Some Background on What Brought us Here

Stephan Oepen, University of Oslo

HPLT & NLPL Winter School, February 4, 2025

NLPL, HPLT, LUMI, OpenEuroLLM

1. Warm-Up: Select Historical Musings (Stephan Oepen)
2. Common Crawl vs. Internet Archive (Nikolay Arefev)
3. FineWeb-Style Ablation Studies (Farrokh Mehryary, Elaine Zosa)
4. LLM Evaluation for Norwegian (Vladislav Mikhailov, David Samuel)

NLPL
Nordic Language Processing Laboratory

NLPL

Nordic Language Processing Laboratory

Network of language technology researchers in Northern Europe;
six university research groups (Denmark, Finland, Sweden, Norway);
national e-infrastructure providers in Finland and Norway;
allocations on Abel and Taito; discipline-specific software & data;
funding from NeIC, matching in-kind contributions from all partners.

So, What's in it for me?

Collaboration Infrastructure

- Distributed team of 25 or so (very) part-timers; mostly a self-help initiative;
- cross-border sharing: everyone can get access to same two superclusters;
- HPC best practices: teaching each other, and also the general support staff.

So, What's in it for me?

Collaboration Infrastructure

- Distributed team of 25 or so (very) part-timers; mostly a self-help initiative;
- cross-border sharing: everyone can get access to same two superclusters;
- HPC best practices: teaching each other, and also the general support staff.

Virtual Laboratory

- Community-maintained repository of discipline-specific software and data;
- modularity, interoperability, uniformity, reproducibilty: `modules` setup;
- common (large) data sets: corpora, embeddings, parsing, translation, ...

So, What's in it for me?

Collaboration Infrastructure

- Distributed team of 25 or so (very) part-timers; mostly a self-help initiative;
- cross-border sharing: everyone can get access to same two superclusters;
- HPC best practices: teaching each other, and also the general support staff.

Virtual Laboratory

- Community-maintained repository of discipline-specific software and data;
- modularity, interoperability, uniformity, reproducibilty: `modules` setup;
- common (large) data sets: corpora, embeddings, parsing, translation, ...

Meeting Place

- Kick-off meeting (2017); Annual winter school; maybe NoDaLiDa workshop.

Community Formation: Annual NLPL Winter Schools

Is the end of academic NLP research in sight?

A discussion moderated by Marco Kuhlmann and Joakim Nivre

With contributions from Ivan Vulić, Emily M. Bender, and Oskar Holmström

Scenario 1: Back to the ivory tower

Ivan Vulić

Academic NLP research in 2050 is confined to research topics that are uninteresting to big tech companies. This includes the use of NLP to understand human language – what some people used to call "computational linguistics", as opposed to NLP – as well as practical applications of NLP under commercially non-viable conditions, such as historical language processing and language technology support for endangered languages.

Scenario 2: NLP as a social science

Emily Bender

Academic NLP research in 2050 is primarily concerned with understanding the application of large language models (and other AI artifacts invented since 2023) in society, partly from a technological perspective but mostly from sociological, psychological and philosophical perspectives. NLP in academia has become a truly interdisciplinary endeavor and most academic NLP groups are now based in social science faculties.

Scenario 3: Return of the Jedi

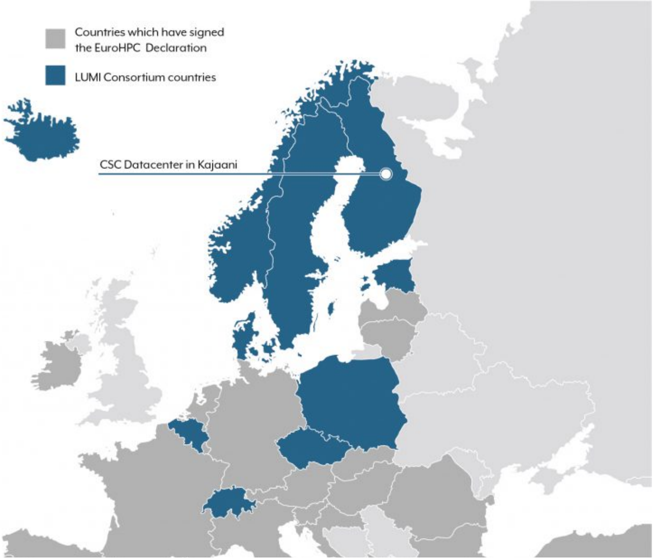
Oskar Holmström

In 2050, the development of new models and algorithms in NLP is dominated by research groups in academia, with big tech companies suffering brain drain as a result. This development was triggered by two important events: the Open AI Act adopted by the United Nations in 2032, requiring all organizations that develop AI models to share both models and training data, and the Universal Turing Machine, the world's largest computer center, sometimes referred to as the CERN of AI, co-founded and jointly owned by all the universities in the world.

LUMI — “The Queen of the North”

https://www.lumi-supercomputer.eu/

Many of us are Members of the LUMI Consortium



LUMI: BERT in an Hour, GPT in a Week

David Samuel and Risto Luukkonen

High Performance
Language Technologies

HPLT Data Sources:

Internet Archive vs. Common Crawl

Nikolay Arefyev, Andrey Kutuzov, Stephan Oepen
University of Oslo

Volunteers who inspected data

Laurie Marta Proyag Ona David Stephan Erik Barry Sampo Bhavitvya Hanna-Mari Nikita Otto Petter Maryam Mateusz Nikolay Jindra Arnisa Tsz Kin Pavel Risto

HPLT v2 Crawl Sources

4.45 PB of crawls (compressed WARCs):

- years 2012-2023
- 18% from CC, 82% from IA

Compare contributions of different crawls to our monolingual datasets:

- the amount of text extracted
- the quality of these texts

Final goal: select additional crawls for HPLT v3!

Name	Years	WARCsize,TB
IA full crawls	**2012-2020**	**3390**
wide5	2012-2012	365
wide6	2012-2013	204
wide10	2014-2014	91
wide11	2014-2014	420
wide12	2015-2015	449
wide15	2016-2017	358
wide16	2017-2018	768
wide17	2018-2020	641
survey3	2015-2016	94
CC full crawls	**2014-2022**	**743**
CC-MAIN-2022-40	2022	83
CC-MAIN-2022-49	2022	93
10 random CC crawls	2014-2022	567
partial crawls	**2013-2023**	**317**
1% of WARCs from the rest 83 CC crawls	2013-2023	46
7% of items from IA ArchiveBot	2013-2023	271

Group of crawls

Splitted crawls **by source (ia/cc) and age (old/medium/new/recent)**. The Survey crawl and the sample from ArchiveBot – separate groups.

cc_o	CC 2013-2014
cc_m	CC 2015-2016
cc_n	CC 2017-2020
cc_r	CC 2021-2023
ia_o	IA WIDE 2012-2014
ia_m	IA WIDE 2015-2016
ia_n	IA WIDE 2017-2020
ia_survey	survey3
ia_archivebot	archivebot

Manual quality inspection

Inspected documents from the deduplicated&cleaned version.

21 languages, 4 groups: ia_o, ia_n, cc_o, cc_n (pilot study)

- 4 groups cover 52% of the whole dataset
- be careful when generalizing results beyond 4 groups or 21 languages

random samples stratified by language and group

- 50 documents per language and group ⇒ 200 documents per language
- for Russian: 150 documents per language and group ⇒ 600 document

Annotation task

Show:

- only the extracted text
- 500/500 characters from the beginning of the fist/second half of each text
- annotators didn’t know which group each text comes from

We asked to provide 3 binary labels for each example:

- porn? empty/1: if the text looks like porn put 1, otherwise leave empty
- unnatural? empty/1: if the most text looks unnatural (e.g. word lists for SEO, mostly boilerplate) put 1, otherwise leave empty
- lang correct? 0/1: always fill this field (otherwise we will not distinguish labeled and unlabeled examples), put 0 if most of the text is not in the target language, otherwise put 1.

