

Invisible on the Web: An Audit of Low-Resource Muskogean Languages in Web-Crawled Corpora

Sharvin Goyal

Independent Researcher

sharvingoyal09@gmail.com

Abstract

Web-crawled corpora now claim that they retain coverage of 1,000–2,000 languages, but what does such coverage mean for a language family at the extreme low-resource end? We audit the collective presence of the Muskogean family (Choctaw, Chickasaw, Muscogee, Alabama, Koasati, Mikasuki) across seven major web corpora. Only two of the six languages appear in any corpus, and their entire kept content is 28 documents comprising 16 distinct texts, dominated by 19th-century Bible translations; a single document comes from a present-day language community. We exhaustively audit both the kept documents and the 15,108 discarded ones, showing that (i) language identification (LID) assigns Muskogean labels to short non-Muskogean fragments with up to 99.7% persistence under re-identification, so quality filters, not LID, determine corpus content; (ii) those same filters discard about twice as many genuine Choctaw documents as they keep, including contemporary tribal-authored text; and (iii) four Muskogean languages are structurally invisible: no LID label exists, and verified Chickasaw text is silently absorbed into Choctaw. We release our audit annotations and seed-text URL list, and recommend family-level label sets, filter audits on the smallest subsets, and engagement with tribal data sovereignty for web corpus builders.

1 Introduction

The largest multilingual web corpora have grown from roughly one hundred languages (Conneau et al., 2020; Raffel et al., 2020) to nearly two thousand (Kargaran et al., 2024; Penedo et al., 2025), driven by broad-coverage language identification (LID) models such as GlotLID (Kargaran et al., 2023). Coverage claims at this scale are usually evaluated in aggregate. For any individual language at the extreme low-resource end, however, the corpus that results from the crawl-identify-filter

pipeline may bear little resemblance to the language as its community uses it, a concern raised forcefully by Kreutzer et al. (2022), whose audit of 205 language-specific corpora found systematic quality failures concentrated in the lowest-resource subsets.

We push this style of audit to its limit for one language family: Muskogean, comprising Choctaw (*cho*), Chickasaw (*cic*), Muscogee (Creek, *mus*), Alabama (*akz*), Koasati (*cku*), and Mikasuki (*mik*), all critically endangered languages of what is currently the southeastern United States and Oklahoma, together having on the order of 15,000 speakers. Because the labeled subsets are tiny, we can do what aggregate audits cannot: read *every* document, kept and discarded, and characterize the pipeline’s behavior completely.

Our contributions: (1) a complete presence census of Muskogean across seven web corpora (§3); (2) an exhaustive precision audit of all 28 kept documents (§4); (3) an analysis of all 15,108 discarded documents showing that quality filters both define the corpus and destroy scarce genuine data (§5); (4) a recall-side stress test on verified text in all six languages (§6); and (5) recommendations for corpus builders (§7). All audit code, annotations, and URL lists are open source.¹

2 Related Work

Kreutzer et al. (2022) manually audited samples from five multilingual web corpora and found that many low-resource subsets were mislabeled or unusable; our audit differs in being exhaustive over a family rather than sampled over many languages, and in auditing the *discarded* documents alongside the kept ones. Corpus-specific audits accompany MADLAD-400 (Kudugunta et al., 2023) and OSCAR (Abadji et al., 2022). GlotLID (Kargaran

¹<https://github.com/Sharvin-coder/muskogean-web-audit>

Corpus	Langs	cho	cic	mus	akz	cku	mik
C4/mC4	103	–	–	–	–	–	–
CC-100	109	–	–	–	–	–	–
OSCAR 23.01	151	–	–	–	–	–	–
MADLAD-400	402	–	–	–	–	–	–
HPLT 2.0	186	–	–	–	–	–	–
GlottCC-V1	1397	✓	–	–	–	–	–
FineWeb-2	1813	✓	–	✓	–	–	–

Table 1: Muskogean subsets across seven web corpora. *Langs* = number of languages the corpus claims.

et al., 2023) explicitly targets low-resource LID and powers both GlotCC (Kargaran et al., 2024) and FineWeb-2 (Penedo et al., 2025); FineWeb-2 additionally publishes the documents its quality filters remove, which makes our filter audit possible. For Choctaw specifically, the only substantial NLP resource is the Chahta Anumpa corpus (Brixey et al., 2018), built from archival and elicited (not web) data. Bird (2020) and Carroll et al. (2020) frame the ethical dimension: Indigenous communities have interests in how their language data is collected and used that web crawling does not currently accommodate.

