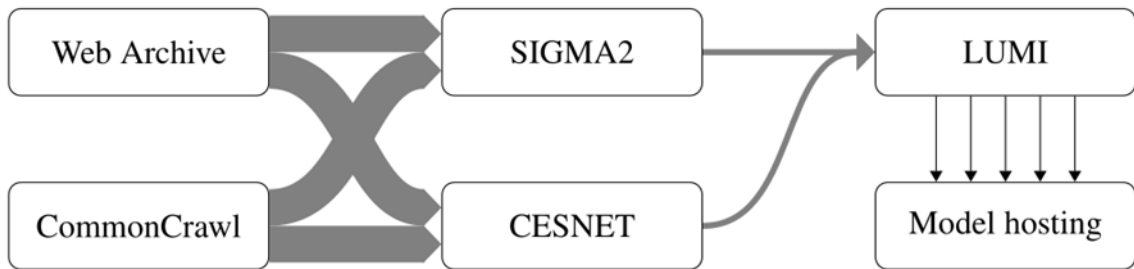


CHARLES UNIVERSITY

<https://hplt-project.org/>

This project has received funding from the European Union's Horizon Europe research 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]

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



- Vast storage allocations in CZ (CESNET) and NO (Sigma2): 7 & 6 petabytes, respectively
- Lightweight text extraction on dedicated compute nodes local to each of the storage sites
- Text refinement, bi-text discovery, model training, evaluation, etc. primarily on LUMI
- Generous national allocations (CZ, FI, NO), for a total of some 10M CPU & 20M GPU hours

2

**Pre-Training
Data**

European Language Hierarchy by Data Availability



Words	#	Languages
>2.5 trillion	1	eng
>100 billion	3	deu, spa, fra
>10 billion	11	ces, dan, ell, fin, hun, ita, nld, pol, por, ron, sve
>1 billion	7	bul, est, hrv, lit, lav, slk, slv
>100 million	2	gle, mlt

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Colin Raffel*

CRAFFEL@GMAIL.COM

Noam Shazeer*

NOAM@GOOGLE.COM

Adam Roberts*

ADAROB@GOOGLE.COM

Katherine Lee*

KATHERINELEE@GOOGLE.COM

Sharan Narang

SHARANNARANG@GOOGLE.COM

Michael Matena

MMATENA@GOOGLE.COM

Yanqi Zhou

YANQIZ@GOOGLE.COM

Wei Li

MWEILI@GOOGLE.COM

Peter J. Liu

PETERJLIU@GOOGLE.COM

Google, Mountain View, CA 94043, USA

<https://arxiv.org/abs/1910.10683v4>

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>



A sparkling update with 1000s of languages

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

Overview of Related Initiatives



Dataset	Size (TB)	Tokens (T)	Languages	% English	Source
English only					
The Pile Gao et al. (2020)	0.8	0.39	1	100	various
C4.en Raffel et al. (2020); Dodge et al. (2021)	0.3	0.16	1	100	Common Crawl
RefinedWeb Penedo et al. (2023)	2.8	0.6	1	100	Common Crawl
Dolma-Web Soldaini et al. (2024)	-	2.28	1	100	Common Crawl
FineWeb Penedo et al. (2024)	-	15	1	100	Common Crawl
Multilingual					
OSCAR-23.01 Suárez et al. (2019)	-	1.1	153	48.43	Common Crawl
CC-100 Conneau et al. (2020)	2.39	0.3	100	18.84	Common Crawl
mC4 Xue et al. (2021)	-	6.3	101	5.67	Common Crawl
ROOTS Laurençon et al. (2022)	1.6	0.4	46	30.03	BigScience, Common Crawl, OSCAR
Glott500-c Imani et al. (2023)	0.6	-	511	*2.16	various
Serengeti Adebara et al. (2023)	0.042	-	517	-	various
CulturaX Nguyen et al. (2024)	27	6.3	167	45.13	OSCAR, mC4
MADLAD-400-clean Kudugunta et al. (2024)	-	2.6	419	50	Common Crawl
MaLA Ji et al. (2024)	-	0.074	939	4	various
monoHPLT v1.2-dedup de Gibert et al. (2024)	11	5.6	75	41	Common Crawl, Internet Archive
HPLT v2 (monolingual, deduplicated) Burchell et al. (2025)	21	7.6	193	44	Common Crawl, Internet Archive

Main Sources of Large-Scale Web Data

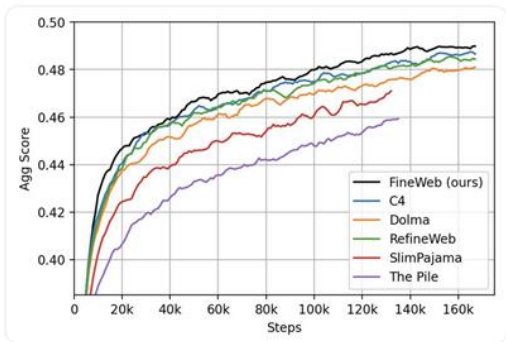


Guilherme Penedo

@gui_penedo

We have just released 🍷 FineWeb: 15 trillion tokens of high quality web data.