Manual quality inspection

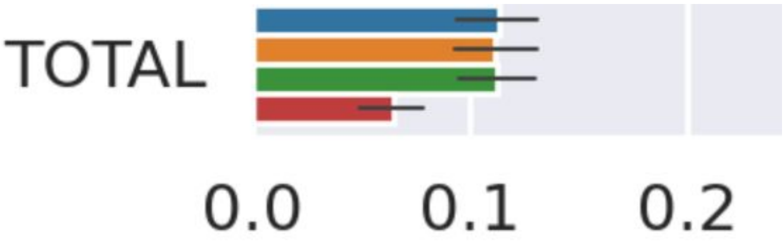
- porn? empty/1: if the text looks like porn put 1, otherwise leave empty
- unnatural? empty/1: if the most text looks unnatural (e.g. word lists for SEO, mostly boilerplate) put 1, otherwise leave empty
- lang correct? 0/1: always fill this field (otherwise we will not distinguish labeled and unlabeled examples), put 0 if most of the text is not in the target language, otherwise put 1.

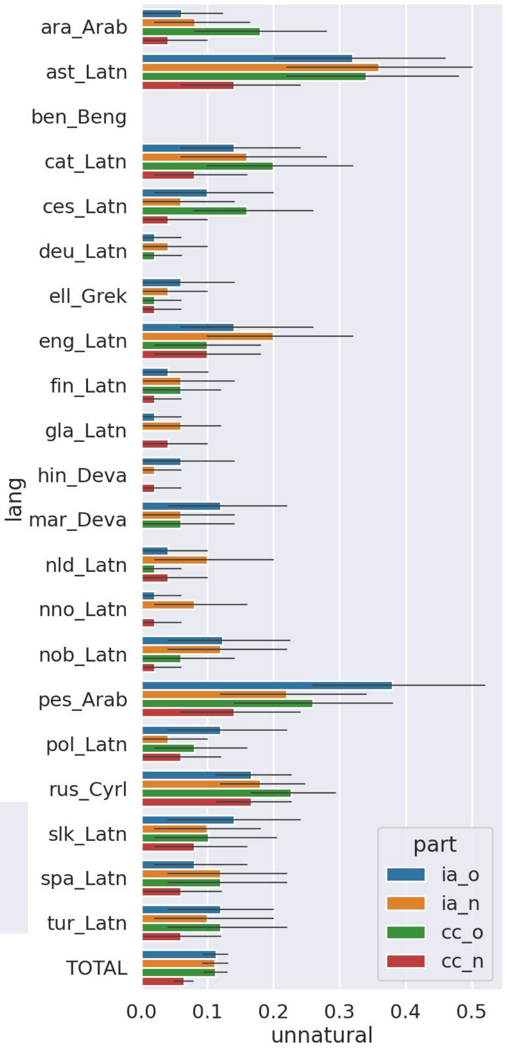
	Language Name	porn	unnatural	lang correct
1	Arabic	0 (---)	9 (5-13)	100 (---)
2	Asturian	0 (---)	28 (22-35)	69 (62-75)
3	Bengali	1 (---)	0 (---)	100 (---)
4	Catalan	0 (---)	14 (9-19)	99 (---)
5	Czech	0 (---)	9 (4-13)	100 (---)
6	Dutch	1 (---)	5 (---)	100 (---)
7	English	1 (---)	13 (8-18)	100 (---)
8	Finnish	1 (---)	4 (---)	100 (---)
9	German	1 (---)	2 (---)	98 (---)
10	Hindi	2 (---)	2 (---)	98 (---)
11	Iranian Persian	0 (---)	25 (18-31)	99 (---)
12	Marathi	0 (---)	6 (---)	97 (---)
13	Modern Greek (1453-)	0 (---)	3 (---)	100 (---)
14	Norwegian Bokmål	2 (---)	8 (4-11)	99 (---)
15	Norwegian Nynorsk	0 (---)	3 (---)	93 (---)
16	Polish	1 (---)	7 (3-11)	100 (---)
17	Russian	2 (1-3)	18 (15-21)	98 (---)
18	Scottish Gaelic	0 (---)	3 (---)	89 (85-93)
19	Slovak	0 (---)	10 (6-14)	100 (---)
20	Spanish	1 (---)	9 (5-13)	100 (---)
21	Turkish	6 (---)	10 (5-14)	99 (---)

Unnatural?

For most individual languages (among annotated) cc_n seems to give much lower prop. of unnatural texts … but within the 95% CI ⇒ no reliable conclusions for individual languages. But if we merge all annotated data together ⇒ the difference is stat. sign.

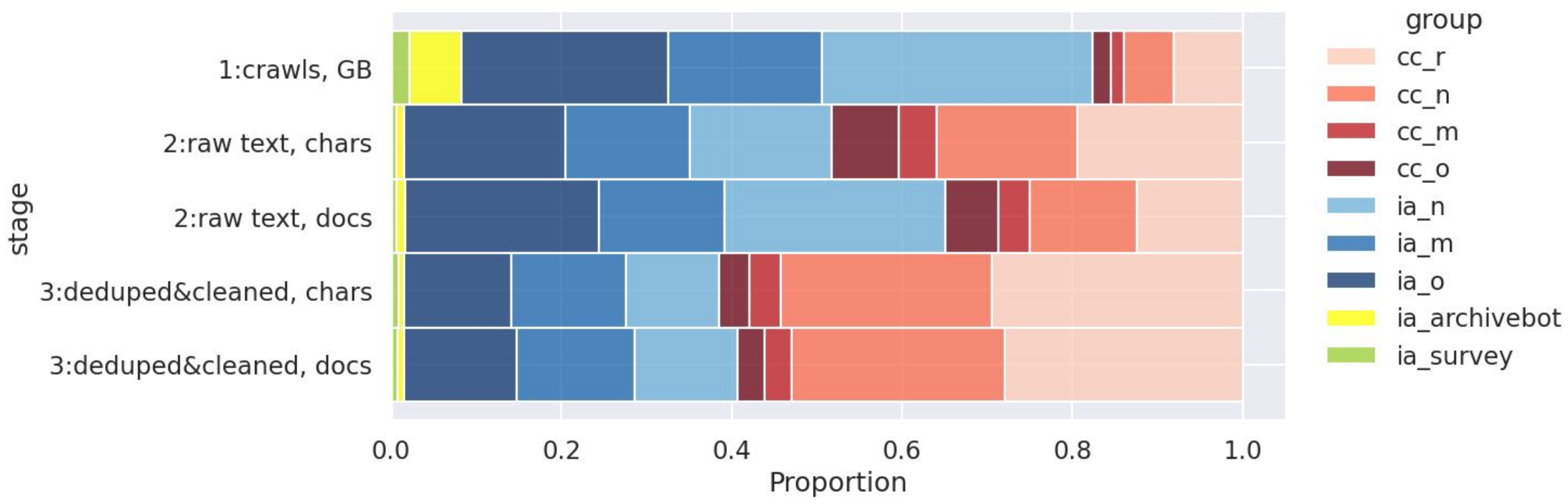
⇒ given a random language (among 21 annotated) the prob. of a random document from cc_n to be unnatural (from the naive human point of view) is lower compared to the other 3 groups.





