

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

Yingli Shen^{1*} Wen Lai^{2,3*} Shuo Wang¹ Xueren Zhang⁴ Kangyang Luo¹
Alexander Fraser^{2,3} Maosong Sun¹

¹Tsinghua University ²Technical University of Munich
³Munich Center for Machine Learning ⁴Modelbest Inc.

Highlight

- We introduce **DCAD-2000**, a large-scale multilingual corpus built from newly extracted Common Crawl data and existing datasets. It covers **2,282 languages**, **46.72 TB** of text, **8.63B documents**, and **159 writing systems**.
- Unlike traditional data-cleaning workflows, we reframe **data cleaning as anomaly detection**, enabling dynamic filtering that automatically removes noisy and anomalous content for higher-quality data.

Dataset	CC Version	#Langs (total)	#Langs (high)	#Langs (medium)	#Langs (low)	#Langs (very low)	Training-Ready
mC4 (Raffel et al., 2020)	CC-MAIN-2020-34	101	0	43	52	6	✗
OSCAR 23.01 (Abadji et al., 2022)	CC-MAIN-2022-49	153	6	42	25	80	✗
Glott500 (Imani et al., 2023)	CC-MAIN-2020-34	511	0	108	79	324	✗
CulturaX (Nguyen et al., 2024)	CC-MAIN-2022-49	167	11	47	27	82	✗
Madlad-400 (Kudugunta et al., 2024)	CC-MAIN-2022-33	419	7	46	39	327	✗
MaLA (Ji et al., 2024)	CC-MAIN-2022-49	939	1	125	78	735	✗
Glottc (Kargaran et al., 2024)	CC-MAIN-2023-50	1331	0	10	52	1269	✗
HPLT-v1.2 (de Gibert et al., 2024)	CC-MAIN-2022-40	191	12	53	38	88	✗
Fineweb-2 (Penedo et al., 2024b)	CC-MAIN-2024-18	1915	10	62	49	1794	✗
DCAD-2000	CC-MAIN-2024-46	2282	13	142	124	2003	✓

