

Same Pages, Cleaner Labels: Language Identification for Eight Low-Resource Languages in mC4 and FineWeb-2

Gary Zhang

William A. Shine Great Neck South High School

zhanggary607@gmail.com

Abstract

Low-resource subsets of multilingual web corpora often carry large amounts of text in the wrong language, and audits keep rediscovering the problem in each new corpus release. Scoring one corpus at a time cannot say whether a later corpus repeats an earlier one’s errors on the same documents. We join mC4 and FineWeb-2, built independently from overlapping Common Crawl snapshots, at the document level for eight low-resource languages. A URL join and a shift-invariant fingerprint sort the first 250k FineWeb-2 documents per language into three overlap classes: *co-selected* (same page, near-identical text), *page-shared* (same page, divergent text), and *unmatched*. Overlap covers 11% to 74% of a FineWeb-2 subset, so both corpora draw heavily on the same pages, and the shared documents are clean anyway. Under a criterion that flags a document only when two classifiers reject its label, 76.5% of sampled mC4 Sundanese and 58.9% of mC4 Kyrgyz are flagged, against at most 2.7% of co-selected FineWeb-2 documents. We attribute the difference to selection at ingestion. Refiltering mC4 through GlotLID still leaves 0.7–17% of survivors disagreeing with fastText for six languages, and barely moves Sundanese or Yoruba. We release the fingerprints, the flags, and the code: <https://github.com/garyzhang1006/wac13-lid-overlap>.

1 Introduction

Web corpora for low-resource languages are usually built from Common Crawl¹ by a language-identification (LID) filter, so the filter’s mistakes become the corpus. Kreutzer et al. (2022) hand-audited the products of this pipeline and found subsets in which most sentences were not in the labeled language. Follow-up audits reported that

33% of OSCAR’s Uyghur subset is Kazakh or Arabic (Zhang et al., 2024). Each new corpus release since then has advertised a better pipeline, and each new audit has found the same problem again.

What no audit reports is what happens to individual misidentified documents. Corpora built at different times draw on overlapping Common Crawl snapshots, so a later corpus may select the very pages that a wrongly labeled earlier subset was built from. If it does, and if those pages come back with the same wrong label, then per-corpus error rates overstate progress and the useful repair is to clean the shared pool of pages once rather than to rebuild pipelines. Whether the same pages return with the same labels is a question about individual documents, and no corpus pair has been measured this way.

In this paper we measure it directly for eight low-resource languages, each paired with a confusable high-resource neighbor (Sundanese/Indonesian, Kyrgyz/Russian, Yoruba/English, and five more), together with two non-confusable controls. We join subsets of mC4 (Xue et al., 2021) and FineWeb-2 (Penedo et al., 2025) on normalized URLs and on shift-invariant word-shingle fingerprints, and every sampled FineWeb-2 document is sorted into one of three overlap classes: co-selected, page-shared, or unmatched. Each class is then scored with a conservative criterion that counts a document only when fastText LID (Joulin et al., 2017) and GlotLID (Kargaran et al., 2023) independently agree that it is not in the labeled language. Every rate is replicated under a second criterion that shares no classifier with either corpus’s pipeline. Section 3 describes the join and the criteria, Section 4 reports the overlap and the flag rates, and Section 5 takes up the rival explanations and the repair that follows.

Contributions. Papers that release a new LID model report accuracy gains on held-out benchmarks, which does not say what a corpus built

¹<https://commoncrawl.org>

with that model contains, nor what it does with the documents an older corpus mislabeled. This paper supplies three things that document-level tracing makes available and single-corpus auditing does not. First, a measurement of how much of a FineWeb-2 low-resource subset consists of pages mC4 also drew on, which turns out to be 11% to 74%. Second, a flag rate for exactly those shared pages, which isolates selection at ingestion from a change of crawl as the explanation for FineWeb-2’s quality. Third, a simulation of the cheap repair available to anyone holding mC4 today, which works for six of eight languages and fails for the two worst ones. We release the fingerprints and every flagged document so that the flags can be checked by hand.