Proportions of data from different crawls

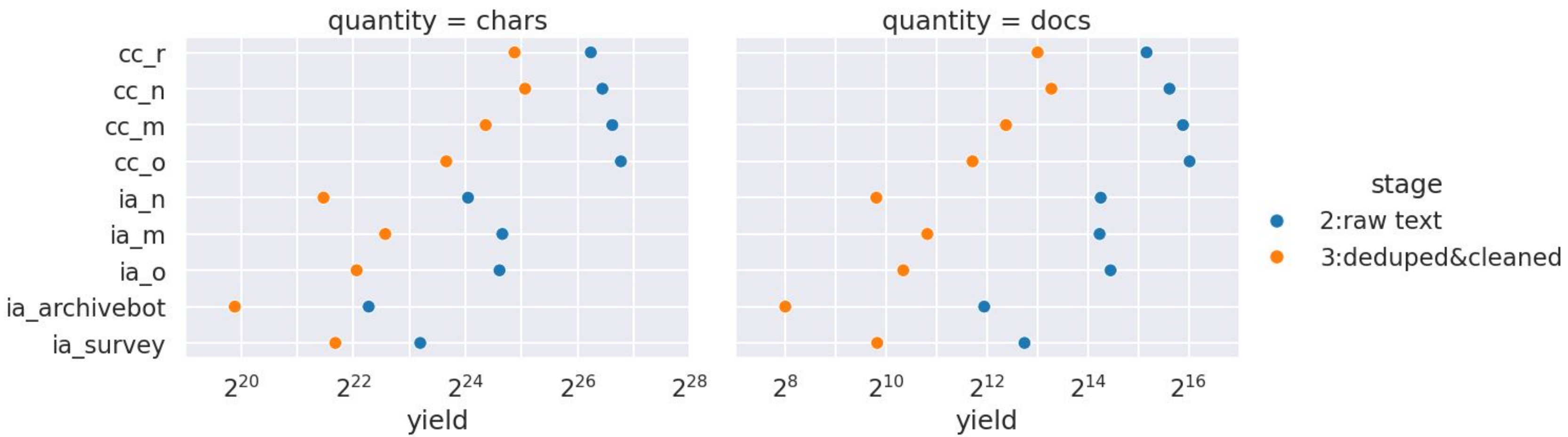
CC contribution is much higher:
~20% of source crawls give ~60% of final texts (measure in chars or docs)



Yields of different crawls

Yields from the new and recent CC crawls (2017 and later) are

- 2-3x larger than the old CC crawls,
- 4-8x larger than most IA crawls
- 32x larger than the IA ArchiveBot crawl



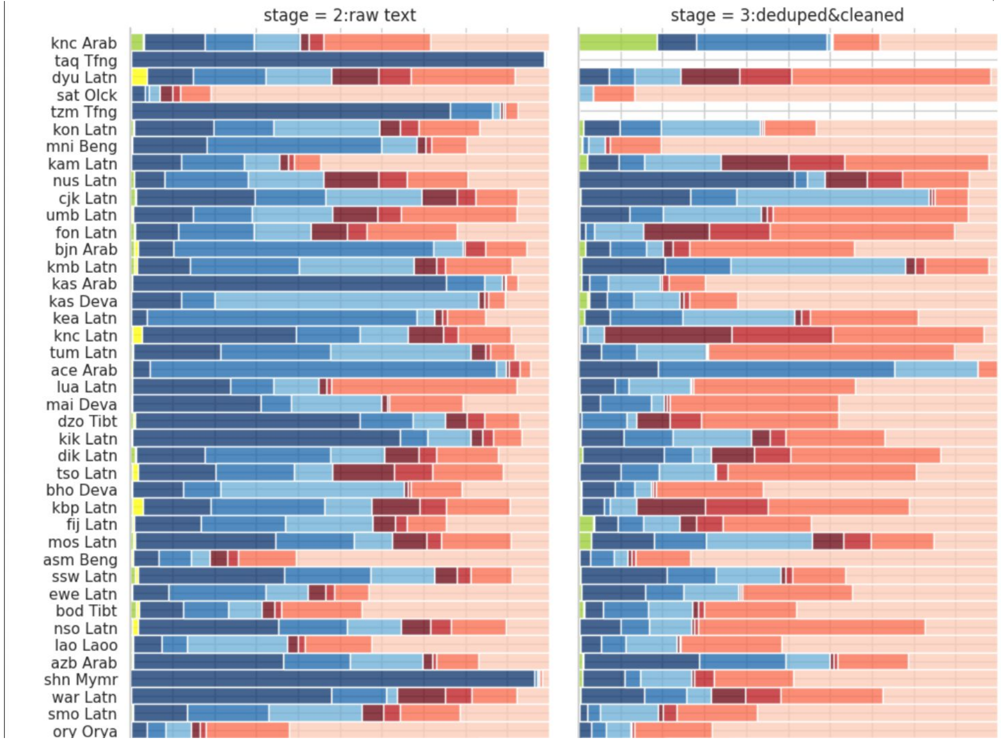
Chars / docs per 1 GB of raw compressed web crawls (WARC files)

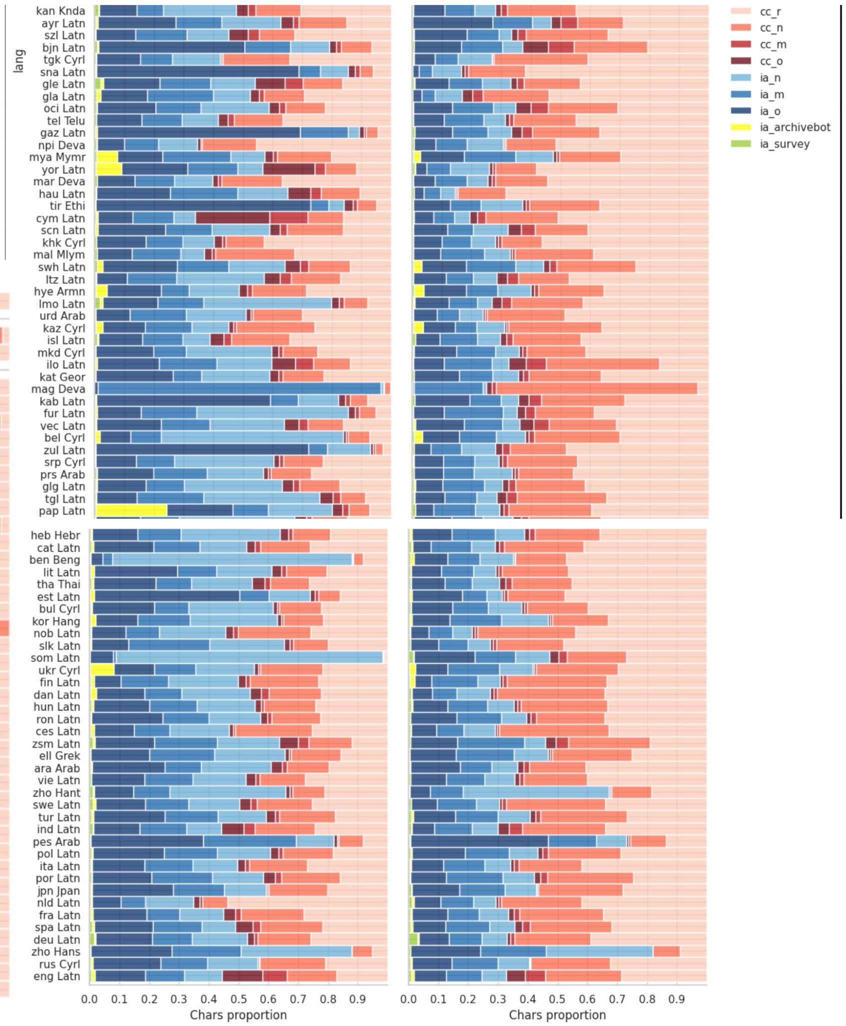
Looks like IA gives much fewer texts with a higher proportion of unnatural texts than the new CC crawls.

