

Extraction Changes the Route: Main-Content Extraction and Language Identification in Web Corpora

Toprak Seda Karaosmanoğlu
Middle East Technical University
Ankara, Türkiye
toprak.karaosmanoglu@metu.edu.tr

Abstract

Main-content extraction commonly precedes language identification (LID) and determines its input in web-corpus pipelines. We measure this dependency on 7,808 WebMainBench pages using three extractors and three LID systems. We compare human-selected main text with annotated-region and full-page extractor outputs. A context flip occurs when annotated-region extraction preserves the route on human-selected text but full-page extraction changes it. For pages with nonempty text and matching human-selected and annotated-region routes, context-flip rates are 1.10–3.20%. On the common eligible set, Magic-HTML has lower paired rates than Trafilatura for all three LID systems, with each corresponding 95% interval excluding zero. All-page route loss counts empty outputs or routes that differ from the route on human-selected text. Magic-HTML’s empty full-page outputs add 5.84 percentage points. The fixed snapshot’s coarse non_en–en difference is positive in all nine configurations and remains positive after standardization to the page-complexity field (`meta.level`). A 100-page audit finds 35 unsupported, 29 supported, and 36 indeterminate routes. Of 35 unsupported flips, 26 move from predicted English to another route, and 17 involve changed content selection alone or in mixed cases. Among supported flips, 21 move to English. Extraction context can change text availability and downstream routing.

1 Introduction

Multilingual web-corpus construction commonly applies main-content extraction before language identification (LID). The extractor converts web-page HTML into text, and the LID system assigns that text a language label used for routing, filtering, storage, or later processing. The two stages therefore form a cascade: the text retained upstream determines the linguistic evidence available to the downstream routing decision.

The HTML context presented to an extractor can change its output. Processing a human-selected main region may preserve the annotated main content, while processing the full page can also retain navigation, interface strings, cookie text, templates, or another page region. It can also change which region is selected as main content. These changes alter the plain text passed to LID and can change the language partition assigned to the same web-page.

This study measures that interaction with WebMainBench, which provides tag-level human main-content annotations for real webpages (Liu et al., 2026; OpenDataLab, 2026b). A *language route* is the normalized base-language label assigned by an LID system to a text representation. For each page, G denotes visible text from the human-selected main content. For extractor e , M_e denotes the extractor output when the same annotated region is processed inside a minimal HTML document, and F_e denotes the extractor output when the full released page is processed. The LID stage receives the normalized plain-text representations G , M_e , and F_e . For LID system m , $\hat{y}_m(X)$ denotes the route assigned to representation X . The central event is

$$\hat{y}_m(M_e) = \hat{y}_m(G) \quad \wedge \quad \hat{y}_m(F_e) \neq \hat{y}_m(G). \quad (1)$$

The first relation selects pages whose route survives extraction of the annotated main region. The second records a route change after the same extractor is exposed to the full page. We call this event a *context flip*.

We cross Trafilatura, Resiliparse, and Magic-HTML with GlotLID, OpenLID-v3, and CommonLingua. Context flips occur in every extractor–LID configuration, with rates of 1.10–3.20% among pages with nonempty G , M_e , and F_e and matching G – M_e routes. Section 3.4 formalizes this eligibility rule. Paired comparisons use the intersection of

extractor-specific eligible sets within each LID system. An all-page route-loss analysis also records empty extraction outputs, making extractor coverage visible in the routing outcome. Context flips span many predicted-route pairs. A blinded human audit evaluates whether sampled full-page routes are supported by the human-selected main text, followed by a descriptive task on visible extraction changes.

Recent extraction work shows that extractor choice changes which webpages survive downstream filtering (Li et al., 2026). We track the same webpage across a controlled annotated-region/full-page contrast and measure a different downstream consequence: movement between language routes. The $G \rightarrow M_e$ condition selects pages whose route is stable under extraction of the annotated region before the full-page context is introduced.

2 Related Work

2.1 Main-content extraction

Main-content extraction maps webpage HTML to the text retained for later corpus processing. Trafilatura was developed for text extraction in web-corpus construction (Barbaresi, 2021). Multilingual evaluation shows that extraction quality varies across languages, layouts, and page structures (Barbaresi and Lejeune, 2020; Bournonville et al., 2025). A comparative study of fourteen extractors and five baselines also reports substantial genre dependence across human-labeled extraction datasets (Bevendorff et al., 2023).

WebMainBench extends this evaluation setting with 7,809 human-annotated pages and tag-level main-content annotations (Liu et al., 2026; OpenDataLab, 2026b). Its released evaluation toolkit supports Trafilatura, Resiliparse, and MagicHTML, the three extractors evaluated here. Li et al. (2026) compare HTML-to-text extractors through a shared filtering pipeline and measure changes in the pages retained downstream. We keep page identity fixed and measure whether changing the extractor’s input context changes the route assigned by a later LID stage.

2.2 Language identification on web data

Language identification assigns language labels to text used in multilingual corpus processing. Web-crawl evaluation introduces domain and class distributions that differ from conventional held-out benchmarks. Caswell et al. (2020) report severe

drops in human-judged accuracy for many languages when systems with strong held-out results are applied to web-crawl text. They trace errors to factors including domain mismatch, class imbalance, and language similarity. CommonLID provides a human-annotated web-domain benchmark covering 109 languages and reports that established evaluation sets overestimate web-domain accuracy for many languages (Ortiz Suarez et al., 2026).

