



TUM

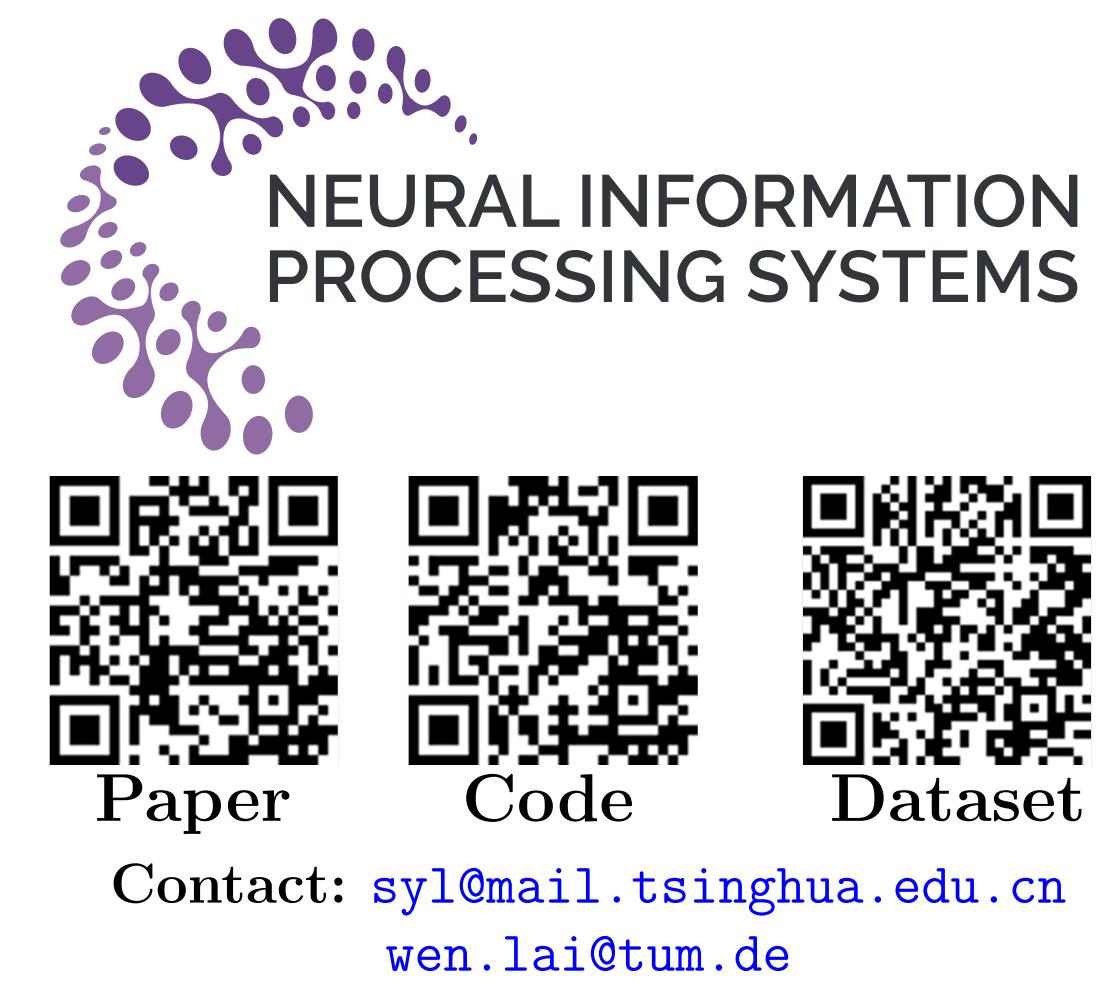
MCML  
Munich Center for Machine LearningMODEL BEST  
面壁智能

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

Yingli Shen<sup>1\*</sup> Wen Lai<sup>2,3\*</sup> Shuo Wang<sup>1</sup> Xueren Zhang<sup>4</sup> Kangyang Luo<sup>1</sup>  
Alexander Fraser<sup>2,3</sup> Maosong Sun<sup>1</sup>