Ideally: take all CC and all IA, improve filtering⇒extract only clean data from everything

Limited budget: just throw IA away and use more CC crawls? Or maybe IA still contributes a lot for some of our languages?

Smallest (left), largest (right bottom), intermediate (right top) langs



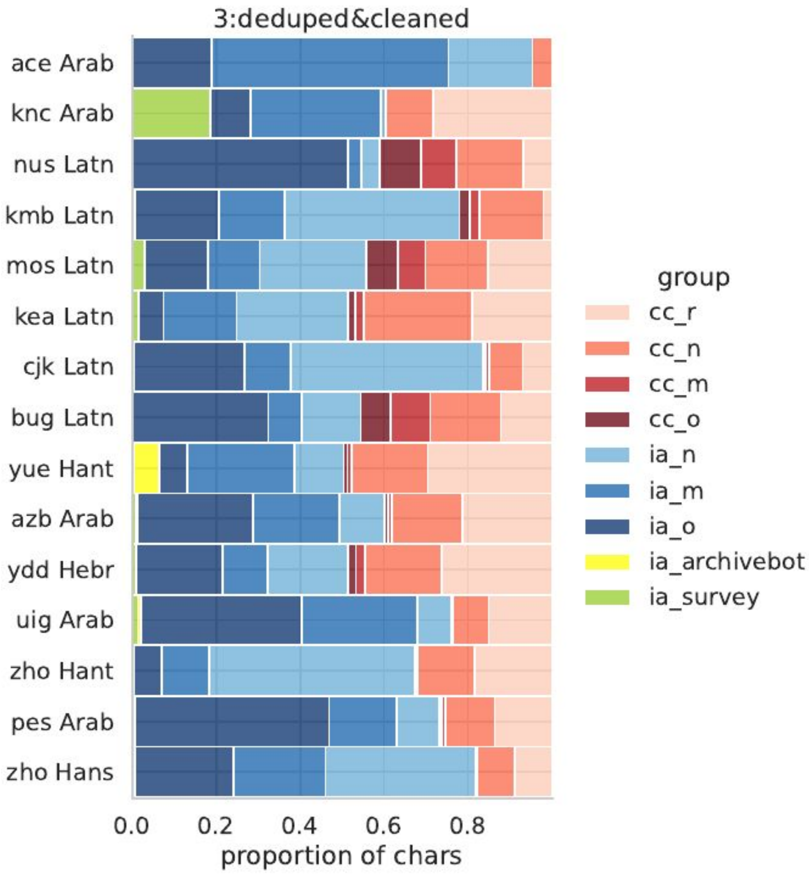


15 languages with the largest contribution of IA

Deduplication&cleaning shift the proportions in favour of CC.

E.g.: langs with >70% of texts from IA:

- 49 langs before dedup&cleaning
- 7 langs after



Conclusions

Quality vs. source crawls.

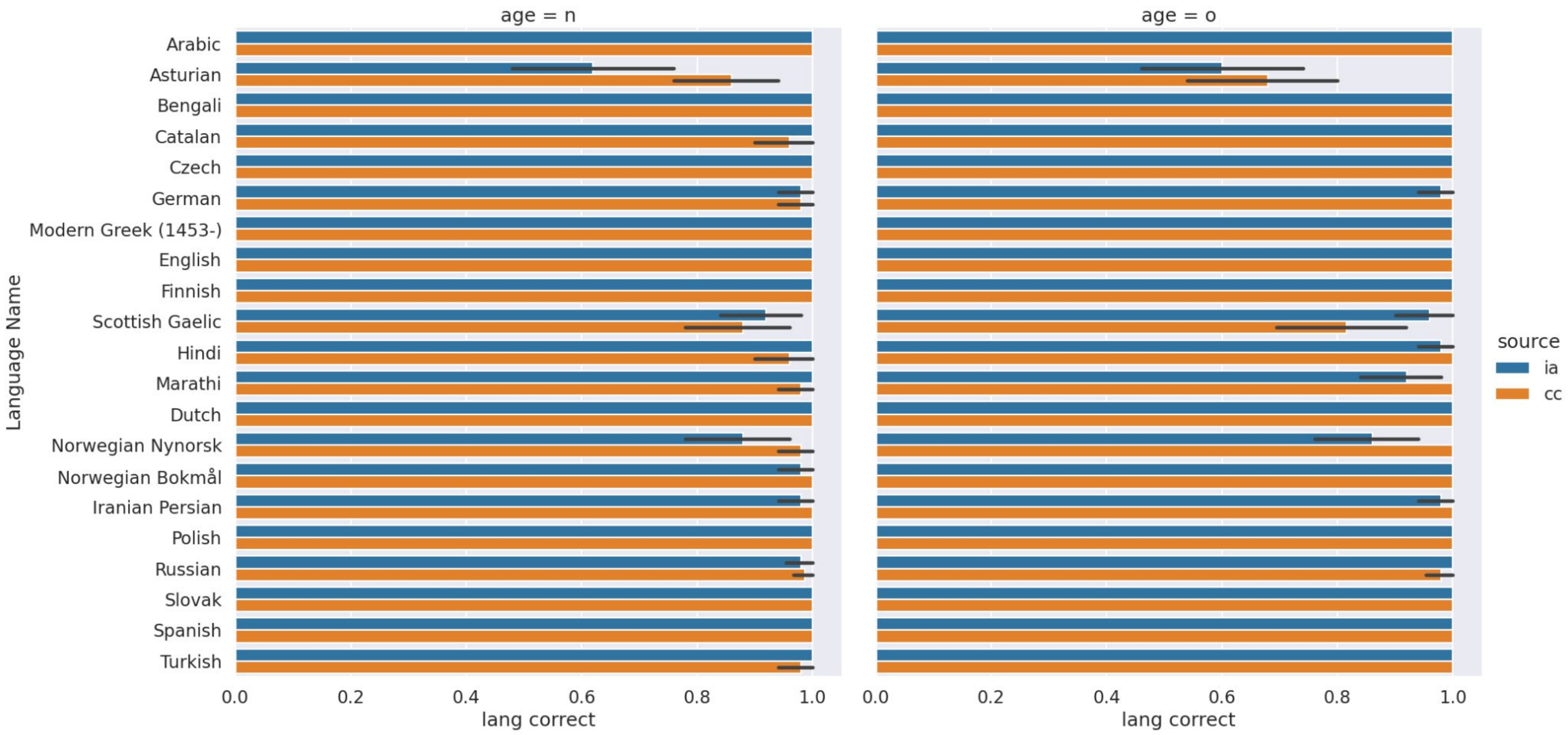
For the 21 inspected language:

1. New CC crawls (2017-2020) give ~2x lower proportion of unnatural texts compared to old CC crawls (2012-2014) and both old and new IA crawls.
2. Low proportion of LID errors for most inspected languages (except for Norwegian Nynorsk, Auturian, Scottish Gaelic). For Low proportion of porn. Couldn't observe consistent dependencies from the source crawls.