2 Related Work

The problem was established by manual audits. Kreutzer et al. (2022) audited 205 language subsets across five crawled corpora and found systematic wrong-language content concentrated in the low-resource subsets, Caswell et al. (2020) traced the failure to LID at web scale, Zhang et al. (2024) quantified it for Chinese minority languages, Dodge et al. (2021) documented analogous issues in English C4, and MADLAD-400 (Kudugunta et al., 2023) self-audited 498 candidate subsets before release and corrected several of them that turned out to be in the wrong language. All of these designs score a single corpus at a time, so they cannot separate a repeated error from a fresh one.

Document overlap between corpora has been measured in other settings. Kuzman Pungeršek et al. (2026) trace near-duplicate documents across two crawl generations of the CLASSLA-web corpora and find 18–20% overlap, and Alrashed and Orabona (2025) use cross-source MinHash agreement between contemporary multilingual corpora as a quality signal. We connect overlap to LID error, which neither of them does. On the construction side, OSCAR (Abadji et al., 2022), CulturaX (Nguyen et al., 2024), HPLT (de Gibert et al., 2024), and ROOTS (Laurençon et al., 2022) each describe a stronger pipeline, variously per-language thresholds, document-level deduplication, and line-level filtering, and FineWeb-2 (Penedo et al., 2025) adapts its pipeline per language with GlotLID. Corpus papers report the yields of their own filters, and none reports what a later corpus does with the

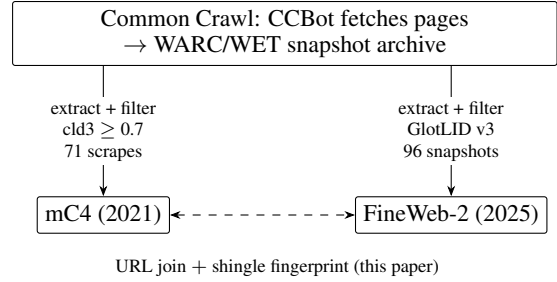


Figure 1: The two corpora are siblings rather than a derivation chain. Each selects and extracts documents independently from the shared Common Crawl archive, under its own LID filter, over snapshot ranges that overlap. Neither corpus performs its own crawl, and neither reads the other, so a document they share was selected twice from one archive. The dashed link is the join we add.

documents an earlier one mislabeled.

3 Method

The two corpora. mC4 and FineWeb-2 are built independently from Common Crawl, which crawls once and archives the result (Figure 1). FineWeb-2 is not derived from mC4 and does not read it, so we avoid lineage terms: a document the two corpora share was selected twice from one archive, not passed from one corpus to the other. We call such a document *co-selected*. Because a corpus does not crawl, a difference in the extracted text of a shared page reflects a difference in snapshot choice or extraction, and we call that case *page-shared* rather than re-crawled. The relation is symmetric: a co-selected pair is co-selected in both directions. We partition FineWeb-2’s sample rather than mC4’s because our question is what the later-built corpus admitted. That choice is an anchoring convention, and no claim of precedence is attached to it. The mC4-side share cannot be read off our numbers, since we scan a prefix of FineWeb-2 and not all of it.

Corpora and languages. For each language we process the first 250,000 documents of the mC4 and FineWeb-2 subsets in stream order, which is a deterministic prefix rather than a random sample. URLs are collected from up to 1.5M mC4 documents, so that the URL join covers most or all of each subset. Subset sizes come from the HuggingFace dataset server. Where the stream contradicts the reported row count, the streamed count is used instead. mC4 Sundanese, for example, holds 280,719 documents. The fingerprint index covers 89% of it. We Horvitz–

Thompson correct every rate by the fraction of the subset our index covers, and where a subset exceeds the 1.5M URL cap we report the corrected rate as a lower bound. The eight languages each have a documented or structurally likely confusable neighbor: Sundanese, Cebuano, Kazakh, Kyrgyz, Azerbaijani, Swahili, Yoruba, and Irish. Finnish and Vietnamese are the non-confusable controls.