We evaluate three multilingual systems with different design goals. GlotLID targets broad language coverage and dataset-creation pipelines (Kargaran et al., 2023). OpenLID-v3 focuses on precision for closely related languages in web-data curation (Fedorova et al., 2026). CommonLingua is a byte-level language identifier designed for corpus curation (PleIAs, 2026). Applying each system to paired extraction representations shows how the assigned language route changes with extraction.

2.3 Extraction followed by language routing

Large multilingual web-corpus pipelines commonly place text cleaning or extraction upstream of language identification. C4Corpus combines boilerplate removal with language detection (Habernal et al., 2016), and HPLT 3.0 exposes a pipeline that includes HTML text extraction and language identification for noisy web text (Oepen et al., 2026). FineWeb2 provides a concrete extraction-to-routing sequence in which Trafilatura processes Common Crawl HTML before GlotLID is applied to the extracted non-English stream (Penedo et al., 2025; Hugging Face, 2025). We measure how that cross-stage routing decision changes when the extractor processes the annotated main region or the full page.

3 Data and Experimental Design

3.1 WebMainBench

We use WebMainBench v1.1 (Liu et al., 2026; OpenDataLab, 2026b) at repository revision 5d a0972e9b58d0c7891ae75053ced97c268f52e3. We refer to this revision-specific file as the fixed snapshot. The webmainbench.jsonl file has SHA-256 85765fe798f07c14eb1c92945046eaa56e0d a59663f70b9c498647d7dfd78884 and contains 7,809 pages. Human-selected main text is usable on 7,808 pages, which form the analysis set. In this fixed snapshot, the released coarse meta.language field marks 6,646 analysis pages as en and 1,162 as non_en. The non_en value aggregates the re-

maining languages and is used only as a coarse descriptive stratum.

The released annotation marks page elements selected as main content and also provides a native `main_html` field. For four analysis pages, native `main_html` has no visible text. We reconstruct their main region deterministically from the maximal elements carrying the released `cc-select="true"` annotation. The human-selected main-content annotation defines the region from which the reference representation G is derived.

3.2 Three text representations

We construct three plain-text representations for each page. G is visible text from the human-selected main region. M_e is the output of extractor e from that same region placed inside a minimal HTML document whose visible body contains only the annotated region. F_e is the output of extractor e from the full released page. Before either extractor input is created, the benchmark’s `cc-select` attribute and `cc-extraStyle` annotation style block are removed. For G , visible-text recovery excludes `script`, `style`, `noscript`, and `template` elements. HTML entity decoding, Unicode Normalization Form KC, and whitespace collapse are then applied consistently to the resulting text. The two processing paths are main-region $\text{HTML} \rightarrow e \rightarrow M_e \rightarrow \text{LID}$ and full-page $\text{HTML} \rightarrow e \rightarrow F_e \rightarrow \text{LID}$.

The $G \rightarrow M_e$ transition measures extractor processing of the human-selected region. The $M_e \rightarrow F_e$ transition introduces the surrounding page context to the same extractor. Equation 1 conditions on route agreement across the first transition and measures the routing change associated with the second.

A representation-level summary precedes LID scoring. Token overlap uses case-folded Unicode word tokens. Multiset precision and recall are combined as F_1 . Table 1 reports extraction coverage and overlap with G . Median G – M_e overlap ranges from .953 to 1.000, while median G – F_e overlap ranges from .875 to .936. Full-page outputs are less similar to G by this overlap measure, consistent with changes in retained text when the extractor processes the full page. The coverage columns separately record pages on which an extractor returns no visible text.

Before applying the primary conditioning step, we also summarize route disagreement with G . Among pages with predictions for both M_e and

F_e , disagreement between G and M_e is 1.59–8.45% across the nine configurations. Disagreement between G and F_e is 2.48–9.99% and exceeds the matched G – M_e rate by 0.55–2.31 percentage points in every configuration. Section 4.1 then restricts the context-flip rate to pages whose G and M_e routes agree.

3.3 Extractors and language identifiers

The extractor set contains Trafilatura 2.1.0 (Barbaredi, 2021), Resiliparse 1.0.9 (Bevendorff et al., 2018), and Magic-HTML 0.1.6 (OpenDataLab, 2026a). Trafilatura returns plain-text output, excludes comments, and receives the page URL through its URL argument. Resiliparse uses main-content extraction with formatting, alternative text, post metadata, and comments disabled. Magic-HTML uses GeneralExtractor with the page URL as its base URL. Visible text is recovered from the returned HTML after excluding `script`, `style`, `noscript`, and `template` elements. These calls follow the interfaces of the released libraries: Trafilatura and Magic-HTML accept URL context, while the evaluated Resiliparse interface receives the HTML input directly. Each extractor uses the same settings for M_e and F_e .