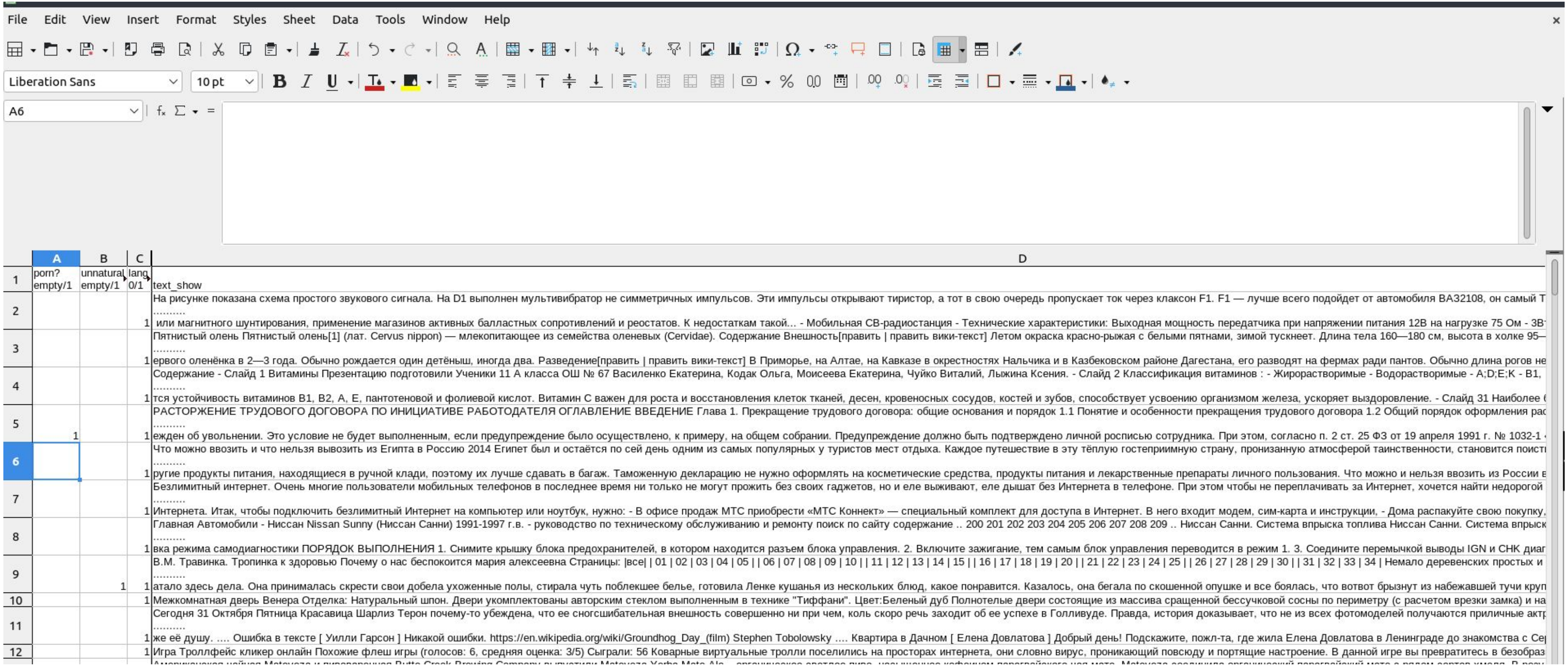
Quantity vs. source crawls.

1. Yields from new CC crawls are 2-3x larger than old CC crawls, 4-8x larger than most IA crawls (32x larger than the IA ArchiveBot crawl).
2. For some languages IA contributes a lot of texts.

Correct language?



Labeling interface



	A	B	C	D
1	porn? empty/1	unnatural empty/1	lang 0/1	text_show
2			1	На рисунке показана схема простого звукового сигнала. На D1 выполнен мультивибратор не симметричных импульсов. Эти импульсы открывают тиристор, а тот в свою очередь пропускает ток через клаксон F1. F1 — лучше всего подойдет от автомобиля ВА32108, он самый Т .... или магнитного шунтирования, применение магазинов активных балластных сопротивлений и реостатов. К недостаткам такой... - Мобильная СВ-радиостанция - Технические характеристики: Выходная мощность передатчика при напряжении питания 12В на нагрузке 75 Ом - 3В
3			1	Пятнистый олень Пятнистый олень[1] (лат. Cervus nippon) — млекопитающее из семейства оленевых (Cervidae). Содержание Внешность[править | править вики-текст] Летом окраска красно-рыжая с белыми пятнами, зимой тускнеет. Длина тела 160—180 см, высота в холке 95— .... ервого оленёнка в 2—3 года. Обычно рождается один детёныш, иногда два. Разведение[править | править вики-текст] В Приморье, на Алтае, на Кавказе в окрестностях Нальчика и в Казбековском районе Дагестана, его разводят на фермах ради пантов. Обычно длина рогов не
4			1	Содержание - Слайд 1 Витамины Презентацию подготовили Ученики 11 А класса ОШ № 67 Василенко Екатерина, Кодак Ольга, Моисеева Екатерина, Чуйко Виталий, Лыжина Ксения. - Слайд 2 Классификация витаминов : - Жирорастворимые - Водорастворимые - A;D;E;K - B1, ... тся устойчивость витаминов B1, B2, A, E, пантотеновой и фолиевой кислот. Витамин С важен для роста и восстановления клеток тканей, десен, кровеносных сосудов, костей и зубов, способствует усвоению организмом железа, ускоряет выздоровление. - Слайд 31 Наиболее
5	1		1	РАСТОРЖЕНИЕ ТРУДОВОГО ДОГОВОРА ПО ИНИЦИАТИВЕ РАБОТОДАТЕЛЯ ОГЛАВЛЕНИЕ ВВЕДЕНИЕ Глава 1. Прекращение трудового договора: общие основания и порядок 1.1 Понятие и особенности прекращения трудового договора 1.2 Общий порядок оформления рас .... ежден об увольнении. Это условие не будет выполненным, если предупреждение было осуществлено, к примеру, на общем собрании. Предупреждение должно быть подтверждено личной росписью сотрудника. При этом, согласно п. 2 ст. 25 ФЗ от 19 апреля 1991 г. № 1032-1
6			1	Что можно ввозить и что нельзя вывозить из Египта в Россию 2014 Египет был и остаётся по сей день одним из самых популярных у туристов мест отдыха. Каждое путешествие в эту тёплую гостеприимную страну, пронизанную атмосферой таинственности, становится поисти .... ругие продукты питания, находящиеся в ручной клади, поэтому их лучше сдавать в багаж. Таможенную декларацию не нужно оформлять на косметические средства, продукты питания и лекарственные препараты личного пользования. Что можно и нельзя ввозить из России в
7			1	Безлимитный интернет. Очень многие пользователи мобильных телефонов в последнее время ни только не могут прожить без своих гаджетов, но и еле выживают, еле дышат без Интернета в телефоне. При этом чтобы не переплачивать за Интернет, хочется найти недорогой .... Интернета. Итак, чтобы подключить безлимитный Интернет на компьютер или ноутбук, нужно: - В офисе продаж МТС приобрести «МТС Коннект» — специальный комплект для доступа в Интернет. В него входит модем, сим-карта и инструкции, - Дома распакуйте свою покупку,
8			1	Главная Автомобили - Ниссан Nissan Sunny (Ниссан Санни) 1991-1997 г.в. - руководство по техническому обслуживанию и ремонту поиск по сайту содержание .. 200 201 202 203 204 205 206 207 208 209 .. Ниссан Санни. Система впрыска топлива Ниссан Санни. Система впрыск .... вка режима самодиагностики ПОРЯДОК ВЫПОЛНЕНИЯ 1. Снимите крышку блока предохранителей, в котором находится разъем блока управления. 2. Включите зажигание, тем самым блок управления переводится в режим 1. 3. Соедините перемычкой выводы IGN и CHK диаг
9		1	1	В.М. Травинка. Тропинка к здоровью Почему о нас беспокоится мария алексеевна Страницы: |все| | 01 | 02 | 03 | 04 | 05 | | 06 | 07 | 08 | 09 | 10 | | 11 | 12 | 13 | 14 | 15 | | 16 | 17 | 18 | 19 | 20 | | 21 | 22 | 23 | 24 | 25 | | 26 | 27 | 28 | 29 | 30 | | 31 | 32 | 33 | 34 | Немало деревенских простых и .... атало здесь дела. Она принималась скрести свои добела ухоженные полы, стирала чуть поблекшее белье, готовила Ленке кушанья из нескольких блюд, какое понравится. Казалось, она бегала по скошенной опушке и все боялась, что вотвот брызнут из набежавшей тучи круп
10			1	Межкомнатная дверь Венера Отделка: Натуральный шпон. Двери укомплектованы авторским стеклом выполненным в технике "Тиффани". Цвет:Беленый дуб Полнотелые двери состоящие из массива сращенной бессучковой сосны по периметру (с расчетом врезки замка) и на
11			1	Сегодня 31 Октября Пятница Красавица Шарлиз Терон почему-то убеждена, что ее сногсшибательная внешность совершенно ни при чем, коль скоро речь заходит об ее успехе в Голливуде. Правда, история доказывает, что не из всех фотомоделей получаются приличные актр .... же её душу. .... Ошибка в тексте [Уилли Гарсон] Никакой ошибки. https://en.wikipedia.org/wiki/Groundhog_Day_(film) Stephen Tobolowsky .... Квартира в Дачном [Елена Довлатова] Добрый день! Подскажите, пожл-та, где жила Елена Довлатова в Ленинграде до знакомства с Се
12			1	Игра Троллфейс кликер онлайн Похожие флеш игры (голосов: 6, средняя оценка: 3/5) Сыграли: 56 Коварные виртуальные тролли поселились на просторах интернета, они словно вирус, проникающий повсюду и портящие настроение. В данной игре вы превратитесь в безобраз