Overlap classes. We normalize each text with NFKC and whitespace collapse, reduce its first 5,000 characters to word 4-gram shingles at every word position, and keep the 128 smallest CRC-32 shingle hashes as a MinHash-style sketch. Shingling at every position makes the fingerprint invariant to the boilerplate and line-level differences that two extraction pipelines produce from one page. A FineWeb-2 document is *co-selected* if some same-language mC4 document’s sketch gives an estimated Jaccard similarity above 0.5, computed on the sketch of the union; *page-shared* if it is not co-selected but its normalized URL appears in the mC4 subset; and *unmatched* otherwise. The matcher is validated inside every run. Held-out mC4 documents were mutated the way a second extraction pipeline would mutate them, with front boilerplate strips and interior line deletions, and 1,846 of 1,846 of them are recovered at the 0.5 threshold, pooled over the eight languages. Against word-shuffled probes the false-match rate is 0 of 1,324. FineWeb-2 documents were also matched against each neighbor’s mC4 subset. A document carrying the Sundanese label in one corpus and the Indonesian label in the other has one of those two labels wrong, with no classifier in the loop. This yields 17 matches (14 Cebuano, 2 Irish, 1 Azerbaijani), which is probably too few to carry statistics, and we report it as a negative result.

Flagging criteria. We sample 20,000 documents per class, or fewer where a class is smaller. A document is *flagged* under the primary criterion only when two classifiers agree against its label. fastText lid.176 must assign some other language, and GlotLID must assign a label outside the target’s ISO 639-3 code. Documents shorter than 40 characters are excluded and their counts reported. A flag is evidence of misidentification rather than proof of it, because no human ground truth enters anywhere in this paper, and we report flag rates on that understanding.

The stack is built around one constraint: mC4

was filtered with cld3² and FineWeb-2 with GlotLID, so no single classifier is neutral toward both corpora. fastText lid.176 is the one widely used web-scale LID that neither pipeline used, which is why it appears in both of our criteria. Its partner in the primary criterion is GlotLID, the model FineWeb-2 itself filtered with, so that criterion asks whether the corpus meets its own gate plus an outside check. That framing favors FineWeb-2, which already passed that gate. We therefore pair fastText with OpenLID (Burchell et al., 2023) as a *symmetric* criterion sharing no classifier with either pipeline, and this one asks whether the corpus meets a standard neither builder chose. We report the two side by side and do not intersect them. Both appear for all ten languages in Table 3, for the two shared classes in the eight target languages in Table 4, and alongside each classifier alone and a classifier-free check on dominant Unicode script. They agree closely, differing by at most 4.5 points on mC4 (Yoruba) and 0.6 points on FineWeb-2 (Sundanese). The designation of one as primary is presentational, since no claim in this paper depends on it.

Because fastText sits in both stacks, a document lid.176 gets right is never flagged by either, so both readings are lower bounds and neither can be pushed up by fastText’s false positives alone. The cost is visible in the GlotLID-alone column: agreement with fastText suppresses 2.4 points of GlotLID’s mC4 flags for Cebuano and under 0.8 points for every other language. GlotLID and OpenLID also draw on overlapping training data, so the symmetric criterion removes the pipeline circularity without making the two readings statistically independent of each other. Every flagged document in every class is archived with all three predictions and a text snippet. Any flag can therefore be checked.

Provenance and release. mC4 filtered with cld3 (threshold 0.7) over 71 Common Crawl scrapes; FineWeb-2 with GlotLID v3 under per-language thresholds over 96 snapshots ending April 2024. We audit with fastText lid.176³, GlotLID v3, and OpenLID, and each model file’s SHA-256 is released with the results. The release at <https://github.com/garyzhang1006/wac13-lid-overlap> holds the pipeline as

²<https://github.com/google/cld3>

³<https://fasttext.cc/docs/en/language-identification.html>

a single script and the per-language result files behind every number in this paper. It also holds one gzipped record per flagged document, carrying its overlap class, its text snippet, its dominant script, and all three classifier predictions, which is what makes the flags checkable by hand. It also ships `verify_paper.py`, which recomputes every figure and table in this paper from those result files without network or model access. Code is under MIT and the data files under CC BY 4.0.