3 Corpus Presence

We checked seven corpora for any subset labeled with a Muskogean ISO 639-3 code: C4/mC4 (Rafel et al., 2020), CC-100 (Conneau et al., 2020), OSCAR 23.01 (Abadji et al., 2022), MADLAD-400 (Kudugunta et al., 2023), HPLT 2.0 (Burchell et al., 2025), GlotCC-V1 (Kargaran et al., 2024), and FineWeb-2 (Penedo et al., 2025). Table 1 shows the result: only the two LID-driven corpora carry any Muskogean subset at all, and only for Choctaw and Muscogee. Chickasaw, Alabama, Koasati, and Mikasuki appear in no labeled subset. Label-level absence is not content-level absence; indeed, §6 finds verified Chickasaw text sitting inside the Choctaw-labeled data, which is precisely why a label census alone cannot close the question.

4 Precision: What Was Kept

The complete kept content is 28 documents: 14 Choctaw and 10 Muscogee in FineWeb-2, and 4 Choctaw in GlotCC-V1 (73.6k characters, roughly 10,800 words, about four times the text of this paper). We annotated every document for language correctness, genre, source era, boilerplate fraction, and content duplication; all annotations were made by a single annotator (the author) against written

Subset	Docs	Texts	Maj. target	Relig. chars	Cont. chars
FW2 <i>cho</i>	14	9	12	88%	6%
FW2 <i>mus</i>	10	7	5	45%	26%
GlottCC <i>cho</i>	4	2	4	100%	0%

Table 2: Exhaustive audit of all kept documents. *Texts* = distinct underlying texts after clustering duplicates; *Maj.* = documents whose main body (excluding navigation and boilerplate) is more than half running text in the labeled language; *Relig.* / *Cont.* = fraction of characters from religious or canonical-translation sources / from sources annotated as contemporary (present-day composition, as opposed to historical or missionary-era texts).

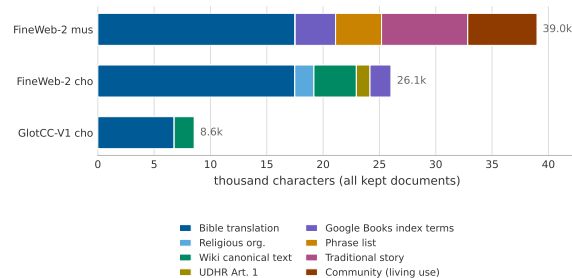


Figure 1: Genre composition (characters) of every kept Muskogean document across the three corpus subsets. Bar labels are rounded to 0.1k characters; totals in the text use exact counts.

criteria (see Limitations), and are released in full for community correction. Documents are clustered as one text when they share the same underlying target-language content (the same translation, article, or excerpt), regardless of URL or surrounding boilerplate. Table 2 and Figure 1 summarize.

Duplication. The 28 documents contain only 16 distinct texts. GlotCC’s entire Choctaw corpus is 4 documents, 3 of which are an identical Tower-of-Babel excerpt from omniglot.com, kept three times because www and non-www URLs defeated deduplication. FineWeb-2 keeps the same Ten Commandments page from two Wikipedia-incubator mirrors. Most strikingly, FineWeb-2’s Choctaw *test split* consists of exactly one document, and it is a near-duplicate of two training documents (the same UDHR Article 1 translation from mirror URLs): train–test contamination of 100% for any model evaluated on this split.