<sup>1</sup>Tsinghua University <sup>2</sup>Technical University of Munich

<sup>3</sup>Munich Center for Machine Learning <sup>4</sup>Modelbest 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	✗
Glot500 (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	✗
Glotcc (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	✓

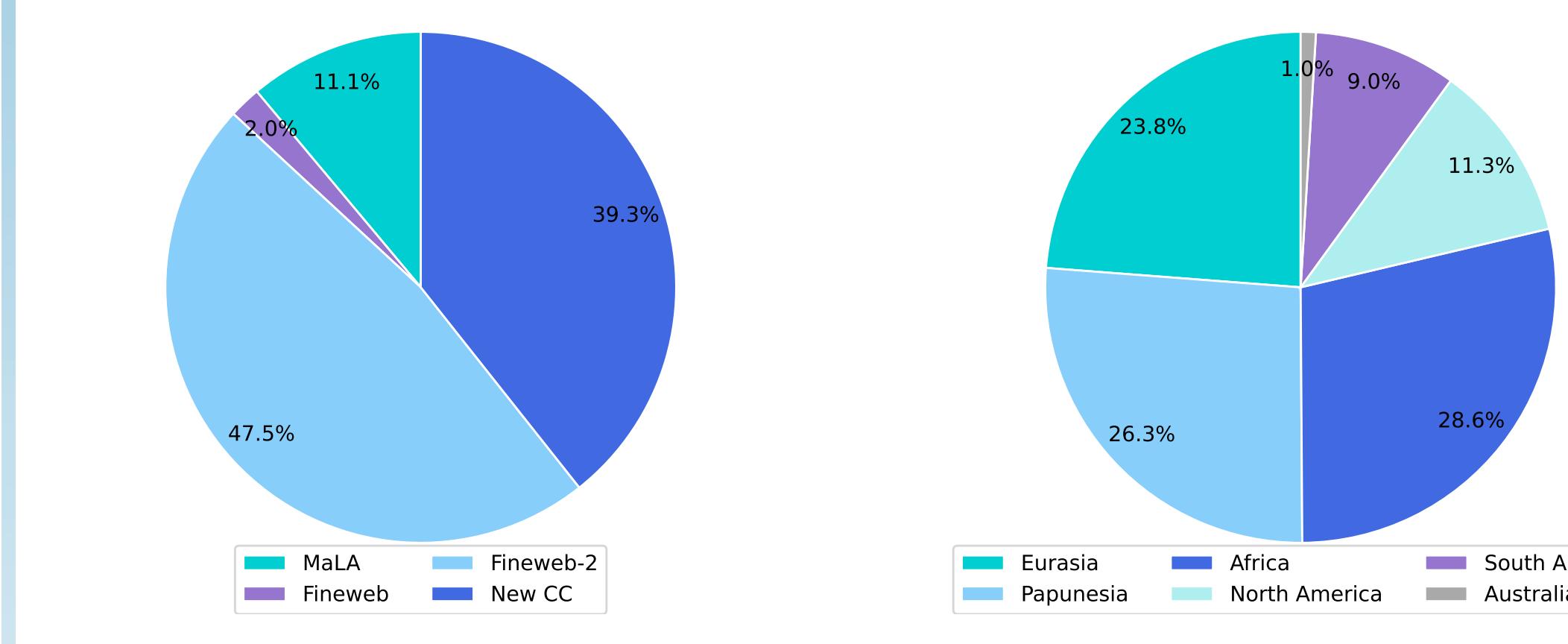
## Pipeline & Dataset

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

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

## Dataset Statistics

### Distribution and Linguistic Diversity



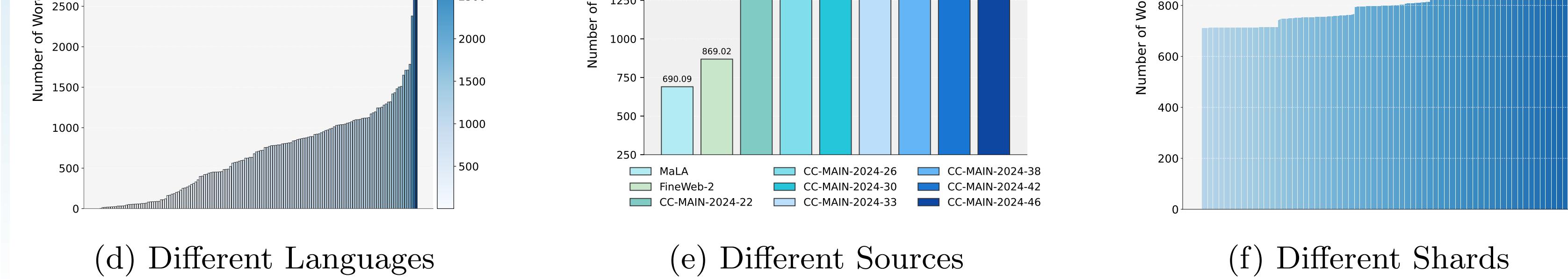
(a) Document Distribution

(b) Geographical Distribution

(c) Script Distribution

(d) Different Languages

(e) Different Sources



(f) Different Shards

## Dataset and Pipeline Evaluation

### Evaluation on DCAD-2000

	Qwen-2.5-7B			Aya-expansive-32B		
	Fineweb-2	New CC	DCAD-200	Fineweb-2	New CC	DCAD-200
<b>SIB-200 (↑)</b>						
H	33.41	34.53	<b>38.26</b> <small>↑4.85</small>	41.72	42.41	<b>47.93</b> <small>↑6.21</small>
M	28.72	29.86	<b>32.65</b> <small>↑3.93</small>	32.25	33.39	<b>38.16</b> <small>↑5.91</small>
L	23.58	24.22	<b>27.12</b> <small>↑3.54</small>	26.87	27.57	<b>33.24</b> <small>↑6.37</small>
VL	13.25	15.43	<b>21.57</b> <small>↑8.32</small>	17.23	19.50	<b>26.38</b> <small>↑9.15</small>

### Glot500-c test (↓)

	H	M	L	VL		