The LID set contains GlotLID (Kargaran et al., 2023), OpenLID-v3 (Fedorova et al., 2026), and CommonLingua (PleIAs, 2026). The repository revisions used are 85cd67164943, b9b09974ffdd, and c6cd15eadc8a, respectively. GlotLID receives normalized text directly. OpenLID-v3 follows its released preprocessing: lowercasing, whitespace collapse, and removal of digits and characters outside Unicode word or space categories. CommonLingua uses its released prediction code with a maximum input length of 512 UTF-8 bytes. All systems return one top label. Route comparison removes model-specific label wrappers and maps script-specific suffixes to ISO-639-3 base-language identity.

3.4 Context-flip rate and all-page route-loss rate

For page i , extractor e , and LID system m , write the three representations as G_i , M_{ie} , and F_{ie} and define the context-flip indicator C_{iem} by

$$C_{iem} = \mathbf{1}[\hat{y}_m(M_{ie}) = \hat{y}_m(G_i) \wedge \hat{y}_m(F_{ie}) \neq \hat{y}_m(G_i)]. \quad (2)$$

Eligibility requires visible text in G , M_e , and F_e and route agreement between G and M_e . Empty

Extractor	M_e text	F_e text	$G-M_e$ token F_1		$G-F_e$ token F_1	
			Mean	Median	Mean	Median
Trafilatura	7,792	7,805	.858	.953	.756	.875
Resiliparse	7,567	7,723	.900	1.000	.763	.887
Magic-HTML	7,192	7,352	.796	.982	.741	.936

Table 1: Representation coverage and token overlap with human-selected main text. Text counts report pages with visible extractor output. Token F_1 summarizes multiset overlap of case-folded Unicode word tokens across all 7,808 pages. Empty extractor outputs contribute zero overlap.

extractor outputs are treated as extraction failures and are summarized through the coverage counts in Table 1. The primary estimand is

$$P(\hat{y}_m(F_e) \neq \hat{y}_m(G) \mid \hat{y}_m(M_e) = \hat{y}_m(G), G, M_e, F_e \neq \emptyset). \quad (3)$$

Each eligible page receives equal weight. Pages sharing the same normalized hostname form a URL-host cluster. All 7,808 analysis pages yield a hostname. Primary 95% confidence intervals use 2,000 nonparametric URL-host-cluster bootstrap replicates. The primary endpoint, primary eligibility rule, and bootstrap procedure were specified before LID scoring.

An *all-page route-loss rate* uses all 7,808 analysis pages. For $R \in \{M_e, F_e\}$, route loss occurs when extractor output is empty or when the predicted route on R differs from the route on G . The M_e and F_e rates compare each representation directly with G , so a page may lose the G route at M_e and recover it at F_e . This all-page rate combines extraction availability with preservation of the G route.

The context-flip estimand is primary. Paired extractor, metadata-stratified, and cross-LID analyses are secondary. All-page route loss, meta.level standardization, predicted-route transitions, and the audit-direction cross-tabulation are exploratory. The human annotation study then assesses whether sampled full-page routes are supported by the benchmark’s human-selected main text.

3.5 Additional analyses

Primary denominators can differ across extractors because extraction coverage and $G-M_e$ route agreement vary. Within each LID system, the paired analysis uses the intersection of the three extractor-specific eligible sets. We call this the *common eligible set*. Each extractor pair is evaluated on that identical page set. Percentage-point differences use 5,000 URL-host-cluster bootstrap replicates.

A secondary descriptive analysis reports context-flip rates separately for the fixed snapshot’s en and non_en metadata values and estimates the non_en minus en risk difference with 5,000 URL-host-cluster bootstrap replicates. The snapshot also provides a three-level complexity field, meta.level (simple, mid, hard). An exploratory direct standardization reweights the two metadata strata to the pooled meta.level distribution among eligible pages within each extractor-LID configuration. The pooled weights are held fixed at their observed values across 5,000 URL-host-cluster bootstrap replicates for the standardized risk-difference intervals.

A secondary cross-LID analysis identifies pages that satisfy the primary eligibility rule for all three LID systems for a given extractor and counts how many systems flip on each page. An exploratory predicted-route transition analysis records the route on G and the route on F_e for every context flip. It reports the number of distinct predicted-route pairs, concentration by route pair and URL-host cluster, and three broad directions: eng→other, other→eng, and other→other. Here, other denotes any normalized route other than eng. The page remains the unit of observation. Confidence intervals in the paired and secondary analyses are descriptive and unadjusted for multiplicity. Primary, paired, and metadata-stratified bootstraps use seed 20260815+ k , where k is a fixed analysis/configuration offset. Exploratory all-page intervals use seed 20260920+ k , with k indexing the LID system, extractor, and endpoint. The meta.level standardization separately uses seed 20260920.

3.6 Human annotation study

The annotation frame contains 737 distinct pages with at least one context flip across the nine configurations. A simple random sample selects 100 unique pages from this frame using seed 202608152122. When a sampled page contains several flip events, one event is selected uniformly.

Each sampled page contributes one annotation item, giving every sampled page equal weight. We refer to this as a page-balanced audit. The two three-person panels work independently. Each panel assigns one joint panel-level label per item, drawing on members’ relevant language knowledge. Reliability therefore compares the two panel outputs.

