DCAD-2000: A Multilingual Dataset across 2000+ Languages with Data Cleaning as Anomaly Detection

Yingli Shen[1]*, Wen Lai[2]*, Shuo Wang[1], Xueren Zhang[3]
Kangyang Luo[1], Alexander Fraser[2], Maosong Sun[1]

[1]Tsinghua University, [2]Technical University of Munich, [3]Modelbest Inc.

Motivation

- Multilingual LLMs still biased toward high-resource languages.
- Existing datasets (CulturaX, Madlad-400, MaLA, GlotCC, Fineweb-2) have issues:
  1. Outdated Common Crawl → stale knowledge.
  2. Limited coverage of medium/high-resource languages.
  3. Insufficient data cleaning → difficult to directly employ in training multilingual LLMs.
- Existing data cleaning pipeline have issues:
  1. Rely on document-level features and fixed thresholds → difficult to extend to multilingual setting.

Goal

- Build the largest and cleanest multilingual dataset for training LLMs.
- Propose a novel data cleaning pipeline that works for multilingual setting.

Data Collection

- Our dataset integrates four major sources:
  1. **MaLA Corpus:** Covers 939 languages, aggregating data from Bloom, CC100, Glot500, and others.
  2. **FineWeb:** High-quality English web corpus (15T tokens) from Common Crawl, updated monthly.
  3. **FineWeb-2:** Multilingual extension of FineWeb covering 1,915 languages, built from 96 Common Crawl dumps (2013–Apr 2024).
  4. **New Common Crawl Data:** Freshly extracted multilingual data from May–Nov 2024 (CC-MAIN-2024-22 to CC-MAIN-2024-46).

Data Cleaning as Anomaly Detection

Reframing the problem

Traditional: manual thresholds on document features.
Ours: treat data cleaning as an *anomaly detection* task.

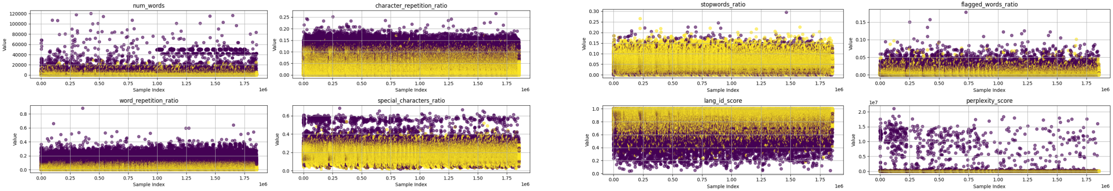
Data Cleaning as Anomaly Detection

Reframing the problem

Traditional: manual thresholds on document features.
Ours: treat data cleaning as an *anomaly detection* task.

- Extract 8 statistical features per document (length, repetition, LID score, perplexity, etc.).
- Train a language-agnostic model (Isolation Forest) to detect anomalies.
- Removes noisy / irrelevant content automatically.

Data Cleaning as Anomaly Detection

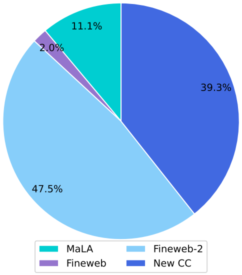


Takeaway

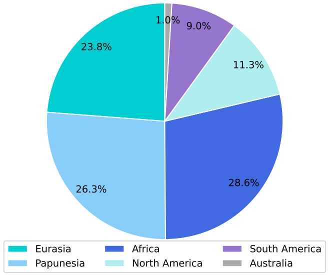
- Clear clustering distinguishes anomalous from normal data, with anomalies showing distinct patterns.
- Language ID and perplexity scores serve as key indicators of linguistic irregularities.

DCAD Analysis

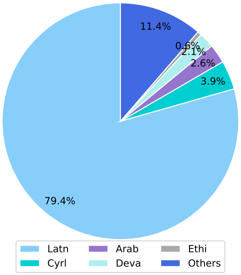
- **Total Statistics:** 2,282 languages, 46.72TB of data, and 8.63 billion documents, spanning 155 high- and medium-resource languages and 159 writing scripts.



(a) Document Distribution



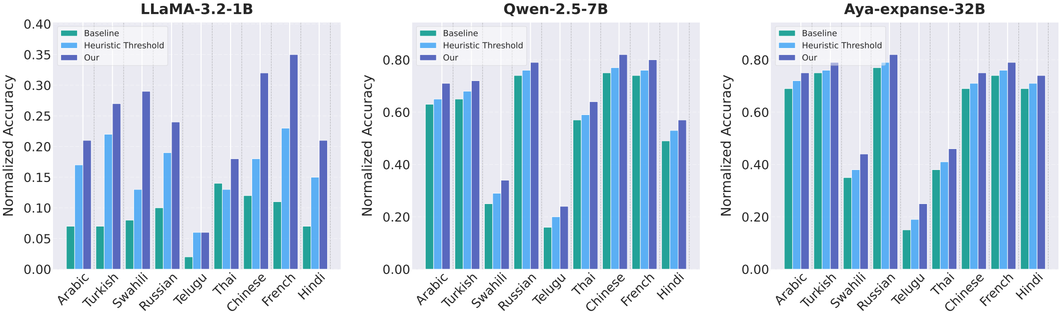
(b) Geographical Distribution



(c) Script Distribution

Effectiveness of DCAD Pipeline

- Compare with Heuristic Threshold based Cleaning Pipeline
  - Evaluate on FineTask Benchmark across LLaMA-3.2-1B / Qwen-2.5-7B / Aya-expanse-32B



Evaluation

Comparison of Anomaly Detection Algorithms

- Anomaly Detection Algorithm using Isolation Forest, One-Class SVM, Local Outlier Factor and K-Means