	347.21	334.18	<b>303.38</b> <small>↓43.83</small>	273.85	257.24	<b>225.28</b> <small>↓48.57</small>
	385.72	389.24	<b>369.15</b> <small>↓16.57</small>	326.92	321.16	<b>302.53</b> <small>↓24.39</small>
	426.33	419.25	<b>404.28</b> <small>↓22.05</small>	372.62	367.26	<b>341.34</b> <small>↓31.28</small>
	479.04	463.36	<b>433.48</b> <small>↓45.56</small>	396.33	392.33	<b>385.86</b> <small>↓10.47</small>

### FLORES-200 (↑) – Eng→X

	H	M	L	VL		
	15.24	16.07	<b>18.47</b> <small>↑3.23</small>	23.45	24.33	<b>26.33</b> <small>↑2.88</small>
	12.83	13.46	<b>15.49</b> <small>↑2.66</small>	19.36	20.21	<b>21.62</b> <small>↑2.26</small>
	8.94	9.28	<b>10.25</b> <small>↑1.31</small>	16.61	17.24	<b>18.36</b> <small>↑1.75</small>
	6.33	7.25	<b>9.05</b> <small>↑2.72</small>	12.51	13.16	<b>14.77</b> <small>↑2.26</small>

### FLORES-200 (↑) – X→Eng

	H	M	L	VL		
	16.31	16.92	<b>18.84</b> <small>↑2.53</small>	23.86	24.13	<b>26.94</b> <small>↑3.08</small>
	13.65	14.05	<b>16.27</b> <small>↑2.62</small>	20.45	20.36	<b>22.53</b> <small>↑2.17</small>
	9.47	10.22	<b>11.48</b> <small>↑2.01</small>	17.67	17.82	<b>18.93</b> <small>↑1.26</small>
	7.28	7.81	<b>9.65</b> <small>↑2.37</small>	13.25	13.56	<b>15.88</b> <small>↑2.63</small>

### 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%
<b>Avg</b>	<b>83.4%</b>	<b>12.2%</b>	<b>4.4%</b>	<b>4.4%</b>	<b>5.2%</b>	<b>16%</b>	<b>78.8%</b>	<b>5.2%</b>

### Evaluation on DCAD Pipeline