4 Results

Table 1 carries the three findings.

Overlap. Between 9.5% and 41% of the sampled FineWeb-2 documents are co-selected with mC4 once we correct for the fraction of mC4 our index covers, and a further 1.9% to 33% are page-shared. Taken together these two classes, which we call the shared pool, make up 11% to 74% of a subset. For Swahili the shared pool is 74% of the subset. For Kyrgyz, Kazakh, and Irish it runs from 40% to 52%. In those languages FineWeb-2 most probably worked over much the same web that mC4 did, and its filter is nearly the only thing that differs. Yoruba sits at 11%, where the two corpora genuinely diverge in what they drew on, so a claim about the shared pool there speaks for a minority of the subset. The co-selected class is large enough to measure in every language, 5,901 documents at the smallest, which is what the next two findings need.

Flag rates. mC4 flag rates under the dual criterion range from 7.4% (Swahili) to 76.5% (Sundanese), with Kyrgyz at 58.9% and Yoruba at 44.0%. Unmatched FineWeb-2 text ranges from 0.18% to 2.98%, a ratio between 8 (Cebuano) and more than 200 (Kyrgyz). Under the symmetric criterion the same comparison gives 8.1–74.0% for mC4 against 0.29–2.40% for FineWeb-2 (Table 3). Three of every four sampled documents in the mC4 Sundanese subset, which is still distributed through the HuggingFace hub, is still cited as Sundanese training data, and is still merged into later corpora, fail two separately trained language classifiers that were built from different data and by different groups. The competing labels are divided. Of the flagged documents, 30.5% are sent by GlotLID to Indonesian or Standard Malay and 38.0% to the wider Malayic family, against 20.5% to English, so the confusable neighbor takes the

plurality and ordinary English web text takes most of the rest. The two non-confusable controls sit at 1.6% (Finnish) and 3.4% (Vietnamese) in mC4. This rules out our measurement apparatus as the source of the high rates, though it does not by itself establish neighbor confusion as their cause.

The shared classes. The co-selected class localizes the repair. Those documents sat in mC4’s heavily flagged pool and were selected again by FineWeb-2, and they come back almost clean, at most 2.66% (Irish) and as low as 0.05% (Azerbaijani), against mC4 pool rates of 7.4–76.5%. Part of that cleanliness is definitional, since these documents passed FineWeb-2’s GlotLID gate and our primary criterion includes GlotLID. This is where the symmetric criterion matters, because it puts the co-selected Sundanese class at 1.11% where the mC4 pool sits at 74.0%. Page-shared documents, the same pages under a different snapshot or extraction, run 0.15–2.84% (Table 4). They sit close to unmatched text, above it by one to one-and-a-half points in Yoruba, Irish, and Cebuano and below it elsewhere, and far under the mC4 pool in every language. FineWeb-2 therefore did not avoid mC4’s pages; it selected them again, at up to 74% of its volume, and kept only what its filter accepted. We attribute the difference between the two corpora to selection and filtering at ingestion rather than to a change in the underlying web, with the caveat that for Yoruba this argument probably rests on too little of the Yoruba subset, 11%, to carry the language on its own.

5 Discussion

Measurement circularity. FineWeb-2 filters with GlotLID and our primary criterion includes GlotLID, so the low FineWeb-2 rates under that criterion partly reflect the corpus agreeing with its own gatekeeper. The symmetric criterion answers this, since it is independent of both pipelines. Under it the contrast holds at 74.0% against 2.40% for Sundanese, and in every other language as well (Table 3; all ten languages, all five readings, no exclusions). Single-classifier readings bracket the absolute FineWeb-2 level (Sundanese 3.0–16.7%), and lid.176’s coverage is itself unreliable for some of these languages. As a uniform caveat, wherever fastText alone flags more than 10% of unmatched FineWeb-2 text (Sundanese 16.7%, Irish 10.9%, Yoruba 78.7%), the fastText-only reading says more about lid.176 than about the corpus.