All six participants were proficient in English. Across the group, additional language knowledge included Turkish, Portuguese, French, Japanese, Chinese, and German at varying proficiency levels. We distributed complementary language backgrounds across the two panels. Translation tools were permitted as auxiliary checks. Before annotation, all participants read the same short guideline and completed five calibration examples drawn outside the evaluation sample.

The first task evaluates support for the full-page route in the human-selected main text. Each item shows the target route assigned to F_e , a G excerpt of up to 3,000 normalized characters, and an excerpt-truncation indicator. Extractor identity, LID identity, URL, M_e , F_e , and the route assigned to G are hidden. *supported* marks a target route whose language is dominant in G or forms a substantial coherent portion of it. *unsupported* marks a target route whose language is absent or incidental in G . *indeterminate* covers excerpts with insufficient linguistic evidence or uncertain language distinctions. We call *supported* and *unsupported* judgments determinate. For target languages outside a panel’s direct language knowledge, a determinate judgment requires visible evidence that G is dominated by another identifiable language. Items lacking that evidence receive *indeterminate*. Truncated excerpts with little visible evidence for the target route also receive the indeterminate label.

The same panels complete a second task that describes the extraction change across G , M_e , and F_e . Items show the first-task consensus label, the extractor and LID identities, the three route labels, and the three text excerpts. *added surrounding text* marks full-page output that preserves substantial main-content material and adds surrounding page material. *changed content selection* marks full-page output that drops or replaces substantial main-content material and selects a different region. *mixed* combines both patterns. *unclear* covers excerpts with insufficient evidence for a stable assignment to the other three descriptions.

Panel-level reliability is reported with exact agreement and Cohen’s κ . Differences between

the two panel outputs are resolved by consensus. For the extraction-change task, differing panel judgments are adjudicated item by item with all four extraction-change labels available. The proportion of unsupported routes among determinate support judgments uses a Wilson 95% confidence interval. An exploratory cross-tabulation relates consensus support labels to the three predicted-route directions defined above. It is reported as counts for the page-balanced audit sample.

4 Results

4.1 Context-flip and all-page route-loss rates

Context flips occur across all nine extractor–LID configurations. Table 2 gives the exact numerators, denominators, and URL-host-cluster bootstrap intervals together with the exploratory all-page route-loss rates and a decomposition of full-page route loss.

The context-flip rate is 1.10–3.20% among eligible pages. Magic-HTML has the smallest context-flip point estimate for each LID system. Across all 7,808 pages, F_e route loss is 3.48–10.08%. Magic-HTML returns an empty F_e on 5.84% of pages, compared with 1.09% for Resiliparse and 0.04% for Trafilatura. These empty outputs account for 58.0–66.3% of Magic-HTML’s full-page route loss. Among nonempty F_e outputs, disagreement with G is 4.49–4.50% for Resiliparse and Magic-HTML under GlotLID, 4.05–4.37% for Trafilatura and Magic-HTML under OpenLID-v3, and 3.16–4.40% across the three extractors under CommonLingua. The bootstrap intervals in Table 2 show substantial overlap for the close point estimates. The final column conditions only on nonempty F_e , while the earlier G – F_e disagreement summary also requires an M_e prediction. Their denominators therefore differ.

4.2 Paired extractor comparisons

The paired analysis places all three extractors on the common eligible set within each LID system. These sets contain 6,218 pages for GlotLID, 6,563 for OpenLID-v3, and 6,652 for CommonLingua. Table 3 reports the corresponding rates and paired percentage-point differences.

Across the three LID systems, Magic-HTML has a lower paired rate than Trafilatura on the common eligible set. The Magic-HTML–Resiliparse difference is -0.02 , -0.30 , and -1.44 percentage points for GlotLID, OpenLID-v3, and Com-

LID	Extractor	Flips/ <i>n</i>	Context flip [95% CI]	M_e loss	F_e loss [95% CI]	F_e empty	$\hat{y}(G) \neq \hat{y}(F_e)$ [95% CI]
GlotLID	Trafilatura	228/7,132	3.20 [2.78, 3.65]	8.63	10.03 [9.28, 10.85]	0.04	9.99 [9.22, 10.77]
	Resiliparse	155/7,333	2.11 [1.78, 2.46]	5.85	5.53 [4.98, 6.10]	1.09	4.49 [4.04, 5.01]
	Magic-HTML	95/6,867	1.38 [1.10, 1.69]	10.63	10.08 [9.38, 10.83]	5.84	4.50 [3.98, 5.03]
OpenLID-v3	Trafilatura	196/7,621	2.57 [2.21, 2.97]	2.37	4.09 [3.63, 4.56]	0.04	4.05 [3.59, 4.52]
	Resiliparse	174/7,370	2.36 [2.00, 2.72]	5.37	5.25 [4.72, 5.79]	1.09	4.21 [3.75, 4.71]
	Magic-HTML	79/6,865	1.15 [0.91, 1.43]	10.64	9.95 [9.21, 10.65]	5.84	4.37 [3.87, 4.86]
CommonLingua	Trafilatura	161/7,647	2.11 [1.78, 2.45]	2.04	3.48 [3.08, 3.93]	0.04	3.45 [3.03, 3.88]
	Resiliparse	230/7,428	3.10 [2.68, 3.53]	4.64	5.44 [4.90, 6.01]	1.09	4.40 [3.92, 4.90]
	Magic-HTML	76/6,934	1.10 [0.84, 1.35]	9.73	8.81 [8.16, 9.53]	5.84	3.16 [2.75, 3.58]

Table 2: Context-flip and all-page route-loss rates, in percent. Context-flip denominators contain pages satisfying the primary eligibility rule. Their intervals use 2,000 URL-host-cluster bootstrap replicates. F_e route-loss and nonempty-output disagreement intervals use 2,000 exploratory URL-host-cluster bootstrap replicates. M_e loss and F_e empty are point estimates on all 7,808 pages. The final column is conditioned on nonempty F_e .