We filtered and deduplicated all CommonCrawl between 2013 and 2024. Models trained on FineWeb outperform RefinedWeb, C4, DolmaV1.6, The Pile and SlimPajama!



9:49 AM · Apr 21, 2024 · 249K Views

Two Main Sources of Web Data

Common Crawl



INTERNET ARCHIVE



Worldwide Web Crawls since September 2011 (00–18)

- In total about 5.7 petabytes
- Standardized WARC file format
- “Deeper” than Common Crawl
- Also much larger capture limit
- Little activity in 2022 & 2023

Collaboration Agreement with HPLT

- Entered by UiO; for the full consortium
- Redistribution of derived text & models
- Provides up to 7 petabytes of web data
- Prioritized connection to IA data center
- \$ 100,000 for IA services to be invoiced



Comparison of Common Crawl vs. Internet Archive in Terms of Domains & URLs

- Wide crawls are substantially larger: capture more unique domains and URLs
- 3–4% of unique domains from WIDE15 and WIDE17 are also included in CC40
- Less than 1% of unique URLs from WIDE15 and WIDE17 also occur in CC40

Caveats

- Distinct URLs can capture same page
- Contents of same URL changes over time
- CC40: 2022; Wide15, 17: 2016–17, 2018–20

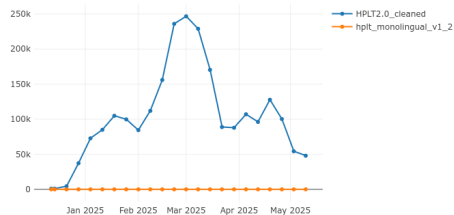
Domains	CC40	WIDE15	WIDE17
CC40	0.041	0.013	0.016
WIDE15	0.013	0.329	0.046
WIDE17	0.016	0.046	0.474
URLs	CC40	WIDE15	WIDE17
CC40	2.793	0.024	0.047
WIDE15	0.024	5.093	0.419
WIDE17	0.047	0.419	10.197

(Billions of unique domains or URLs)

Datasets and models are hosted for public download by Sigma2 (and Hugging Face); (most) text extraction and model training is performed on LUMI.

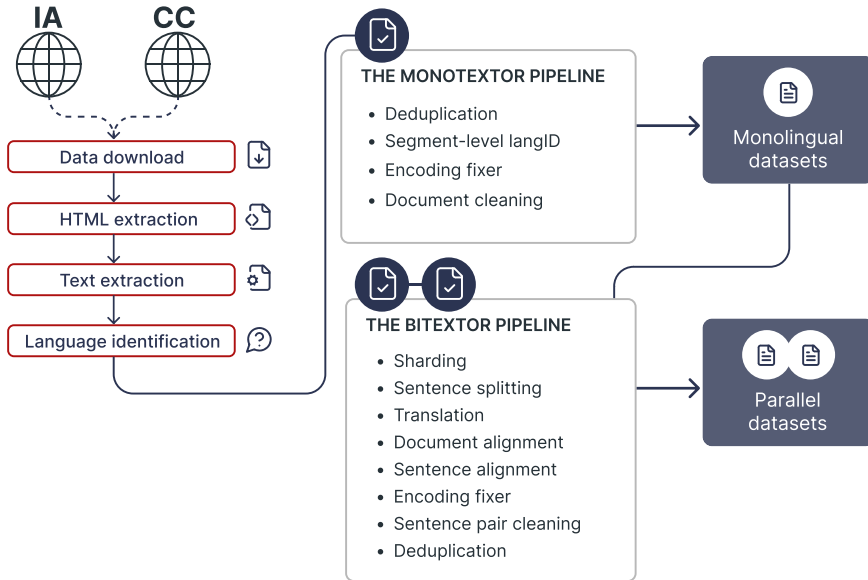
HPLT History

- ▶ HPLT v1 (5.6T deduplicated 'words')
 - ▶ Released in 2023, described in de Gibert et al. (2024)
 - ▶ HPLT v1.2 introduced the **cleaned** version
- ▶ HPLT v2 (7.6T deduplicated 'words')
 - ▶ Released in 2024, described in Burchell et al. (2025)
 - ▶ 15 terabytes of cleaned data.
 - ▶ A larger collection, 'less noise and boilerplate, less duplicates, more unique documents, and generally better quality texts to train language models on'
 - ▶ Turns out this was true!
- ▶ HPLT v3 (about three times larger)
 - ▶ ... coming this summer.