Language	Shared with mC4		Flag rate, % (dual-classifier)				$n_{\text{co-sel.}}$
	% of FineWeb-2 docs		mC4	FineWeb-2			
	co-sel.	page-sh.		all	co-sel.	unmatch.	
Sundanese	16.1	5.8	76.5 [75.9, 77.1]	1.29 [1.12, 1.48]	2.98 [2.75, 3.22]	15,252	
Kyrgyz	30.3	21.7	58.9 [58.2, 59.6]	0.08 [0.05, 0.14]	0.27 [0.20, 0.35]	19,021	
Yoruba	9.5	1.9	44.0 [42.9, 45.2]	0.21 [0.13, 0.34]	1.52 [1.36, 1.70]	7,601	
Irish	24.8	15.5	25.7 [25.1, 26.3]	2.66 [2.45, 2.89]	1.68 [1.51, 1.87]	20,000	
Azerbaijani	≥ 14.8	≥ 8.7	15.9 [15.4, 16.5]	0.05 [0.02, 0.15]	0.18 [0.13, 0.25]	5,901	
Cebuano	15.1	6.0	11.7 [11.3, 12.2]	1.19 [1.04, 1.34]	1.43 [1.27, 1.60]	20,000	
Kazakh	≥ 29.3	≥ 21.9	8.6 [8.2, 9.0]	0.25 [0.18, 0.36]	0.53 [0.44, 0.64]	12,195	
Swahili	41.0	32.8	7.4 [7.0, 7.7]	0.28 [0.22, 0.36]	0.52 [0.42, 0.62]	20,000	

Table 1: Document overlap between mC4 and FineWeb-2, and flag rates within each overlap class. Left: the share of sampled FineWeb-2 documents that are co-sel. (co-selected: a near-duplicate mC4 document exists) or page-sh. (page-shared: the URL is in mC4, the text differs), Horvitz–Thompson corrected for the fraction of mC4 our index covers; Azerbaijani and Kazakh are lower bounds because their mC4 subsets exceed the 1.5M-document URL cap. Right: the percentage of documents that both fastText lid.176 and GlotLID reject, with Wilson 95% intervals, on samples of up to 20,000 documents per class. The final column gives the size of the co-selected sample. Rows are ordered by the mC4 rate. Table 3 repeats the mC4 and unmatched rates under four further criteria, and Table 4 gives both shared classes under both paired criteria.

Absolute levels are therefore treated as criterion-dependent and dual-agreement rates as conservative lower bounds, while the finding we rest on is the cross-corpus ratio, which holds under every criterion. The Unicode-script check catches only cross-script cases, and for Azerbaijani its 9.6% mC4 reading most probably reflects Azerbaijani text written in Cyrillic rather than any other language, so we report it there as a diagnostic rather than a floor.

Rival mechanisms. Three alternatives to the ingestion filter could produce a clean later corpus, and the data speak to each. Deduplication: unmatched documents in MinHash clusters larger than one are dirtier in six of eight languages (Sundanese 4.9% against 2.2%). The differences are a few points against a gap of tens of points, so FineWeb-2’s global dedup helps without carrying the result. The web changed: pre-2021 dumps are dirtier than later ones in six of eight languages (Sundanese 9.6% against 1.8%), yet even the pre-2021 class sits far below mC4’s 76.5%, so the era appears to explain a minority of the gap and the filter the rest. Invisible cleaning: line-level filtering that strips a wrong-language block from a kept page makes the text diverge, and any near-duplicate matcher then reads that page as page-shared instead of co-selected. Our co-selected shares are therefore lower bounds on how often the two corpora hold the same page. The total overlap is unaffected, because such a page still joins on its URL.

Refiltering. The cheap repair is simulated by keeping only the mC4 documents that GlotLID accepts. We then ask how often fastText still disagrees with the label among the survivors. The baseline for that question is mC4’s fastText-alone rate, which runs 9.2–60.2% across six of the eight languages. Those six are brought down to between 0.7% (Kazakh) and 17.0% (Swahili), a reduction of 20-fold for Kyrgyz at the best and 1.4-fold for Swahili at the worst. For Sundanese and Yoruba the rate barely moves, with 88.2% falling to 50.4% and 96.6% falling to 94.2%. Both readings are inflated by lid.176’s unreliability for those two languages, as above, and they remain too large to certify the repair. Refiltering yesterday’s corpus through today’s filter is a real and partial repair, and it is weakest exactly where the problem is worst.