Ablation study for HPLT English data

Farrokh Merhyary, Ville Komulainen, Sampo Pyysalo:
TurkuNLP, University of Turku, Finland

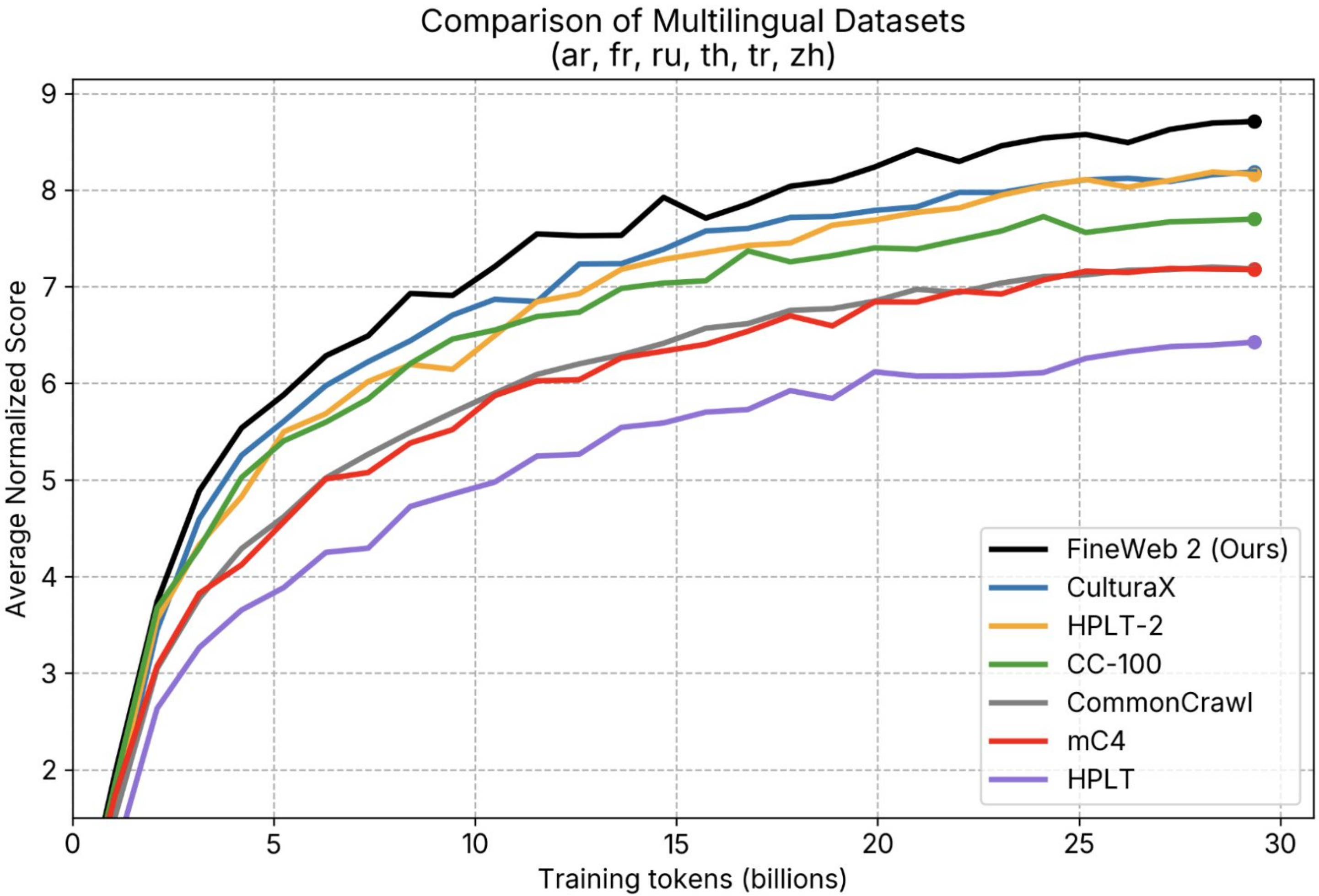
Elaine Zosa
AMD Silo AI (Silo AI)

About TurkuNLP group (https://turkunlp.org/)