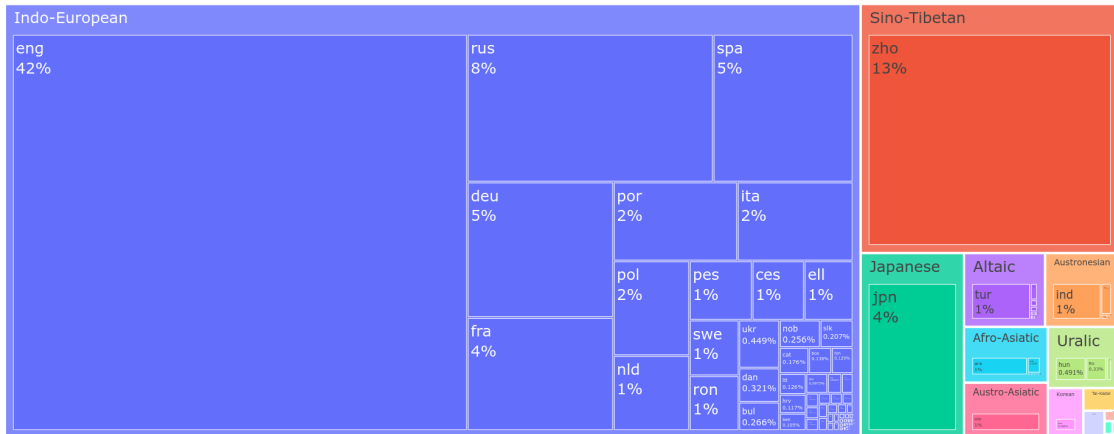


Download statistics from Hugging Face Hub
https://hf.co/datasets/HPLT/HPLT2.0_cleaned

Pipeline for Text Extraction and Enrichment



Distribution of Documents by Language Family



Side-by-Side Comparison for Select Languages

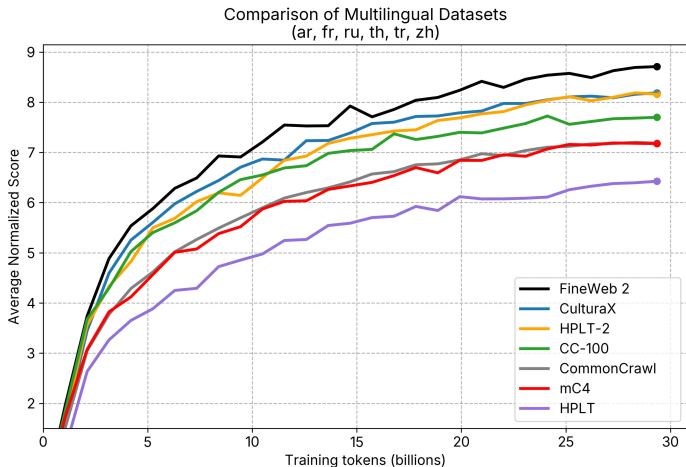


	mC4**	CulturaX**	FineWeb*	HPLT 2.0
Catalan (cat)	13B	12,530M	7,596M	10,020M
Czech (ces)	63B	56,910M	34,180M	42,080M
Finnish (fin)	25B	28,925M	18,558M	18,450M
Maltese (mlt)	5.2B		245M	186M
Norwegian (nor)	27B	18,426M	30,945M	22,390M
Spanish (spa)	433B	373,845M	244,541M	321,950M
Ukrainian (ukr)	41B	38,226M	23,370M	25,230M
	6.3T	6.5T	2.7T	5.6T