LID	Comparison (A–B)	Rate A (%)	Rate B (%)	Δ [95% CI]	<i>n</i>
GlotLID	Magic-HTML–Trafilatura	0.95	1.93	−0.98 [−1.35, −0.61]	6,218
	Magic-HTML–Resiliparse	0.95	0.96	−0.02 [−0.30, 0.27]	6,218
	Trafilatura–Resiliparse	1.93	0.96	+0.96 [0.61, 1.31]	6,218
OpenLID-v3	Magic-HTML–Trafilatura	0.99	1.39	−0.40 [−0.69, −0.09]	6,563
	Magic-HTML–Resiliparse	0.99	1.30	−0.30 [−0.58, −0.03]	6,563
	Trafilatura–Resiliparse	1.39	1.30	+0.09 [−0.22, 0.40]	6,563
CommonLingua	Magic-HTML–Trafilatura	0.95	1.34	−0.39 [−0.71, −0.09]	6,652
	Magic-HTML–Resiliparse	0.95	2.39	−1.44 [−1.87, −1.04]	6,652
	Trafilatura–Resiliparse	1.34	2.39	−1.05 [−1.48, −0.65]	6,652

Table 3: Paired context-flip rates on the common eligible set. Each row compares rate A with rate B. Δ is A minus B in percentage points and is computed from unrounded rates. Confidence intervals use 5,000 URL-host-cluster bootstrap replicates.

monLingua. The Trafilatura–Resiliparse ordering changes across LID systems.

Context-flip rates are lower on the common eligible set than on the extractor-specific pages it excludes. For a given extractor, 282–1,115 extractor-specific eligible pages fall outside the intersection because at least one other extractor fails eligibility. Their context-flip rates are 4.61–11.82%. Rates within the common eligible set are 0.95–2.39%. Table 3 therefore supports paired comparisons on the intersection, while Table 2 retains the extractor-specific and all-page views.

4.3 Released language-metadata strata

The fixed snapshot’s non_en stratum has a higher context-flip point estimate than the en stratum in all nine configurations (Table 4). The unadjusted risk differences are +0.62 to +4.15 percentage points. Six URL-host-cluster intervals lie entirely above zero.

The released meta.level distribution differs across the two strata. Among all 7,808 analysis pages, hard pages account for 30.67% of en and 24.96% of non_en, while simple pages account for 28.41% and 40.02%, respectively. Direct

standardization to the pooled meta.level distribution leaves the non_en–en difference positive in all nine configurations, at +0.83 to +4.34 percentage points. Seven standardized intervals lie entirely above zero.

4.4 Cross-LID overlap

A separate page-level analysis asks whether several LID systems flip on the same extracted representation. Table 5 uses pages satisfying the route-agreement and extraction-coverage conditions for all three LID systems for a given extractor. At least two LID systems flip on 1.76% of Trafilatura pages, 1.36% of Resiliparse pages, and 0.70% of Magic-HTML pages. All three systems flip on 43, 23, and 19 pages respectively.

Most pages with any context flip do so under only one LID system. A smaller subset flips under two or all three after the same full-page extraction.

4.5 Predicted-route transitions

Context flips are distributed across many predicted-route pairs (Table 6). The nine configurations contain 46–115 distinct predicted-route pairs. The largest share attributable to any one pair is 4.6–

LID	Extractor	n (en/non_en)	en %	non_en %	Raw Δ [95% CI]	Standardized Δ [95% CI]
GlotLID	Trafilatura	6,081/1,051	2.78	5.61	+2.83 [1.27, 4.59]	+3.30 [1.66, 5.06]
	Resiliparse	6,261/1,072	1.82	3.82	+2.00 [0.81, 3.26]	+2.50 [1.20, 3.96]
	Magic-HTML	5,866/1,001	1.28	2.00	+0.72 [-0.19, 1.77]	+0.97 [-0.05, 2.07]
OpenLID-v3	Trafilatura	6,503/1,118	2.25	4.47	+2.23 [0.88, 3.75]	+2.55 [1.20, 4.12]
	Resiliparse	6,275/1,095	2.20	3.29	+1.09 [-0.04, 2.28]	+1.31 [0.12, 2.59]
	Magic-HTML	5,853/1,012	1.06	1.68	+0.62 [-0.18, 1.48]	+0.83 [-0.07, 1.85]
CommonLingua	Trafilatura	6,553/1,094	1.59	5.21	+3.62 [2.32, 5.07]	+4.10 [2.69, 5.63]
	Resiliparse	6,327/1,101	2.48	6.63	+4.15 [2.61, 5.78]	+4.34 [2.74, 6.08]
	Magic-HTML	5,930/1,004	0.96	1.89	+0.93 [0.09, 1.91]	+1.26 [0.29, 2.32]

Table 4: Context-flip rates by the fixed snapshot’s coarse meta . language field. Raw Δ is non_en minus en. Both difference columns are in percentage points. Standardized differences reweight both strata to the pooled meta . level1 distribution within each configuration. Intervals use 5,000 URL-host-cluster bootstrap replicates.