Composition. In FineWeb-2’s Choctaw subset, 88% of characters are religious text, predominantly American Bible Society translations (the 1867 Pentateuch and the ABS Mark/Genesis por-

Lang	Removed docs	Med. chars	Still targ. (GlotLID v3)	Genuine lost	Kept docs
<i>cho</i>	853	59	99.7%	27	14
<i>mus</i>	14,255	15	98.5%	5	10

Table 3: The removed piles. *Med.* = median removed-document length in characters (medians because lengths are heavily right-skewed); *Still targ.* = share re-identified as the labeled language by GlotLID v3; *Genuine* = unique URLs of genuine target-language documents found among high-confidence ($p \geq 0.9$) removed docs (a lower bound); *Kept* = kept documents, for comparison.

tions), and only 6.5% come from contemporary sources; GlotCC’s four Choctaw documents are religious in their entirety. Muscogee is more varied (1890s Creek Nation legislative acts, a 1915 traditional story), but 5 of its 10 documents are not majority running Muscogee: three Google Books “common terms” index clouds (alphabetized word salad), an English-scaffolded phrase table, and a genuine story interleaved with platform boilerplate. The single living-community source in the entire collection is the site of Ekvñ-Yefolecv, a Maskoke ecovillage. A model trained on these subsets would learn a 19th-century missionary register, not the language of any present-day Muskogean community.

5 What Filtering Removed

FineWeb-2 publishes what its quality filters discard: 853 Choctaw-labeled and 14,255 Muscogee-labeled documents (Table 3). Our audit of this material has two clearly separated passes: an automated pass re-identifying all 15,108 documents with GlotLID v3, and a manual pass in which we read every document the automated pass re-identified as the target at $p \geq 0.9$ (74 Choctaw and 503 Muscogee documents) and judged it at domain level as genuine target-language text or *junk*, by which we mean text that is not Muskogean at all, typically short foreign-language or non-linguistic fragments.

LID does not reject the junk. 99.7% of removed Choctaw documents and 98.5% of removed Muscogee documents are *still* assigned the same label by the current GlotLID. Persistence under re-identification shows the labels are stable, not that LID played no role in shaping the pool (the pool is itself LID-defined); but it does show that the current model would admit nearly all of this material again. The junk is systematic, not random: for Choctaw,

short fragments rich in pseudo-Choctaw character n-grams, such as Slavic diminutives (*koshka*, *ishchuk*), Japanese names (*Ichiki*, *Itachi*), Hindi song syllables (*chiki chiki*); for Muscogee, strings rich in ⟨v⟩ (Muscogee’s characteristic spelling of the schwa vowel), such as Vietnamese domain names, Turkish pages, and domain-typo generators. With a median of 15 characters (Muscogee), these fragments carry too little signal for any classifier; under an open-world label set they land on the typologically unclaimed label. A cross-check supports this reading: no other public LID system carries any Muskogean label at all (OpenLID has 201 labels, NLLB-LID 218, fastText lid.176 176), so the collision is only observable in GlotLID; rerunning the same $p \geq 0.9$ documents through fastText lid.176 scatters them across major-language labels at low confidence (65–70% below $p = 0.5$). Consequently, in these two pipelines *the heuristic quality filters (Rae et al., 2021), not LID, function as the de facto arbiters of what enters the corpus.*

The filters destroy scarce genuine data.

The manual pass finds 27 unique genuine Choctaw URLs in the removed pile, roughly *twice* the 14 Choctaw documents FineWeb-2 keeps (unique removed URLs vs. kept documents). Losses include Wikipedia-incubator encyclopedia articles, the Choctaw Nation’s own language-school lessons (choctawschool.com), Jehovah’s Witnesses pages, hymn collections, Tatoeba sentences, and, notably, contemporary healthcare non-discrimination notices in Choctaw. The filters responsible for these removals, with genuine-document counts, are list_ratio (12), char_dup_ratio (7), gopher_short_doc (5), gopher_below_alpha (3), and precision_wordlist (1): thresholds tuned for majority-language prose punish exactly the genres (lesson vocabularies, hymn lists, short institutional notices) in which endangered languages actually appear on the web. The Muscogee losses are fewer (5 unique URLs) but of the same character: hymn collections (mvsokkehymns.com), rosary prayers, and a community-history page; ironically, three of the five, including the Muscogee Lord’s Prayer ($p = 1.0$), were removed by a filter named precision_wordlist.