Implications. Progress between these two corpora is real, and it arrived by building a new corpus rather than by fixing the old one. mC4’s low-resource subsets should not be used as monolingual data for the languages measured here. The caution extends to corpora that merge mC4 in, such as CulturaX (Nguyen et al., 2024), where a Sundanese subset that fails two classifiers three times out of four arrives as an ingredient. Anyone holding mC4 can price the cheap repair from our simulation: re-filtering is enough for moderate rates and not for the worst subsets.

Limitations

Our results cover one corpus pair, eight languages, and two controls. Adding a third and fourth corpus would test whether the ingestion-filter account generalizes, and CulturaX and OSCAR are the natural next pair, since a merge-based construction makes overlap certain rather than something to be measured; both are gated, and we leave them to future work.

The 250k-document prefixes are deterministic slices of each subset in stream order, not random samples. Stream order in these distributions is not randomized, so the flag rates we report describe the head of each subset and may not describe its tail. This bias falls on the per-class rates rather than on the overlap shares, since the URL join covers up to 1.5M mC4 documents and in six of eight languages the whole subset. We sample both corpora and both controls the same way, so the comparisons among them are internally consistent even where the absolute level is not representative.

Our URL join does not require the two corpora to have drawn the same Common Crawl record. We hold fetch timestamps on the mC4 side and dump identifiers on the FineWeb-2 side, and we use them for the temporal analysis in the rival-mechanism discussion, but requiring record identity in the join would reclassify as page-shared any page that both corpora selected unchanged from different snapshots, which is the common case for static pages. Separating snapshot choice from extraction difference within the page-shared class needs that stricter join and would sharpen the result.

Because both criteria require fastText to agree, every rate we report is a lower bound on misidentification, and a language that lid.176 systematically mishandles will be understated in both. Kazakh’s corrected shares are lower bounds for a second reason: the dataset server’s row count is contradicted by the stream itself, so we use the streamed count and mark the result accordingly. The matcher finds near-duplicates rather than paraphrases, so text rewritten beyond Jaccard 0.5 joins only through its URL.

We score whole documents, so pervasive code-mixing, Sundanese with Indonesian in particular, is pooled with wrong-language text rather than separated from it. Irish’s co-selected documents are the least clean of the eight at 2.66%, and Irish pages are heavily bilingual, so a document-level label is partly ill-defined there. Finally, classifier-based

estimates inherit classifier error for exactly these languages, and no human ground truth appears anywhere in this paper. A native-speaker audit of a few hundred flagged documents would convert our flag rates into misidentification rates, and it is the single most valuable addition to this design. The archived flags exist so that readers, including native speakers, can check any number we report.

Ethics Statement

The release quotes the first 300 characters of each flagged document. Those excerpts come from pages that Common Crawl collected and that both mC4 and FineWeb-2 already redistribute in full, and we kept them short so that a reader can check a flag without us republishing whole documents. We made no attempt to identify the author of any page, and the release documentation gives a contact route for removal requests.

A high flag rate is a statement about a filter. We would not want it read as a statement about a language or about the people who write the web in it, particularly since our own rates come from classifiers with known weaknesses on exactly these languages. That is why we archive the evidence and ask for human verification instead of presenting the rates as settled.

Acknowledgements

We thank the four anonymous WaC-13 reviewers. One of them rejected the lineage vocabulary that the submitted version rested on, and the terminology of this version, its title included, follows from that objection. Others asked why the two classifier stacks are built as they are, and asked for a diagram of how two corpora arrive at the same page, and both are now in the paper.