Extractor	n	Any	≥ 2	All 3
Trafilatura	6,999	5.04%	1.76%	0.61%
Resiliparse	7,211	5.46%	1.36%	0.32%
Magic-HTML	6,706	2.33%	0.70%	0.28%

Table 5: Cross-LID overlap on pages eligible for all three LID systems. Columns report the share with at least one, at least two, or all three LID systems producing a context flip.

13.2% of flips, and the ten most frequent pairs together account for 27.0–48.7%. The direction of the transition also varies by pairing: eng→other accounts for 28.2–60.0% of flips, other→eng for 15.2–52.3%, and other→other for 19.5–24.8%. Transitions involving eng account for 75.2–80.5% across configurations. The English-heavy page distribution contributes to this concentration.

The leading predicted-route pair changes across extractor–LID combinations, and the largest pair share is at most 13.2% within a configuration. Examples include eng→glg (Galician), dag→eng (Dagbani), dan→eng (Danish), eng→ita (Italian), eng→rus (Russian), and eng→spa (Spanish). Flips are also dispersed across URL-host clusters: the largest cluster contributes 1.2–3.2% of flips across configurations, and the five largest clusters together contribute 5.6–10.5%.

4.6 Human analysis of sampled flips

Table 7 summarizes the two panel-level annotation tasks. Route-support judgments agree on 96 of 100 items, with Cohen’s $\kappa = .940$. The four disagreements are between an indeterminate label and a determinate label and resolve to indeterminate in consensus. Consensus yields 35 unsupported, 29 supported, and 36 indeterminate routes. Among the 64 determinate judgments, 54.7% are unsupported by the human-selected main text, with a Wilson 95% confidence interval of 42.6–66.3%.

Extraction-change judgments agree on 86 items, with $\kappa = .721$. Consensus assigns 64 pages to added surrounding text, 22 to changed content selection, 9 to mixed behavior, and 5 to unclear. Among the 35 unsupported routes, 18 are added surrounding text, 10 are changed content selection, and 7 are mixed. Changed content selection therefore appears either alone or as part of a mixed case in 17 of the 35 unsupported flips.

Table 8 gives one audited context flip. The human-selected text and the annotated-region extraction contain the same English description. Full-page extraction contains a shorter English account description together with Chinese interface strings, and CommonLingua changes its predicted route from eng to wuu (Wu Chinese). The panels judge the full-page route unsupported by G , and the extraction change is labeled mixed. The Chinese interface strings are *tiězi* “Post”, *duìhuà* “Conversation”, *X de xīn yònghù?* “New to X?”, *lìjǐ zhùcè* “Sign up now”, *chūcuò le* “Something went wrong”, and *qǐng chángshì chóngxīn jiǎzài* “Please try reloading.”

A context flip records a changed prediction. The audit then evaluates whether the route on F_e is supported by G . Among the 35 unsupported routes, 26 move from predicted eng in G to another route in F_e , one moves from another route to eng, and eight move between routes other than eng. Among the 29 supported routes, 21 move from another route to eng, two move from eng to another route, and six move between routes other than eng. These counts describe the page-balanced audit sample and are separate from the corpus-level metadata strata in Table 4.

LID	Extractor	Pairs	Leading pair(s), n (%)	eng→other	other→eng	other→other
GlotLID	Trafilatura	112	eng→glg, 26 (11.4)	49.6	28.1	22.4
	Resiliparse	80	dag→eng, 18 (11.6)	30.3	49.7	20.0
	Magic-HTML	67	dag→eng/eng→spa, 5 (5.3)	47.4	31.6	21.1
OpenLID-v3	Trafilatura	115	dan→eng, 9 (4.6)	28.6	51.0	20.4
	Resiliparse	101	dan→eng, 13 (7.5)	28.2	52.3	19.5
	Magic-HTML	56	eng→spa, 5 (6.3)	48.1	30.4	21.5
CommonLingua	Trafilatura	88	eng→ita, 9 (5.6)	46.6	31.7	21.7
	Resiliparse	99	eng→rus, 26 (11.3)	60.0	15.2	24.8
	Magic-HTML	46	eng→spa, 10 (13.2)	59.2	19.7	21.1

Table 6: Predicted-route transitions among context flips. Source is the route on G and destination is the route on F_e . **Pairs** is the number of distinct predicted-route pairs. The GlotLID–Magic-HTML row has two co-leading pairs, each with five flips. Direction columns are percentages of flips. Route codes are normalized ISO-639-3 model predictions.

Task	Panel A	Panel B	Consensus	Agreement	κ
Route support	29 / 35 / 36	30 / 38 / 32	29 / 35 / 36	96%	.940
Extraction change	65 / 24 / 5 / 6	67 / 24 / 5 / 4	64 / 22 / 9 / 5	86%	.721

Table 7: Panel-level annotation results for 100 sampled context flips. Route-support counts are supported / unsupported / indeterminate. Extraction-change counts are added surrounding text / changed content selection / mixed / unclear.