	LLaMA-3.2-1B					Qwen-2.5-7B					Aya-expanse-32B				
	Baseline	**Iso_Forest**	**OC_SVM**	**LOF**	**K-Means**	**Baseline**	**Iso_Forest**	**OC_SVM**	**LOF**	**K-Means**	**Baseline**	**Iso_Forest**	**OC_SVM**	**LOF**	**K-Means**
Arabic	0.07	**0.21**	0.18	**0.21**	0.14	0.63	**0.71**	0.68	0.65	0.68	0.69	**0.75**	0.70	0.71	0.69
Turkish	0.07	0.27	**0.29**	0.17	0.15	0.65	0.72	**0.73**	0.67	0.68	0.75	**0.79**	0.77	0.76	0.77
Swahili	0.08	**0.29**	0.25	0.19	0.19	0.25	0.34	0.27	**0.35**	0.27	0.35	**0.44**	0.36	0.37	0.41
Russian	0.10	**0.24**	0.19	0.18	0.15	0.74	**0.79**	0.75	0.75	0.76	0.77	**0.82**	0.79	0.80	0.79
Telugu	0.02	**0.06**	0.05	0.04	0.04	0.16	0.24	**0.26**	0.20	0.21	0.15	0.25	0.19	0.21	**0.27**
Thai	0.14	**0.21**	0.18	0.18	0.15	0.57	**0.64**	0.59	0.59	0.61	0.38	**0.46**	0.42	0.43	0.40
Chinese	0.12	**0.32**	0.28	0.25	0.21	0.75	**0.82**	0.77	0.76	0.78	0.69	**0.75**	0.71	0.71	0.73
French	0.11	0.35	**0.37**	0.30	0.23	0.74	**0.80**	0.76	0.76	0.75	0.74	**0.79**	0.76	0.76	0.76
Hindi	0.07	**0.21**	0.17	0.16	0.14	0.49	**0.57**	0.52	0.53	0.52	0.69	**0.74**	0.72	0.73	0.72

Evaluation

Comparison with Other Multilingual Datasets

- Compare with Fineweb-2 / New CC / DCAD-2000
- Evaluate on SIB-200 / Glot500 / FlORES-200

		LLaMA-3.2-1B			**Qwen-2.5-7B**			**Aya-expanse-32B**		
		Fineweb-2	New CC	DCAD-200	Fineweb-2	New CC	DCAD-200	Fineweb-2	New CC	DCAD-200
SIB-200 (↑)	H	8.24	8.86	**10.37** ↑2.13	33.41	34.53	**38.26** ↑4.85	41.72	42.41	**47.93** ↑6.21
	M	7.31	7.92	**9.15** ↑1.84	28.72	29.86	**32.65** ↑3.93	32.25	33.39	**38.16** ↑5.91
	L	6.06	6.45	**7.83** ↑1.77	23.58	24.22	**27.12** ↑3.54	26.87	27.57	**33.24** ↑6.37
	VL	3.68	4.27	**5.24** ↑1.56	13.25	15.43	**21.57** ↑8.32	17.23	19.5	**26.38** ↑9.15
Glot500-c test (↓)	H	426.37	403.58	**373.14** ↓53.23	347.21	334.18	**303.38** ↓43.83	273.85	257.24	**225.28** ↓48.57
	M	446.28	436.94	**423.75** ↓22.53	385.72	389.24	**369.15** ↓16.57	326.92	321.16	**302.53** ↓24.39
	L	503.38	493.27	**473.96** ↓29.42	426.33	419.25	**404.28** ↓22.05	372.62	367.26	**341.34** ↓31.28
	VL	584.55	569.34	**532.86** ↓51.69	479.04	463.36	**433.48** ↓45.56	396.33	392.33	**385.86** ↓10.47
FLORES-200 (↑) (Eng–X)	H	3.14	3.82	**5.26** ↑2.12	15.24	16.07	**18.47** ↑3.23	23.45	24.33	**26.33** ↑2.88
	M	2.75	2.94	**3.89** ↑1.14	12.83	13.46	**15.49** ↑2.66	19.36	20.21	**21.62** ↑2.26
	L	2.27	2.41	**3.14** ↑0.87	8.94	9.28	**10.25** ↑1.31	16.61	17.24	**18.36** ↑1.75
	VL	1.85	2.05	**2.35** ↑0.50	6.33	7.25	**9.05** ↑2.72	12.51	13.16	**14.77** ↑2.26
FLORES-200 (↑) (X–Eng)	H	3.94	3.98	**4.26** ↑0.32	16.31	16.92	**18.84** ↑2.53	23.86	24.13	**26.94** ↑3.08
	M	3.52	3.66	**3.80** ↑0.28	13.65	14.05	**16.27** ↑2.62	20.45	20.36	**22.53** ↑2.17
	L	3.05	3.12	**3.24** ↑0.19	9.47	10.22	**11.48** ↑2.01	17.67	17.82	**18.93** ↑1.26
	VL	2.73	2.83	**3.14** ↑0.41	7.28	7.81	**9.65** ↑2.37	13.25	13.56	**15.88** ↑2.63

Summary

- **DCAD-2000:** A large-scale multilingual dataset covering 2,282 languages and 159 scripts, offering broad geographic and linguistic diversity, with expanded coverage of 155 high- and medium-resource languages.
- **Framework:** We reformulate data cleaning as an *anomaly detection* task, eliminating the need for manual threshold tuning.

Dataset: https://huggingface.co/datasets/openbmb/DCAD-2000

Pipeline: https://github.com/yl-shen/DCAD-2000

Thank you!
Questions & Comments?

Paper

Code

Dataset

Contact: wen.lai@tum.de syl@mail.tsinghua.edu.cn