References

- Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. [Towards a cleaner document-oriented multilingual crawled corpus](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355.
- Sultan Alrashed and Francesco Orabona. 2025. [Mix, MinHash, and match: Cross-source agreement for multilingual pretraining datasets](#). *arXiv preprint arXiv:2512.18834*.
- Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. [An open dataset and model for language identification](#). In *Proceedings*

- of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 865–879.
- Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. [Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608.
- Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. [A new massive multilingual dataset for high-performance language technologies](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128.
- Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. [Documenting large webtext corpora: A case study on the colossal clean crawled corpus](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. [Bag of tricks for efficient text classification](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431.
- Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. [GlotLID: Language identification for low-resource languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218.
- Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2022. [Quality at a glance: An audit of web-crawled multilingual datasets](#). *Transactions of the Association for Computational Linguistics*, 10:50–72.
- Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. [MADLAD-400: A multilingual and document-level large audited dataset](#). In *Advances in Neural Information Processing Systems 36 (Datasets and Benchmarks Track)*, pages 67284–67296.
- Taja Kuzman Pungeršek, Peter Rupnik, Vit Suchomel, and Nikola Ljubešić. 2026. [The growing gains and pains of iterative web corpora crawling: Insights from South Slavic CLASSLA-web 2.0 corpora](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 7545–7555, Palma de Mallorca, Spain. ELRA Language Resource Association.
- Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. [The BigScience ROOTS corpus: A 1.6TB composite multilingual dataset](#). In *Advances in Neural Information Processing Systems 35 (Datasets and Benchmarks Track)*, pages 31809–31826.
- Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2024. [CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2025. [FineWeb2: One pipeline to scale them all: Adapting pre-training data processing to every language](#). In *Second Conference on Language Modeling (COLM)*.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.
- Chen Zhang, Mingxu Tao, Quzhe Huang, Jiuheng Lin, Zhibin Chen, and Yansong Feng. 2024. [MC²: Towards transparent and culturally-aware NLP for minority languages in China](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8832–8850.

A Flagged-Document Examples

Table 2 reproduces archived evidence snippets for documents both classifiers flagged, for six of the eight languages, drawn from the unmatched FineWeb-2 class. Each release record carries the document’s first 300 characters together with all three predictions.

B All Criteria, All Languages

Table 3 puts every criterion side by side, so a reader can see which conclusions depend on the choice of classifier pair and which hold under all five.

Labeled	Detected (fastText / GlotLID / OpenLID)	Opening of the archived record, with transliteration and translation
Sundanese	id / ind_Latn / ind_Latn	<i>Suku sunda adalah salah satu suku yang ada di Indonesia. Ciri khas dari suku sunda adalah orangnya yang ramah da...</i> Indonesian: 'The Sundanese are one of the peoples found in Indonesia. What is distinctive about the Sundanese is that they are friendly and...'
Kyrgyz	ru / rus_Cyrl / rus_Cyrl	<i>Команда кафедры «Телематика» Кыргызского государственного технического университета им.И.Раззакова и ОФ «Технолэ...</i> <i>Komanda kafedry «Telematika» Kyrgyzskogo gosudarstvennogo tekhnicheskogo universiteta im. I. Razzakova i OF «Tekhnole...</i> Russian: 'The team of the Telematics department of the I. Razzakov Kyrgyz State Technical University and the Technolo... public foundation'
Kazakh	ru / rus_Cyrl / rus_Cyrl	<i>Таразский государственный педагогический институт, г. Тараз В физкультурно-спортивной деятельности, как обоснова...</i> <i>Tarazskiy gosudarstvennyy pedagogicheskiy institut, g. Taraz V fizkul'turno-sportivnoy deyatel'nosti, kak obosnova...</i> Russian: 'Taraz State Pedagogical Institute, Taraz. In physical-education and sports activity, as substantiated...'
Yoruba	es / spa_Latn / spa_Latn	<i>Nueva Versión Internacional (Castilian) Todavía no hay audio para esta traducción. Un destino común IA todo esto...</i> Spanish: 'New International Version (Castilian). There is no audio yet for this translation. A common destiny. 1 To all this...'
Irish	en / eng_Latn / eng_Latn	<i>Applications are NOW open for JUNIOR EUROVISION 2020 Do you have a great singing voice? Are you between the ages...</i>
Azerbaijani	en / eng_Latn / eng_Latn	<i>Arsenal leaves out Mkhitaryan from Europa League final amid safety concerns Armenian national Henrikh Mkhitaryan...</i>