Rep.	Route	Extracted text (abridged)
G	eng	Founded in 1898, #PekingUniversity pioneers intellectual and technological advancements ... #PKU continues to cultivate world-class innovators and connect global minds.
M_e	eng	Founded in 1898, #PekingUniversity pioneers intellectual and technological advancements ... #PKU continues to cultivate world-class innovators and connect global minds.
F_e	wuu	帖子对话X的新用户? 立即注册... Peking University was established in 1898. This account is dedicated to #Peking University's global outreach and communications. 出错了。请尝试重新加载。

Table 8: A sampled context flip (CommonLingua with Resiliparse). The route sequence is eng→eng→wuu (Wu Chinese). Consensus labels the full-page route as unsupported by G and the extraction change as mixed.

5 Discussion

5.1 How coverage affects extractor comparisons

The context-flip rate asks how often full-page extraction changes the route after annotated-region extraction has preserved the route on G . The all-page route-loss rate also records extraction failure. Magic-HTML illustrates the distinction: its context-flip estimates are low, while its lower coverage contributes most of its all-page route loss. Across LID systems, the nonempty-output intervals overlap at least one other extractor, leaving no uniform extractor ranking on this endpoint. The

$G \rightarrow M_e$ disagreement summary also shows that route changes can arise before full-page context is introduced. The context-flip rate conditions those cases away.

This pattern complements extraction studies that track downstream filtering. Li et al. (2026) show that extractor choice changes which webpages survive a filtering pipeline. Here, the same page can also move between language routes after extraction changes the text presented to LID. Language routing assigns pages to language-specific corpus partitions, so a context flip can move a fixed webpage between partitions and change the material available downstream.

5.2 Eligibility restricts the paired comparison population

The common eligible set excludes pages on which any extractor fails the primary eligibility rule. Context-flip rates are 0.95–2.39% inside this intersection and 4.61–11.82% among extractor-specific eligible pages left outside it, so the paired differences describe a more restricted population than the extractor-specific estimates in Table 2.

5.3 LID input processing

CommonLingua consumes at most the first 512 UTF-8 bytes of extracted text. Full-page extraction can change the retained content and which material occupies that fixed input window, so both effects contribute to the CommonLingua results.

OpenLID-v3 applies its released normalization before prediction, while GlotLID receives the normalized extraction directly. The factorial design therefore measures each extractor–LID combination using each LID system’s released input processing.

5.4 Coarse metadata and human evidence

The coarse non_en contrast remains positive after standardization to the shared meta.level distribution. The result is aggregate because the metadata field pools multiple languages.

The audit separates route changes by support in G . Among the 29 supported flips, 21 move from another predicted route to English, and the destination English route is supported by the human-selected main text. The audit labels the F_e route, leaving the status of the original G route unassessed. This pattern is consistent with corrective route changes in part of the sample. Unsupported flips concentrate in the opposite direction: 26 of 35 move from English to another predicted route. The extraction-change task describes visible changes in the extracted text: added surrounding material is frequent, and changed content selection appears in 17 of the 35 unsupported cases. These labels are descriptive. Causal attribution would require controlled interventions.

6 Conclusion

Changing the HTML context seen by a main-content extractor can change downstream language routing even when extraction of the annotated main region preserves the route on G . Context flips occur in all nine extractor–LID configurations. The all-page analysis shows how extraction coverage contributes to route loss when some pages yield no text for LID.

The human audit shows that full-page route changes are heterogeneous: supported, unsupported, and indeterminate cases all occur, with visible extraction changes that include added surrounding material and shifts in selected content. The fixed snapshot’s coarse non_en contrast also remains positive after standardization to the pooled meta.level distribution. The context-flip and all-page analyses show that extraction context can affect text availability and page-level language routing.

Limitations

The analysis set is English-heavy under the fixed snapshot’s coarse metadata: 6,646 pages carry en and 1,162 carry non_en. The non_en field aggregates multiple languages, so the metadata contrast supports only stratum-level estimates. Individual-language context-flip rates require finer reference labels. The transition analysis uses model-predicted routes and characterizes routing behavior. LID accuracy estimates require reference language labels aligned with the evaluated inventories.

The extractor comparison covers Trafilatura, Resiliparse, and Magic-HTML. Neural extractors such as Dripper (Liu et al., 2026) form an additional model family for future cascade evaluation. Context-flip rates depend on extractor-specific eligibility, and the paired analysis further restricts evaluation to the common eligible set. The all-page route-loss rate counts an empty output as route loss, so extraction availability contributes directly to that endpoint. CommonLingua’s 512-byte input limit also means that extraction changes can alter which text falls inside the model’s observed prefix.

The human study samples 100 pages from the 737 pages with at least one context flip. Each page receives equal sampling weight and contributes one selected flip event. The resulting direction counts describe this page-balanced audit sample. Corpus event-weighted direction rates fall outside the sampling design. Support judgments use excerpts of the human-selected main text, and 36 items remain indeterminate after consensus. The later extraction-change task exposes the system identities, route labels, and all three excerpts, so its labels are descriptive accounts of visible extraction changes. Controlled HTML interventions would support finer attribution to individual page regions.