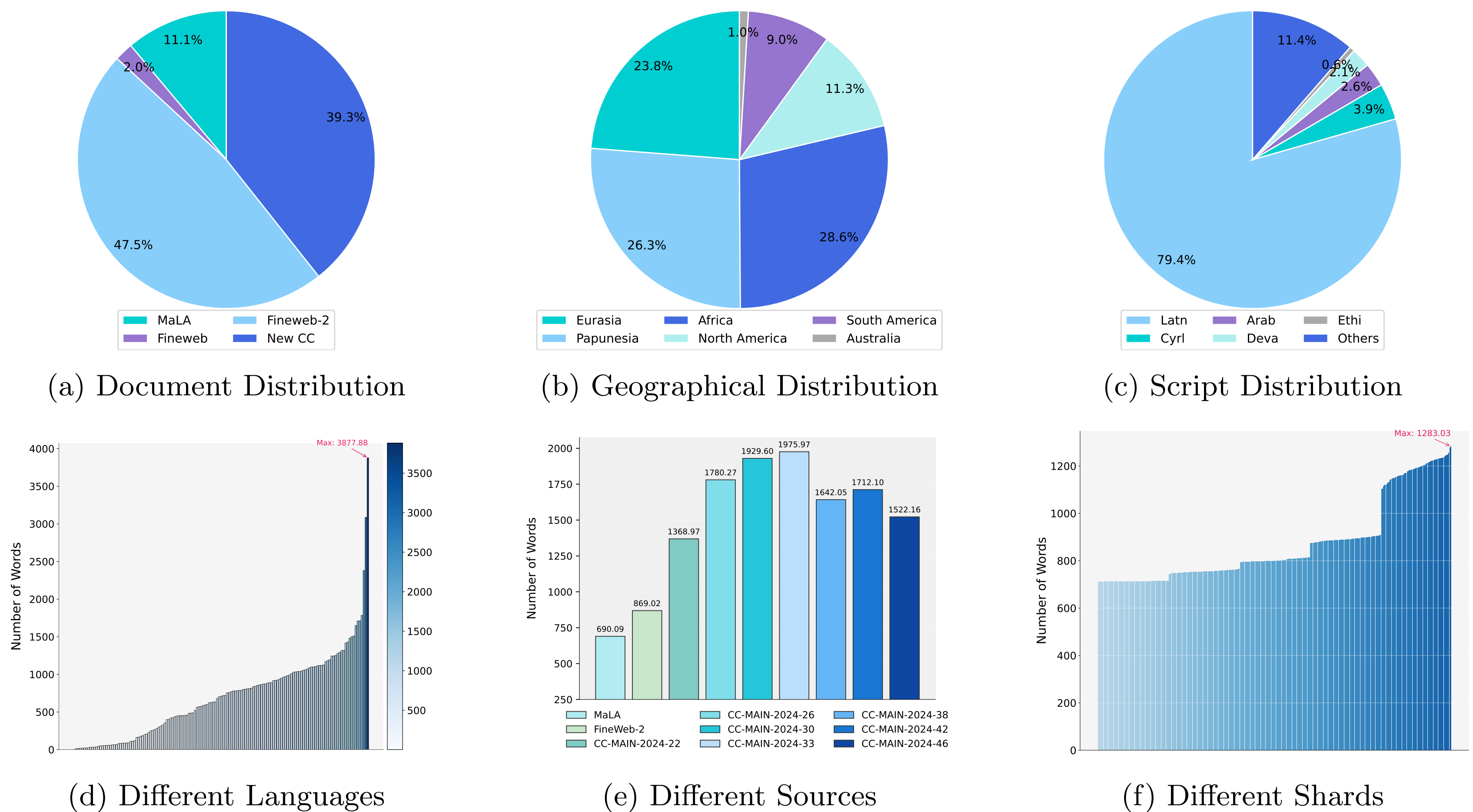
Pipeline & Dataset

GitHub: <https://github.com/yl-shen/DCAD-2000>

Hugging Face: <https://huggingface.co/datasets/openbmb/DCAD-2000>

Dataset Statistics

Distribution and Linguistic Diversity



Dataset & Pipeline

① **Data Collection:** DCAD combines multiple sources to ensure broad linguistic coverage and data freshness:

Source	Coverage	Key Features
MaLA	939 languages	Aggregates data from <i>Bloom</i> , <i>CC100</i> , and <i>Glott500</i> ; deduplicated via MinHashLSH.
Fineweb	English	15T tokens (Nov 2024 release), cleaned and filtered using <i>Datatrove</i> .
Fineweb-2	1,915 languages	Covers 96 Common Crawl dumps (2013–Apr 2024) with multilingual deduplication.
New Common Crawl	Recent data	21.54TB processed (May–Nov 2024) through the Fineweb-2 pipeline.

② **Feature Extraction:** For each document t , we extract 8 quality features

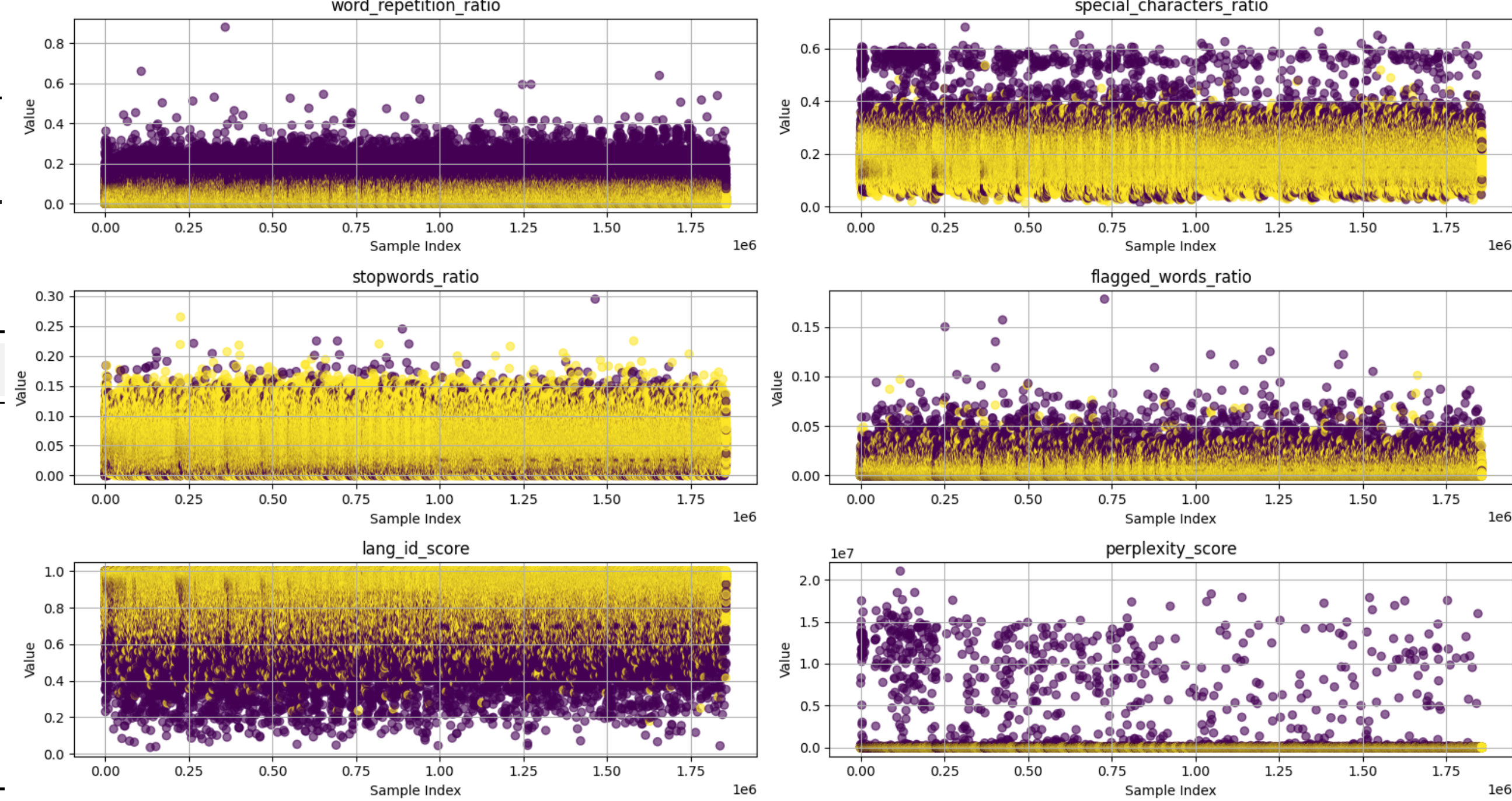
Category	Feature	Description
Structure	Word count	Detects outlier lengths.
Repetition	Character / Word repetition	Flags spam and low-information text.
Symbols	Special character ratio	Detects noisy or adversarial inputs.
Linguistic	Stopword ratio	Assesses fluency and naturalness.
Safety	Flagged word ratio	Detects toxic or profane content.
Identification	LID score	Detects mislabeled or mixed-language text.
Fluency	Perplexity score	Evaluates grammatical soundness.