Table 2: Evidence snippets for dual-classifier flags in unmatched FineWeb-2 text, for six of the eight languages. Each row is the first 112 characters of a record in `flags_<lang>.jsonl.gz` in the release, whitespace normalized, so a reader can find the full document by searching that file for the opening words. All three classifiers agree in each of these six. Transliterations and English translations of our own follow the two Cyrillic and the two other non-English records; the Irish and Azerbaijani records are already English, which is the point of listing them. The Yoruba subset also contains Amharic-script Bible text, which GlotLID labels `amh_Ethi` in 47 archived records.

C Co-Selected and Page-Shared Rates

Table 4 separates the two shared classes, which the main text reports only as a range, and gives the sample size behind each rate.

Language	mC4, all documents (%)					FineWeb-2, unmatched (%)				
	dual	sym	fastText	GlottLID	script	dual	sym	fastText	GlottLID	script
Sundanese	76.5	74.0	88.2	76.8	0.03	2.98	2.40	16.68	2.98	0.01
Kyrgyz	58.9	57.0	60.2	59.2	0.52	0.27	0.29	0.65	0.27	0.01
Yoruba	44.0	39.6	96.6	44.2	0.78	1.52	1.49	78.66	1.52	0.27
Irish	25.7	24.1	35.6	25.8	0.07	1.68	1.58	10.90	1.69	0.03
Azerbaijani	15.9	16.1	16.7	16.2	9.64	0.18	0.36	0.67	0.18	0.01
Cebuano	11.7	9.5	17.6	14.1	0.01	1.43	1.26	8.45	1.44	0.09
Kazakh	8.6	8.3	9.2	9.3	1.23	0.53	0.50	0.75	0.55	0.01
Swahili	7.4	8.1	23.5	8.1	0.01	0.52	0.53	5.08	0.55	0.02
Finnish (ctrl.)	1.6	1.5	2.1	2.2	0.02	0.40	0.39	0.53	0.42	0.00
Vietnamese (ctrl.)	3.4	3.4	3.9	3.5	0.28	0.15	0.15	0.22	0.15	0.01

Table 3: Flag rates under every criterion, all languages, no exclusions: dual (fastText $\wedge$ GlottLID), symmetric (fastText $\wedge$ OpenLID), fastText alone, GlottLID alone, and the classifier-free Unicode-script check. The two control rows have no overlap classes, so their right-hand block covers the whole FineWeb-2 sample rather than an unmatched subset. The two paired criteria agree within 4.5 points on mC4 and 0.6 points on FineWeb-2 in every language. Comparing the dual column with GlottLID alone shows what requiring fastText’s agreement costs: 2.4 points for Cebuano on mC4, under 0.8 points everywhere else.

Language	Co-selected (%)			Page-shared (%)		
	dual	sym	<i>n</i>	dual	sym	<i>n</i>
Sundanese	1.29	1.11	15,252	2.30	2.07	6,222
Kyrgyz	0.08	0.08	19,021	0.16	0.18	20,000
Yoruba	0.21	0.17	7,601	2.62	2.49	1,488
Irish	2.66	2.46	20,000	2.84	2.48	20,000
Azerbaijani	0.05	0.07	5,901	0.15	0.24	20,000
Cebuano	1.19	1.06	20,000	1.96	1.74	12,324
Kazakh	0.25	0.24	12,195	0.53	0.48	20,000
Swahili	0.28	0.24	20,000	0.48	0.46	20,000

Table 4: Flag rates for the two shared classes under both paired criteria, with sample sizes. Compare against the mC4 pool rates in Table 3: the same pages, selected by the later pipeline, are flagged 6 to 700 times less often than the pool they came from, with the smallest gaps in Cebuano and Irish and the largest in Kyrgyz. The symmetric criterion, which shares no classifier with either corpus, gives the same picture.