6 Recall: Six Languages vs. the Pipeline

We assembled a verified seed set of Muskogean web text (14 segments across all six languages,

Language	Label	Segs	Full	No diac.	Frag.
Choctaw	<i>cho</i> _Latn	5	5/5	5/5	0.75
Muscogee	<i>mus</i> _Latn	4	4/4	4/4	1.00
Chickasaw	<i>none</i>	2	→ <i>cho</i> _Latn (0.74)		
Alabama	<i>none</i>	1	→ <i>eng</i> _Latn (0.33)		
Koasati	<i>none</i>	1	→ <i>rel</i> _Latn (0.88)		
Mikasuki	<i>none</i>	1	→ <i>bea</i> _Latn (0.42)		

Table 4: GlotLID v3 on verified Muskogean text. For Choctaw and Muscogee: correct top-1 identifications (full text and diacritic-stripped) and mean recall on 8-word fragments. The four languages without labels are shown with their top-1 assignment (as-crawled text).

drawn from tribal, religious, community, and reference sources) and ran GlotLID v3 under three conditions: full segment, diacritics stripped, and 8-word fragments, the last constructed as consecutive non-overlapping 8-word windows over each segment (Table 4). With one or two segments per language, these results characterize a floor, not a distribution; per-segment outputs are released with the annotations.

On genuine running text, GlotLID is excellent: 9/9 segments correct with full context, robust even to complete diacritic stripping. Failures concentrate exactly where the web is fragmentary: the mean 8-word-fragment recall of 0.75 for Choctaw (Table 4) averages over running-text segments at 0.75–1.0 and a single vocabulary-list segment that falls to 0.25, mirroring the junk analysis: short Choctawish strings are unreliable in both directions.

The deeper problem is structural. GlotLID v3’s 2,102 labels include only *cho* and *mus* from the family. Both verified Chickasaw segments (UDHR Article 1, as published and with its HTML spacing artifacts merged programmatically) are absorbed into Choctaw, at 0.74 and 0.53 respectively (with Blackfoot second at 0.41 on the cleaned variant), and we found exactly this in the wild: the Chickasaw UDHR from unicode.org sits in FineWeb-2’s Choctaw removed pile, identified as Choctaw at $p = 0.98$. The single Koasati and Mikasuki segments are assigned to unrelated languages of Kenya (Rendille) and Canada (Beaver); with one segment each, these specific targets are illustrative data points, but the underlying fact that no Muskogean label exists to receive them is deterministic and does not depend on sample size. The pattern is not GlotLID-specific: under fastText lid.176, which also lacks all six labels, our 14 verified segments scatter across unrelated major-language labels at a median confidence of 0.15. For four of six

Muskogean languages there is no path by which any amount of web text could reach a correctly-labeled corpus.

7 Discussion and Recommendations

For corpus builders. (1) *Audit the smallest subsets exhaustively*: at this scale a human can read the entire corpus in an hour, and the audit changes the headline claim (FineWeb-2’s effective Choctaw content is nine texts, not “1,800+ languages including Choctaw”). (2) *Report content-level, not URL-level, deduplication*: mirror URLs produced 3-of-4 duplicate corpora and a fully contaminated test split. (3) *Family-aware LID*: when a label set contains one member of a closely related family, sister languages are silently annexed; label sets should either cover the family or calibrate open-world rejection. (4) *Filter thresholds are register policies*: Gopher-style thresholds encode assumptions about prose that systematically reject the wordlists, hymns, lessons, and notices where endangered languages live online. We do not advocate dropping quality filters wholesale; rather, thresholds should be subset-size-aware (relaxed or human-reviewed below a few hundred documents, where exhaustive review is cheap), and publishing removed documents, as FineWeb-2 does, should become standard so such losses stay recoverable. Nothing in these mechanisms is Muskogean-specific: any small family with one corpus-visible member and several label-absent sisters, a common configuration world-wide, is exposed to the same absorption and filter dynamics, though confirming this across further families remains future work.