③ **Anomaly Detection:** Features are standardized and evaluated via statistical anomaly detection algorithm (e.g., **Isolation Forest**)

$$f(\tilde{\mathbf{x}}) = \begin{cases} 1, & \text{keep (clean)} \\ -1, & \text{remove (anomalous)} \end{cases}$$

The result partitions data into clean ($\mathcal{X}_{\text{keep}}$) and anomalous ($\mathcal{X}_{\text{remove}}$) subsets.

④ Visualization



Dataset and Pipeline Evaluation

Evaluation on DCAD-2000

	Qwen-2.5-7B			Aya-expanse-32B		
	Fineweb-2	New CC	DCAD-200	Fineweb-2	New CC	DCAD-200
SIB-200 (↑)						
H	33.41	34.53	38.26 ↑4.85	41.72	42.41	47.93 ↑6.21
M	28.72	29.86	32.65 ↑3.93	32.25	33.39	38.16 ↑5.91
L	23.58	24.22	27.12 ↑3.54	26.87	27.57	33.24 ↑6.37
VL	13.25	15.43	21.57 ↑8.32	17.23	19.50	26.38 ↑9.15
Glott500-c test (↓)						
H	347.21	334.18	303.38 ↓43.83	273.85	257.24	225.28 ↓48.57
M	385.72	389.24	369.15 ↓16.57	326.92	321.16	302.53 ↓24.39
L	426.33	419.25	404.28 ↓22.05	372.62	367.26	341.34 ↓31.28
VL	479.04	463.36	433.48 ↓45.56	396.33	392.33	385.86 ↓10.47
FLORES-200 (↑) – Eng→X						
H	15.24	16.07	18.47 ↑3.23	23.45	24.33	26.33 ↑2.88
M	12.83	13.46	15.49 ↑2.66	19.36	20.21	21.62 ↑2.26
L	8.94	9.28	10.25 ↑1.31	16.61	17.24	18.36 ↑1.75
VL	6.33	7.25	9.05 ↑2.72	12.51	13.16	14.77 ↑2.26
FLORES-200 (↑) – X→Eng						
H	16.31	16.92	18.84 ↑2.53	23.86	24.13	26.94 ↑3.08
M	13.65	14.05	16.27 ↑2.62	20.45	20.36	22.53 ↑2.17
L	9.47	10.22	11.48 ↑2.01	17.67	17.82	18.93 ↑1.26
VL	7.28	7.81	9.65 ↑2.37	13.25	13.56	15.88 ↑2.63

Manual Evaluation on DCAD

Language	Retained Documents (Kept by filter)				Deleted Documents (Removed by filter)			
	Good	Borderline	Bad	Residual Noise	Good	Borderline	Bad	False Positives
English	86%	10%	4%	4%	5%	14%	81%	5%
Chinese	82%	13%	5%	5%	6%	18%	76%	6%
German	84%	12%	4%	4%	5%	16%	79%	5%
Japanese	81%	12%	7%	7%	6%	17%	77%	6%
French	84%	14%	2%	2%	4%	15%	81%	4%
Avg	83.4%	12.2%	4.4%	4.4%	5.2%	16%	78.8%	5.2%

Evaluation on DCAD Pipeline