Ethical Considerations

The study uses webpages and human-selected content distributed through WebMainBench (Open-DataLab, 2026b). Six student volunteers participated without compensation, and all provided informed consent before taking part. The human evaluation was conducted using short benchmark excerpts, with URLs withheld during the route-support task.

References

- Adrien Barbaresi. 2021. [Trafilatura: A web scraping library and command-line tool for text discovery and extraction](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131, Online. Association for Computational Linguistics.
- Adrien Barbaresi and Gaël Lejeune. 2020. [Out-of-the-box and into the ditch? multilingual evaluation of generic text extraction tools](#). In *Proceedings of the 12th Web as Corpus Workshop*, pages 5–13, Marseille, France. European Language Resources Association.
- Janek Bevendorff, Sanket Gupta, Johannes Kiesel, and Benno Stein. 2023. [An empirical comparison of web content extraction algorithms](#). In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2594–2603, New York, NY, USA. Association for Computing Machinery.
- Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2018. [Elastic ChatNoir: Search engine for the ClueWeb and the Common Crawl](#). In *Advances in Information Retrieval – 40th European Conference on IR Research (ECIR 2018)*, volume 10772 of *Lecture Notes in Computer Science*, pages 820–824, Cham. Springer.
- Aurélien Bournonville, Gaël Dias, Thomas Largillier, Emmanuel Marchand, Fabrice Maurel, Guillaume Pitel, and François Rioult. 2025. [Multilingual evaluation of main content extractors for web pages](#). In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’25*, pages 2853–2857, New York, NY, USA. Association for Computing Machinery.
- 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, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Mariia Fedorova, Nikolay Arefyev, Maja Buljan, Jindřich Helcl, Stephan Oepen, Egil Rønningstad, and Yves Scherrer. 2026. [OpenLID-v3: Improving the precision of closely related language identification – an experience report](#). In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 275–292, Rabat, Morocco. Association for Computational Linguistics.
- Ivan Habernal, Omnia Zayed, and Iryna Gurevych. 2016. [C4Corpus: Multilingual web-size corpus with free license](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 914–922, Portorož, Slovenia. European Language Resources Association (ELRA).
- Hugging Face. 2025. [FineWeb 2 processing pipeline](#). GitHub repository. Source code accompanying FineWeb2.
- Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. [GlotLID: Language identification for low-resource languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.
- Jeffrey Li, Joshua P. Gardner, Doug Kang, Fangping Shi, Karanjeet Singh, Chun-Liang Li, Herumb Shandilya, David Leo Wright Hall, Oncel Tuzel, Percy Liang, Ludwig Schmidt, Hadi Pouransari, and Fartash Faghri. 2026. [Beyond a single extractor: Re-thinking HTML-to-Text extraction for LLM Pre-training](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5836–5861, Rabat, Morocco. Association for Computational Linguistics.
- Mengjie Liu, Jiahui Peng, Wenchang Ning, Pei Chu, Jiantao Qiu, Ren Ma, He Zhu, Rui Min, Lindong Lu, Linfeng Hou, Kaiwen Liu, Yuan Qu, Zhenxiang Li, Chao Xu, Zhongying Tu, Wentao Zhang, and Conghui He. 2026. [Dripper: Token-efficient Main HTML extraction with a Lightweight LM](#). In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, KDD ’26*, pages 3258–3269, New York, NY, USA. Association for Computing Machinery.
- Stephan Oepen, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie V. Burchell, Lucas Georges Gabriel Charpentier, Pinzhen Chen, Mariia Fedorova, Ona de Gibert, Barry Haddow, Jan Hajič, Jindřich Helcl, Andrey Kutuzov, Veronika Laippala, Zihao Li, Bhavitvya Malik, Vladislav Mikhailov, Amanda Myntti, and 10 others. 2026. [HPLT 3.0: Very large-scale multilingual resources for LLMs and MT. mono- and bi-lingual data, multilingual evaluation, and pre-trained models](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 1409–1434, Palma de Mallorca, Spain. ELRA Language Resource Association.
- OpenDataLab. 2026a. [Magic-HTML](#). Software repository. Version 0.1.6.
- OpenDataLab. 2026b. [WebMainBench](#). Version 1.1 benchmark and evaluation toolkit.
- Pedro Ortiz Suarez, Laurie Burchell, Catherine Arnett, Rafael Mosquera, Sara Hincapié Monsalve, Thom Vaughan, Damian Stewart, Malte Ostendorff, Idris Abdulmumin, Vukosi Marivate, Shamsuddeen Hassan Muhammad, Atnafu Lambebo Tonja, Hend Al-Khalifa, Nadia Ghezaiel Hammouda, Verrah Akinyi Otiende, Tack Hwa Wong, Jakhongir Saydaliev, Melika Nobakhtian, Muhammad Ravi Shulthan Habibi, and 78 others. 2026. [CommonLID: Re-evaluating](#)

state-of-the-art language identification performance on web data. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33063–33080, San Diego, California, United States. Association for Computational Linguistics.

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*. Spotlight.

PleIAs. 2026. [CommonLingua: Byte-level language identification for 334 languages](#).