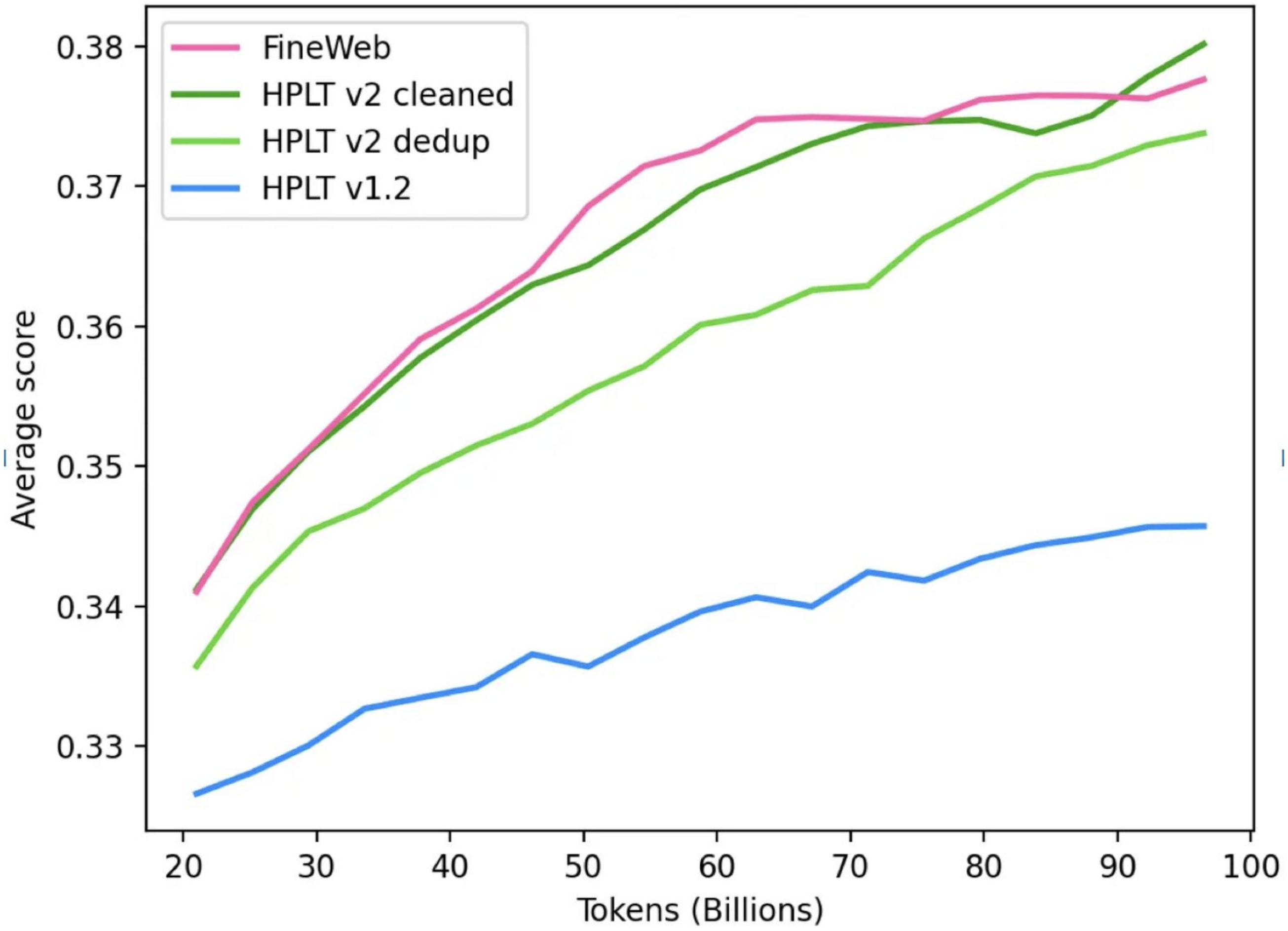
About AMD Silo AI (https://www.silo.ai/)

TurkuNLP + AMD Silo AI collaboration:

- FinBERT (TurkuNLP)
- FinGPT (TurkuNLP)
- GPT 3.5 technical report release → TurkuNLP + Silo AI (extreme scale call - CSC Lumi)

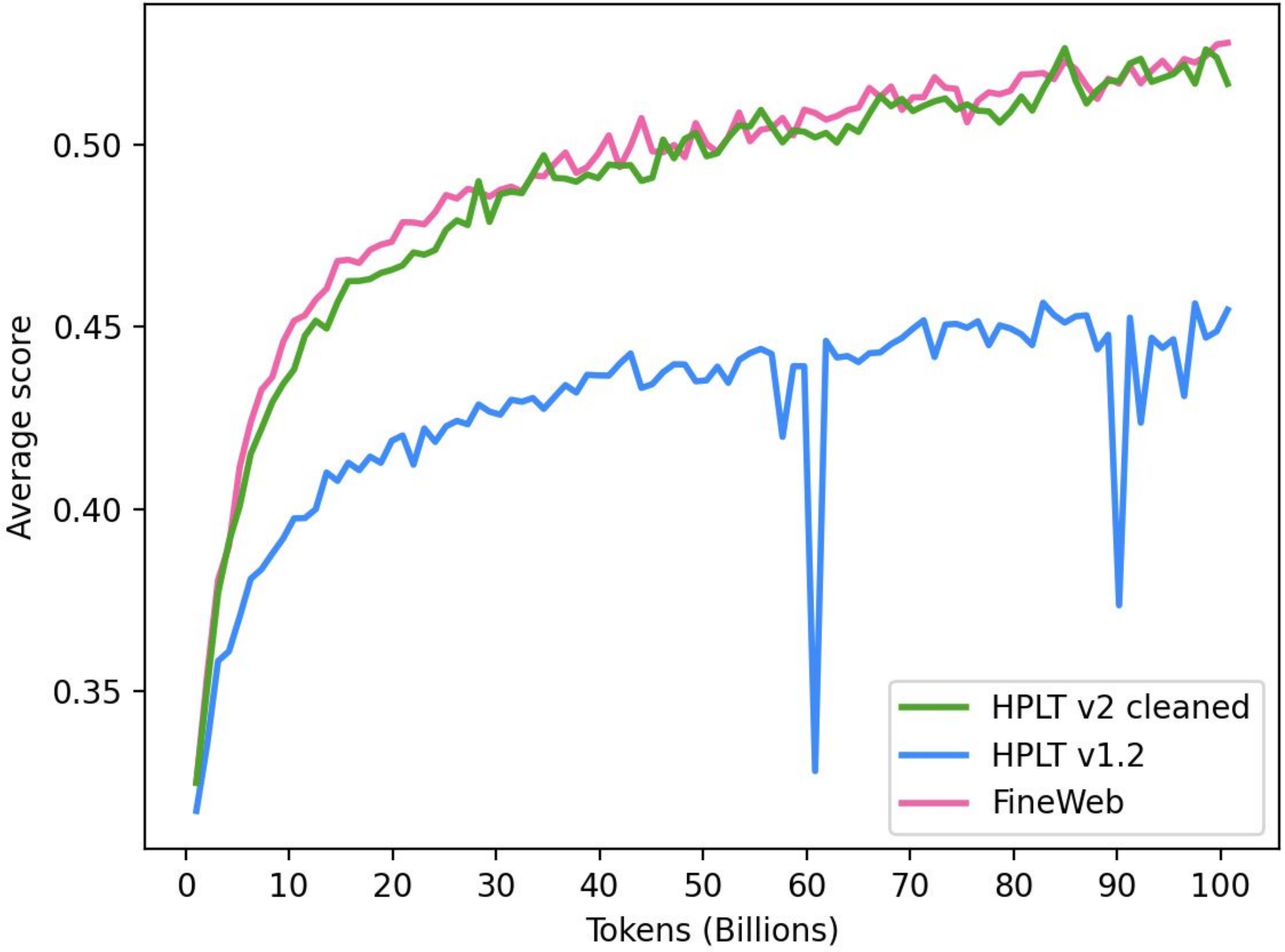


Comparison of Multilingual Datasets
(ar, fr, ru, th, tr, zh)
Average Normalized Score
Training tokens (billions)
9
8
7
6
5
4
3
2
0
5
10
15
20
25
30
FineWeb 2 (Ours)
CulturaX
HPLT-2
CC-100
CommonCrawl
mC4
HPLT



GPT-NeoX framework on 8 nodes on the LUMI cluster, where each node has 4 MI250X GPUs.

For evaluation, we use the HuggingFace LightEval in a zero-shot setting with the tasks ARC (Easy and Challenge), Hellaswag, PICA, and OpenbookQA.



Megatron framework on 16 nodes on the LUMI cluster, where each node has 4 MI250X GPUs.