For Muskogean NLP. The usable web-crawled resource is effectively: nine Choctaw texts and seven Muscogee texts, plus at least 27 recoverable Choctaw documents in the removed pile. Any serious corpus effort for these languages must instead build on community partnership (Bird, 2020) and archival resources (Brixey et al., 2018).

8 Conclusion

For the Muskogean family, web-scale “coverage” dissolves under exhaustive inspection: two of six languages present, sixteen distinct texts, a missionary-era register, a contaminated test split, filters that discard twice what they keep, and four languages with no representational path into any corpus. The web is not yet a corpus for these languages, but the pipeline changes we recommend

are concrete, and the communities’ own web presence (language schools, ecovillages, official notices) shows where respectful collection, done with consent, could actually begin.

Limitations

Our audit covers one language family and the specific corpus releases named; findings on junk composition and filter losses are lower bounds, since we manually verified only removed documents re-identified at $p \geq 0.9$: this threshold introduces a selection bias, and genuine documents scoring lower would only add to the losses. Likewise, our presence census establishes absence from *labeled subsets*, not absence of the languages’ content from the underlying crawls, as the Chickasaw-under-Choctaw case demonstrates. All annotations were made by a single annotator (the author), who is neither a speaker nor a specialist scholar of any Muskogean language, against written criteria released with the artifact: orthographic markers (e.g., Choctaw’s underlined vowels and upsilon-style vowel letters; Muscogee’s ⟨v⟩ schwa and ⟨ē⟩), verification of high-frequency lexical items against published dictionaries and community learning materials, and identification of each document’s source text. With one annotator no agreement statistics are possible; every judgment is released with its evidence for community correction, and validation by Muskogean-language scholars or speakers is invited and would strengthen the annotations. The recall-side seed set is small (14 segments; one or two per language); this is itself a finding about the scarcity of verifiable open-web text, but it limits its statistical claims, and the specific misassignment targets reported for Koasati and Mikasuki are single observations. We evaluate GlotLID v3 because it underlies both corpora studied; our fast-Text lid.176 cross-check probes generality, but no public alternative carries Muskogean labels, so the collision behavior itself is testable only in GlotLID.

Ethics Statement

The languages studied belong to sovereign tribal nations. Following the CARE principles (Carroll et al., 2020) and Bird (2020), we do not redistribute any scraped text: we release document-level annotations, URLs, and derived statistics only, all pointing to material already public. We recognize that even a URL list can facilitate further collection; the released artifact therefore documents a

correction-and-removal process, and we commit to honoring removal or correction requests from the nations or from site stewards. We flag that several tribal-authored pages (language school lessons, community sites) were crawled into corpora without any consent mechanism beyond robots.txt, which cannot express community-level data governance; we encourage corpus maintainers to establish opt-out and consultation channels with tribal nations.

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 (LREC 2022)*, pages 4344–4355.
- Steven Bird. 2020. [Decolonising speech and language technology](#). In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*, pages 3504–3519.
- Jacqueline Brixey, Eli Pincus, and Ron Artstein. 2018. [Chahta anumpa: A multimodal corpus of the Choctaw language](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, et al. 2025. [An expanded massive multilingual dataset for high-performance language technologies \(HPLT\)](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 17452–17485.
- Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, et al. 2020. [The CARE principles for indigenous data governance](#). *Data Science Journal*, 19(1):43.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.
- 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.

- Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. 2024. [GlotCC: An open broad-coverage CommonCrawl corpus and pipeline for minority languages](#). In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024) Datasets and Benchmarks Track*.
- 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 (NeurIPS 2023) Datasets and Benchmarks Track*.
- 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](#). *arXiv preprint arXiv:2506.20920*.
- Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. [Scaling language models: Methods, analysis & insights from training Gopher](#). *arXiv preprint arXiv:2112.11446*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.