Enhancing Multilingual LLM Pretraining with Model-Based Data Selection

Bettina Messmer, Vinko Sabolčec, Martin Jaggi
firstname.lastname@epfl.ch
Machine Learning & Optimization Laboratory, EPFL

Data Curation Primarily Targets the English Language

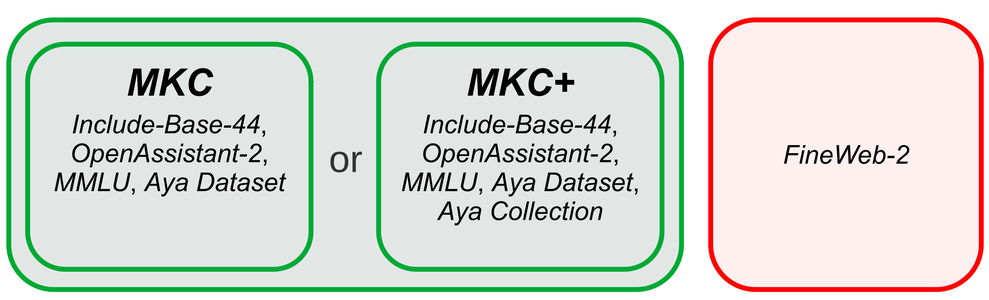
- High-quality data is a basis for strong LLMs
- Top English datasets (FineWeb-Edu and DCLM) demonstrate performance of model-based methods over heuristic methods
- However, the leading multilingual dataset (FineWeb-2) uses heuristic methods

- **Challenge:** how to scale model-based methods to other languages?

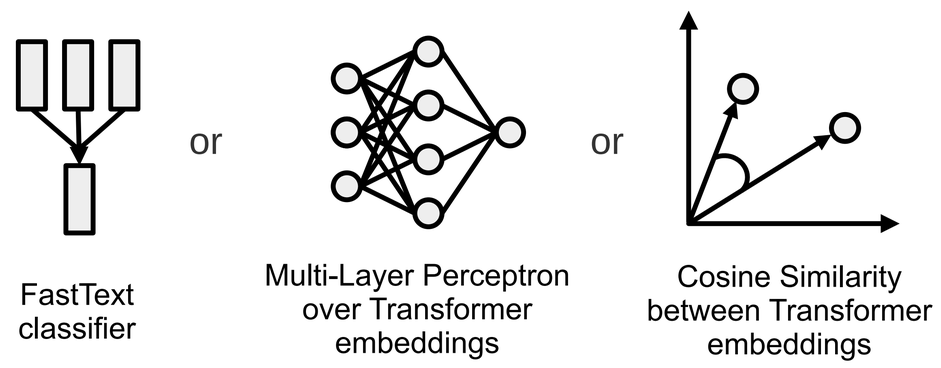
Scaling Model-Based Data Curation Across Diverse Languages

- **Goal:** identify structured and knowledge-rich documents

- Use a set of positive (Multilingual Knowledge Collection) and negative (FineWeb-2) representative samples:



- Select highest scoring data using model-based methods:



Selecting the Right Dataset and Methods

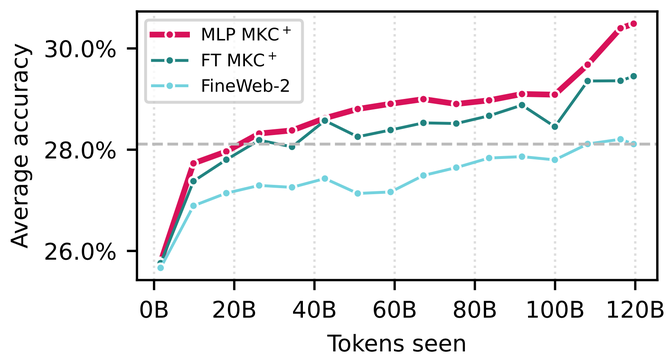
- **Left:** Multi-Layer Perceptron with *MKC+* dataset provides highest performance
- **Right:** Diverse classifier training datasets are crucial for selecting high-quality data

Approach	Average Rank
MLP MKC$^+$	4.35
MLP MKC	6.11
FT MKC$^+$	7.17
FT MKC	8.04
CS MKC	8.10
Baseline	8.72
CS MKC$^+$	8.79

Dataset	Average Rank
MKC$^+$	2.52
Aya Collection	2.91
Aya Dataset	3.17
MMLU	3.57
Baseline	4.09
OpenAssistant-2	4.53
Include-Base-44	5.42

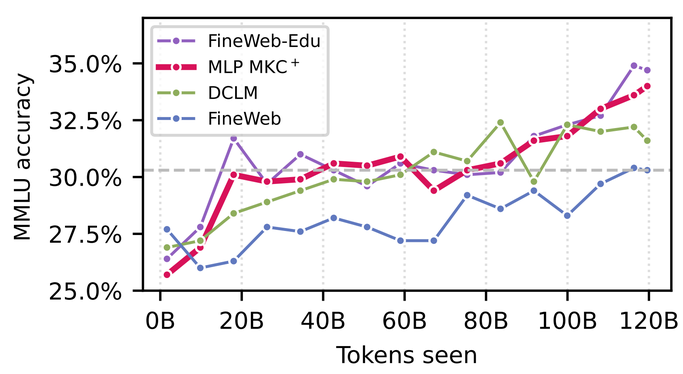
Comparing the Approach to Existing Datasets

- 6x fewer tokens needed to match FineWeb-2 performance and higher performance when fully trained



- Approach transfers to English and outperforms FineWeb-Edu and DCLM

Dataset	Ours	DCLM*	FW-Edu*	FW*
Average Rank	1.8333	2.3889	2.4444	3.3333
ARC (Challenge)	0.3550	0.3530	**0.3850**	0.3010
ARC (Easy)	0.6670	0.6470	**0.6970**	0.5880
CommonsenseQA	0.3870	**0.4100**	0.3770	0.3850
HellaSwag	**0.6040**	0.5960	0.5700	0.5930
MMLU	0.3400	0.3160	**0.3470**	0.3030
OpenBookQA	0.3860	0.3840	**0.4180**	0.3560
PIQA	0.7510	0.7510	0.7410	**0.7620**
WinoGrande	**0.5720**	0.5610	0.5660	0.5550
TriviaQA	0.0820	**0.1240**	0.0320	0.0370



Enhancing Multilingual LLM Performance

- Multilingual pretraining performance in 5 languages is higher compared to monolingual models under the same token budget

Dataset	Ours$_M$	Ours	FW-2	FW-2$_M$
Average Rank	1.8333	2.0556	3.0000	3.1111
Belebele	**0.3667**	0.3533	0.3444	0.3511
HellaSwag	0.5270	**0.5380**	0.5180	0.4970
X-CSQA	0.2740	0.2740	**0.2870**	0.2750
XNLI 2.0	**0.7660**	0.7400	0.7180	0.7330
FQuAD	**0.3212**	0.2803	0.2401	0.2459
MMLU	0.2841	**0.2895**	0.2706	0.2735
Mintaka	0.0456	0.0438	**0.0712**	0.0579
X-CODAH	**0.2900**	0.2667	0.2633	0.2567
ARC (Challenge)	0.2970	**0.3180**	0.2850	0.2670

More Ablations, Code, and Dataset in 20 Languages

Paper

Code

Dataset