3

**Empirical
Validation**

FineWeb-Style Ablation Studies



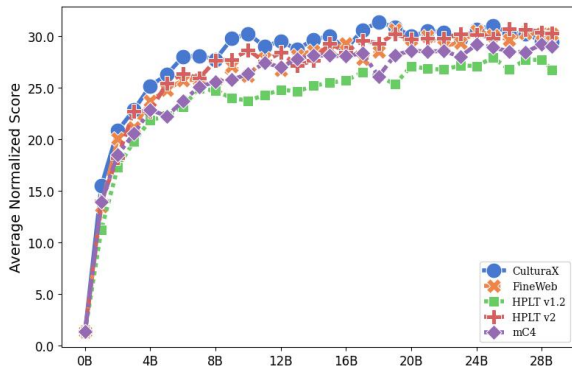
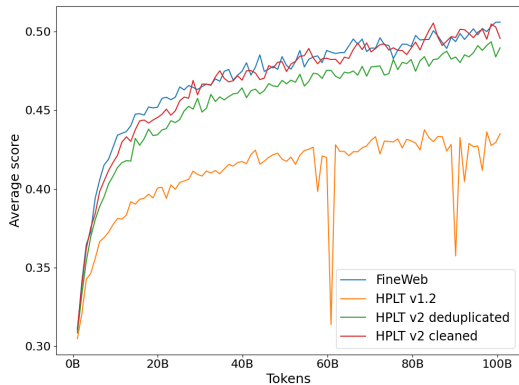
FineWeb2: Hugging Face effort to evaluate multilingual corpora for training generative LLMs

Penedo et al. (2024), <https://huggingface.co/datasets/HuggingFaceFW/fineweb-2>

FineWeb-Style Ablation Studies



Our own evaluations on **English** (left) and **Norwegian** (right):



HPLT v2 seems **much better** than version 1.2,
and performs on par with the currently most fashionable datasets.

Outlook: Multilingual Evaluation (HPLT-E, 'EuroFineTasks')



Final Project Year: Data Quality & Mixture Studies

Presupposes ability to perform in-depth, 'meaningful' multilingual end-to-end experimentation: ablation studies, evaluation.

Coming together these Days: HPLT-E (Role Model: NorEval)

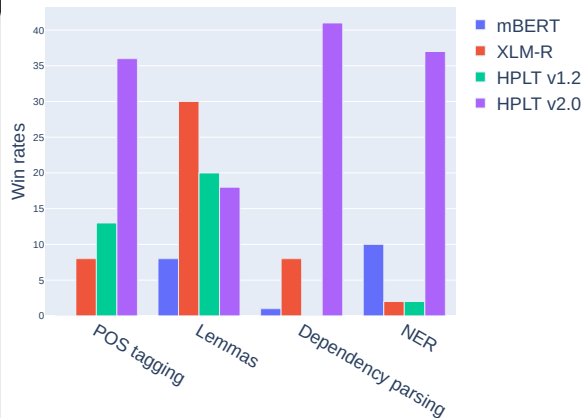
- ▶ Focus languages: Czech, Finnish, French, Norwegian, Ukrainian, Basque, Catalan, Galician, Spanish;
- ▶ strong emphasis on evaluation datasets curated by native speakers;
- ▶ seek to push for common practise with multi-prompt evaluation;
- ▶ 'hunt & scavenge': currently close to 140 tasks in uniform taxonomy.

4

**Validation with
masked language
models**

LTG-BERTs

- ▶ LTG-BERTs Samuel et al. (2023) on 53 languages, including Bokmål, Nynorsk, Luxembourgish, Galician, Croatian etc.
- ▶ evaluation: on part-of-speech tagging, lemmatization and dependency parsing of Universal Dependencies (UD) treebanks de Marneffe et al. (2021) and (NER) on WikiAnn Pan et al. (2017)
- ▶ all models available at <https://hf.co/collections/HPLT/hplt-20-bert-models-67ba52ae96b1fb8aae673493>



5

**Outlook:
OpenEuroLLM**

An Extended Consortium



Universities and Research Organizations



Companies



HPC Centres





The EU is on track
to doubling the
compute capacity
of the EuroHPCs.



Co-funded by
the European Union



- Adebara, I., Elmadany, A., Abdul-Mageed, M., and Alcoba Inciarte, A. (2023). SERENGETI: Massively multilingual language models for Africa. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, Findings of the Association for Computational Linguistics: ACL 2023, pages 1498–1537, Toronto, Canada. Association for Computational Linguistics.
- Burchell, L., de Gibert, O., Arefyev, N., Aulamo, M., Bañón, M., Chen, P., Fedorova, M., Guillou, L., Haddow, B., Hajič, J., Helcl, J., Henriksson, E., Klimaszewski, M., Komulainen, V., Kutuzov, A., Kytöniemi, J., Laippala, V., Mæhlum, P., Malik, B., Mehryary, F., Mikhailov, V., Moghe, N., Myntti, A., O'Brien, D., Oepen, S., Pal, P., Piha, J., Pyysalo, S., Ramírez-Sánchez, G., Samuel, D., Stepachev, P., Tiedemann, J., Variš, D., Vojtěchová, T., and Zaragoza-Bernabeu, J. (2025). An expanded massive multilingual dataset for high-performance language technologies.