For evaluation, we use the HuggingFace LightEval in a zero-shot setting with the tasks ARC (Easy and Challenge), Hellaswag, PICA, and OpenbookQA.

About LLMs:

- Poro Model
- Viking Models
- Europa Models

Ablation studies on NorEval

Preliminary results for Norwegian

David Samuel and Vladislav Mikhailov
Language Technology Group (LTG)
University of Oslo

Background

Benchmarks for Norwegian

Text embedding evaluation

Scandinavian Embedding Benchmark (SEB)

10 tasks for Norwegian Bokmål & Nynorsk

NLU evaluation

NorBench / ScandEval

8 / 4 tasks mostly for Norwegian Bokmål

NLG evaluation

NLEBench

9 tasks mostly for Norwegian Bokmål

Background

Benchmarks for Norwegian

Text embedding evaluation

Scandinavian Embedding Benchmark (SEB)

10 tasks for Norwegian Bokmål & Nynorsk

NLU evaluation

NorBench / ScandEval

8 / 4 tasks mostly for Norwegian Bokmål

NLG evaluation

NLEBench

9 tasks mostly for Norwegian Bokmål

Limitations

(no) coverage of Norwegian Nynorsk

standard NLP tasks, with a high overlap

machine-translated data 👾

NorEval

A Norwegian language understanding and generation evaluation suite

Large-scale multi-task evaluation

Zero- and few-shot evaluation on 24 tasks across 10 categories, ranging from Norwegian-specific knowledge to rewriting

Reliable data quality

Only human-annotated, -translated, and -localized examples

Diverse evaluation design

17 novel tasks, higher coverage of Norwegian Nynorsk, and a pool of 100+ prompts

Fully open & public leaderboard

Benchmarking 20+ Norwegian language models against one another and human baselines

Ablation studies

Experimental setup

Norwegian-specific tokenizer

- We train a new tokenizer for Norwegian
  - realistic fertility, higher efficiency, no “dead” embedding vectors
- A single shared tokenizer trained on equal number of random samples from the evaluated corpora
- Byte-level BPE with 50K tokens

LM pretraining

- Separate training runs for 5 evaluated corpora:
  - HPLT v1.2
  - HPLT v2.0
  - FineWeb 2.0
  - CulturaX
  - mC4
- 1.8B Llama-like models trained on 30B tokens (a corpus is repeated if necessary)

Ablation studies

Experimental setup

Zero-shot evaluation of 150 LM checkpoints on 12 tasks using a single prompt

- Ranking sentence pairs (knowledge of the Norwegian language)
- Sentence completion (knowledge of the Norwegian language)
- Multiple-choice QA (Norwegian-specific & world knowledge, commonsense reasoning, truthfulness)
- Generative QA (machine reading comprehension)

NorCommonsenseQA (Bokmål)

`Spørsmål: {{question}}\n\nSvar:`

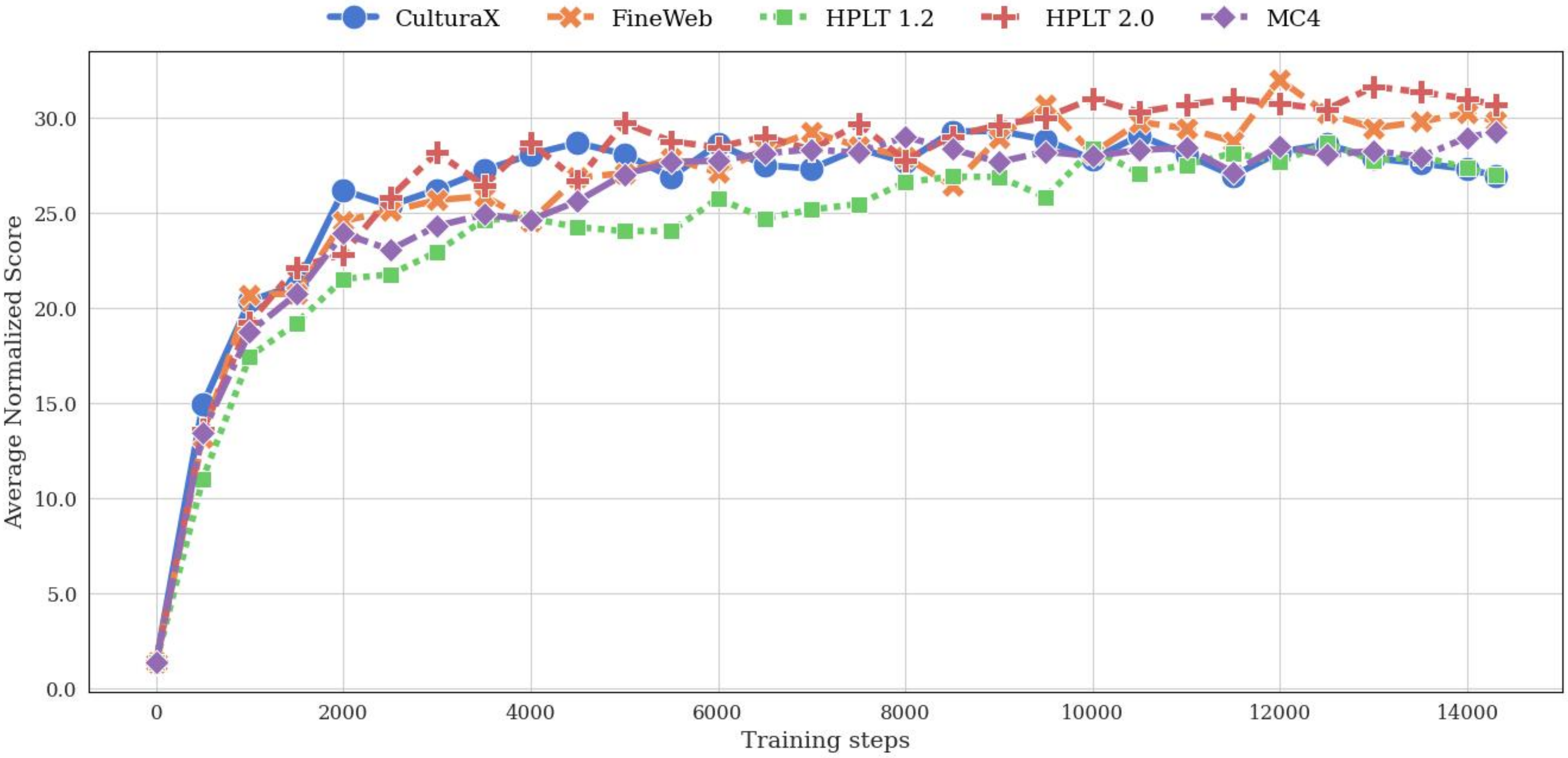
Hvis statsministeren ønsket å forby slanger, hvor ville han foreslått lovforslaget?

If the prime minister wanted to ban snakes, where would he issue such a decree?

A. *På gata* *(In the street)*
B. *I en tropisk skog* *(In a tropical rainforest)*
C. *I Edens hage* *(In the garden of Eden)*
D. *På Eidsvoll* *(At Eidsvoll)*
E. *I Stortinget* *(At the parliament)*

Ablation studies

Preliminary results



Ablation studies

Other considerations

- Prompt sensitivity — “noise”
  - There is no single best prompt for LMs, even of the same pretraining corpus composition but of different size
- Task selection sensitivity
  - What happens if we add or discard “fine” tasks, which do not pass stricter criteria choices?
- Rank aggregation methods
  - There are various aggregation methods besides Borda and multi-stage rank aggregation procedures