References II

- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.
- de Gibert, O., Nail, G., Arefyev, N., Bañón, M., van der Linde, J., Ji, S., Zaragoza-Bernabeu, J., Aulamo, M., Ramírez-Sánchez, G., Kutuzov, A., Pyysalo, S., Oepen, S., and Tiedemann, J. (2024). A new massive multilingual dataset for high-performance language technologies. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and

References III

Evaluation (LREC-COLING 2024), pages 1116–1128, Torino, Italia. ELRA and ICCL.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. Computational Linguistics, 47(2):255–308.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., and Gardner, M. (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1286–1305.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. (2020). The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027.

References IV

- Imani, A., Lin, P., Kargaran, A. H., Severini, S., Jalili Sabet, M., Kassner, N., Ma, C., Schmid, H., Martins, A., Yvon, F., and Schütze, H. (2023). Glot500: Scaling multilingual corpora and language models to 500 languages. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.
- Ji, S., Li, Z., Paul, I., Paavola, J., Lin, P., Chen, P., O'Brien, D., Luo, H., Schütze, H., Tiedemann, J., and Haddow, B. (2024). EMMA-500: Enhancing massively multilingual adaptation of large language models. arXiv preprint arXiv:2409.17892.

References V

- Kudugunta, S., Caswell, I., Zhang, B., Garcia, X., Xin, D., Kusupati, A., Stella, R., Bapna, A., and Firat, O. (2024). Madlad-400: A multilingual and document-level large audited dataset. Advances in Neural Information Processing Systems, 36.
- Laurençon, H., Saulnier, L., Wang, T., Akiki, C., Villanova del Moral, A., Le Scao, T., Von Werra, L., Mou, C., González Ponferrada, E., Nguyen, H., et al. (2022). The bigscience roots corpus: A 1.6 tb composite multilingual dataset. Advances in Neural Information Processing Systems, 35:31809–31826.
- Nguyen, T., Nguyen, C. V., Lai, V. D., Man, H., Ngo, N. T., Dernoncourt, F., Rossi, R. A., and Nguyen, T. H. (2024). CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language

References VI

Resources and Evaluation (LREC-COLING 2024), pages 4226–4237, Torino, Italia. ELRA and ICCL.

- Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., and Ji, H. (2017). Cross-lingual name tagging and linking for 282 languages. In Barzilay, R. and Kan, M.-Y., editors, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.
- Penedo, G., Kydlíček, H., allal, L. B., Lozhkov, A., Mitchell, M., Raffel, C., Werra, L. V., and Wolf, T. (2024). The FineWeb datasets: Decanting the web for the finest text data at scale. In The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

References VII

- Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Alobeidli, H., Cappelli, A., Pannier, B., Almazrouei, E., and Launay, J. (2023). The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data only. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, Advances in Neural Information Processing Systems, volume 36, pages 79155–79172. Curran Associates, Inc.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140):1–67.

References VIII

- Samuel, D., Kutuzov, A., Øvrelid, L., and Velldal, E. (2023). Trained on 100 million words and still in shape: BERT meets British National Corpus. In Vlachos, A. and Augenstein, I., editors, Findings of the Association for Computational Linguistics: EACL 2023, pages 1954–1974, Dubrovnik, Croatia. Association for Computational Linguistics.
- Soldaini, L., Kinney, R., Bhagia, A., Schwenk, D., Atkinson, D., Authur, R., Bogin, B., Chandu, K., Dumas, J., Elazar, Y., Hofmann, V., Jha, A., Kumar, S., Lucy, L., Lyu, X., Lambert, N., Magnusson, I., Morrison, J., Muennighoff, N., Naik, A., Nam, C., Peters, M., Ravichander, A., Richardson, K., Shen, Z., Strubell, E., Subramani, N., Tafjord, O., Walsh, E., Zettlemoyer, L., Smith, N., Hajishirzi, H., Beltagy, I., Groeneveld, D., Dodge, J., and Lo, K. (2024). Dolma: an open corpus of three trillion tokens for language model pretraining research. In Ku, L.-W., Martins, A., and Srikumar, V., editors, Proceedings of the 62nd Annual Meeting

References IX

of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Suárez, P. J. O., Sagot, B., and Romary, L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7). Leibniz-Institut für Deutsche Sprache.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 483–498, Online. Association for Computational Linguistics.